EXPLORING WAYS TO IMPROVE STI'S RECOGNITION OF THE EFFECTS OF POOR SPECTRAL BALANCE ON SUBJECTIVE INTELLIGIBILITY

Glenn Leembruggen Acoustic Directions Pty Ltd, ICE Design Pty Ltd, Australia
University of Sydney, Australia
Marco Hippler University of Applied Sciences Cologne
Peter Mapp Peter Mapp and Associates UK

1 INTRODUCTION

Over their many years of designing and commissioning sound systems, the authors have acquired considerable field evidence showing that an unbalanced frequency response can greatly affect subjective speech intelligibility. Relatively small changes to the response, sometimes as small as 1 dB, can noticeably affect the intelligibility of conversational speech and the degree of listening concentration that is required.

The speech transmission index (STI) (1) has gained international acceptance as a useful measure of the ability of a transmission path to faithfully transmit speech intelligibility. However, recent work (2), (3) indicates that changes in subjective intelligibility due to poor frequency response do not appear to be reflected in STI measurements of those systems.

The improvements to intelligibility that equalisation can provide suggest that either the speech spectrum or the model of psychoacoustic upward-masking used by STI might not correctly reflect the subjective process of listening to speech. Other mechanisms are also likely to be contributing to this situation.

The work presented in this paper draws upon prior work (2), (4). In (2), Leembruggen and Stacey compared measured STI and subjective word scores and noted a considerable mismatch between those scores, especially when various filters were applied to the speech.

This paper examines the effects on STI of a range of spectra that occur during regular speech and explores the use of alternative masking models with the STI method.

2 PRIOR WORK IN 2003

2.1 Measurement Procedure

A loudspeaker and dummy head with binaural microphones were set up in an anechoic chamber. The response of the speaker was then measured at each ear with binaural microphones at a distance of 1.5 m from the speaker on axis and processed by MLSSA v10w to yield the loudspeaker's anechoic frequency response of the speaker and the system STI. The system was then relocated to a reverberation chamber. Again the system STI was measured at a distance of 1.5 m from the speaker and using acoustic absorption material, the reverberation time of the chamber was adjusted so that the measured STI was approximately 0.5.

Seven different frequency response shaping filters (Filter shapes 3 to 9) were then sequentially inserted into the drive chain to change the speaker's frequency response. For each filter, the impulse response was captured and the frequency response and STI of the system measured with a speech-weighting filter connected in series with the response-shaping filter.

2.2 Subjective Procedure

A CD of anechoically recorded female speech was prepared and consisted of 1000 carrier sentences with single-syllable, phonetically-balanced (PB) words situated at the end of each sentence. Three groups of 50 words were then played through the speaker in the anechoic chamber (Filter shape 1) and recorded on the dummy head at a distance of 1.5 m from the loudspeaker. The system was relocated to a reverberation chamber and another thee groups of words played through the loudspeaker and recorded binaurally at a distance of 1.5 m (Filter shape 2). For each of the seven response-shaping filters, three lists of 50 words were replayed and recorded for filter shapes 3 to 9. When the groups were exhausted, a reshuffled version of the lists was used.

The recordings of the nine shapes were then distributed to listeners in the UK and Australia. In the UK, seven listeners evaluated all or part of the three lists for each of the nine shapes. In Australia, three listeners evaluated all of the three lists for each of the nine shapes. The sentences were presented to listeners through headphones, and the listener wrote down the word at the end of the sentence. The playback level of the recordings was approximately 70 dBA at the listener’s ear for all filter shapes.

2.3 Filter Shapes

The frequency responses of the tonal filters were chosen to exaggerate subjective listening difficulties. Figure 1 shows the relative frequency responses of those filters, and to allow easier comparison, each response is normalised to its value at 1 kHz.

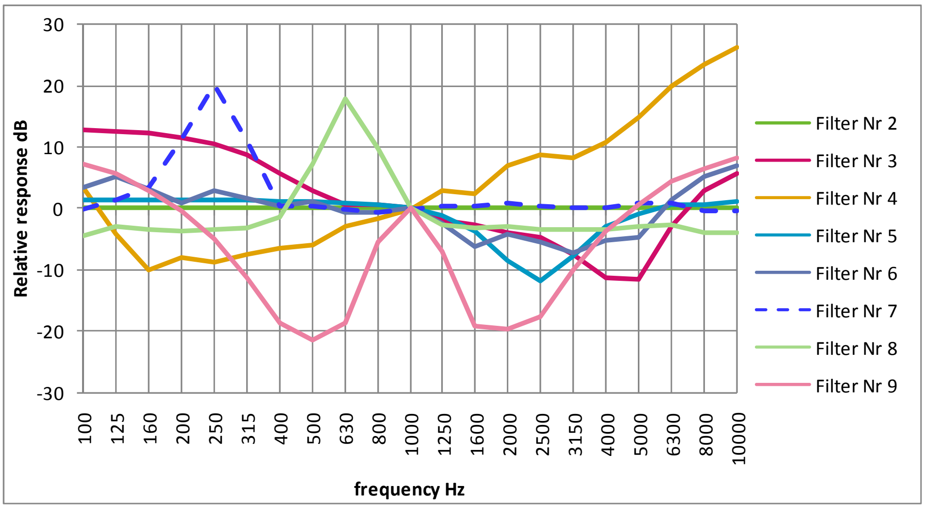

Figure 1 Relative frequency responses of filters used to modify the speech spectrum. Each response is normalised to its value at 1 kHz.

2.4 Word Score Results

Figure 2 gives the word score results for each filter shape. The following comments are made.

a) Although the word score testing was not carried out rigorously in accordance with the ISO TR 4870 standard, and there was a wide range in the results, the trends were clear.

b) The average Australian scores for each filter shape were generally lower than the corresponding UK scores. This was likely to result from accent differences.

c) The UK and Australian average scores showed a similar trend over the range of filter shapes.

d) There was a noticeable reduction in the word score with the filters inserted.

e) Even though the test words were well-articulated, each of the Australian listeners found it necessary to concentrate while listening, in order to discern the test words. More concentration was required for the filtered words. If this concentration had not been applied, the scores would have been lower.

f) The Australian listeners found the process to be tiring, and yet the measured STI was of the order of 0.5, which is a value that is typically specified for sound systems.

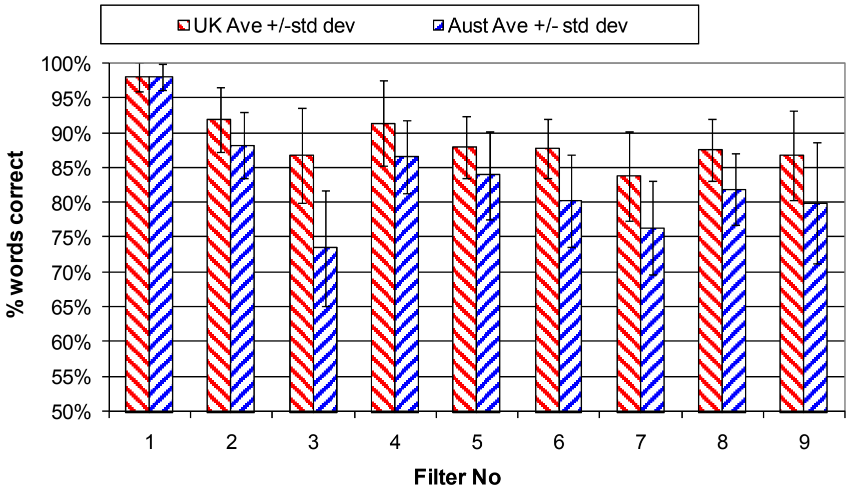

Figure 2 Word scores for the different filter shapes. Note that Filters 1 and 2 have flat responses and Filter 1 is anechoic, while filter shapes 2 to 9 are reverberant. The error bars show the standard deviations.

2.5 Comparison with Measured STIs

The word scores were converted to an equivalent STI value using the common intelligibility score (CIS). Figure 3 shows those word-equivalent STI values and measured values of male STIr, calculated according to the 2003 STI IEC standard. Although the talker was a female and the STI measurements are male, that difference does not change the results appreciably.

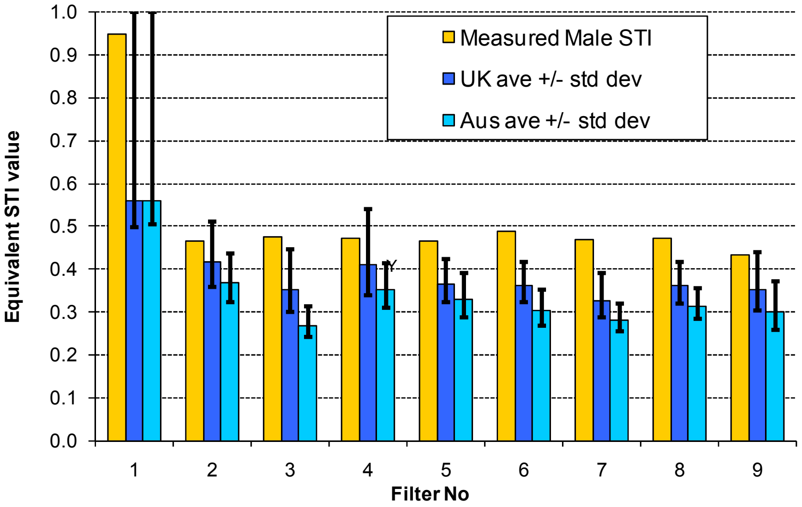

Figure 3 Comparison of equivalent STIs of PB word scores with measured Male STIrs. The error bars show the range of standard deviations.

The following comments are made.

a) In filter shape 1 (anechoic and flat response), the error bars extend up to an STI of 1. This is caused by the CIS conversion which amplifies the STI when word scores exceed 97%. At this value, a 3% change in PB score results in a STI change from 0.55 to 1.0.

b) The word scores are always lower than the measured STIs.

c) The effect of talker accent on intelligibility can be seen. Wijngaarden et al (5) noted that non-native talkers and listeners require a higher STI score for similar intelligibility with native listeners.

2.6 Conclusions

From the work described above, we conclude that the STI method does not satisfactorily account for the loss of subjective speech intelligibility that accompanies speech with poor spectral balance. Possible reasons for this mismatch are:

a) the use of a long term spectrum rather than a short term spectrum

b) an inaccurate masking algorithm

c) a combination of a) and b) above

The work conducted for this paper investigates these aspects.

3 SPEECH SPECTRA

Six talkers (5 male, 1 female) were recorded anechoically and a 10 second segment of each talker extracted. The anechoic data was then reverberated using FIRverb with a reverberation time of approximately 2 s in each octave bandwidth. The spectra of each talker in specific time slices was then found using scan analysis provided by the waterfall function in WinMLS2004.

The following spectra were found using a Hanning window with 50% overlap for each talker for both the anechoic and reverberant environments.

- 10 slices of 1 s length
- 40 slices of 250 ms length
- 200 slices of 50 ms length

3.1 Preparation of Spectra for Analysis

The speech spectra were then prepared for analysis as follows:

a) All spectra were bundled into one-third octave bands.

b) The total rms level of the ten one-second slices was computed for each talker to form the long-term L_{eq} in each one-third octave band.

c) The long-term L_{eq} levels were A weighted and summed to give the long term LA_{eq} of each talker and normalised to the long-term operational speech level of 75 dBA. The resulting normalization factor D was stored for subsequent use.

d) Each of the 1 second, 250 ms and 50 ms time-slice spectra was then adjusted by the normalization factor D.

3.2 Spectral Data

Figure 4 compares the anechoic long term L_{eq} spectrum of the six talkers and their average with the IEC spectrum (6) which has been interpolated into 1/3rd octave bands and normalised to 75dB (for comparison to the time-slice spectra). Figure 5 shows the corresponding data for the reverberant environment.

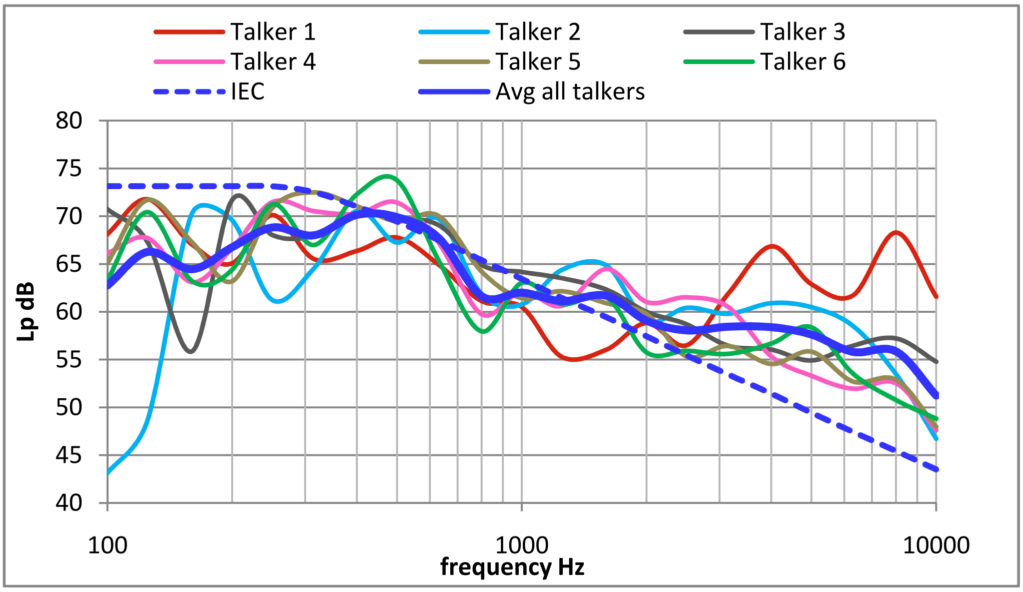

Figure 4 Comparison of long term L_{eq} anechoic spectra of 6 talkers and their average with IEC spectrum

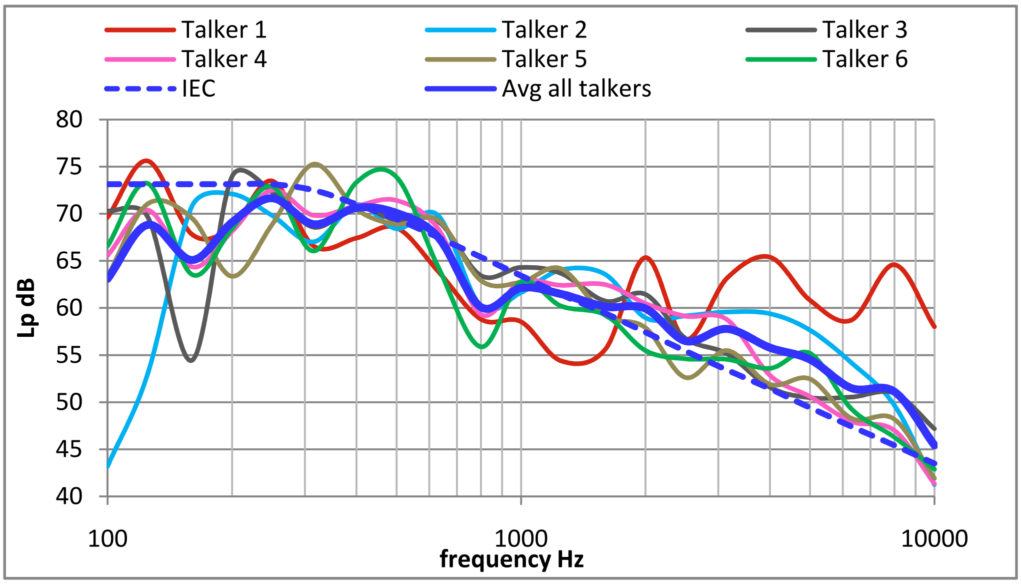

Figure 5 Comparison of long term L_{eq} reverberated spectra of 6 talkers and their average with IEC spectrum

Figure 6 compares the IEC spectrum with spectral data in 1/3rd octave bands for Talker 1 in the anechoic environment for the three time-slices, while Figure 7 shows the equivalent data for the reverberant environment. The following data is shown:

- normalised IEC spectrum
- mean level of each 1/3rd octave band for the stated length of time-slice
- 10th percentile of each 1/3rd octave band for the stated length of time-slice
- 90th percentile of each 1/3rd octave band for the stated length of time-slice
- spectrum of an individual time-slice that shows strong differences with the IEC spectrum

To ensure that the statistics were not skewed by spectra representing soft syllables or gaps between words, any spectrum whose total level was less than 50 dBA was removed from the analysis.

Spectral data for the remaining five talkers in both anechoic and reverberant environments is given in the Appendix.

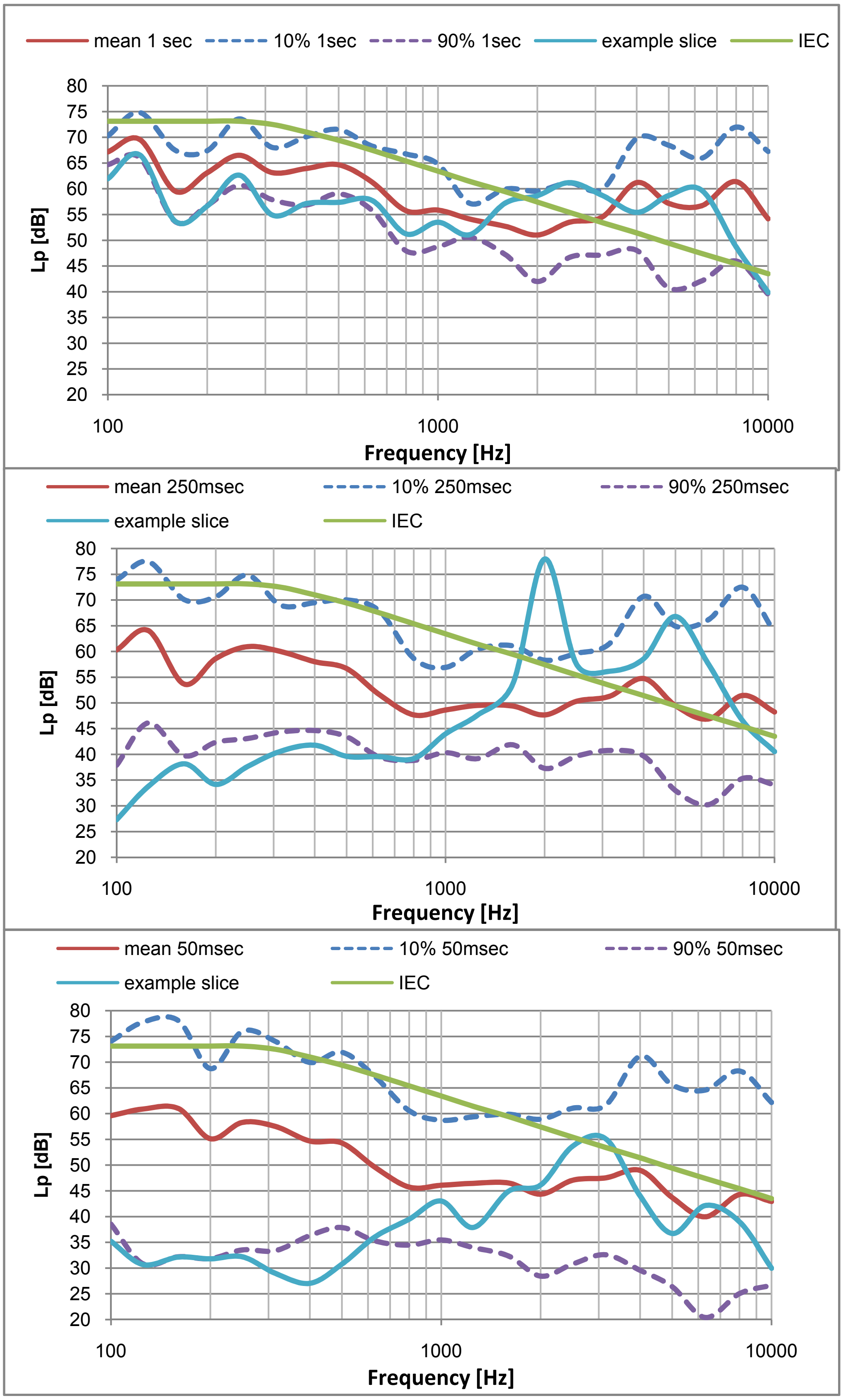

Figure 6 Statistics of speech spectra in 1/3rd octave bands for 1 s (top), 250 ms (middle) and 50 ms (bottom) time slices in anechoic environment for Talker 1.

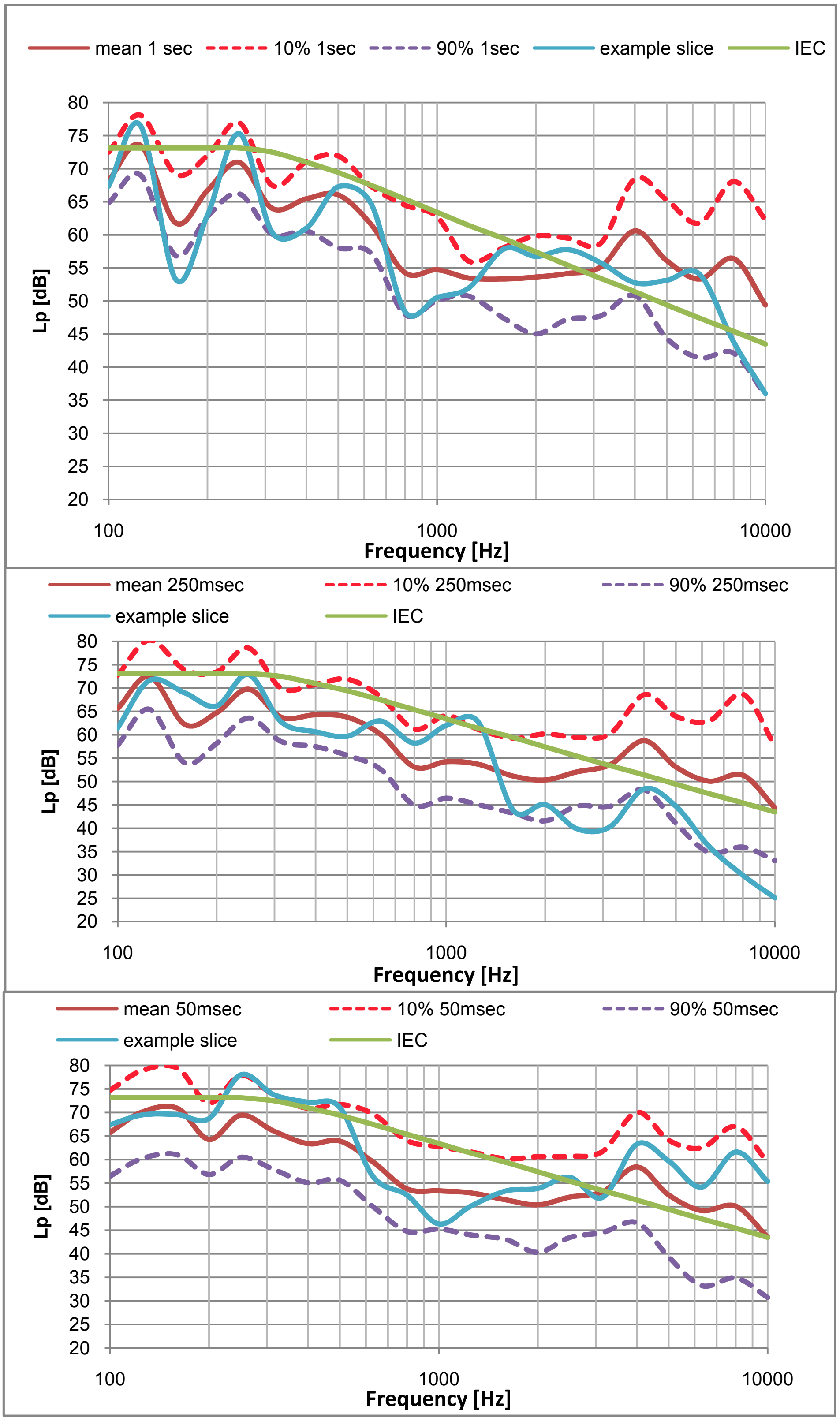

Figure 7 Statistics of speech spectra in 1/3rd octave bands in reverberant environment for three same time slices as the anechoic spectra of Figure 6.

4 MASKING ALGORITHMS

4.1 Loss of SNR due to masking

Three masking models have been used to compare their relative impacts on STI values with the range of filtered speech spectra described in 4:

a) Model used in STI draft IEC 60268-16 standard, slated for release in 2010.

b) Model used in the Speech Intelligibility Index SII as per the standard S3.5-1997 (7).

c) Models derived from Excitation Pattern as developed by Moore and Glasberg et al (8), (9), (10), (11), (12), (13).

Each of the models produces an equivalent noise term to quantity the effect of masking in each octave band. This term is then used to adjust the measured modulation index.

4.2 STI Masking Algorithm

In the STI method, only the speech in octave band *i* is deemed to produce masking noise in the adjacent octave band *i*+1 immediately above band *i*. This masking noise adds to other noise sources (such as background noise, threshold of hearing) that are also present in octave band *i*+1, further degrading the SNR of the speech in that band.

The amount of masking in octave band *i*+1 depends on the level of the signal in octave band *i.* The STI standard specifies a slope of the masking curve according the level in each octave band.

Figure 8 illustrates the formation of masking noise and the signal to noise ratios in the 1 kHz octave band (SNR) for male speech with the IEC spectrum adjusted to levels of 71 dBA and 97 dBA. For these two speech levels, the apparent signal to noise ratios in the 1 kHz octave band due to masking by speech in the 500 Hz octave band are 19 dB and 4 dB respectively.

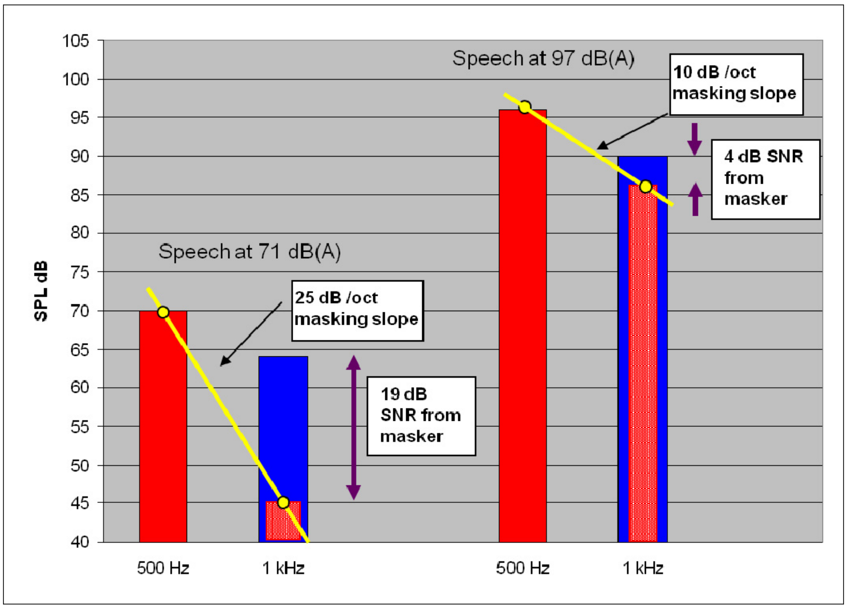

Figure 8 Simplified illustration of STI's masking of wanted signal at 1 kHz by a signal at 500 Hz.

4.2.1 Calculating the masking in STI

To determine the auditory masking level in say octave band *k*, the sound pressure level of the speech and ambient noise in the preceding octave band *k*-1 must first be found. Using the relationships between the acoustic level and the associated masking level given in Table 1, the equivalent masking noise *am*dB is found for band *k*.

Item	Range 1	Range 2	Range 3	Range 4
Octave band level L_{k-1} dB SPL	< 63	≥ 63 and < 67	≥ 67 and < 100	≥ 100
Auditory masking *am*dB	0,5 × L_{k-1} - 65	1,8 × L_{k-1} – 146,9	0,5 × L_{k-1} – 59,8	-10

Table 1 Auditory masking levels as a function of the acoustic octave band level.

As the auditory masking factor $\boldsymbol{amf}$ is an intensity parameter, Eq 1 is used to convert the *am*dB into that form. Eq 2 is then used to calculate the intensity of the audio masking signal in each octave band.

$$amf = 10^{(amdB/10)} \qquad \text{Eq 1}$$

$$I_{am,k} = I_{k-1} * amf \qquad \text{Eq 2}$$

where:
$I_{am,k}$ is the audio masking intensity in octave band k
I_{k-1} is the intensity of the signal in octave band k-1

The masking Intensity $I_{am,k}$ is then used to adjust each modulation index m_{kf} as per Eq 3.

$$m'_{kf} = m_{kf} \frac{I_k}{I_k + I_{am,k} + I_{rs,k}} \qquad \text{Eq 3}$$

where
$I_k$is the intensity of the signal in octave band k
$I_{rs,k}$is the absolute reception threshold which is not discussed further.

4.3 SII Masking Model

A detailed description of the computation of SII is given in (7). This standard claims that SII is “highly correlated with the intelligibility of speech under a variety of adverse listening conditions such as noise masking, filtering, and reverberation”.

SII is intended to reflect the proportion of total speech cues available to the listener and its values range between 0 and 1. An SII of 0.5 indicates that half of the speech cues are delivered to the listener.

As per STI, SII is completely based on the signal to noise ratios in specific frequency bands, with every parameter contributing to intelligibility loss being converted to an equivalent noise level.

To calculate the SII, the long-term averaged spectrum levels of speech and noise are used. Both speech and noise signal are determined separately in specific frequency bands (critical bands, one-third octave bands and octave bands). Adjustments are applied to the measured speech level to take into account for effects such as upward spread of masking, hearing threshold for pure tones, and distortion caused by high speech levels.

Additional background information is given in (14).

In contrast to the STI algorithm, the SII algorithm allows a given band of speech *i* to produce masking noise in all other bands above band *i*. This process is illustrated in Figure 9 in which the 500 Hz speech band produces equivalent masking noise in the other bands (shown in brown) reducing the signal to noise ratio of the other bands (light blue).

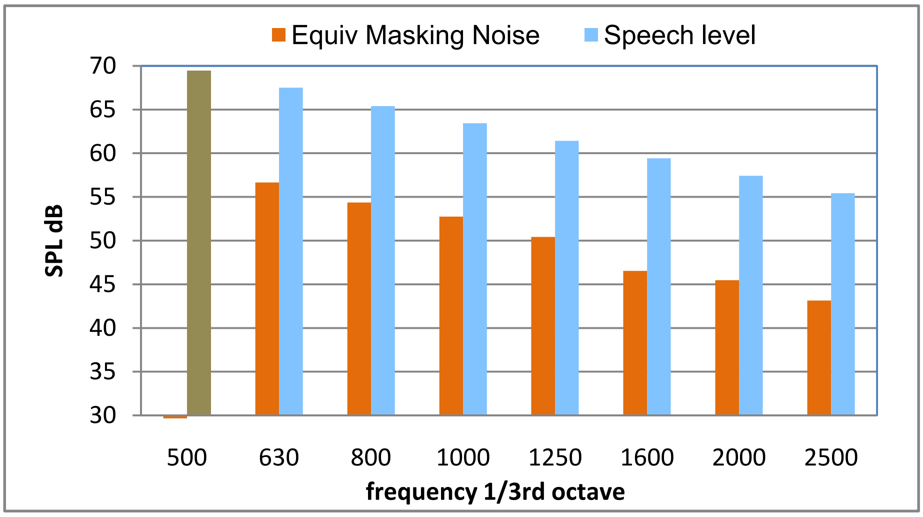

Figure 9 Conceptual illustration of masking noise predicted by the SII algorithm in the 1/3rd octave bands above 500 Hz due to speech in the 500 Hz band. (Data is for Talker 1 reverberant, 50 ms spectra time-slice 71)

As per the STI method, the SII method computes the total equivalent masking noise spectrum level Z_i for the i^{th} calculation band (see Eq 5). The equivalent noise comprises of two components:

a) A term N_i representing ambient noise in that band. As the background noise is included separately in our calculations of STI, this term is not used.

b) A summed term representing the upward spread of masking resulting from the speech signal in other bands.

4.3.1 Computing upward masking for SII

The process commences by finding the slope C_i of the upward masking function for the *i*-th frequency band according to Eq 4,

$$\mathrm{C_i} = -80 + 0.6[\mathrm{B_i} + 10\log \mathrm{F_i} - 6.353] \qquad \text{Eq 4}$$

where:
B_i is the larger of N_i (the background noise) or the self speech masking spectrum V_i expressed as a spectrum level

Eq 4 is derived from Ludvigsen (15) which in turn has been distilled from masking curves found in Zwicker (16), (17). The total level of masking in the *i*-th calculation band is found from the second (summed) term in Eq 5.

$$Z_i = 10\log\left\{10^{0.1N'_i} + \sum_{k}^{i-1} 10^{0.1[B_k + 3.32C_k \log(0.89F_i/F_k)]}\right\} \qquad \text{Eq 5}$$

where:
N'_i is the noise spectrum level in band *i*
$\mathrm{C_k}$ is the slope per octave of the upward spread of masking from band *k* below band *i*
B_k is the larger of $\mathrm{N_i}$ (the background noise level) or the self-speech masking spectrum $\mathrm{V_i}$
k is the index for each one-third octave frequency band

The level of upward masking in each frequency band is calculated by multiplying the ratio of the frequency being masked F_i with the masking frequency F_k (expressed in octaves) by the slope C_i of for the frequency band *i*.

There are two important differences between the SII model and the STI model of masking:

- ➢ STI uses the total sound pressure level as the masker.
- ➢ The self-speech masking level V_i is defined as 24 dB below the speech level, and surprisingly, this level is used to determine the masking slope C_i .

This adjustment factor of -24 dB was proposed by French and Steinberg (18 p. 110) as the speech level that is available to cause masking and was used to prevent the articulation from exceeding unity if the effective sensation level exceeded 36 dB[1]. To achieve this, the available level is deemed to be 24 dB below the long-term average intensity of speech in each band. In essence, the factor of 24 dB delays the onset of masking until the spectrum level in each band exceeds the threshold of hearing by more than 50 dB. To give context to this level, a spectrum level (1 Hz bandwidth) of 50 dB is present in the 1 kHz third-octave band when speech is approximately 81 dBA.

Interestingly, Ludvigsen (15) makes no mention of the 24 dB factor is his prediction of masked thresholds. Regardless of the correctness or otherwise of French and Steinberg's use of the 24 dB factor, we believe that the slope of the masking characteristic should be directly determined by the method of (15). We note also that the 24 dB factor has no equivalent in the STI model.

Work conducted in (4) and also for this paper has showed that when the 24 dB factor is used, self-masking of speech produces little equivalent masking noise.

4.4 Excitation Pattern Model

The excitation pattern for a given sound refers to the distribution of the excitation of neurons in the ear evoked by that sound. The pattern is derived from the model describing the series of bandpass auditory filters that are present in the inner ear (8) to (13).

4.4.1 Response of the Outer and Middle Ear

The frequency response of the acoustic transmission system from a free-field to the cochlear must be included in calculations of excitation patterns (19). The ANSI standard S3.4-2007 (20) gives the frequency response shown in Figure 10 for the ear's transmission system.

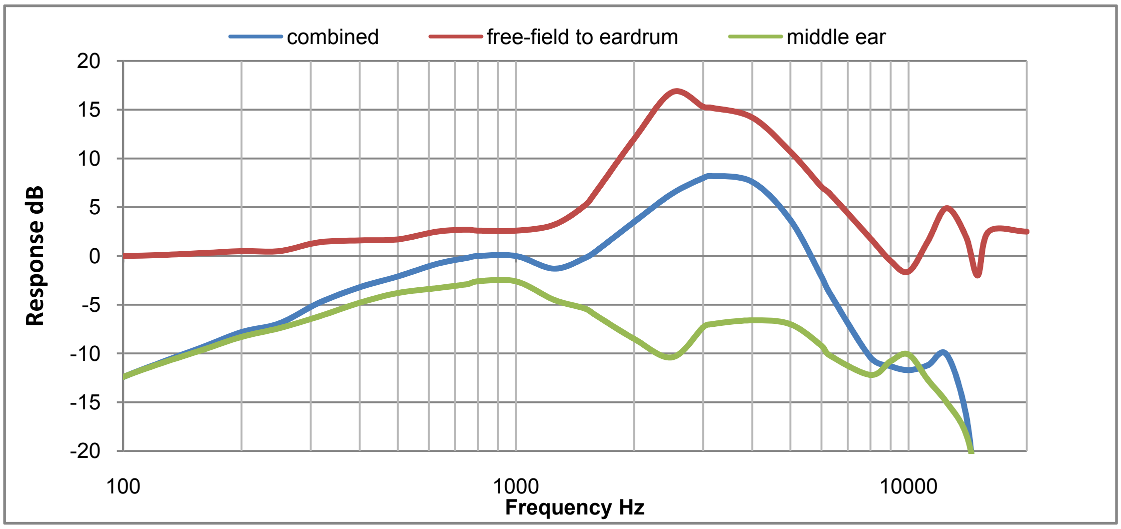

Figure 10 Frequency responses of outer and middle ear sections

[1] The effective sensation level E-M where E is the level of speech above hearing threshold and M is the masking from all noise sources. The lower limit of E is 6 dB giving 0% articulation, and the range from 0 to 100% articulation is 30 dB. Hence W=(E-M-6)/30. Under noise less conditions, the articulation exceeds 100% if E exceeds 36 dB, so a term Bs-24 is introduced as a self-noise term to limit the articulation.

4.4.2 Auditory Filters

The auditory filter model estimates the frequency selectivity of the hearing system at a particular centre frequency and is believed to correspond to the filtering that occurs at a particular place along the basilar membrane. Integral to the filter model is the concept of the filter's Equivalent Rectangular Bandwidth (ERB).

The generic equation describing the overall shape of the auditory filter is the rounded exponential ROEX(p,r) developed by Patterson (21) and given in Eq 6.

$$W(g) = (1 + pg)e^{-pg} \qquad \text{Eq 6}$$

where:
$(1 + pg)$ rounds the top of the exponential and flattens the filter at its centre frequency f_c.
p determines the shape of the passband of the filter (bandwidth and the slope of the filter skirts). The higher the value of p the more sharply tuned the filter is.
g is the normalised deviation in frequency from the filter centre frequency in Eq 7:

$$g = |f - f_c|/f_c \qquad \text{Eq 7}$$

where:
f_c is the centre frequency of the auditory filter
f is the frequency

As the upper and lower skirts of the bandpass filters are asymmetrical, the term p in Eq 6 is split into an upper $p_u$and lower skirt p_l.

The model of the auditory filter has undergone considerable refinement in the last twenty five years, especially the equations describing the ERB and the slope of the filter skirt below the centre frequency. Eq 8, Eq 9 and Eq 10 give the most recent versions of the ERB; slopes of the lower filter skirt p_l and upper skirt respectively. (11)

$$ERB = 24.7(4.37F + 1) \qquad \text{Eq 8}$$

$$p_l(x) = p_{l(51)} - 0.35(p_{l(51)}/p_{l(51.1k)})(X - 51) \qquad \text{Eq 9}$$

$$p_u(x) = p_{u(51)} + 0.118(X - 51) \qquad \text{Eq 10}$$

where:
X is the sound pressure level in each ERB . This term changes the filter shape according to the level of the SPL at the filter's bandwidth level relative to 51 dB.

$p_{l(51)}$ and $p_{u(51)}$ are both calculated by Eq 11 :

$$p_{l(51)} = p_{u(51)} = \frac{4f_c}{ERB} \qquad \text{Eq 11}$$

An example of the model's computation of auditory filters is shown in Figure 11, in which the filters at 1 kHz are computed for input sound pressure levels ranging from 20 dB to 90 dB. The decreasing slope of the lower filter skirt with increasing level is readily seen.

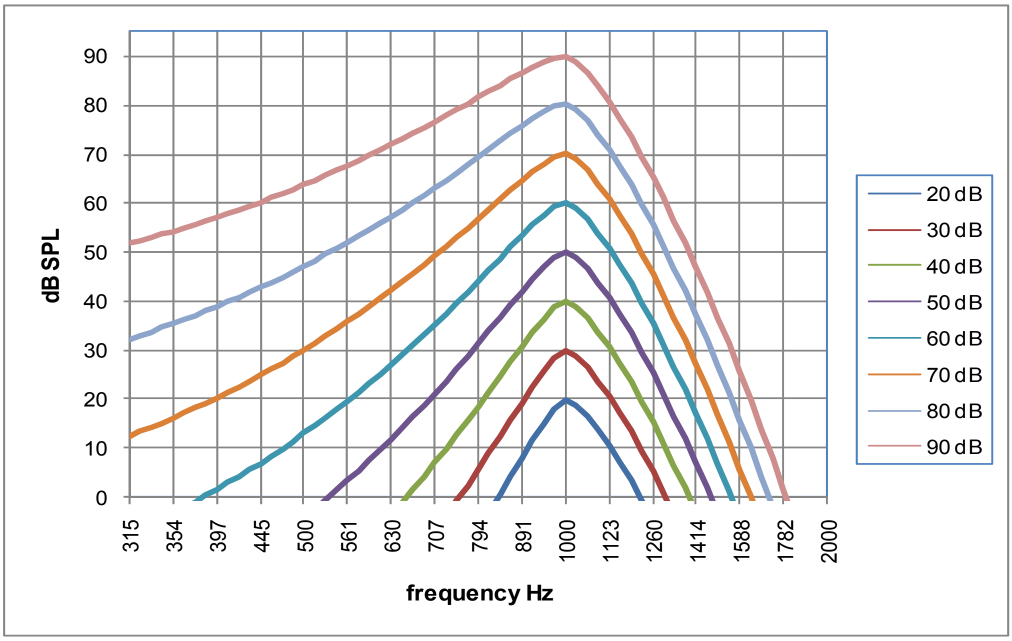

Figure 11: Auditory filter with sine wave input at 1 kHz for input levels ranging from 20 to 90 dB.

4.4.3 Calculating the Excitation Pattern

The excitation pattern of a signal is the auditory stimulation provided to the brain. It represents the spread of neural activity (or excitation) along the basilar membrane in the cochlea. Each point on the excitation pattern results from the output of a different auditory filter in the cochlea.

An example of the calculation of an excitation pattern is given in Figure 12 which shows the way a 1 kHz tone produces outputs over a range of auditory filters. The combination of these auditory filters primarily distributes the neural activity above the 1 kHz auditory filter, producing an upward spread of masking noise. In comparison, the downward spread of masking is much less.

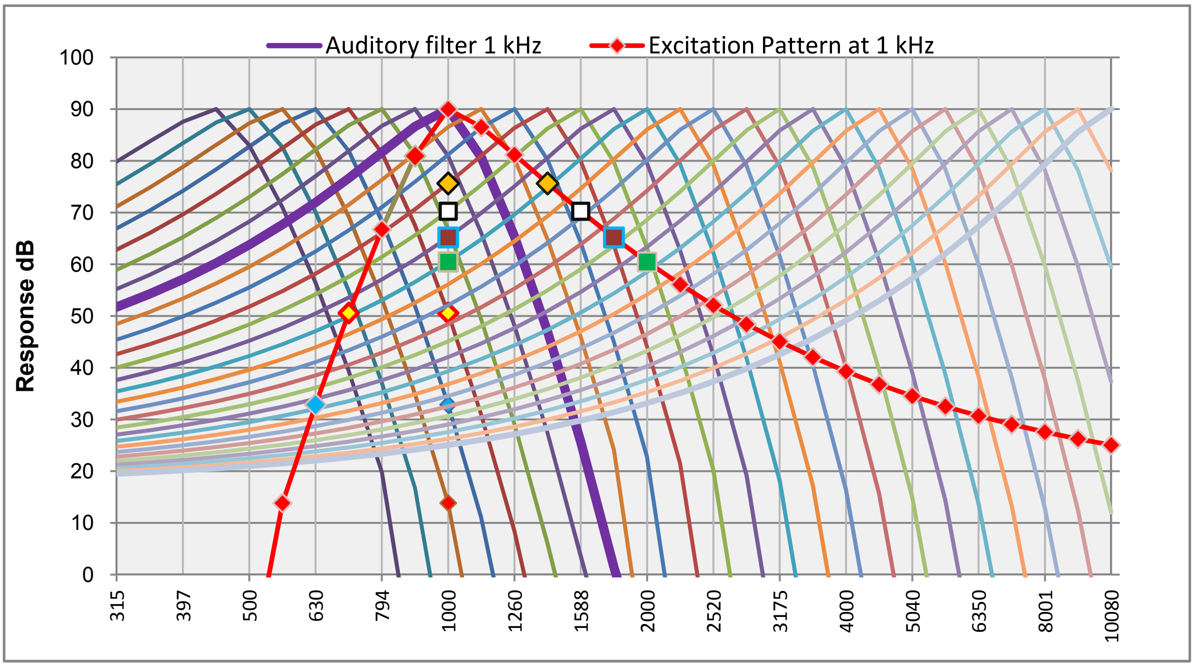

Figure 12 Illustration of the method to calculate the excitation pattern at 1 kHz from the responses of auditory filters. The coloured markers at various frequency points on the excitation pattern graph (red) correspond with the same markers on the 1 kHz ordinate.

Another illustration of an excitation pattern is given in Figure 13. This figure shows the EP for a sound consisting of a fundamental at 200 Hz and its first nine harmonics, all present at equal levels, ranging from 40 dB to 100 dB SPL. The upward trend of each pattern is due to the frequency response of the outer and middle ear. As the level increases, the gaps between the harmonics progressively "fill in" representing additional neural activity due to the upward spread of masking.

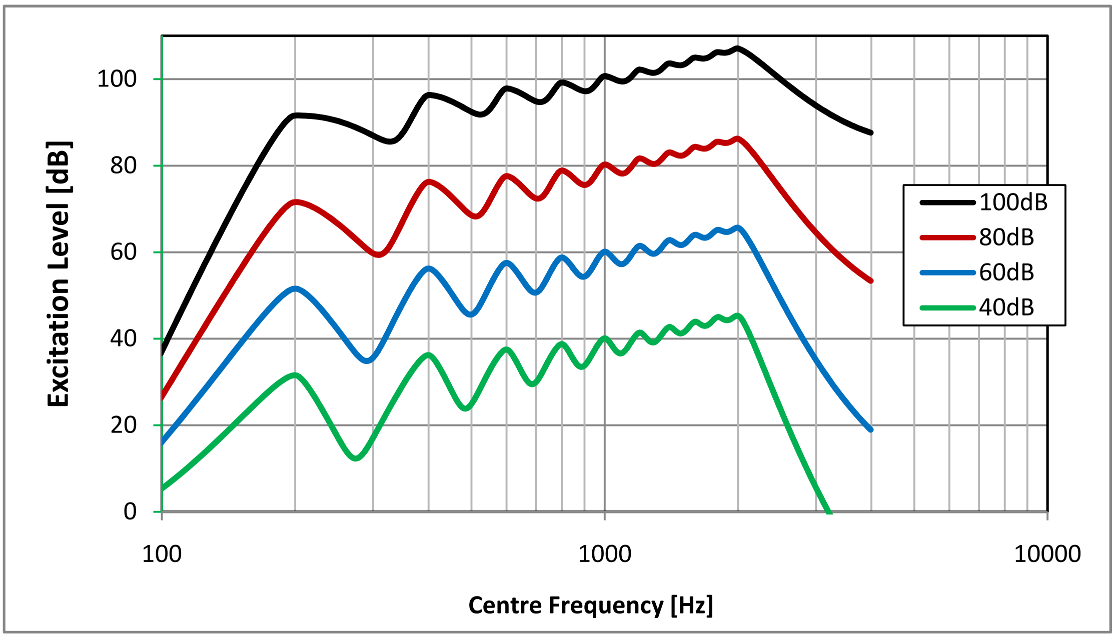

Figure 13 Calculated excitation pattern for a sound consisting of the first nine harmonics of 200 Hz.

4.4.4 Methods of finding the effective signal to masking ratio

The masking noise levels produced by each speech spectrum were found using three methods, each of which gave different results.

Method 1

The effective signal to noise ratio SNR in a given frequency band can be determined from the difference between two excitation patterns calculated using the method described in Section 4.4.5. This method assumes that the speech has spectral lines at 0.1 ERB intervals.

This method yielded unexpectedly low values for the SNRs, and was therefore not used for the STI analysis. Two alternative methods were developed.

Method 2

Method 2 is similar to Method 1, but assumes the speech has spectral lines at 1/3rd octave intervals, as used in the SII model.

Method 3

The speech signal is regarded as being lumped into spectral lines at 1/3rd octave intervals, and the slope of the relevant masking curve is used to compute the masking level. This method is similar to the method used in SII masking model.

4.4.5 Method 1 Difference of EPs with spectral lines at 0.1 ERB intervals

The difference is found between the following two excitation patterns:

- The perceived level of the signal in the *i* th one-third-octave band is the integrated intensity of the excitation pattern (EP) in the *i* th band when only that band is used for calculation of the EP; i.e., all other bands are removed.
- The perceived level of the masking noise in the *i* th one-third-octave band is the integrated intensity of the EP in the *i* th band when all bands other than the *i* th band are used to calculate the EP.

Although Moore's EP method uses a signal's fundamental frequency and its harmonics, we have assumed that speech is a noise-like signal with energy distributed consistently throughout 1/12 octave bands extending from 100 Hz to 10 kHz.

Our method of calculating the EPs of the speech signal and that of the masking signal follows that given in a Fortran programme in (10). The following steps were used:

1. Adjust the selected time-slice of the speech spectrum to account for i) the frequency response for the frequency-response modifying filter and ii) the filtering by the outer and middle ears.
2. Allocate the adjusted 1/3rd octave levels into 1/12th octave band levels.
3. Compute the 0.1 ERB band numbers corresponding to each 1/3rd and 1/12th octave bands.
4. Allocate the 1/12th octave speech levels into 0.1 ERB wide bands using the appropriate corrections for bandwidth. This process provides an equivalent to the frequency harmonics of (10).
5. Allocate the 1/12th octave band levels into ERB wide bands to determine the level of the masking signal X in Eq 10 for subsequent use in the associated 0.1ERB bands.
6. Calculate the excitation pattern with the speech energy in the desired *i*th 1/3rd octave band at 0.1ERB steps.
7. Logarithmically sum the level in each 0.1ERB band included in the *i*th 1/3rd octave band to yield the wanted speech level.
8. Calculate the excitation pattern at 0.1ERB steps with speech energy present in all but the *i*th 1/3rd octave bands.
9. Logarithmically sum the masking noise energy present in the *i*th 1/3rd octave band.

Figure 14 illustrates the process described above, showing the following parameters:

- The input spectrum before ear filtering in 1xERB wide bands at 0.1 ERB intervals for Talker 1 anechoic, unfiltered, 50 ms slice 171.
- The input spectrum after ear filtering in 1xXERB wide bands at 0.1 ERB intervals.
- The input spectrum in 0.1 ERB wide bands at 0.1 ERB intervals (As this spectrum is not integrated into ERB bands, the levels are 10dB (10*log[0.1ERB]) lower than the input spectrum which is in ERB wide bands.
- The EP with only the *i*th band at 500 Hz present.
- The EP with only the *i*th band at 500 Hz removed.

The output of the complete process is shown in Figure 15, which is a set of values at 1/3rd octave intervals of the speech EP value and the masking EP value. The difference between the speech and masking values at each frequency point represents the SNR at each frequency. Figure 16 shows the complete EPs for the 50 ms time-slice 37.

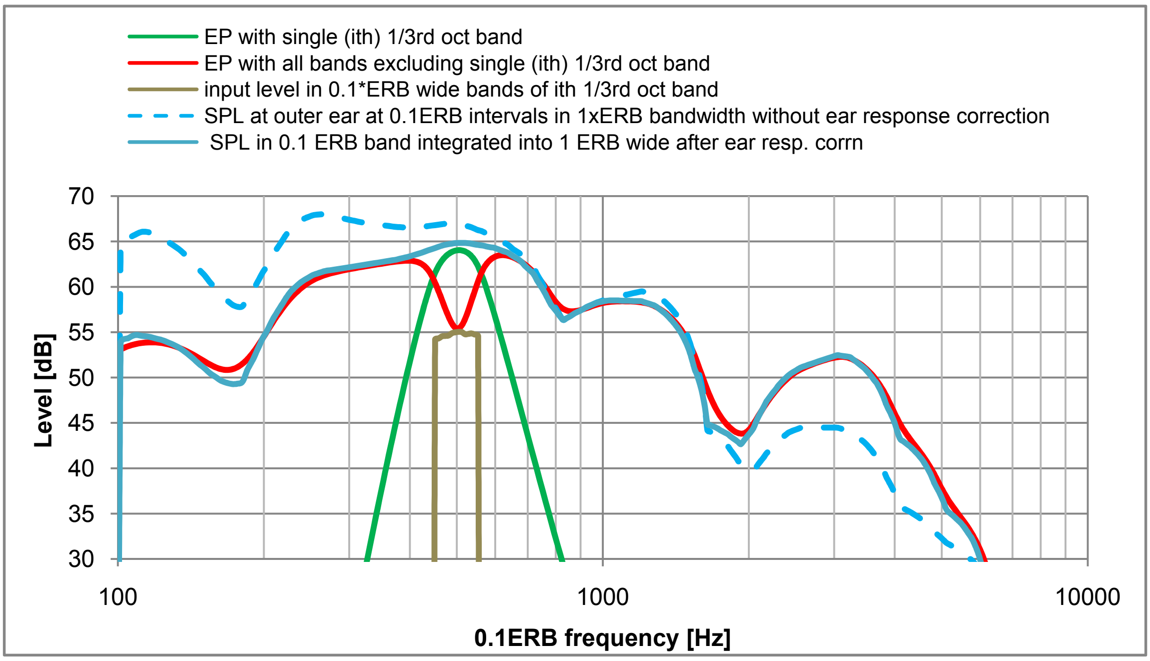

Figure 14 Input SPLs and excitation patterns for Method 1

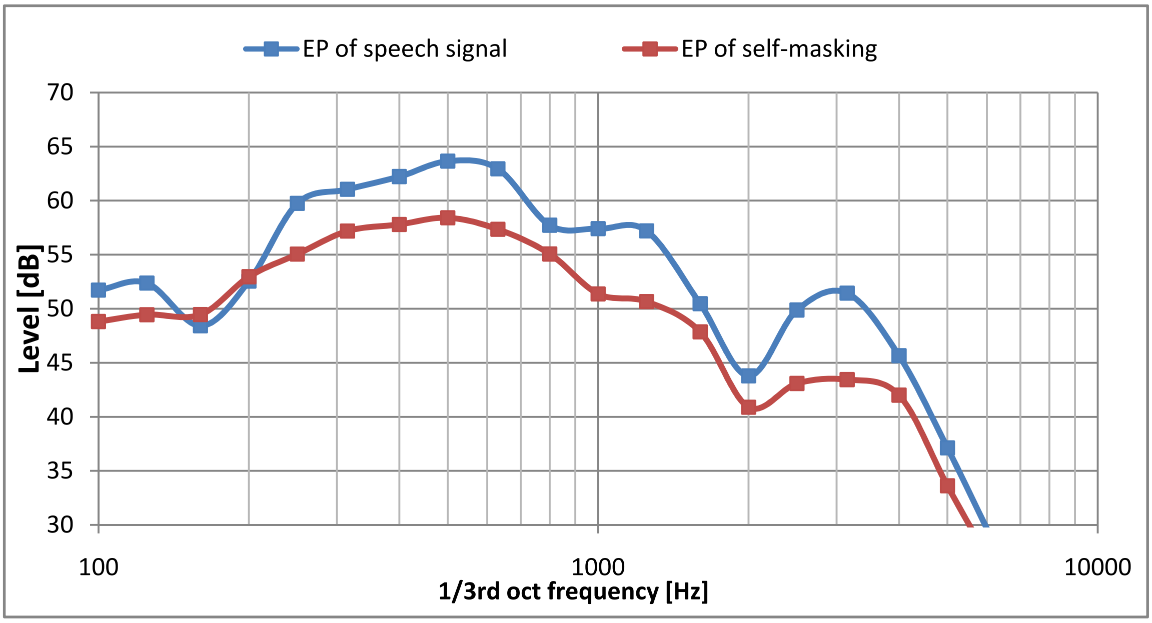

Figure 15 Total excitation patterns of speech and masking signals for Talker 1 anechoic 50 ms time-slice 37.

The results in Figure 15 and Figure 16 show typical SNRs of 5 dB or less, which suggest that the self-masking from normal everyday speech would be sufficient to substantially degrade intelligibility. Clearly this is not the case, and therefore this model appears unsuitable for this process.

Checks of the calculations were made to confirm that the method was being implemented correctly. A literature search yielded (22), in which the authors assumed that the masking pattern calculated from the EP using Moore and Glasberg's model should be parallel to the excitation pattern of the input signal, but shifted vertically downwards by a small amount. The authors used a factor of 6dB to represent the level that the masking level was shifted below the signal and this bears some similarity to our observations.

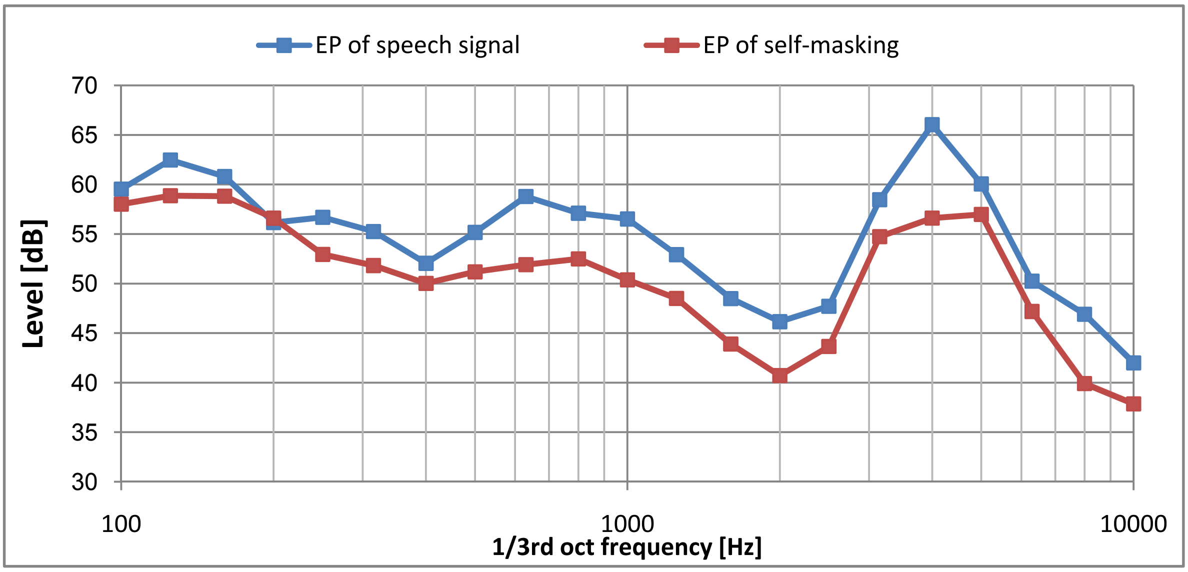

Figure 16 Total excitation patterns of speech and masking signal for Talker 1 anechoic 50 ms time-slice 171.

4.4.6 Method 2 - Difference of EPs with spectral lines at 1/3rd octave intervals

When the spacing of the effective speech spectral lines was progressively decreased from 0.1ERB intervals through ERB and finally to 1/3rd octave intervals, the SNRs progressively increased. From this we conclude that the EP of speech should not be computed with narrow spectral lines that effectively simulate noise. Figure 17 shows the input spectra and the calculated EPs using this method with the 1 kHz band as the i^{th} band.

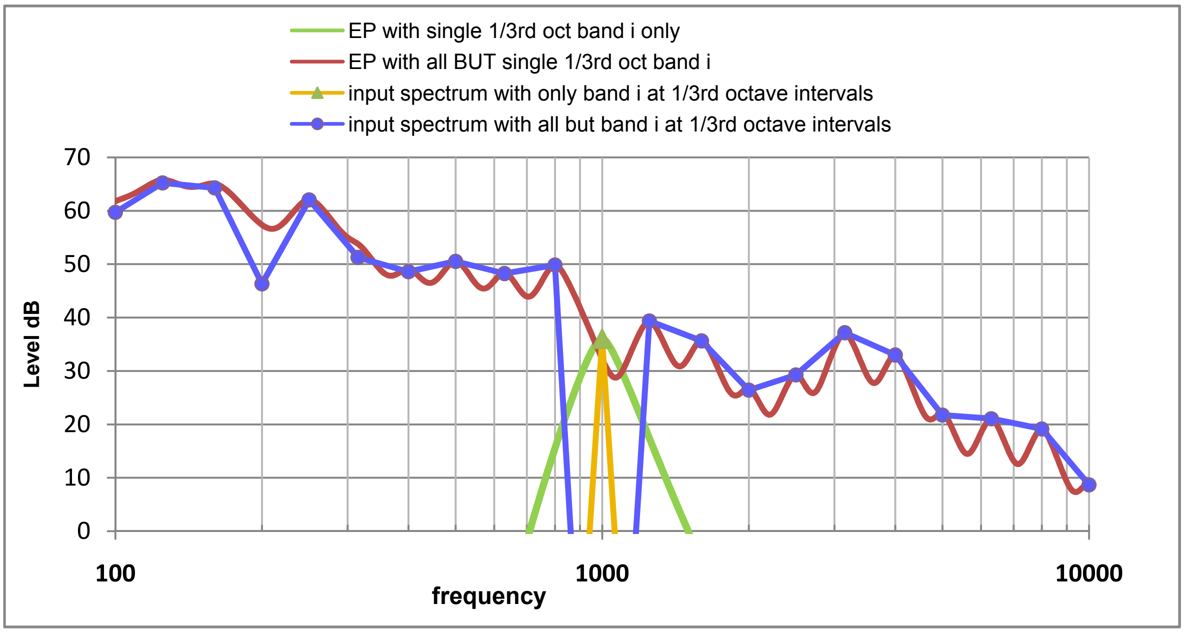

Figure 17 Input SPLs and excitation patterns for Method 2

As the SNRs resulting from this method were more in line with those of the STI and SII methods, Method 2 was used for the STI calculations, with and without ear filtering.

4.4.7 Method 3 - Slope of Excitation Pattern

The excitation pattern for pure tones at octave intervals ranging from 125 Hz to 8 kHz were computed for levels ranging from 20 dB to 90 dB SPL using the method described in Section 4.4.5. From each tone's excitation pattern, equations were developed for lines matching the first part of the excitation pattern (above the tone being examined). Figure 18 shows an example of the lines that were matched to range of EPs at 1 kHz from 20dB to 90 dB SPL. Note that the effect of downward masking is not included in this method

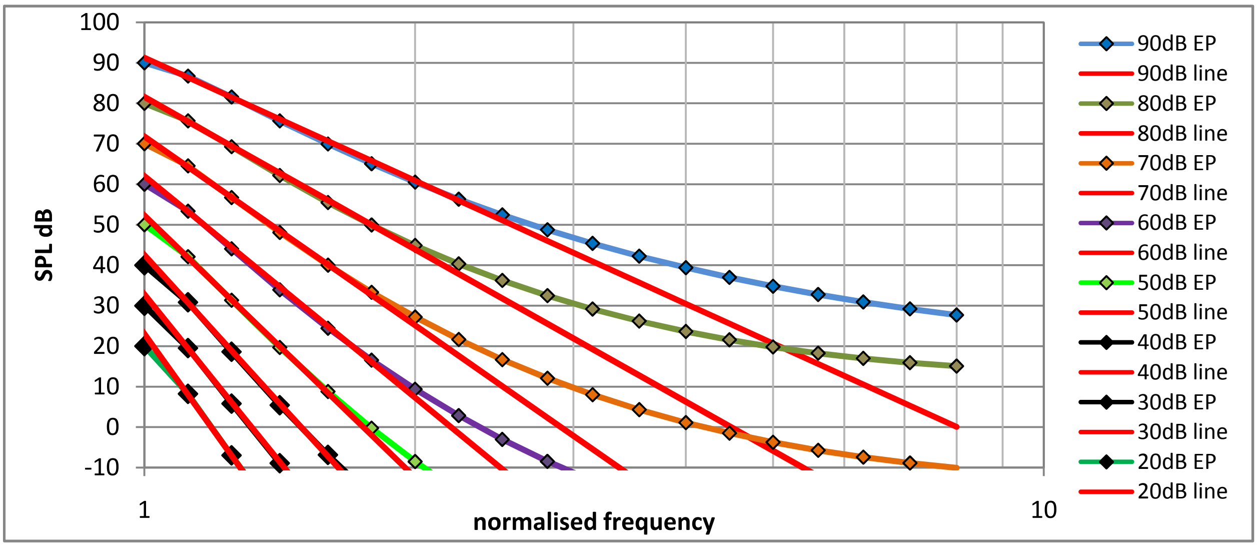

Figure 18 Excitation patterns for a 1 kHz tone presented at different levels and associated lines matched to the first part of each pattern. Note that the frequency scale is normalised to the input frequency.

From the 49 equations (7 frequencies with 7 levels), the equation parameters were interpolated at 1/3rd octave intervals between the octave-based filters and for each level between the 10 dB steps. These equations were then used to compute the masking produced by speech frequency *i* at each 1/3rd octave interval above frequency *i*.

As the EP method proposed by Moore includes the acoustic filtering produced by the outer and middle ear, it is instructive to examine the differences in the masking resulting from this filtering. Figure 19 shows an illustration of masking levels above 500 Hz calculated in 1/3rd octave bands (using the slope method) from speech in the bands 500 Hz and above with and without ear-filtering.

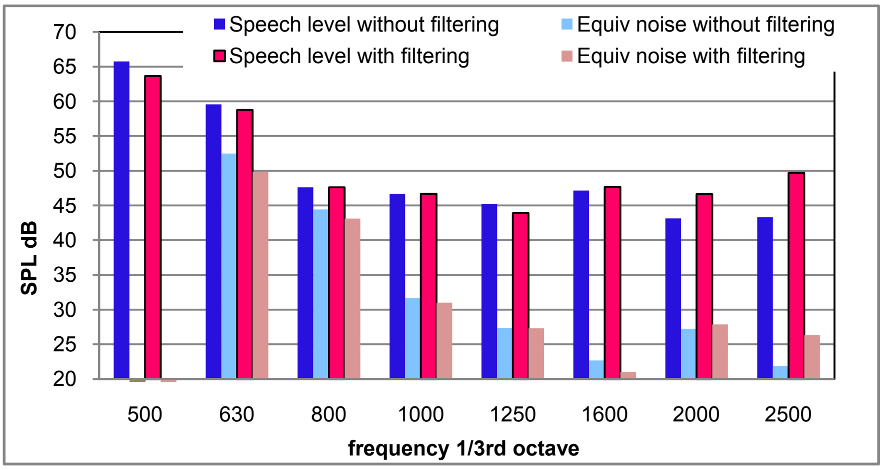

Figure 19 Conceptual illustration of equivalent masking noise predicted by the EP slope model in the 1/3rd octave bands above 500 Hz due to speech in the bands 500 Hz and above. Data is presented with and without the filtering of the outer/middle ear. Data for Talker 1 reverberant, 50 ms spectra time-slice 71.

A comparison was made of the masking levels computed by the EP slope method with those of the SII model. Figure 20 compares the masking due to speech in the 500 Hz and 1 kHz bands associated with a total L_{eq} level of 75 dBA and the IEC spectrum, while Figure 21 compares the masking at a total L_{eq} level of 95 dBA.

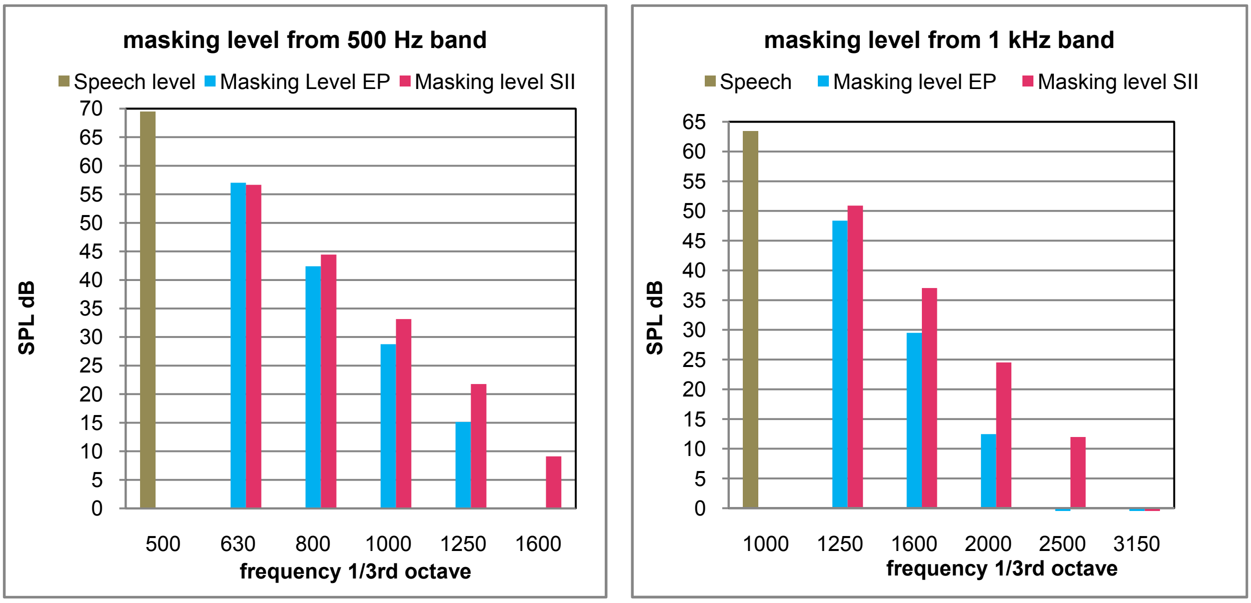

Figure 20 Comparison of EP (slope method without outer/middle ear filtering) and SII masking levels with the component in the 500 Hz or 1 kHz 1/3rd octave bands only of the IEC spectrum with a total level of 75 dBA. Note the change of vertical scale.

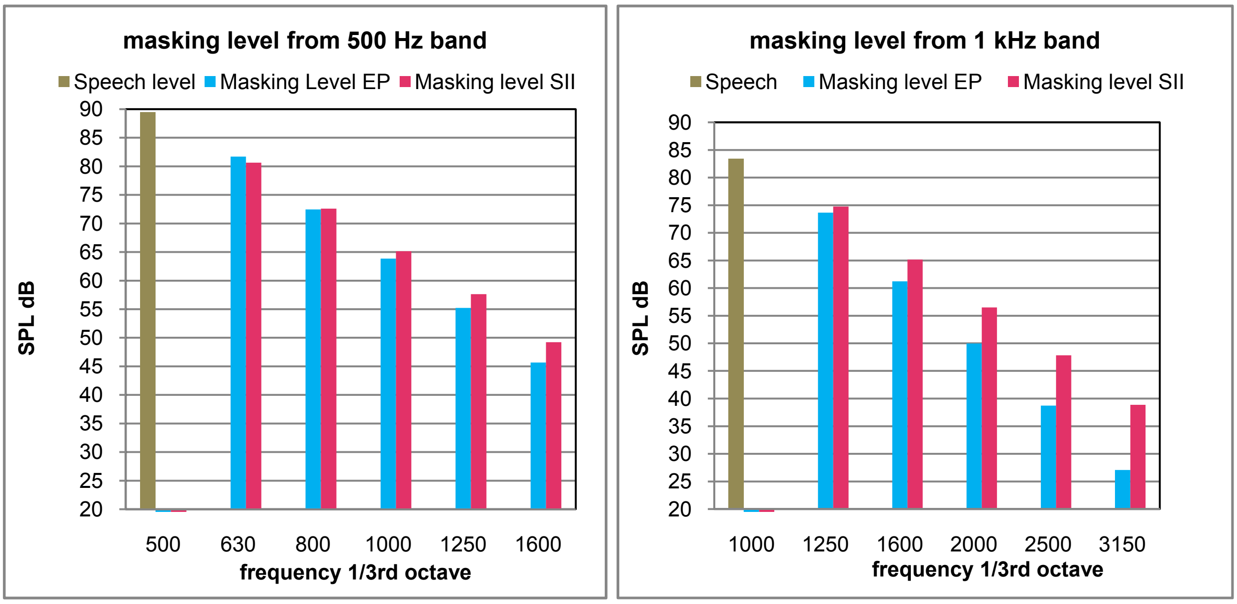

Figure 21 Comparison of EP (slope method without outer/middle ear filtering) and SII masking levels with speech in the 500 Hz or 1 kHz bands only. The total long term Leq of the speech containing the selected 1/3rd octave segment was 95 dBA. Note the change of vertical scale.

When the 1/3rd octave equivalent masking levels for EP and SII methods are bundled into octave bands and compared with those computed by STI masking, the data in Figure 22 results.

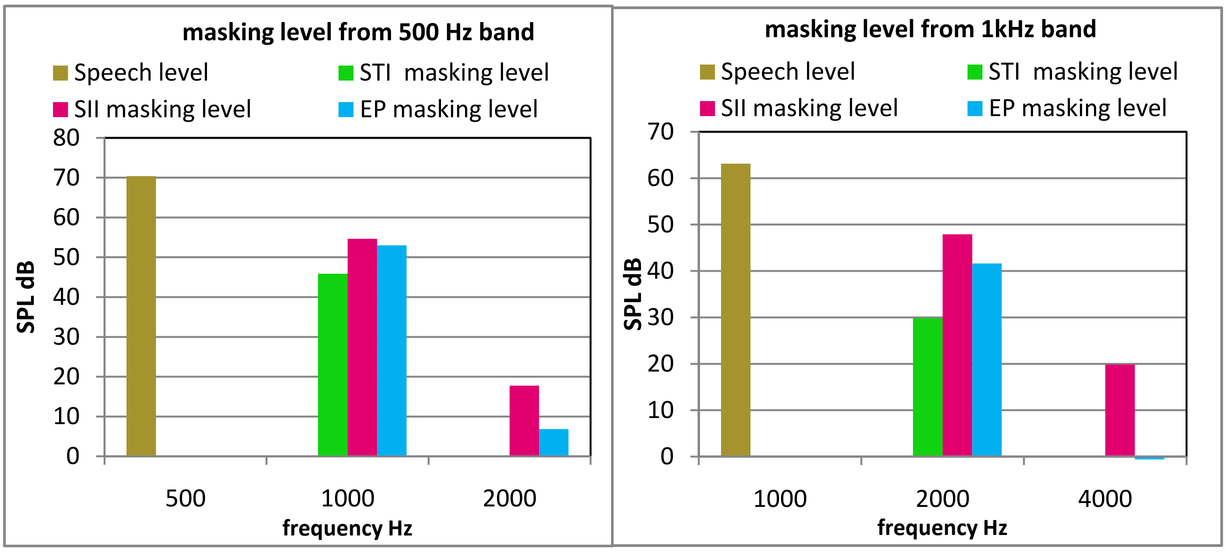

Figure 22 Comparison of masking levels computed by the STI, SII and EP slope methods with speech in the 500 Hz or 1 kHz octave bands only. Note that filtering by the outer/middle ear is not used for the EP calculations.

5 COMPUTATION OF STI WITH THREE MASKING MODELS

The STI values were computed with the three different masking methods and the range of filters and speech spectra. The following six masking models were used:

1. STI
2. SII
3. EP using 0.1ERB resolution with 1/3rd octave spectral lines with ear filtering
4. EP using 0.1ERB resolution with 1/3rd octave spectral lines without ear filtering
5. EP slope method with 1/3rd octave spectral lines with ear filtering
6. EP slope method with 1/3rd octave spectral lines without ear filtering

The computation steps listed below were followed.

5.1 Adjustments to the measured MTF Matrix

As i) the MLSSA analyser was used to measure the MTF matrices in (2) had applied masking to those matrices, and ii) some SNRs were less than 30 dB, the MTF matrices were de-noised and then de-masked, by applying the inverse of the specified masking adjustments.

5.2 Preparation of Spectra for STI Calculations

The 1/3rd octave speech spectra were prepared for insertion into the STI calculations as follows:

a) The total rms level of the ten one-second slices was computed for each talker to form the long-term L_{eq} in each 1/3rd octave band.

b) The long term L_{eq} in each 1/3rd octave band was then adjusted by the frequency responses of the eight filter shapes.

c) The filtered long-term L_{eq} levels were A weighted and summed to give the long term LA_{eq} of each talker and normalised to the nominated long-term operational speech level of 75 dBA. The resulting normalization factor D was stored for later use.

d) Each of the 1 second, 250 ms and 50 ms time-slice spectra was then adjusted by the response of the eight filter shapes and the normalization factor D. The data from this stage is termed the "processed time-slice spectra".

e) All spectra with total levels less than 45 dBA were discarded.

f) All adjusted spectra were logarithmically summed into octave bands for inputting into the STI algorithm as the Speech Signal.

5.3 Inclusion of background noise

Noting Steinbrecher's (23) concerns, a realistic amount of background noise was introduced into the calculations of STI. A noise spectrum corresponding to NR20 was used to ensure that under operational situations where background noise is almost universally present, the reduction in signal to background noise ratio due to a depressed frequency response was accounted for.

5.4 Computing STI using STI masking

The STI was calculated using the STI masking model for each processed time-slice spectra and talker.

5.5 Computing STI using SII masking

The spectrum level of each processed time-slice spectra was computed, and the total masking levels were computed in 1/3rd octave bands using Eq 4 and the summed term in Eq 5.

The masking spectrum levels were converted back to 1/3rd octave speech and masking levels which were summed into octave bands to yield the octave band SNR's. These SNRs were subtracted from the octave band levels of each processed time-slice spectra to yield the masking noise in octave bands. Those noise levels were converted to intensity $I_{am,k}$.and using Eq 3 to insert the masking noise, the STI was calculated for each processed time-slice spectrum and talker.

5.6 Step 3 Computing STI using Excitation Pattern masking

The speech and masking noise levels in 1/3rd octave bands were found for each processed time-slice spectra using EP Methods 2 and 3.

The process produced output 1/3rd octave speech and masking levels that were summed into octave bands to yield the octave band SNR's. These SNRs were subtracted from the octave band levels of each processed time-slice spectra to yield the masking noise in octave bands. Those noise levels were converted to intensity $I_{am,k}$.and using Eq 3 to insert the masking noise, the STI was calculated for each processed time-slice spectrum and talker.

6 RESULTS

6.1 STIs with IEC Speech Spectrum

Figure 23 compares the STIs of the six methods for the IEC speech spectrum given in (6). The following trends are observed:

a) The differences in the STI values approximately range from 0.02 to 0.1.

b) SII masking method yields the lowest STI values.

c) Of the four EP methods, the slope method without ear-filtering yields the lowest STI values.

d) The EP 0.1ERB method with ear filtering yields STI values that are similar or exceed the STI method.

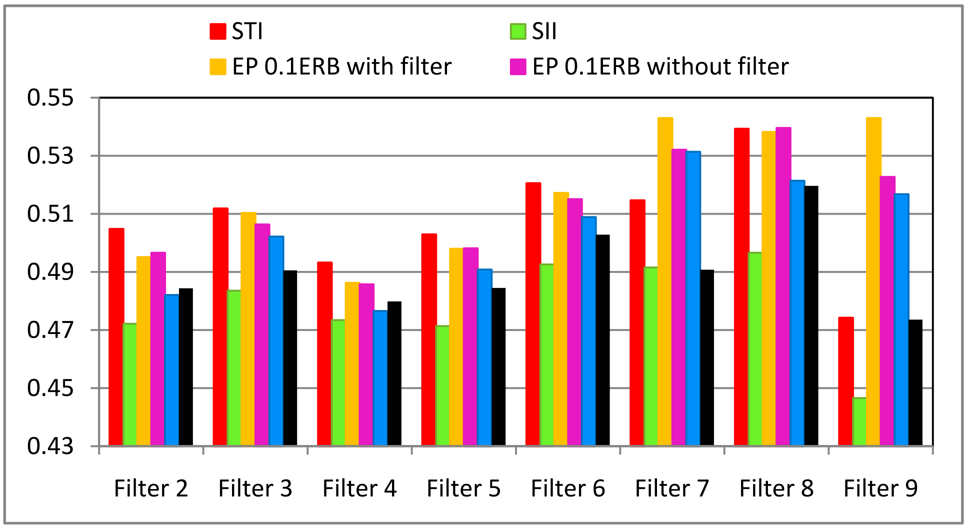

Figure 23 Comparison of STIs predicted by the six masking models with the IEC speech spectrum.

The octave band MTI values were examined for the eight filter shapes to help understand the contribution of each octave band to the overall STI value. Comparisons of the MTIs for Filters 7 and 9 are given in Figure 24 which shows some of the extremes of the overall behaviour.

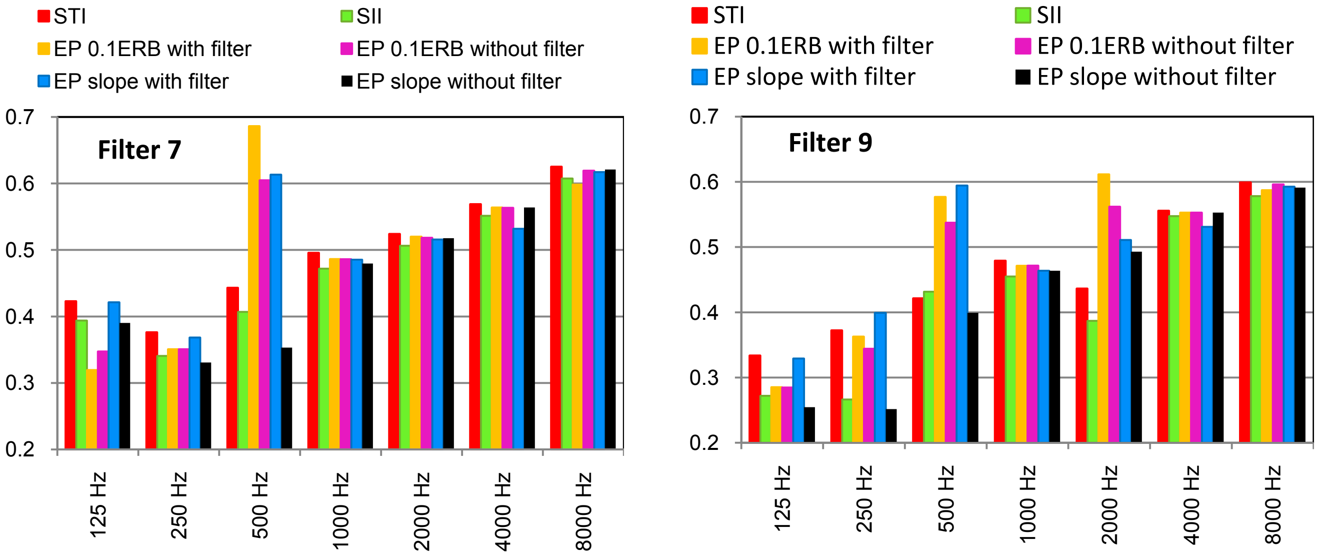

Figure 24 Octave band MTI values of two filter shapes for six masking methods with IEC spectrum

6.2 STIs with spectra of talkers

6.2.1 Histograms of STI values for Talker 1

Using the range of spectra obtained for Talker 1 reverberated, the STI values for each filter and masking model were examined for their distribution of STI values. Twenty bin-ranges were formed between the maximum and minimum values of STI for each masking model. Figure 25 shows histograms of the STI values in bin sizes equal to (max-min)/20.

The following trends are observed:

a) The bulk of the STI values taken over the filters and short-term spectra lie in a remarkably narrow range, varying mostly by only 0.03, with filter 8 showing a range of 0.05.

b) Although the bulk of the SII values can be up to 0.1 lower than the STI, the majority are within 0.05 of the STI. This is just larger than the generally accepted JND of 0.03 STI.

c) The shape of the distribution of the STI and EP in 0.1ERB intervals is generally narrower than with SII or EP slope models.

d) The shape of the distribution of the SII and EP slope methods has some similarity.

e) Although the shape of the distributions of EP slope method with and without filtering is generally similar, they often differ significantly in value.

6.2.2 Mean STIs for all talkers

The mean STI value for each talker, situation, and filter was computed for the six masking methods with all the short-term spectra. Comparisons of the mean STI values for the anechoic and reverberated spectra are shown in Figure 26 and Figure 27 respectively.

The following trends are observed in the anechoic speech data:

a) The STI values with SII masking are universally the lowest and are typically 0.02 to 0.04 below those with STI masking.

b) Depending on the filter shape, the highest STI values occur with either STI masking or EP with 0.1ERB masking.

c) The STI values with SII masking are generally 0.1 to 0.2 below those with the EP slope method.

d) The STI values with EP slope with and without ear filtering do not show a consistent trend. Whether or not the EP slope with ear-filtering is greater than without ear-filtering depends on the filter shape.

The following trends are observed in the reverberant speech data:

e) The STI values with SII masking are universally the lowest and are typically 0.02 to 0.06 below those with STI masking.

f) Depending on the filter number, the highest STI values occur with either STI masking or EP with 0.1ERB masking.

g) The STI values with SII masking are generally 0.1 to 0.4 below those with the EP slope method.

h) The STI values with EP slope with and without ear filtering do not show a consistent trend. Whether or not the EP slope with ear-filtering is greater than without ear-filtering depends on the filter shape.

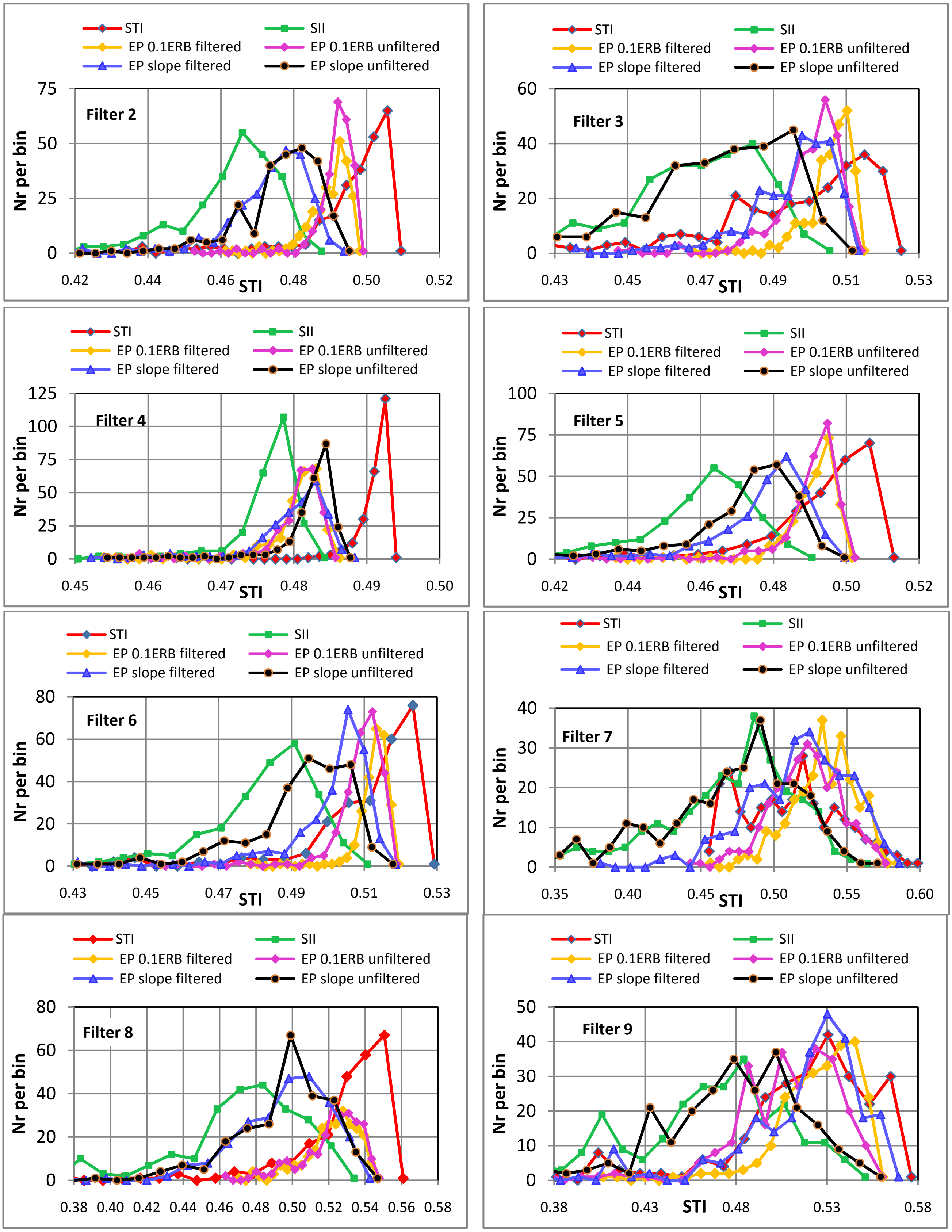

Figure 25 Histogram of STI values for six masking methods for the nine filters with Talker 1 reverberated and all time slices. Note the differences in scales between graphs.

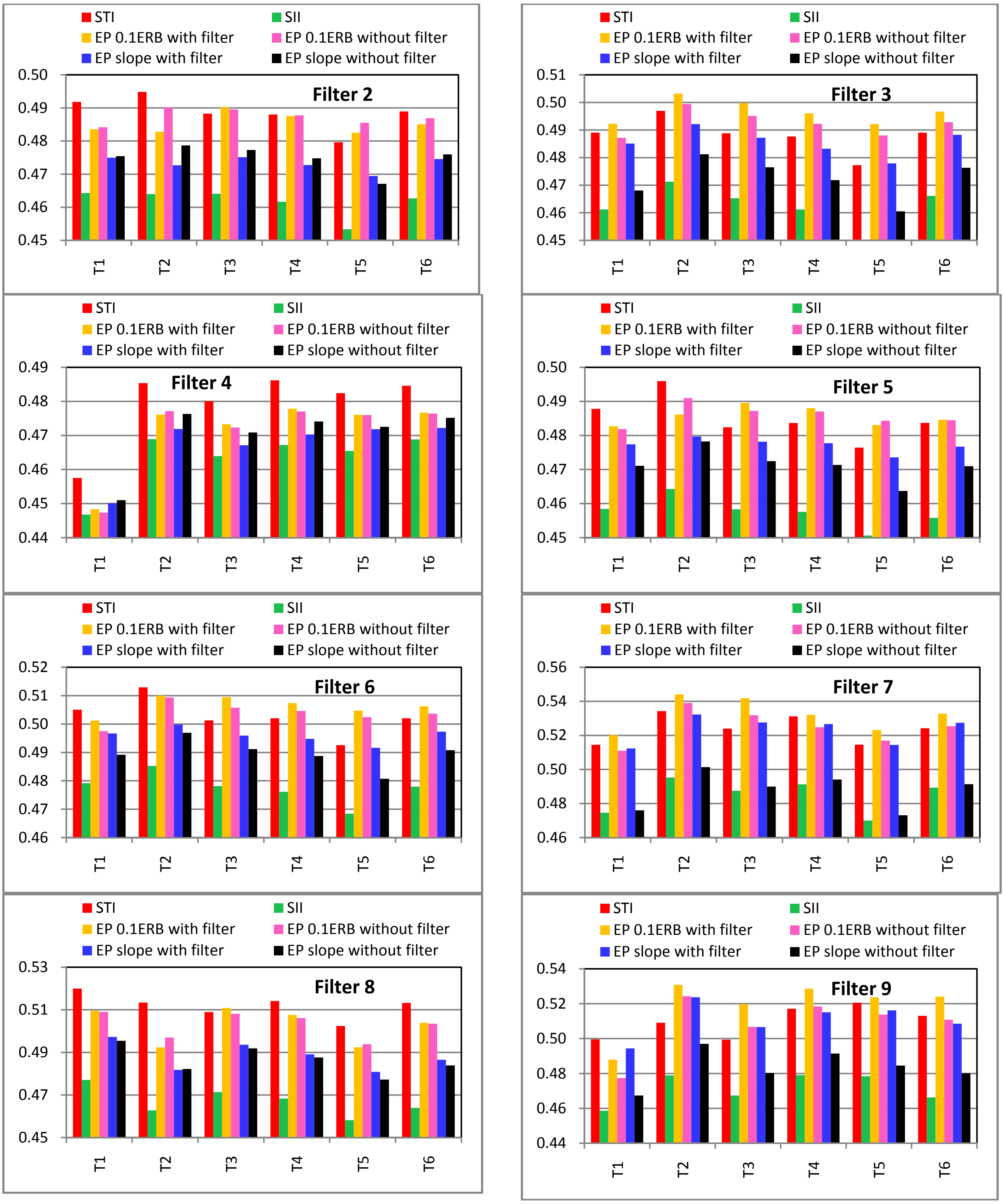

Figure 26 Mean values of STI with the STI, SII and EP slope masking methods. Data is for anechoic speech and Tn indicates Talker n.

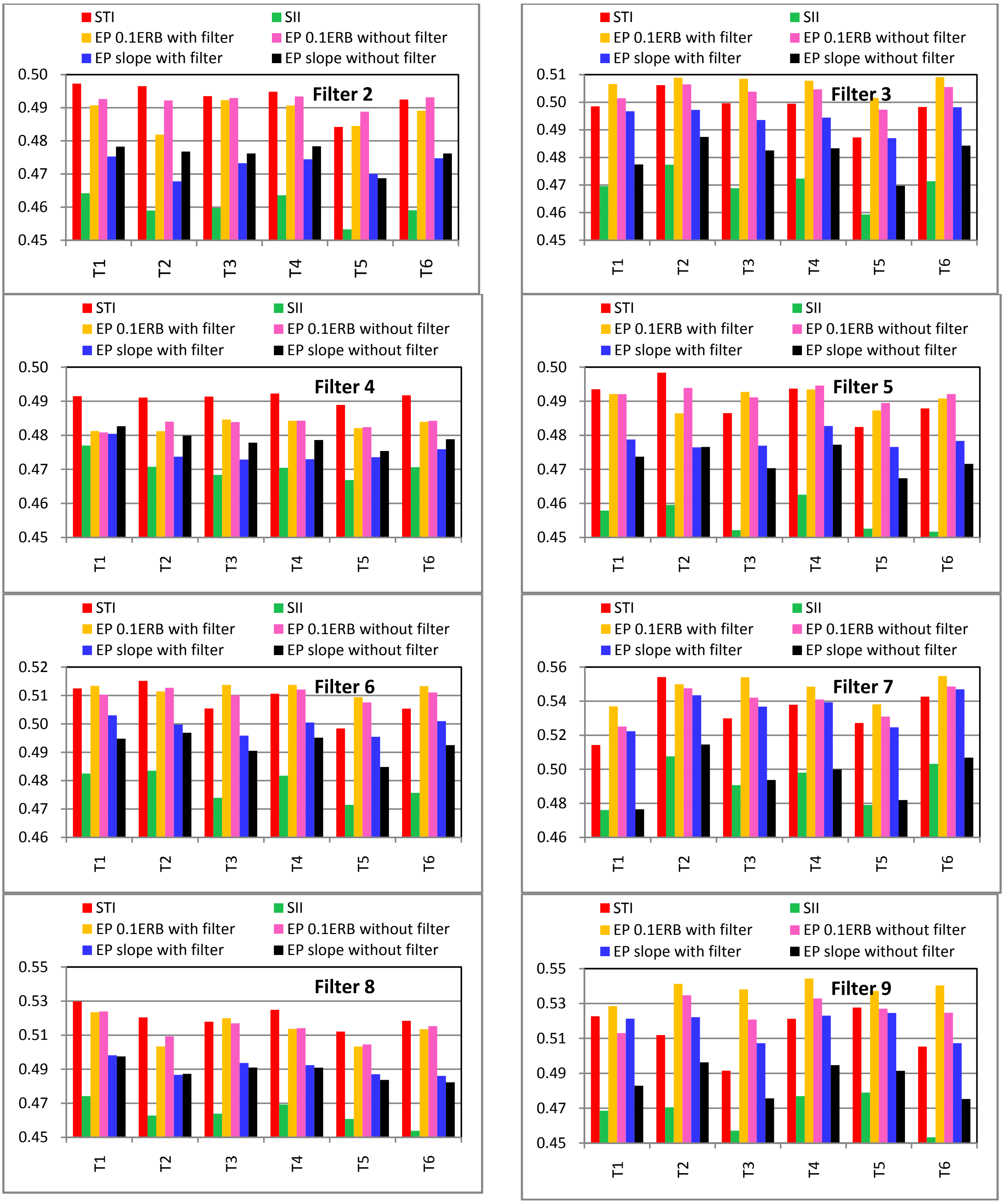

Figure 27 Mean values of STI with the STI, SII and EP slope masking methods. Data is for reverberated speech and Tn indicates Talker n.

7 SUMMARY AND CONCLUSIONS

This paper has investigated the extent of changes to the values of the STI resulting from modifications to two key parameters of that metric. These two aspects are i) the spectrum of speech and ii) the model of the ear's upward masking mechanism. These two parameters are used to predict the level of equivalent noise resulting from the self-masking of speech. .

7.1 Speech Spectra

The standard STI methodology uses a specific long-term spectrum of speech. To investigate changes to the STI values resulting from short-term speech spectra, short-term spectra of six talkers were found using time intervals of 1 s, 250 ms and 50 ms with both anechoic and reverberated speech.

Analysis of these spectra showed variations of up to +12 and -40 dB relative to the IEC spectrum. Compared to the IEC spectrum, the average spectrum of the six talkers in the anechoic environment shows approximately 10 dB less energy at low frequencies and 8 dB more energy at high frequencies. With reverberated speech, the average spectrum has approximately 6 dB less energy at low frequencies and 5 dB more energy at high frequencies than the IEC spectrum.

7.2 STI Values with six Masking Methods

The effects of five alternative methods of psychoacoustic masking on STI values were calculated for a large range of speech spectra and compared to the STI values obtained with the specified STI masking method. Table 7.1 summarises the six masking methods.

The basis for the calculation was a measured MTF matrix for which the STI value was 0.5. The STI values were computed for six talkers, each with eight filter shapes. The filter shapes had severe frequency response aberrations, and were intended to reflect the response of a sound system that has an extremely poor frequency response.

The long-term L_{eq} level of each talker with the applied filter shape was normalised to 75 dBA, and the resulting short-term spectra computed with this normalisation. A background noise level of NR20 (approximately 33 dBA) was also applied to the STI calculations.

Method	Type and source	Comment	Ear filtering	Assumes speech spectral lines at	Calculation interval
1	STI Specified in IEC standard 60268-16	Uses a defined equation to predict masking	no	octave intervals	octave
2	SII Specified in ANSI S3.5-1997	Uses defined equations to predict masking. Important modification was made; the specified attenuation of 24 dB for the speech level was not used.	no	1/3rd octave intervals, which are ultimately integrated into octave bands.	1/3rd octave
3	Difference of two excitation patterns in the inner ear. Moore, Glasberg et al	Computes difference between EP with only one band and the EP with all bands other than that band.	yes	1/3rd octave intervals, ultimately integrated into octave bands.	0.1xERB
4			no		
5	Slopes derived from excitation pattern responses	Uses equations that we developed to predict masking.	yes	1/3rd octave intervals, ultimately integrated into octave bands.	1/3rd octave
6			no		

Table 7.1: Description of the six masking models

All these masking models are based on masking with stationary signals, and do not consider temporal masking mechanisms. Only Methods 3 and 4 take the ear's downward masking into account, with the STI and slope methods not considering this mechanism at all. However, the EP slope method that we developed could be extended to include downward masking if deemed appropriate.

7.3 Primary Findings and Conclusions

Our principal finding is that when the STI values are calculated with the six masking models using the range of short-term spectra and filter shapes, the resulting values do not differ significantly from the value obtained with the STI masking method and the long-term IEC speech spectrum. None of the calculated STI values using the range of spectra and masking models approached the equivalent STI value associated with the subjective word-score for each filter shape.

Our principal conclusions follow from this result:

a) When incorporated into the STI method using octave bands, the six steady-state masking models do not produce STI values that satisfactorily reflect the subjective reduction in intelligibility that occurs in practice with poor spectral balance.

b) A different masking model that also includes the temporal effects of pre and post masking is required if STI is to satisfactorily reflect the subjective experience of listeners under conditions of poor spectral balance.

c) As masking can occur in bandwidths that are narrower than an octave, we conclude that the concept of octave bands used in STI may be contributing to this result.

d) The range of measured spectra found in this study suggests that the standardised IEC spectrum may not necessarily the reflect the speech spectrum of a given individual talker. Further study is required to review typical, contemporary speech spectra.

   Care should therefore be exercised when using the standard spectrum – as it may differ substantially from that actually relevant to a particular system or circumstance. This is particularly the case with pre-recorded messages whose spectral or dynamic properties may have been deliberately modified in order to enhance the perceived intelligibility.

7.4 Other Findings and Conclusions

a) The STI values for the different filter shapes exhibit a variation of 0.02 to 0.1 over the range of spectra. Noting that a change of 0.1 STI is a significant change in intelligibility, this range of values suggests that using a long-term speech spectrum for STI may not be appropriate.

b) Of the masking models examined in this thesis, the excitation pattern method gives the most detailed calculation of masking noise. Both upward and downward masking is considered and speech signals can be considered using their actual spectral lines.

c) However, if the speech is lumped into specific bandwidths as per the (SII and STI methods), the ratios of speech to masking noise (SNR) obtained from the excitation patterns are sensitive to the spacing of the frequencies (spectral lines) into which the speech energy is lumped.

d) With speech lumped into spectral lines at 0.1 ERB intervals, the resulting SNRs are typically 5 dB or less. This suggests that the self-masking from normal everyday speech would be sufficient to substantially degrade intelligibility. Clearly, this is not the case, and therefore this interval is unsuitable for this process.

e) If excitation patterns are to be predicted with a frequency resolution of 0.1 ERB intervals, the analysis should be conducted using the actual spectral lines of a given talker, and not integrated as it is done with SII (1/3rd octave) and STI (octave) methods.

f) In a large number of cases, the unfiltered EP slope method shows similar STI histograms to the SII masking method; although the SII method generally produces lower masking SNRs.

g) Processes such as those discussed by Goldsworthy and Greenberg (24) incorporating temporal effects might be useful in narrowing the gap between subjective experience and the objective measure of STI.

8 APPENDIX

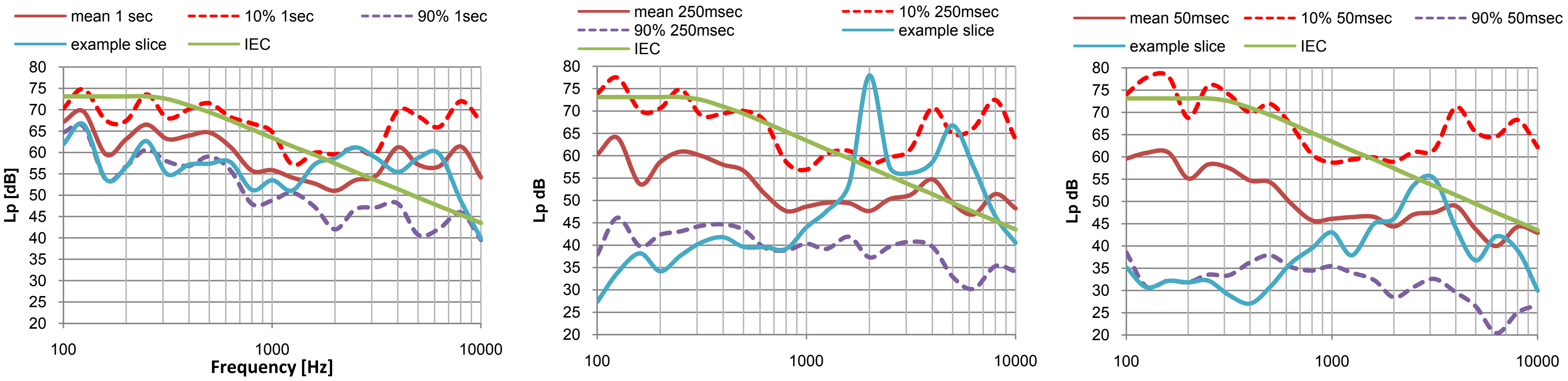

Figure A- 1 Time slice data for Talker 1 anechoic

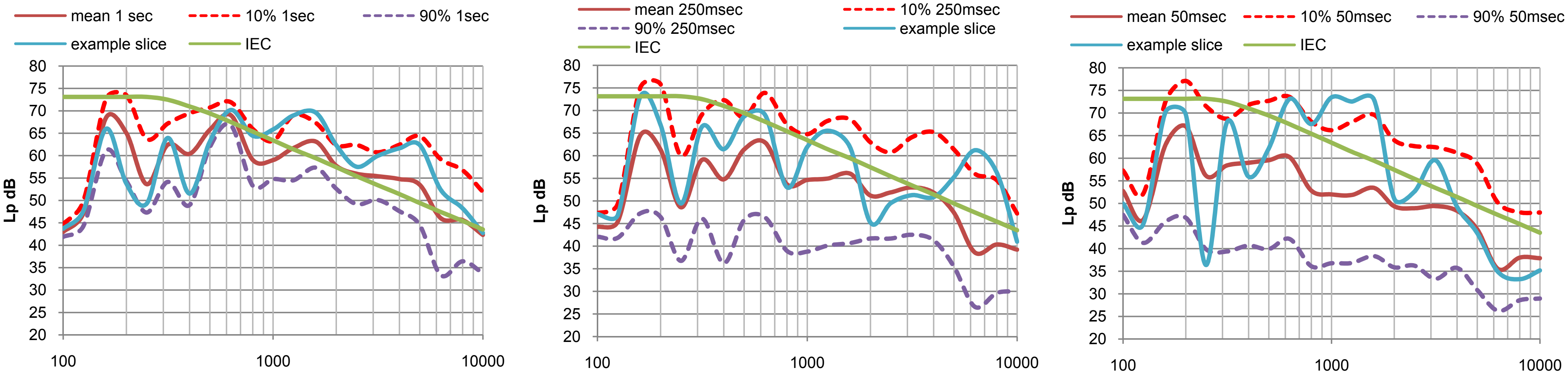

Figure A- 2 Time slice data for Talker 2 anechoic

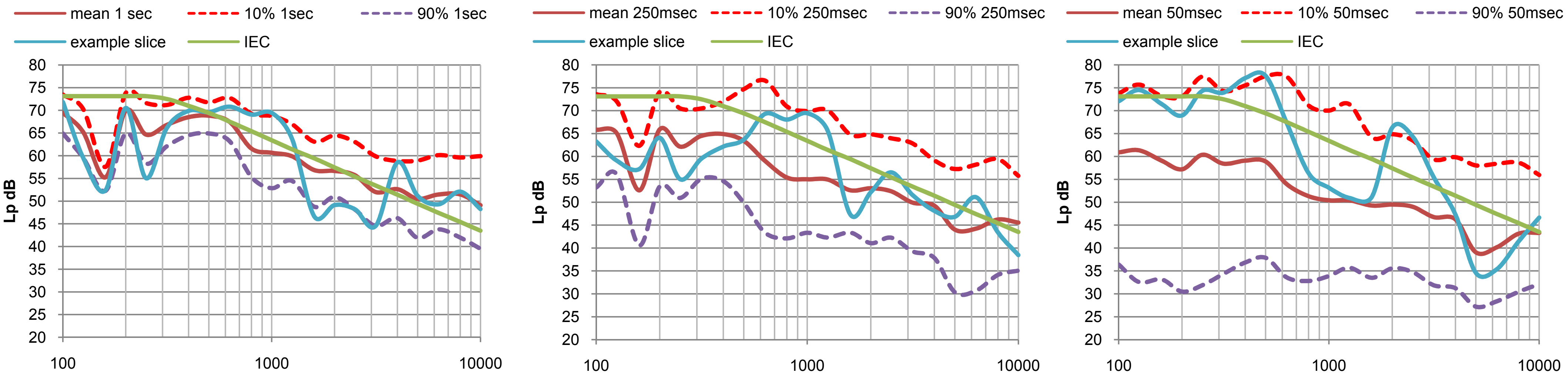

Figure A- 3 Time slice data for Talker 3 anechoic

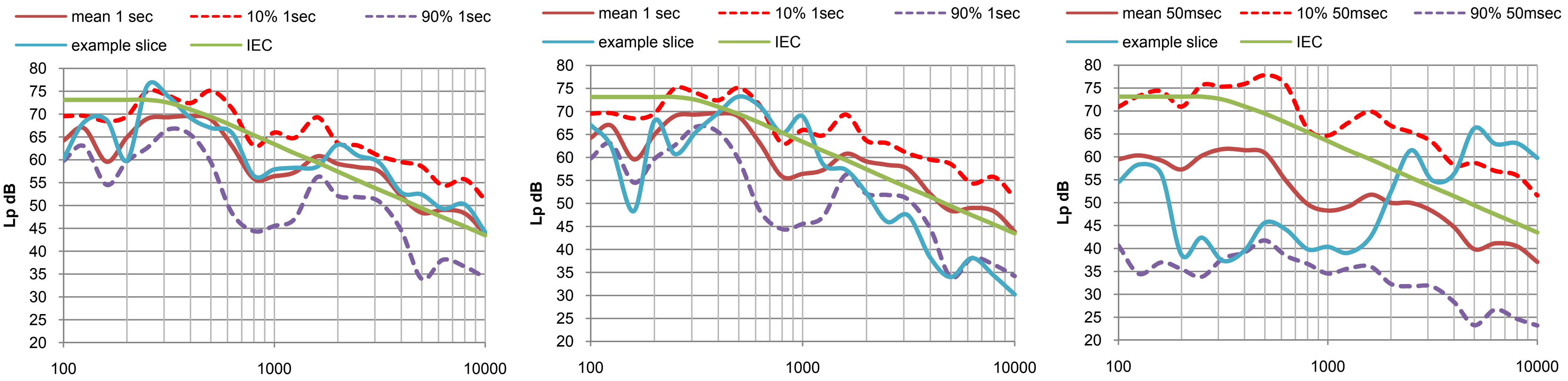

Figure A- 4 Time slice data for Talker 4 anechoic

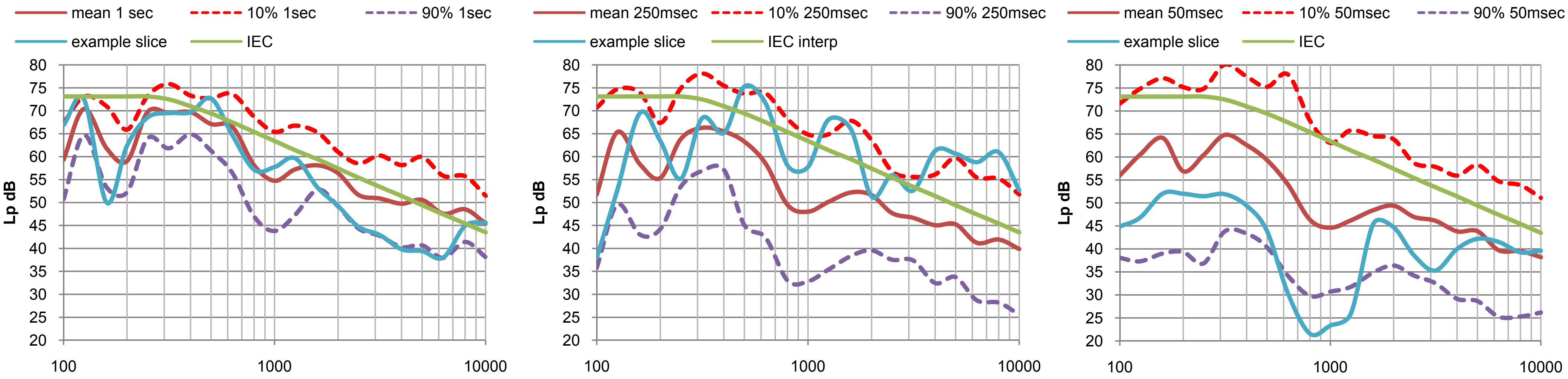

Figure A- 5 Time slice data for Talker 5 anechoic

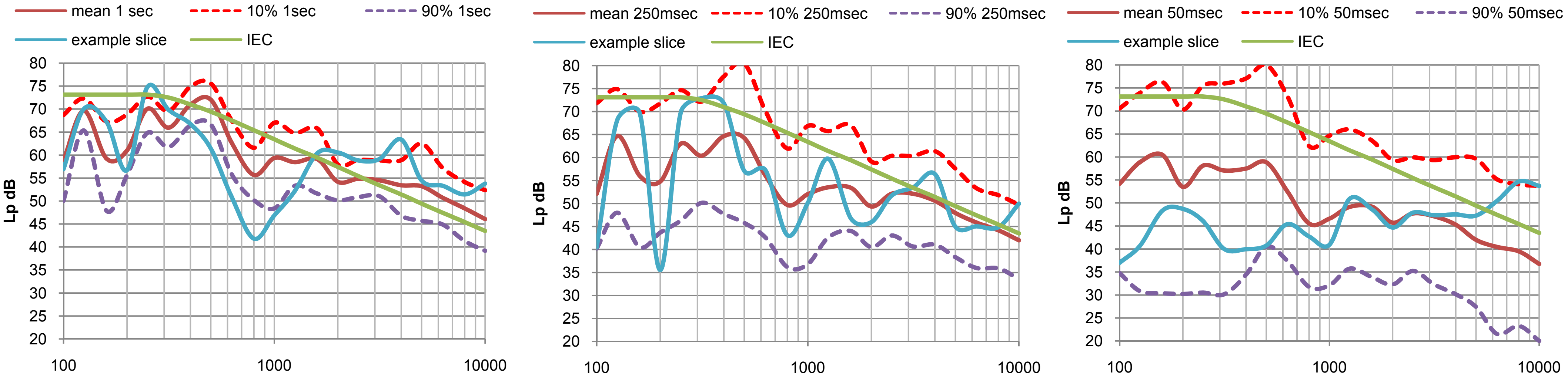

Figure A- 6 Time slice data for Talker 6 anechoic

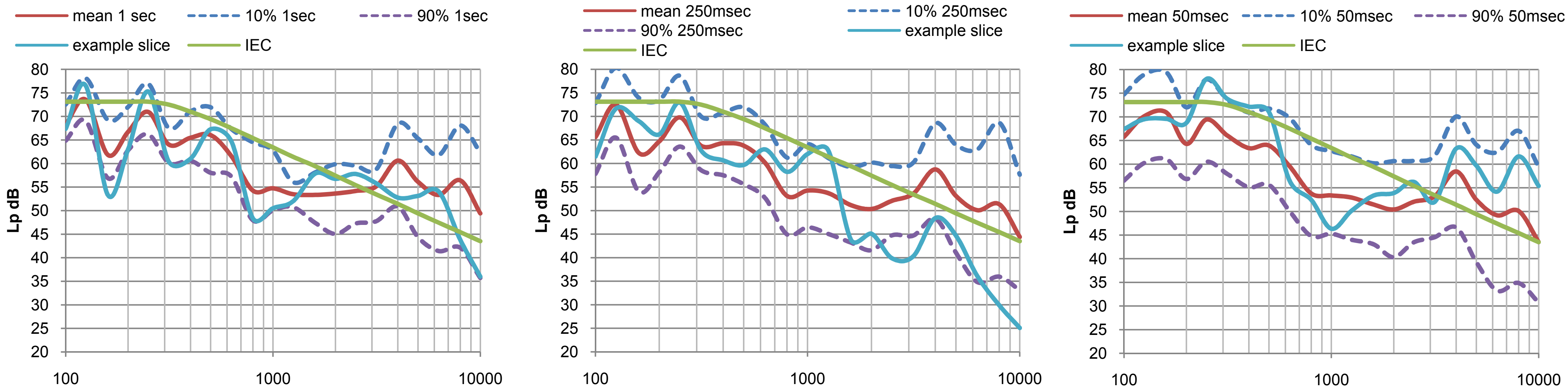

Figure A- 7 Time slice data for Talker 1 reverberated

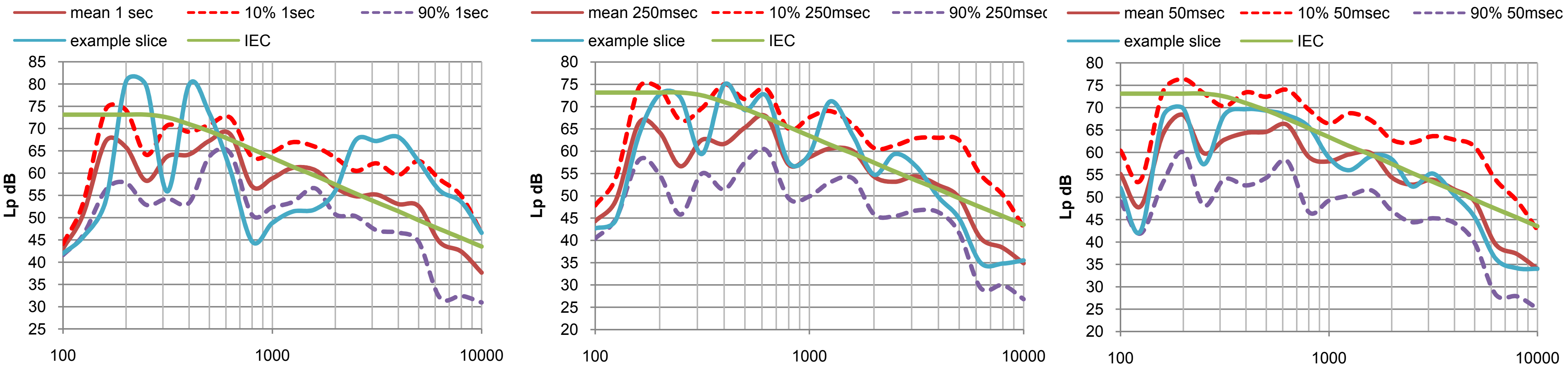

Figure A- 8 Time slice data for Talker 2 reverberated

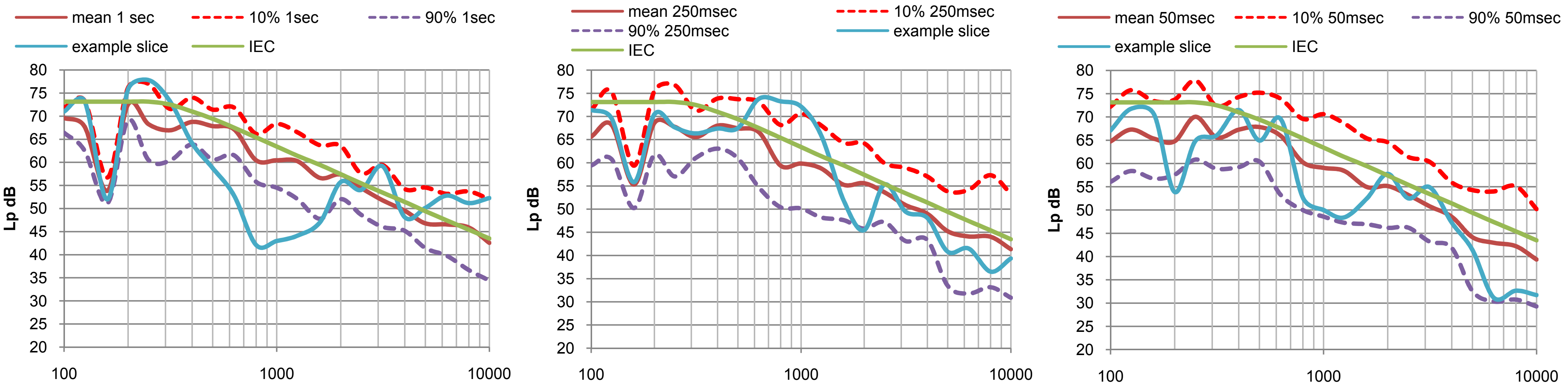

Figure A- 9 Time slice data for Talker 3 reverberated

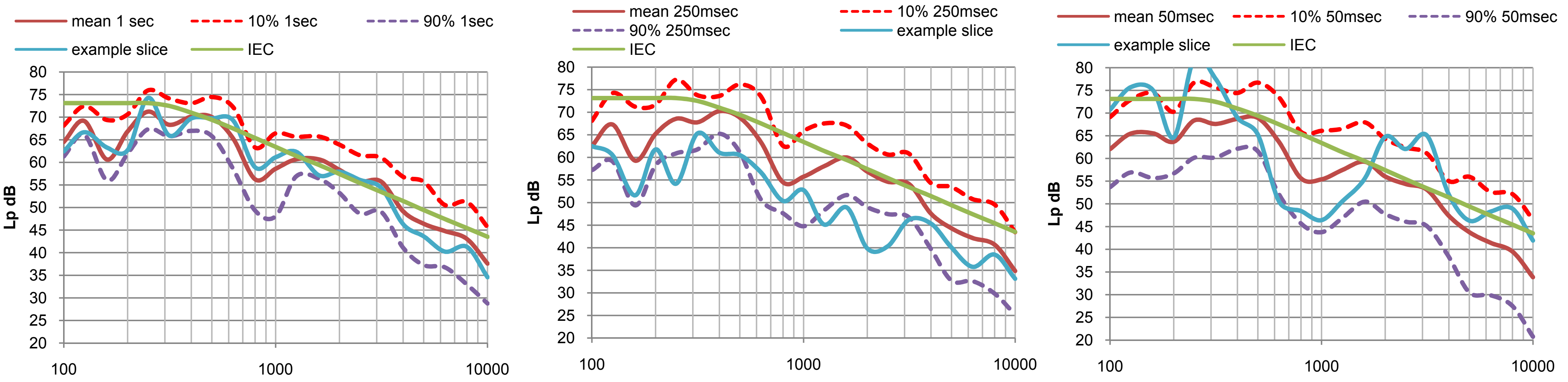

Figure A- 10 Time slice data for Talker 4 reverberated

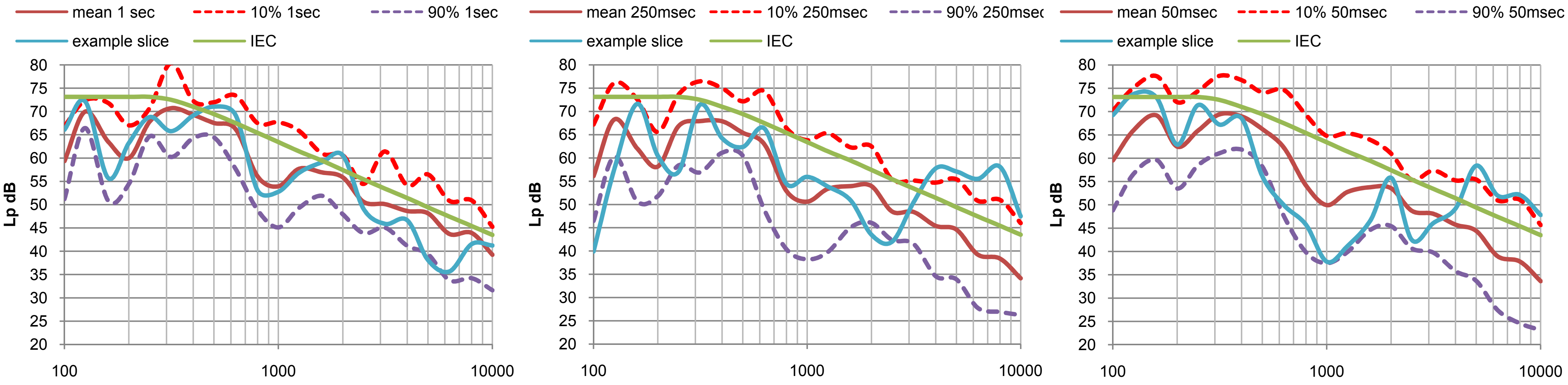

Figure A- 11 Time slice data for Talker 5 reverberated

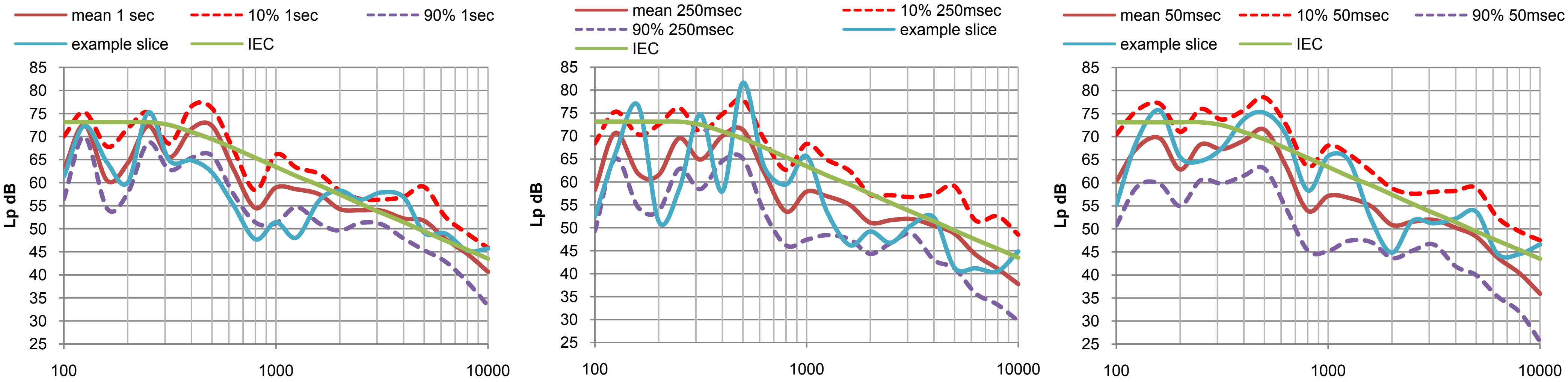

Figure A- 12 Time slice data for Talker 6 reverberated

9 REFERENCES

1. H. J. M. Steeneken, T. Houtgast. A physical method for measuring speech-transmission quality. Journal of the Acoustical Society of America. 1980, Vol. 67, 1, pp. 318-326.

2. Leembruggen, G.A, Stacy A. Should the Matrix be Reloaded? Proc IOA. 2003.

3. Mapp, P. Some Effects of Equalisation on Sound System Intelligibility and Measurement. Preprint AES 115th Convention . 2003.

4. Leembruggen, G. Is SII better than STI at recognising the effects of poor tonal balance on intelligibility? Proc IOA. 2006, Vol. 28, Part 6.

5. Wijngaarden, S.J., Steeneken, H.J.M. and Houtgast, T. Quantifying the intelligibility of speech in noise for non-native listeners. J. Acoust. Soc Am. 2002, Vols. 112 p 3004-3013.

6. IEC. Sound System Equipment Part 16: Objective rating of speech intelligibility by Speech Transmission Index. 2nd Edition 2003. International Standard No. 60268-16.

7. American National Standards Institute. Methods for calculation of the Speech Intelligibility Index. New York : s.n., 1997. ANSI S3.5-1997.

8. Moore, Brian C.J. and Glasberg, Brian R. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. Journal of the Acoustical Society of America. 1983, Vol. 74, 3, pp. 750-753.

9. Moore, Brian C. J. and Glasberg, Brian R. Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns. Hearing Research. 1987, Vol. 28, pp. 209-225.

10. Moore, Brian C. J., Glasberg, Brian R and Baer, Thomas. A Model for the Prediction of Thresholds, Loudness, and Partial Loudness. Journal of the Audio Engineering Society. 1997, Vol. 45, 4, pp. 224-239.

11. Glasberg, Brian R. and Moore, Brian C.J. Derivation of auditory filter shapes from notched-noise data. Hearing Research. 1990, Vol. 47, pp. 103-138.

12. Glasberg, Brian R.; Moore, Brian C.J. Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments. Journal of the Acoustical Society of America. 1986, Vol. 79, pp. 1020-1033.

13. Moore, Brian C. J. An Introduction to the Psychology of Hearing. 5th Edition. Bingley : Emerald Group, 2008.

14. Palvovic, Chaslav v. Derivation of primary parameters and procedures for use in speech intelligibility predictions. Journal of the Acoustical Society of America. 1987, Vol. 82, 2, pp. 413-422.

15. Ludvigsen, Carl. Relations among some psychoacoustic parameters in normal and cochlearly impaired listeners. Journal of the Acoustical Society of America. 1985, Vol. 78, 4, pp. 1271-1280.

16. Zwicker, Eberhard. Ueber die Lautheit von ungedrosselten und gedrosselten Schallen. Acustica. 1963, Vol. 13, pp. 194-211.

17. Zwicker, Eberhard and Fastl, Hugo. Psychoacoustics Facts and Models. 3rd Edition. Berlin : Springer, 2007.

18. French N.R and Steinberg J.C. Factors Governing the Intelligibiliity of Speech Sounds. Journal of the Acoustical Society of America. 1947, Vol. 19, 1, pp. 90-119.

19. Glasberg, Brian R.; Moore, Brian C.J. Prediction of absolute thresholds and equal loudness contours using a modifed loudness model (L). Journal of the Acoustical Society of America. 2006, Vol. 120, August 2006.

20. American National Standards Institute. Procedure for the Computation of Loudness of Steady Sounds. ANSI S3.4-2007.

21. Patterson, Roy D, et al. The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold. Journal of the Acoustical Society of America. 1982, Vol. 72, 6, pp. 1788-1803.

22. Wang, Ye, Vilermo Miikka. An Excitation Level Based Psychoacoustic Model for Audio Compression. Proceedings of the seventh ACM international conference on Multimedia. 1999, Vols. Pages: 401 - 404 .

23. Steinbrecher, T. Speech Transmission Index: Too weak in time and frequency? Proc IOA. 2008, Vol. 30, Part 6.

24. Goldsworthy, Ray and Greenberg, Julie. Analysis of speech-based transmission index methods with implications for non-linear operations. Journal of Acoustical Society of America. 2004, Vol. 116 pp 3679 to 3689, Dec 2004.