Audio Engineering Society

Convention Paper

Presented at the 114th Convention
2003 March 22–25 Amsterdam, The Netherlands

What You Specify Is What You Get (part 1)

Johan van der Werff,(johanvanderwerff@cs.com) [1], and Dick de Leeuw (mook@peutz.nl) [2]

[1] Peutz & Associes, PBox 66, 6585ZH Mook, The Netherlands, mook@peutz.nl.

[2] Peutz & Associes, PBox 66, 6585ZH Mook, The Netherlands, mook@peutz.nl.

ABSTRACT

The Peutz prediction algorithms for speech intelligibility as published in 1971 in the J.A.E.S. vol. 19 are still valid and are found remarkably accurate considering the simplicity. However, some revision is necessary for adaptation to the contemporary room simulation and sound system design programs. This paper will deal with the prediction of the Articulation Loss of consonants (AL_{cons}) based on data usually available in the design phase of a project. Special attention will be given on how to deal with multiple sources, nearer and further apart. For the attendees to the presentation of this paper there will be an Excel® spreadsheet available for a quick calculation according to the proposed method.

1. INTRODUCTION

To show how complex the speech intelligibility process is, the subsequent steps will be described below:

Speech intelligibility starts of course with the speech sounds of a speaking person, but...

- Some speakers articulate better than others, which means that speech sounds are better separated in time and frequency
- Speakers adapt to the environment:
  - They speak louder when there is more environmental noise (listen to someone with a headphone with music on)
  - They speak slower when in a room with a lot of reverberation

In the transporting medium (air) there can already be some competing sounds. Although the medium is completely linear at normal sound levels and can transport sound without mutual influence, the ear however cannot detect a softer sound in the presence of a louder sound with an overlapping spectral content.

- Environmental noise is not related to the speech and can be overcome by a louder speech level.
- Reverberant sound is caused by the speech sound itself and can only be overcome by reducing the reverberant sound level by a higher directivity of speaker and/or listener (cupping the hands around the mouth and behind the ears), by moving closer to each other or reducing reverberant sound level by acoustic absorption.

The ear can discriminate sounds in frequency and time at the same time, but in both domains the resolution is limited, even more so when the listener is older and/or suffers from some hearing damage. To complicate things further the resolution is also level dependant, the speech intelligibility decreases with speech levels above 80 db(A) (with equal signal to noise and direct to reverberant ratios).
Although speech is carried by speech sound, the information is carried by the modulation of the sound. In every frequency-slot at the size of the resolution, sufficient modulation space (in level) is necessary to carry the information.
Each of the smallest parts of speech, phonemes, have a more or less well defined set of frequencies which sound at the same short time. In order to decode these, it is necessary to separate a sufficient number of these modulations from the competing noises, otherwise the wrong phoneme may be decoded. Luckily most words have several phonemes, and only a limited set of phonemes make a valid word in the language of the speaker. The listener may be aware of the masking noises and can (unconsciously) guess which frequencies are not trust worthy and which other phonemes can be possible. The message is decoded:

- At word level (usually not consciously),
- At sentence level (not every set of words make a valid sentence),
- At message level (not every sentence makes sense in the context of the message)
- The eye can see the mouth and the facial expression of the speaker, which often helps much more than most people are aware of in decoding the message. Seeing a speaker clearly can make up for 5 dB worse signal to noise ratio. Some deaf people can decode the message completely based on visual clues only.

Perhaps even more parameters play a role. The above mentioned are obvious ones and are quantified in research, in the future we may discover other processes in our brains to decode speech.
All in all, if 10% or less of the phonemes are wrongly understood they are corrected unconsciously and not noticed by the listener. The speech intelligibility will subjectively be rated as good or very good. A lot of people (especially young and some with hearing loss) can function very well in daily life (even theater visits) with a loss rate of 20% of the phonemes. They will rate this subjectively as average or a little below average but not as bad. The subjective rating is however not always a good guideline. In an experiment [1] it has been found that at some place where the gap between direct sound and the reflections was a little big compared to the level of the reflections, the subjects found that they had to put in more work to understand the message and subsequently rated the speech intelligibility lower, but when the results where analyzed the objective rating was actually higher than on the “good” seats.

Philosophical observation:
Since so much guessing is going on in the decoding of the speech, it is not difficult to see that a lot of misunderstandings can take place because one is likely to understand what one is biased towards to hear and not what is really said and what would be understood when one is listening with an open mind.

1st Conclusion:
Any system that would measure speech intelligibility should incorporate all the above, the adaptation of the speaker and the decoding possibilities of the listener. The most obvious and complete is the use of live speakers and listeners. By electronic means only assessing speech intelligibility is almost impossible. What can be done is to find a set of data that can be assessed on the spot and find empirically the statistical relationship between speech intelligibility measured with test persons and the assessed data. All known methods of electronically measuring or calculating speech intelligibility are based on the statistical relation between one or more physical parameters and the real speech intelligibility. The user of these systems should be aware of its limitations. Parameters not incorporated in the measurement are a blind spot in the system. Data shall never be interpreted outside the boundaries that underlay the statistical relationship.

2. MEASURING SPEECH INTELLIGIBILITY

There should be a clear distinction between the measures of speech intelligibility and the indicators of speech intelligibility. Measures of speech intelligibility always incorporate the human ear and brains, indicators use electronic equipment to quantify speech intelligibility.

2.1. Measures

Measures of speech intelligibility are

Word tests. A meaningful word is embedded in a carrier sentence and written down by the listeners. The sentences are spoken in the room and over the sound system under test. When a sufficient number of sentences and listeners are used to be statistically significant, the correctly understood words are a measure for intelligibility.

Modified rhyme tests. As above but the listeners have a limited choice of meaningful words that rhyme.

Nonsense word tests. As with word tests, but now logatoms are used in the format Consonant-Vowel-Consonant (CVC). With Articulation Loss of Consonants (AL_{cons}) only the wrongly understood consonants are counted. Peutz found that for speech in rooms the vowels are much easier understood than consonants and hence the loss of consonants are the deciding factor in speech intelligibility. AL_{cons} is expressed in %. Under perfect conditions (speech direct on headphone) a combination of a very good speaker and a very good listener will have an AL_{cons} of 2.5%. In excellent room acoustical conditions they can have on top of that 5% AL_{cons} or less. An extra 5% loss is still considered as good and another 5% extra loss is still considered fair and sufficient for most messages. The initial 2.5% is considered the zero correction or proficiency factor. Which target to set for a certain situation depends on the proficiency (to be expected) of the talker and the listener. Excluding the zero correction factor a target of $\leq$ 10% AL_{cons} for $\geq$ 80% of the audience area and for the other part $\leq$ 15% AL_{cons} is usually sufficient for most purposes, even for non-standard messages and non-first language listeners.

The in our opinion best choice is AL_{cons} because:

- non-significant information (for speech intelligibility in rooms) is weeded out
- proficiency in the language is not measured
- there is almost no contamination by right guesses because even if the consonants are phonemically balanced (same occurrence of phonemes in the test words as in the language) there is no way to guess the right consonant because in a nonsense word the other phonemes have by definition no meaningful relationship in any language

The outcome of the tests depend strongly on the procedures used. For instance in an anechoic room with 0 dB signal to noise ratio the modified rhyme test finds 85% intelligibility and the Al_{cons} method 55% intelligibility (100% is perfect and equivalent with 0% AL_{cons}). See ISO TR 4870 for a discussion about speech intelligibility tests.

2.2. Indicators

Some widely known indicators of speech intelligibility are:

Articulation Index (AI) as described in ANSI S 3.5-1969. According to the AI the speech intelligibility is correlated with the weighted signal to noise ratio in 20 frequency bands. AI is blind for direct to reverberant ratio. On a scale from 0 to 1, 0.8 or higher means a good intelligibility and 0.2 or lower a good privacy. Since reverberant sound is not accounted for, it is a reasonable indicator for speech intelligibility in outdoor situations without significant reflections and a good indicator of privacy in all circumstances.
When a consultant specifies AI he actually specifies signal to noise ratio and hence audibility of speech, not intelligibility as such.

C50, C80, D50, etc. Are measures of 'clarity' or 'deutlichkeit'. They are a measure of early to late ratio in a certain frequency band. They are blind for noise and blind for the reverberation time, although the level of the reverberant sound is accounted for. They can equally well be used for predicting speech intelligibility although different numbers yield a 'good' speech intelligibility. The measures necessarily assume that the early part contains all the information and the late part none. This is only true in situations with a long reverberation time. From the Peutz equations we know that AL_{cons} will be less than $9*RT_{60}$. So in a room with a reverberation time of 1 second the AL_{cons} will be approx. 9% in the total absence of direct sound.
When a consultant specifies one of the 'C' or 'D' values he actually specifies direct to reverberant ratio and hence clarity of speech (or music) in a room with 'average' reverberation time in the absence of noise, assuming other octaves as the specified are 'in balance' with the specified ones, not intelligibility as such.

STI, STIPA, STITEL or **RASTI** are based on the assessment of the Modulation Transfer Function (MTF) of a room + sound system in the presence of background noise. The STI method is described in IEC 60268-16. The STI is a full function based on 98 data points (14 modulating frequencies in 7 octave bands). The others are subsets which reduce measurement time somewhat at the cost of losing the chance to catch odd behavior of a sound system or environment. STI is blind for narrow band coloration of the sound by the sound system. As long as the signal to noise ratio is not significantly compromised, odd equalizer settings will have no influence on the measured STI. This may be correct for 'dry' circumstances like telephone lines or outdoor but in a

concert hall even a slightly wrong equalizer setting or a inferior microphone can have a big influence on the speech intelligibility, which will not show up in measured STI. STI is also blind for level dependant distortions (i.e. bad connections in connectors) and enhancements by speech-processors. When a consultant specifies STI, he actually specifies MTF, which may have a good correlation with speech intelligibility under the right circumstances but it is not intelligibility as such.

2nd Conclusion:
It makes sense to specify a real speech intelligibility measure and accept STI, one of the subsets, or another indicator as an assessment tool. If a contractor has cut some corners in a blind spot of the assessment tool, it is always possible to demand a proper assessment with live listeners. This will always show, if conducted properly, the right intelligibility, even if effects are involved that are not known at that moment.

3. CALCULATING SPEECH INTELLIGIBILITY

Predicting speech intelligibility by calculation at the design phase of a sound system is much less complex, anyway after 1971. The designer can calculate the system performance just for one octave (i.e. 1 kHz) and make all other octaves 'in balance'. Peutz presented in an article in the J.A.E.S. a set equations to predict AL_{cons} from a few easily assessable acoustical parameters [2]. Up to a critical distance for intelligibility (D_{ci}), which is actually 3.16*D_c,

$$D_{ci} = 0.2\sqrt{\frac{QV}{RT_{60}}} \quad (m) \qquad (1)$$

$$AL_{cons} = \frac{200D^2RT_{60}^2}{QV} + a \quad (\%) \quad (2)$$

Where:
D is the distance to the source in meter
Q is the 1.4 kHz directivity of the source
RT_{60} is the 1.4 kHz reverberation time of the room in seconds
V is the volume of the room in m^3
a is the zero correction factor for a certain speaker-listener combination and lies between 1.5% (almost perfect) and 12.5% (still with normal hearing). It is the measure for proficiency of speaker and listener.

Beyond the D_{ci} the AL_{cons} is:

$$AL_{cons} = 9RT_{60} + a \quad (\%) \qquad (3)$$

Also a graph is given which shows the relationship between signal to noise ratio and AL_{cons}. A combination of this graph and graphical expression of equations 1-3 is shown in figure 1.

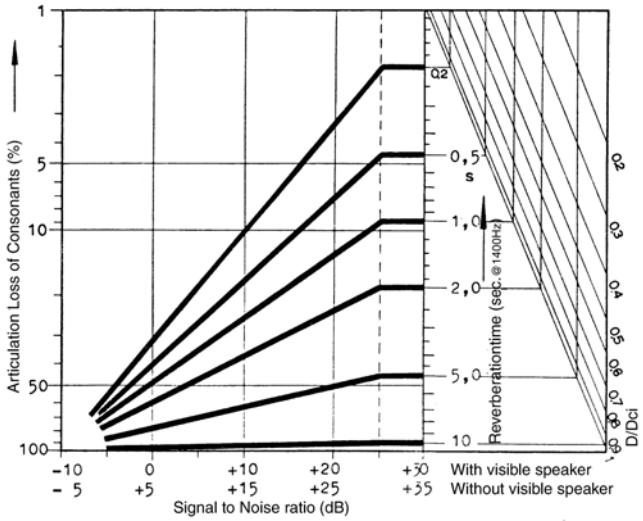

Figure 1: AL_{cons} in relation to signal to noise ratio, reverberation time, and distance relative to D_{ci}. At D/D_{ci} is 1, the direct to reverberant ratio is –10 dB, 0 dB direct to reverberant ratio (D_c) is at D/D_{ci} is 0.316. Use: Calculate D_{ci} with (1), use the average in the 1kHz and 2kHz octave bands for reverberation time and Q, find ratio between listener distance and D_{ci}, for distances larger than D_{ci} take 1, follow diagonally ratio line up to reverberation time, go horizontally to the left and follow signal to noise(PSIL) ratio line downwards until the expected signal to noise ratio is reached, go horizontally to the left and read the expected AL_{cons} (exclusive the zero correction factor a).

It is good to know which methods Peutz has used to establish the relationships, in order to find the boundaries of the valid area of the statistical model.
Peutz used:
A sound source with the directivity of the human head, in well behaving rooms without discrete reflections with:

- An almost flat reverberation time characteristic over frequency,
- Reverberation time was the average between the 1 and 2 kHz octaves,
- The average absorption coefficient apparent to the source is not given in his equations or

elsewhere in his work but it is reasonable to assume it is between 20% and 30%.

The sound levels involved can be reconstructed at:

Direct sound without significant early reflections (also due to the position of the source at approx. ear height) measured or calculated from the wide band or A weighted speech level (for natural speech they have almost the same value).

$$AL_{cons} = 100 * 10^{\left(\left[\frac{Lp_s - Lp_n + 10}{35}\right] * \left(\left[\frac{Lp_r - Lp_d}{10}\right] + \log_{10}(0.009 * RT_{60})\right)\right)} + a \ (\%) \qquad (4)$$

Reverberant sound which approximates a homogeneous isotropic sound field, calculated from direct sound level and the average of the 1kHz and 2kHz octave acoustical parameters.

Signal, sum of the wideband direct and reverberant sound.

Noise, babble and traffic noise evenly distributed over the room, level taken as the average of the 500 Hz, the 1kHz and the 2kHz octaves. This level is usually 5 to 6 dB lower than the dB(A) value

Since this statistical model for predicting speech intelligibility is based on:

Directivity of the speaker,

Distance to the speaker and

Room parameters like:

Reverberation time and

Volume,

the model assumes that:

- The frequency response of the sound system is fairly flat.
- The directivity of the source does not change drastically outside the 1kHz-2kHz octave bands
- The reverberation times in the octave bands are the same or slowly decreasing with increasing frequency.
- The noise has a wide spectral content without narrow peaks which may harm speech intelligibility more than is expected by the level alone.

These equations and graphs have proven to be very valuable since 1971, but also have their limitations. It is not easy to predict the speech intelligibility when loudspeakers of different makes and types, with all different driving powers are used in one room. But it is not difficult to convert the equations to levels of direct sound, reverberant sound and noise, which are relatively easy to calculate or measure, even when the systems are complex with different speakers types and drive levels. The graph and the equations however can be rewritten in a different equation (4) with parameters:

- Direct sound level (Lp_d),
- Reverberant sound level (Lp_r),
- Signal level (Lp_s), which is the combined level of the direct and the reverberant sound,
- Noise level (Lp_n),
- RT_{60} @1400 Hz.

Constraints:

Lp_s-$Lp_n \leq 25$ dB: if the signal to noise ratio is more than 25 dB, the speech intelligibility is not degraded by noise. If noise is not a significant issue the signal to noise ratio term between [] can be set to 1, the equation will be as (2) and will be described by the right hand side of the graph.

Lp_r-$Lp_d \leq 10$ dB: if direct and reverberant sound level differ more than 10 dB, the lower level is not influencing speech intelligibility noticeably. For worst case situations (no significant direct sound) the reverberant to direct term between [] can be set to 1, the equation will be described in the left hand side of the graph.

Whole equation: can of course never be more than 100 %

If both terms between [] are set to 1, the equation will be essentially as (3)

For use in situ, RT_{60}, Lp_s, Lp_r and Lp_n can be measured directly, Lp_d can easily be calculated from Lp_s and Lp_r. For calculation purposes the following set can be used: For calculating the reverberant sound level for each loudspeaker:

$$Lp_{rLsi} = Lp_{1W-1m} + 10\log_{10}\left(\frac{300 RT_{60} P_{el}(1-\alpha)}{QV}\right) \ (dB) \quad (5)$$

where:

Lp_{rLsi} is reverberant sound level caused by loudspeaker i

$Lp_{1w\text{-}1m}$ is sound level at 1 Watt input at 1 meter distance

P_{el} is input power in Watts

α is the absorption coefficient apparent to loudspeaker i (usually between 0.2 and 0.3), in (4) it is implicitly 0.2667, if this is taken, the set equations are ‘α neutral’ as in (2).

Q is the directionality factor of loudspeaker i

The combined reverberant sound level is the sum of the individual sound pressures.

$$Lp_r = 10\log_{10}\left(\sum_i 10^{\frac{Lp_{rLsi}}{10}}\right) \quad (dB) \quad (6)$$

For calculation with the aid of a computer program like EASE:
Calculate direct sound level in the 1 kHz band including interference by (mis)alignments in time, all loudspeakers at the right level. If not compensated for in the program, it will be wise to allow a reduction of the direct sound level from long line arrays by 1 to 2 dB, because:

- Acoustical centres of sources may not be as perfectly aligned as in the program,
- Sources will not have a perfect equality in phase and output level
- Air will not be perfectly still, with a perfectly even temperature and so….

Amplitude and phases of the output of the individual sources may not be exactly as the program assumes. If the alignment is less than perfect the sound pressure at the ear of the listener (where everything adds up) will be lower.

Calculate reverberant sound level from (5) and (6), using the same parameters as used in the program for calculating direct sound. When calculating arrays it is sufficient to know (or estimate) the parameters of one of the loudspeakers in the array and take the electrical power of all the loudspeakers together. For all of the other loudspeakers of a kind in the room the same applies. Take the parameters of the type and the power of all loudspeakers together of this type. Do this for each type with (5) and sum it all with (6). Using the 1kHz parameters instead of the 1400 Hz or the average of the 500 Hz, 1kHz and 2kHz bands will yield values that can be a little on the safe side when the reverberation time Characteristic is not flat and the noise spectrum is much different from the NC curves.

Although the above equations have proven to be reasonably accurate and likely to give the correct values, it must be remembered that it remains a prediction of AL_{cons}, statistically based on a few physical parameters.

4. CONCLUSIONS

When a sound system is properly engineered and the acoustics of the room behave as expected in the calculations, it is likely that the AL_{cons} value and hence the speech intelligibility realized in the room will be according to the values calculated in the design state. Deviations (measured with converted MTF values) of more than 2% are not necessary, and are certainly a reason for the author to see ‘what’s gone wrong’. However this is not the reason for choosing the title of this paper. It is to point out that it is not reasonable to expect more than is specified. If clarity is specified, you may get a direct to reverberant ratio in a certain octave only. If Articulation Index is specified, you may get signal to noise ratio only. If STI is specified you may get a weighted Modulation Transfer Function only. These parameters do not represent true speech intelligibility. Speech intelligibility is far more complex. If speech intelligibility is wanted it is better to specify Articulation Loss of Consonants. Of course every other measure or indicator that is easily measurable and suitable for the situation may be used as a provisional measure that can be replaced immediately when the ears tell a different story than the numbers.

5. REFERENCES

[1] Johan van der Werff, J.A.S.A. vol. 101, iss. 5, pp. 2401-3203, May 1997.

[2] V.M.A Peutz and W. Klein, J.A.E.S. vol. 19, nr. 11, pp 915-922, Dec. 1971.