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 2)

Johan van der Werff, (johanvanderwerff@cs.com) [1] and R.A.Metkemeijer, (zoetermeer@peutz.nl) [2]

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

[2] Peutz & Associes, P.Box 696, 2700 AR Zoetermeer, The Netherlands, zoetermeer@peutz.nl

ABSTRACT

The Peutz prediction algorithms for the Articulation Loss of consonants (AL_{cons}) as published in 1988 (85th convention in Los Angeles) did not seem to get the attention they deserved in the acoustical society. Perhaps this is due to the confusion it may have stirred because of the totally different set of algorithms compared to the 1971 set, or perhaps due to the more complicated calculations. But most likely how and where to get the physical quantities needed for input. This paper will deal with the underlying principles, how to extract the data from an impulse response and how to calculate the AL_{cons} from that. It is thought that this will be a valuable addition to the well known STI measurements. The data can be narrow band (one octave wide) and is in the gathering not sensitive for signal processors in the signal chain or for the type of filters used in the post processing of the data. For the attendees to the presentation of this paper there will be a computer program available which reads a set of measured or calculated impulse responses, extracts the data, calculates the AL_{cons} and presents the results.

1. INTRODUCTION

The chapters 1 and 2 of 'What You Specify Is What You Get, part1' are the perfect introduction to this paper. Pleas read that first, I wait until you finished.

It should be clear by now that speech intelligibility is far more complicated than signal to noise ratio, direct to reverberant ratio or even modulation transfer function. Measures based on these parameters come in very handy most of the time, but can never be the final answer. Even what comes below will not be very different in that respect, but it is a different approach which not only is reasonably accurate but also surprisingly flexible and expandable.

2. SPEECH INTELLIGIBILITY MODEL BASED ON INFORMATION THEORY

In [1] Peutz gave an extensive insight in the relations and equations regarding hearing and speech recognition. He concludes that the processing of information regarding speech is a stochastic process which has to be described in statistical terms: the chance that a person will understand the phoneme,

word or sentence right. Certain physical conditions have to be met before recognition is possible. Recognition is however not a conditional process, because speech recognition is a parallel process and not a serial process. This means that not all the speech cues are necessary for recognition, if sufficient information is passed, recognition is possible. So in fact not-recognition is a conditional process.
He finds that speech intelligibility is the product of two parameters:

- The remaining speech information available to the listener, quantifiable in an information index (i) and a
- Recognition measure (m) which combines the emitted speech cues by the speaker and the personal ability to decode the speech cues by the listener. It is the combined proficiency factor of speaker and listener.

In room acoustics vowels are much easier transferred than consonants. Consonants are defining speech intelligibility and therefore i_c as the mean information index for consonants and m_c as the mean recognition measure for consonants are the best choice in quantifying speech intelligibility.
The speech cues are information in the way it is defined in the information theory and can be treated as such. Recognition depends, except on the proficiency of the talker to articulate well, also on the number of words a listener has at his immediate disposal, if a word is not in that range, he has to look further and has less processing power and time left for decoding speech cues and therefore this is at the cost of information. Intelligibility is therefore influenced by the speech rate and speech pauses especially when addressed in a foreign language.
The calculation rules for information transfer are:

- $i_1 * i_2$ for information losses in a serial channel.
- $i_1 + i_2$ for information transferred trough independent channels, which do not carry the same information.
- $i_1 + i_2 - i_1 * i_2$ etc. for common information transferred trough independent channels

For every type of loss like masking by noise, frequency filtering, time distortions etc. a recognition function can be defined with the shape of an integrated Gausian probability distribution. Since this function is a little complicated to calculate Peutz defined an equation which is much easier to calculate but almost identical in outcome.
It is possible to define a recognition function for every frequency band separately, also is it possible to define one for every phoneme. These combined with noise, time and frequency distortions would lead to enormous complex functions, but would yield the intelligibility of any text spoken by a good speaker and perceived by a good listener. And there lies the snag: the system assumes good speakers and listeners. If a limited number of speakers are using the system it would be possible to incorporate them in the calculations but it will be difficult to incorporate the abilities of the listeners. The spread in an arbitrary group of listeners will be large. It makes not much sense to multiply a very precise number with an inherently imprecise number. But nevertheless the calculation scheme allows a very high precision in almost every aspect of the recognition of speech. It could be used for special purposes or optimization processes, especially because the results would be highly diagnostic in nature.
In order to keep the procedures practical in this paper, the masking of speech by noise and reverberation will be developed and quantified based on the

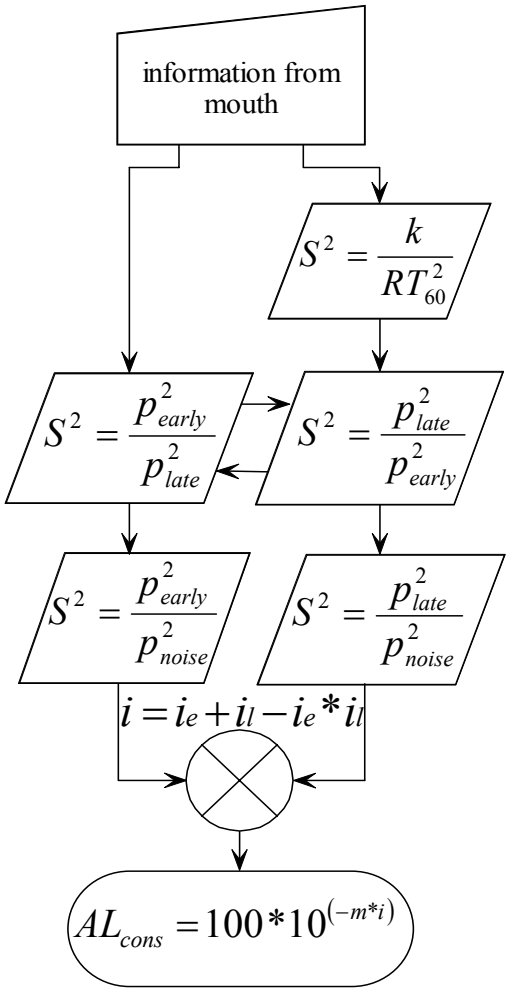

Figure 1: Information flow diagram

acoustical properties of the transmission chain in one octave only. Other octaves will be assumed to be in balance.

3. CALCULATING WITH THE PEUTZ '88 ALGORITHMS

The calculation scheme that is going to be implemented is shown in the information diagram in figure 1. After the information leaves the mouth it is separated in two paths, a direct path or actually an "early" path because also the early energy contains lots of speech cues, and a reverberant path or actually a late part where the information transmitted is dependant on the reverberation time of the exponential decaying diffuse sound field. This means that calculation of the AL_{cons} with this method has to be done with a measured or calculated impulse response. The early and the late sound mask each other, if one is more than 10dB louder than the other the softer has no influence on the speech intelligibility anymore. Each of the paths are independently masked by noise. All the indices in one path are combined by multiplying. Because the paths carry independently the same information they are combined with the rule: $\boldsymbol{i_e+i_l-i_e*i_l}$. The equations used have the general form of:

$$i = \frac{0.5\log_{10}\left(\frac{1+aS^2}{1+\frac{1}{a}S^2}\right)}{\log_{10} a} \qquad (1)$$

Where:

a	is a constant
S	is the ratio between the information carrier and its masker in the path. For:
i_{el}	a=10, $S^2=p^2_{early}/p^2_{late}$
i_{en}	a=13, $S^2=p^2_{early}/p^2_{noise}$
i_{le}	a=10, $S^2= p^2_{late}/p^2_{early}$
i_{ln}	a=13, $S^2=p^2_{late}/p^2_{noise}$
i_t	a=10, $S^2=k/RT^2_{60}$, where k=5

See figure 2.

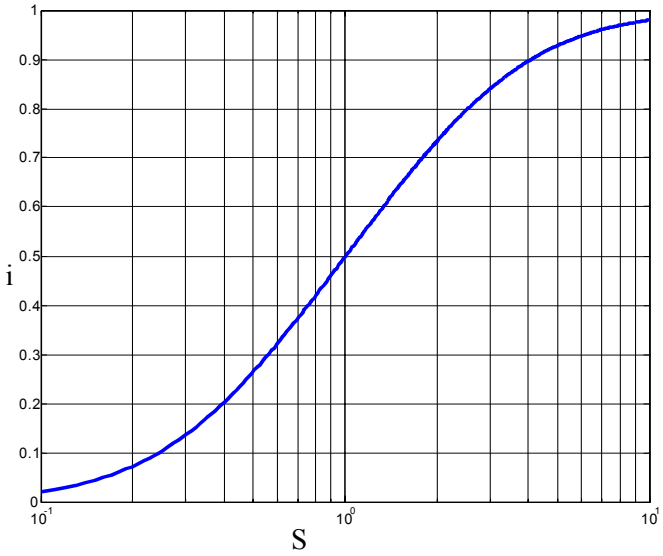

Figure 2: i as a function of S, with a=10, calculated with (1)

To combine them:

$$i_e = i_{el} * i_{en}$$

$$i_l = i_t * i_{le} * i_{\ln}$$

$$i = i_e + i_l - i_e * i_l \qquad (2)$$

To convert to AL_{cons}:

$$AL_{cons} = 100 * 10^{(-m*i)} \qquad (3)$$

Where:

m	is recognition measure for consonants: 1.7
i	is the information index for consonants

Inside the m factor is the zero correction factor a as used in the early equations (see part 1). In practical use however, the a factor is often omitted and the speech intelligibility criterion for the sound system adjusted because the a factor is a factor for the talker and listener, not for the sound system. If is chosen for m=1.9, the values will be more comparable with the early equations without a.

3.1. Defining early-late windows:

It is difficult to define the exact early window time, especially if square windows are taken. If a strong reflection is present it can mean that in one row in the seating area the reflection is inside the early window and one row further away it is inside the late window. The numbers would be jumping but the ear would hardly notice anything. This is often the case with the familiar Clarity and Deutlichkeit measures. It is proposed in this paper to use a half Hanning window for the early sound and the other half for the beginning of the late window. Both windows start at the arrival time of the direct sound. They cross at the effective window length, in this case 50ms. The total window length to generate is 4 times as long (200ms). The second half of the window is used for windowing the early part of the impulse response and the first half of the window is used for windowing the first 100 ms of the late part of the impulse response. After 100ms the late window stays 1 until the end of the valid part of the impulse response. See figure 3 for an example. After windowing the data is filtered with in this case a 1kHz octave filter. Filtering after the data is windowed is done because in that case the precision in time that comes with the wide band impulse response measurement is kept. Frequency filtering means that the response is smeared out in time, depending on the filter used, making it ambiguous to define the proper starting and stopping points.
After filtering the data is squared and summed, the values can be used as p^2_{early} and p^2_{late} in the equations.

3.2. Noise

It is difficult to bring the right amount of noise in to the measured impulse response. Impulse response measurements take a fixed short time and it is difficult to measure on that very time at which the noise is at its average value. The author prefers to:

- measure the impulse response wide band, with as little noise as possible,
- measure the noise over a significant length of time,
- perform statistical analysis,
- calculate the most likely signal to noise ratio and
- use this to calculate the information indices concerning noise.

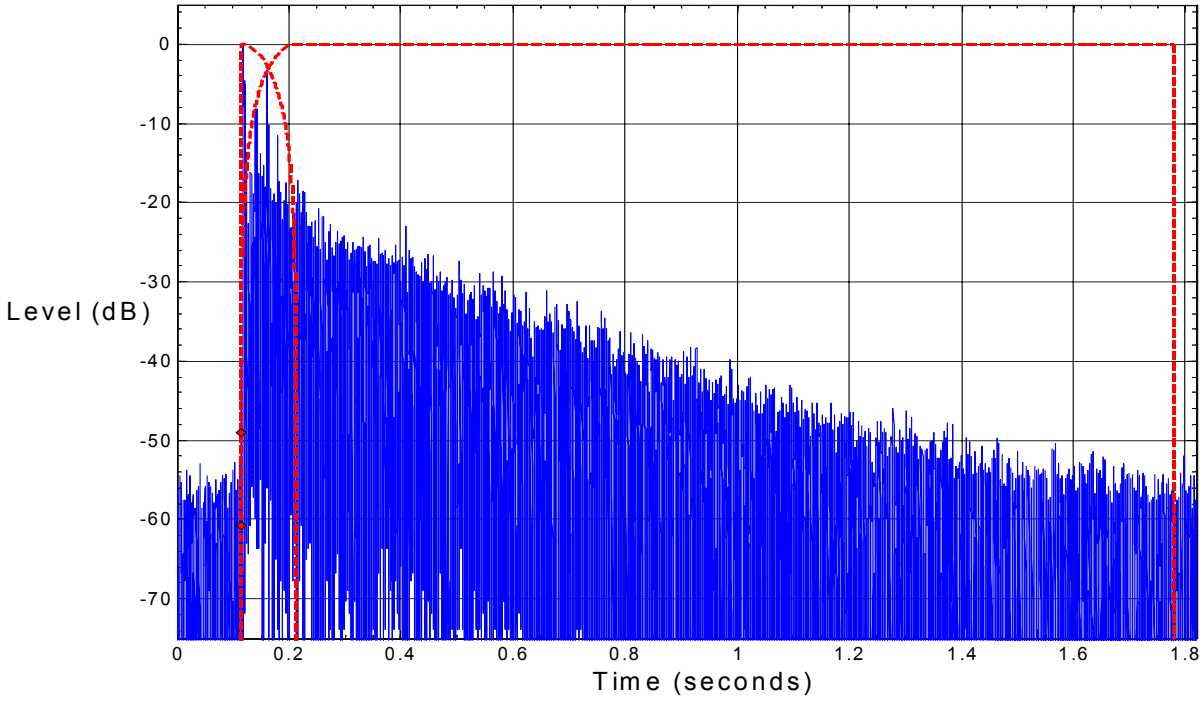

Figure 3: ETC and weighting windows for early and late sound

4. PERFORMANCE

There are two (and a half) ways to compare these algorithms with other methods:

- Comparison with other predicting algorithms
- Comparison with AL_{cons} assessed with (nonsense)wordlists,

- Comparison with estimations based on personal experience

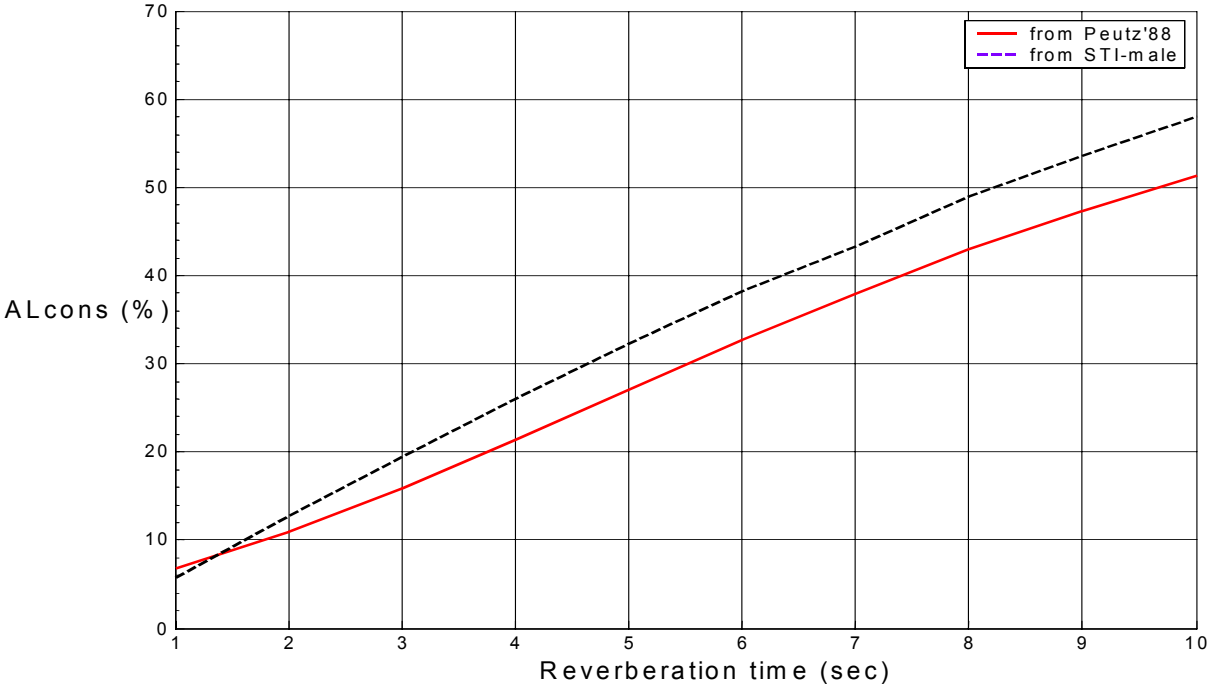

Figure 4 Comparison ALcons from STI-male with Peutz'88

The latter is of course a little ambiguous but an experienced designer with a lot of measuring experience should be able to estimate the AL_{cons} with an accuracy of approx. 2%. This is about the same deviation that can be found in a group of "good" listeners.

In the preparation of writing this paper the algorithms where at first implemented as Peutz stated them in his paper and notes. The accuracy was remarkably good, especially when processing impulse responses where subjective listening experience differed significantly from the AL_{cons} calculation using the STI algorithms [2] and [3]. Most of the time the values calculated with the Peutz'88 algorithms are within 2 % AL_{cons} from the values calculated using the STI algorithms, so it is difficult to say which one is closer to the truth, if it has any significance at all. But for some responses it becomes quite clear that the Peutz'88 algorithms have their merits. For instance:

In a theater situation speech intelligibility was assessed with (nonsense)wordlists and impulse responses where measured with and without noise. The real AL_{cons} was 6 to 7%. The values calculated from STI with 15dB signal to noise ratio where 3% and with the Peutz'88 algorithms 6% (with 12 dB s/n 7%).

A contractor sent us some measured impulse responses and recorded speech in a Mosque with 8 seconds reverberation time, asking us why the intelligibility was so bad. Calculating Alcons with the STI algorithms showed values between 18 and 39% AL_{cons} which means: totally unintelligible. When listening to the speech recordings the speech intelligibility was certainly not good, bud not as bad as the calculated values suggested. When processing the data according to the Peutz'88 algorithms at least 20 of the 60 responses where between 10 and 15% AL_{cons}. This comes much closer to the subjective impression.

A question was: how does this set equations compare with other algorithms like the equations in part 1 and STI. To make a principal comparison possible with STI a series of artificial impulse responses was generated of perfect homogeneous exponentially decaying sound fields with reverberation times between 1 and 10 seconds. From these responses the ALcons was calculated with the Peutz'88 formulae and from the STI with the new male weighting. The results are shown in figure 4. The early to late ratio for these files range from 0.25 dB for the 1 second file to –11.6 dB for the 10 second file. Comparison of the Peutz'88 equations with the Peutz'71 equations is not directly possible, anyway they shall not coincide, because of the principle differences. The Peutz'71 equations use direct and reverberant sound, the Peutz'88 equations early and late sound. Even if there is no direct sound there will be early sound, otherwise the impulse response would start later. For

the signal to noise ratio there is also a problem. The Peutz'71 equations are based on speech level and PSIL, the Peutz'88 equations on signal to noise ratio in the same frequency band.
It is quite possible to adjust the ***a, k*** and ***m*** factors to match the STI based curves in figure 4 or the $9*RT_{60}$ curve from the Peutz'71 algorithms. However when comparing measured AL_{cons} (with nonsense wordlists) with the Peutz'88 algorithms with the adjusted parameters it showed that the algorithms where now much less accurate. Although visually in graphs the compatibility with other algorithms is much better, the actual predicting capability of AL_{cons} however was much worse. It is clear that the set equations are a unity where it is hardly possible to find a one to one comparison with statistical parameters of other algorithms.

5. CONCLUSIONS

This calculation scheme is reasonably accurate, and in at least some cases more accurate than other electronic means to calculate speech intelligibility. It is however by no means the final answer. It is a statistical based indicator with a limited number of parameters to predict AL_{cons}. Its value lies in the different approach, the easy assessable data, the quick calculation and its accuracy, which all may be beneficial. It is not out of the question that it is possible that this calculation scheme can be fine tuned to an even higher degree of accuracy, even with data in a single frequency band. It will however not be simple and straight forward like matching the constituent equations with other known algorithms. The algorithms calculate very quick so it will hardly be a calculation time problem to incorporate more physical parameters or making it multi band (2 or 3). This opens the possibility to incorporate the negative effects of strong coloration in situations with a long reverberation time. Before these improvements can be implemented, sufficient data needs to be gathered and evaluated in order to be sufficiently accurate. From this data it will also be possible to find an algorithm to convert STI to ALcons which is more accurate than the contemporary algorithm. Based on the currently available data it looks that it needs some adjusting.

The conclusions in part 1 concerning "What You Specify Is What You Get" are of course equally valid in this part 2.

6. REFERENCES

[1] V.M.A. Peutz, paper 2732 (E-4) presented at the 85th convention of the AES, November 1988.

[2] International standard IEC 60268-16: Objective rating of speech intelligibility by speech transmission index.

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