Acoustic and Biomedical Engineering

Influence of Modulation Detection Threshold on Speech Intelligibility

K. Leo*

Faculty of Physics and Applied Mathematics, Gdańsk University of Technology

G. Narutowicza 11/12, 80-233 Gdańsk, Poland

Speech intelligibility in classrooms and lecture halls was investigated. Acoustical measurements in the rooms have been performed, with special concern about 125 ms of early decay phenomenon. Weighting function of early reflections relative energy derived from modulation threshold and speech modulation spectrum of signal is proposed.

PACS: 43.55.Hy

1. Introduction

1.1. Reverberation phenomena in rooms vs. speech intelligibility

Speech intelligibility in rooms is affected by masking caused by external noise and reverberation generated by speech itself. This paper deals with speech sounds interacting with reverberation.

Speech intelligibility depends on both early and late reflections, with special stress on duration of speech sounds compared to the delay and energy of reflected sound. In this paper an influence of early reflections on speech intelligibility is analyzed, with attention paid to modulation threshold of human ear.

1.2. Early reflections vs. speech intelligibility

Reflected sounds coming to the listener in the first 50–80 ms after the direct sound are considered as positively contributing to speech intelligibility. Due to that, echogram can be divided into useful and detrimental parts, which consist of early and late reflections, respectively. Parameters describing the energy of the early part of echogram are usually good predictors of speech intelligibility.

A typical parameter comparing both parts of echogram is C_y, where y is boundary expressed in ms between early and late reflections

$$C_y = 10\log\left(\int_0^y p^2(t)\,\mathrm{d}t \Big/ \int_y^\infty p^2(t)\,\mathrm{d}t\right)\ [\mathrm{dB}]\,. \qquad (1)$$

Kürer [1] proposed parameter T_s [s] (*Schwerpunktzeit*), defined as a point on time axis, which divides the echogram into two parts of equal energy. Bradley [2] proposed early reflection benefit (ERB), describing the level of energy for first 50 ms reflections, relative to the first 10 ms reflections. He concluded that in speech recognition process, the listener can have the benefit from either very early reflections (up to 10 ms) and still useful later reflections (up to 50 ms)

$$\mathrm{ERB} = 10\log\left(E_{50}/E_{10}\right)\ [\mathrm{dB}]\,. \qquad (2)$$

1.3. Time weighting of parameters

Formulae (1), (2) are constructed in such a manner that strong reflection, when shifted slightly in time, can be considered an early or late one. It can significantly change the value of the parameter. To avoid a "sharp border effect" in time domain, Lochner and Burger assigned a weighting factor to energy of reflection [3]. The values of weighting function depend on the reflection arrival time (see Fig. 6). This concept is also suitable for estimating speech intelligibility

$$\eta = E_{\mathrm{useful}}/E_{\mathrm{detrimental}} = \int_0^{95\mathrm{ms}} p^2(t)a(t)\,\mathrm{d}t \Big/ \int_{95\mathrm{ms}}^\infty p^2(t)\,\mathrm{d}t\,, \qquad (3)$$

η — ratio of useful to detrimental energy, $a(t)$ — weighting function. Similar approach is presented in Peutz's algorithm for calculating ALLCONS parameter. Two overlapping Hanning half windows weight both early and late reflections energy in function of their arrival time. The point of crossing the windows is 50 ms [4].

1.4. Influence of speech envelope fluctuations on speech intelligibility

Approach presented in this paper follows Shannon's idea, in which bands of noise have been modulated by speech envelope [5]. Shannon showed that the increasing number of frequency bands from 1 to 4 refers to the increase of speech intelligibility from 50% to 90%. Drullman [6] investigated the contribution of different modulation frequencies in speech recognition. He found that the consonants are more sensitive to the envelope filtering than vowels. Spectral content of vowels is well preserved due to the longer duration and relatively higher intensity. The presence of high-frequency modulations help to preserve short duration consonants.

* e-mail: krzysztof.leo@gmail.com

Modulation frequency 8 Hz divides modulation spectrum into two equally intelligible frequency regions. Another characteristic modulation frequency for speech is crossover frequency between consonants and vowels. It lies between 3 and 6 Hz (average 4 Hz).

1.5. Modulation spectrum of speech and speech transmission index

Speech can be represented as the sum of modulated signals in different modulation frequency bands and different depth of modulation. This concept is a base for speech transmission index (STI) procedure. Modulation spectrum of testing signal used in STI procedure ranges from 0.63 Hz to 12.5 Hz, and imitates the average modulation spectrum of speech signal. STI measures speech intelligibility in function of reverberation and external noise. Valleys of modulated signal are fulfilled with reverberation or noise. Measured reduction of modulation factor at given modulation frequency is transposed into apparent signal-to-noise ratio [7, 8].

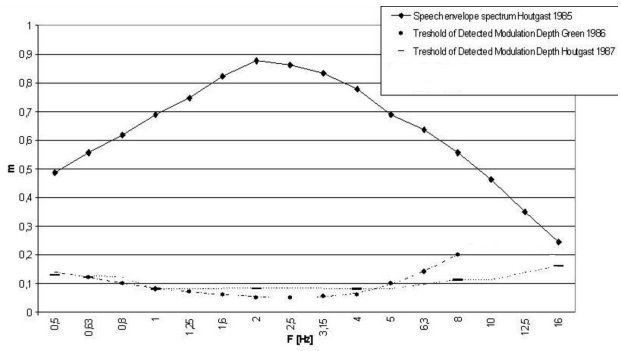

Fig. 1. The modulation index as a function of modulation frequency; average speech modulation spectrum, analyzed speech was recorded in anechoic room [7], thresholds of detected modulation depth [9, 10]. Values of threshold varies among authors due to the type of testing signal.

Notes about subjective threshold of modulation perception can be found in the literature (Fig. 1). The modulation threshold has not been considered as an element of instrumental assessment of speech intelligibility. Therefore, the threshold values have not been included in the assessment procedure. The aim of this paper is to show that this problem is worth to be reconsidered.

2. Temporal weighting of early energy

2.1. Modulation frequencies in speech signal

According to STI procedure, speech can be represented as a set of modulated signals differing in values of modulation index and modulation frequencies. They are represented by speech modulation spectrum (Fig. 1) [7]. Periods of modulated signal can be divided into four equal parts representing different sensitivity to early reflections. Reflections occurring after the peak of modulated signal are interfering with second and fourth quarter part of modulated signal period — Fig. 2. This process creates coloration in spectral domain.

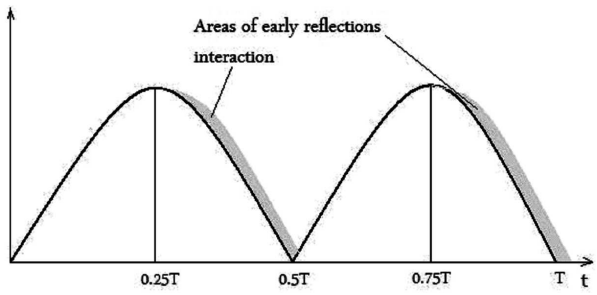

Fig. 2. Four quarters of modulated signal period, second and fourth quarter interfering with early reflections — the preview presentation. Abscissa — time expressed by parts of period, ordinate — squared sound pressure.

It is assumed in this paper that early reflections interfere with speech amplitude envelope in the second quarter of modulated signal period ($0.25T$–$0.5T$, Fig. 2). They are crucial in determining speech intelligibility.

Early energy measure C_y (time division y is equal to the second quarter period of modulated signal) can measure early energy influence on speech intelligibility. Time division y can be calculated using Eq. (4):

$$y = T/4 = 1/(4F)\,. \tag{4}$$

y — time boundary for early and late reflections, T — modulation period of speech component, F — modulation frequency in one third octave bands, as shown in Fig. 1.

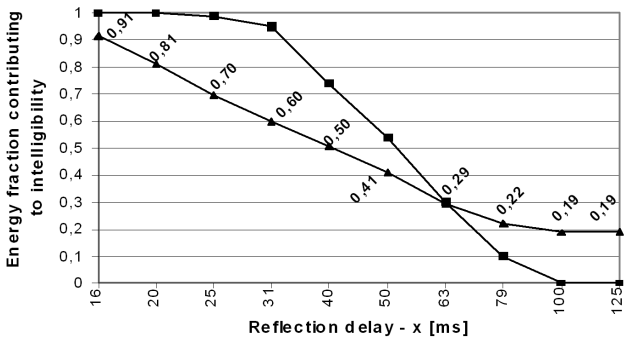

Fig. 3. ▲ Proposed weighting function calculated from differences between average speech spectrum and modulation detection threshold. ■ Weighting function used by Lochner–Burger [3].

For further considerations ten values of y are used, as in the Lochner–Burger case [3] ($y = 16, 20, \ldots, 125$ ms, see Fig. 3).

2.2. Temporal weighting of early energy

In this paper, temporal weighting of relative energy arriving to the listener is proposed, similarly as in the

Lochner and Burger criterion [3]. Weighting procedure proposed in this paper has the advantage, because it ties signal properties to acoustic properties of the room. Modulation properties of speech signal, as well as subjective modulation reception threshold, are taken into account in the proposed procedure.

Initial 125 ms of impulse response is chosen for further analysis due to the largest difference between modulation spectrum of speech and modulation threshold for modulation frequency occurring at 2 Hz. Quarter of period for modulation frequency 2 Hz is 125 ms due to Eq. (4). The remaining part of impulse response should be evaluated depending on weights for later parts of the echogram. These can give rating of late reflections influence on speech intelligibility.

To include the influence of the modulation threshold and speech modulation spectrum on the assessment of speech intelligibility via early reflections measurement, weights of early reflections are proposed. Weights are equal to one minus the difference between modulation index of speech and modulation detection threshold. The average value of two threshold curves is used, see Fig. 1. Input data are taken from [7, 9, 10].

Proposed weighting curve refers to discrete time intervals equal to the quarters of period of speech modulation frequencies. The curve has shape similar to the Lochner and Burger curve, which is based on intelligibility experiments in simulated sound fields [3]. The new element of the proposed approach is that speech modulation spectrum and modulation detection threshold has been incorporated into weights calculation.

To predict speech intelligibility taking into account modulation threshold of human ear, the following procedure is proposed:

— calculate 10 values C_y ($y = 16, 20, \ldots, 125$ ms, see Fig. 3) from Eq. (1),

— calculate weighted energy average of these values (weights marked ▲ in Fig. 3),

— logarithm of this average, called $C_{16–125}$, is interpreted as a speech intelligibility predictor.

3. Experimental verification

To verify the above procedure, five halls were tested, in which acoustic measurements and intelligibility logatoms tests were performed. The halls are listed in Table I. The halls were empty.

In acoustic measurements a pair of directional loudspeakers GENELEC 8020 A was used. Impulse responses of all rooms were obtained by means of maximum length sequence technique. The aim of measurements was objective assessment of speech intelligibility.

Speech intelligibility was also assessed subjectively by reading CVC logatom lists, consisting of 100 logatoms each. Two lists were read, each for different group of listeners. SPL value at 1 m in front of the loudspeaker and signal to noise ratio in a hall was 60–70 dB and 40–60 dB, respectively. Subjectively assessed speech intelligibility was averaged in circles of radius of approximately 2 m around each observation point used in acoustic measurement with average standard deviation 5.35%.

TABLE I

Dimensions and volumes of the measured halls.

Hall	Dimensions [m]	Volume [m^3]	No. of measurement points
classroom 1	$7.62 \times 7.75 \times 3.3$	195	5
classroom 2	$6.5 \times 12.5 \times 3.3$	264	6
lecture hall 1	$8.3 \times 9.07 \times 4.2$	317	10
lecture hall 2	$14 \times 10 \times 6$	840	6
hall 3	$12.5 \times 15.5 \times 7.5$	1356	5

4. Results

Acoustic parameters correlation with subjective speech intelligibility in each room have been measured. Values of squared Pearson's R^2 coefficient describing linear correlation of speech intelligibility with acoustic parameters are listed in Table II. Early energy parameters with coefficient R^2 exceeding 0.65 are marked in italic, corresponding parameters with smaller correlation are marked in bold.

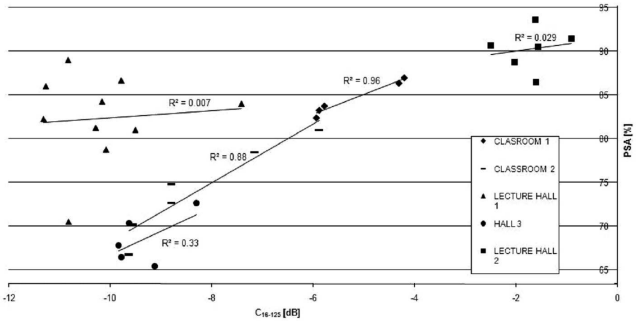

Fig. 4. Speech intelligibility PSA in function of $C_{16–125}$. Results for 5 rooms. For each room linear regression line with corresponding local R^2 is shown.

Figure 4 shows dependence of speech intelligibility on parameter $C_{16–125}$. These data create global tendency for 4 rooms, one of the rooms (lecture hall 1) falls off. Supposedly it is due to the lack of early reflections — the room is very wide. For each of 5 rooms' data linear correlation coefficient and trend line are shown in Fig. 4.

TABLE II

Values of squared Pearson's (R^2) coefficients measured between speech intelligibility data (PSA) and objective acoustic parameters (including STI). PSA — percentage syllable articulation, STI — speech transmission index, $C_{16–125}$ [dB] — sum of weighted energy ratio clarity coefficients.

	class-room 1	class-room 2	lecture hall 1	lecture hall 2	hall 3
volume [m^3]	195	317	1356	1.9	0.6
T_{30} [s]	264	840	0.95	3.2	2.3
			R^2		
G	0.91	0.85	0.01	0.28	0.11
T_{20}	0.06	0.41	0.03	0.54	0.49
T_{30}	0.10	0.11	0.10	0.27	0.00
EDT	0.05	0.13	0.28	0.01	0.07
C_{40}	*0.95*	*0.91*	**0.02**	**0.08**	**0.42**
C_{50}	*0.85*	*0.93*	**0.03**	**0.07**	**0.07**
C_{80}	*0.76*	*0.88*	**0.06**	**0.19**	**0.26**
D_{50}	*0.85*	*0.93*	**0.03**	**0.06**	**0.07**
T_s	*0.96*	*0.92*	**0.10**	**0.24**	**0.18**
PSA	1.00	1.00	1.00	1.00	1.00
STI	0.27	*0.79*	**0.05**	**0.08**	**0.05**
$C_{16–125}$	*0.96*	*0.89*	**0.01**	**0.03**	**0.33**

5. Conclusions

- Because early decay parameters (C_{40}, C_{50}, C_{80}, D_{50}, $C_{16–125}$) show the highest correlation with speech intelligibility in the CLASSROOM 1, CLASSROOM 2, they are good predictors of speech intelligibility in such rooms.
- Parameter $C_{16–125}$ follows statistical trends of C_{40}, C_{50}, C_{80}, D_{50}. Thus it can be regarded as predictor in speech intelligibility instrumental assessment.
- In larger and reverberant halls, the lack of significant correlation between early energy parameters and speech intelligibility has been stated (lecture hall 1, lecture hall 2 and hall 3). This is due to the high degree of diffusion caused by large volumes and long reverberation. Higher diffusion means less dependence of intelligibility on the position of the listener (see lower correlation in Table II).
- This survey shows the need for more measurements. Concern should be given to the: source–listener geometry, reverberation and diffusion. Parameter $C_{16–125}$ and weighting curve is a field for further investigation. It is hoped that weighting of impulse response by means of parameter $C_{16–125}$ can give advantage in more precise and stable predictor of early energy influence on speech intelligibility.

References

[1] R. Kürer, in: *7th Int. Congress on Acoustics*, Akademiai Kiadó, Budapest 1971.

[2] J.S. Bradley, *J. Acoust. Soc. Am.* **80**, 846 (1986).

[3] J.P.A. Lochner, J.F. Burger, *Acustica* **11**, 195 (1961).

[4] J. van der Werff, in: *114 AES Convention*, 2003, Paper No. 5763.

[5] R.V. Shannon, F.G. Zeng, V. Kamath, J. Wygonsky, M. Ekelid, *Science New Series* **270**, 5234 (1995).

[6] R. Drullman, *J. Acoust. Soc. Am.* **97**, 101 (1995).

[7] T. Houtgast, H.J.M. Steeneken, in: *RASTI, a Tool for Evaluating Auditoria*, Brüel & Kjær, 1985, p. 13.

[8] P. Larm, V. Hongisto, *J. Acoust. Soc. Am.* **119**, 2 (2006).

[9] D. Green, in: *Auditory Frequency Selectivity*, Eds. B.C. Moore, R. Patterson, Plenum Press, Cambridge 1986, p. 351.

[10] T. Houtgast, *J. Acoust. Soc. Am.* **85**, 1676 (1989).