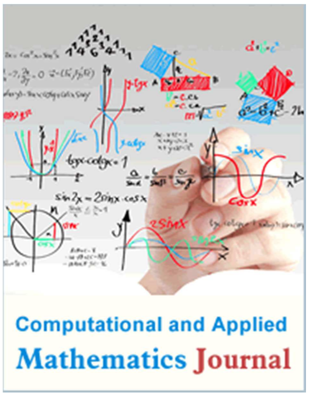
Computational and Applied Mathematics Journal
2015; 1(3): 60-66
Published online April 20, 2015 (http://www.aascit.org/journal/camj)

AASCIT American Association for Science and Technology



Keywords
Late Reverberation Suppression,
Optimal Parameters Values,
Speech Quality,
Speech Recognition Accuracy

Received: March 28, 2015
Revised: April 9, 2015
Accepted: April 10, 2015

Speech Recognition Performance as Measure of Speech Dereverberation Quality

Arkadiy Prodeus

Acoustic and Electroacoustic Department, Faculty of Electronics, NTUU KPI, Kyiv, Ukraine

Email address
aprodeus@gmail.com

Citation
Arkadiy Prodeus. Speech Recognition Performance as Measure of Speech Dereverberation Quality. *Computational and Applied Mathematics Journal.* Vol. 1, No. 3, 2015, pp. 60-66.

Abstract
Optimal, in the sense of automatic speech recognition (ASR) accuracy maximum, parameters of the late reverberation suppression technique have been proposed in this paper. It was shown that the value 50 ms as boundary between early reflections and late reverberation, which usually is used when problems of speech quality and intelligibility is studied, isn't best for ASR systems, for which optimal value is 100 ms. It was shown also that, when estimating late reverberation power spectrum, an optimal value of averaging parameter should be associated with statistical speech constants such as phoneme and stationary durations. Several speech quality indicators were used, and it was found that recognition accuracy is the best indicator in the sense of ability to inform the user about reached compromise between reverberation suppression and speech distortion.

1. Introduction

Modern telecommunications and ASR systems operate sometimes in very difficult acoustic environments. For example, when speaker is in room characterizedwith room impulse response (RIR) $h(t)$ and microphone is located at considerable distance from the speaker's mouth which is the source of speech signal $x(t)$, the reverberated speech signal is observed at the output of the microphone

$$y(t)=\int_{-\infty}^{\infty} h(v)x(t-v)dv = x(t)\otimes h(t)\ .$$

Here $\otimes$ is convolution symbol. Signal's $y(t)$ quality becomes degraded as compared with one for signal $x(t)$. Moreover, ASR systems recognition accuracy decreases when signal $y(t)$ is input, because of these systems are usually trained with undistorted speech signals $x(t)$ [1-2].

When decomposing RIR (Fig. 1) into early reflections and late reverberation

$$h_i(t)=\begin{cases} h(t), & 0\le t\le T_l; \\ 0, & \partial p.\,t \end{cases} \quad h_l(t)=\begin{cases} h(t+T_l), & t\ge 0; \\ 0, & \partial p.\,t \end{cases},$$

signal $y(t)$ can be represented as

$$y(t)=h_i(t)\otimes x(t)+r(t)\ . \tag{1}$$

Here $r(t) = h_l(t) \otimes x(t - T_l)$ is component due to late reverberation, and T_l is time corresponding to boundary between early reflections and later everberation. When assuming that the terms of (1) are statistically independent, it has become clear why late reverberation may be interpreted as kind of noise. Unfortunately, strong non-stationarity of late reverberation makes ineffective traditional techniques of noise suppression, because these techniques are designed for stationary or slow non-stationary noise.

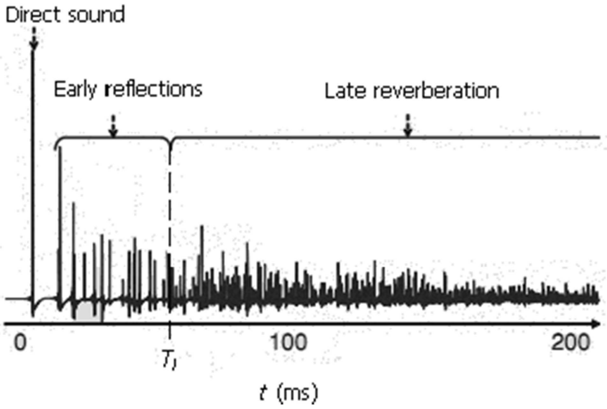


Fig. 1. *RIR structure*

Speech correction in frequency domain is one of the most widely used approach to noise suppression, and it was used in [3-4] for later everberation suppression. Analytically the technique is described as

$$\hat{\lambda}_x^{1/2}(l,k) = G(l,k)\lambda_y^{1/2}(l,k) \text{ ,}$$

where $\lambda_y(l,k)$ is l -th frame power spectrum of signal $y(t)$ at frequency $f_k = kF_s / N_{fft}$; F_s is sampling rate; N_{fft} is FFT parameter; k is number of frequency sample; $\hat{\lambda}_x(l,k)$ is l -th frame power spectrum estimator of signal $x(t)$; $G(l,k)$ is correction filter gain. Phase of distorted signal $y(t)$ is used as enhanced signal $\hat{x}(t)$ phase.

When distance between sound source and microphone is more than critical one, late reverberation power spectrum $\lambda_r(l,k)$ need be estimated as follows [3-4]:

$$\lambda_r(l,k) = e^{-2\delta(k)T_l} \cdot \lambda_y(l - N_l, k) \text{ ,} \tag{2}$$

where $N_l = T_l / T_{inc}$; T_{inc} denotes the frame shift value; $\delta(k) = 2\ln 10 / T_{60}(k)$; $T_{60}(k)$ is reverberation time.

It was proposed in [3] use running averaging when power spectrum $\lambda_y(l,k)$ is estimated. Averaging parameter η_z ($0 \le \eta_z < 1$) is used for controlling of the time interval duration of averaging spectrogram $Y(l,k)$ of signal $y(t)$:

$$\hat{\lambda}_y(l,k) = \eta_z \hat{\lambda}_y(l-1,k) + (1-\eta_z)\left|Y(l,k)\right|^2 . \tag{3}$$

Equation (3) was modernized in [4] as follows:

$$\hat{\lambda}_y(l,k) = \eta_z(l,k)\hat{\lambda}_y(l-1,k) + (1-\eta_z(l,k))\left|Y(l,k)\right|^2 \text{ ,} \tag{4}$$

$$\eta_z(l,k) = \begin{cases} \eta_z^d(k), & \left|Y(l,k)\right|^2 \le \hat{\lambda}_y(l-1,k); \\ \eta_z^a(k), & \textit{in other cases,} \end{cases} \tag{5}$$

$$\eta_z^d(k) \le \frac{1}{1 + 2\delta(k)T_{inc}} \text{ ,} \tag{6}$$

$$0 \le \eta_z^a(k) < \eta_z^d(k) . \tag{7}$$

A lot of claims can be made to the results of [3-4]. Firstly, there was not made an attempt to optimize parameter T_l used in (2). Secondly, there was not took into account the statistical characteristics of the speech in the averaging procedures (3)-(7). As a result, the constant $\eta_z = 0.9$ proposed in [3] for the range of reverberation time $T_{60} = 0.4 - 1.7$ s looks not well-founded. In addition, this proposal is not consistent with (6), according to which $\eta_z^d(k)$ must be chosen taking into account the reverberation time $T_{60}(k)$. Finally, it was not experimentally confirmed in [4] the insistent need of parameter $\eta_z(l,k)$ dependence on variable l , and the ratio (7) looks too uncertain.

Attempts to eliminate some of these shortcomings had been made in [6-9]. The optimum value of T_l had been experimentally evaluated in [6], and the results were refined in [7]. In addition, the possibility of optimization of the averaging procedures had been studied in [7] and it was shown experimentally that a simple averaging (3) can lead to better results than the cumbersome procedure (4)-(7). Recommendations on the deriverberation algorithm optimization in the absence of a priori information about RIR characteristics were obtained in [8]. Analysis of features of some objective indicators of deriverberation algorithm quality had been performed in [9].

In this paper, the results of [6-9] are refined, expanded and equipped with extended comments.

Significant research efforts have been directed to the area of speech quality and intelligibility assessment. Results of investigation of the use of automated speech recognition technology as a means to evaluate coding algorithms for digital speech had been presented in [10-11].The word recognition ratio as performance metric was used in [11], whereas phoneme recognition ratio was used in [10]. It was shown in [12] that ASR systems recognition accuracy and speech quality indicator PESQ are two of the best measures for intelligibility estimation, though ASR fails at high noise condition. The results support the idea that no measure works universally well, and it was emphasised that before choosing a measure for system evaluation, the suitability should be assessed.

That’s why another important objective of this paper is comparing the ability of different objective quality measures to inform user about reaching compromise between reverberation suppression and speech distortion.

2. Dereverberation Optimality Criteria

It is natural to assume that, when changing the dereverberator parameters T_l and η_z^d , it is possible find their optimal values which maximize the quality of speech signals and the accuracy of automatic speech recognition (Fig. 2).

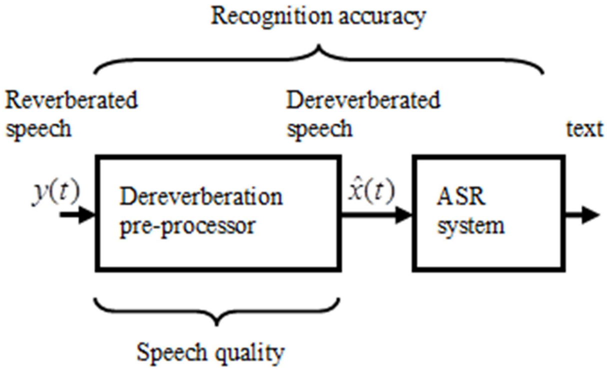


Fig. 2. *Dereverberation quality indicators*

When estimating speech quality, segmental Signal-to-Reverberation Ratio (SRR)

$$SRR = \frac{1}{L}\sum_{l=1}^{L} 10\lg\left(\sum_{n=Rl}^{Rl+N-1} x^2(l,n) \Big/ \sum_{n=Rl}^{Rl+N-1} [x(l,n)-\hat{x}(l,n)]^2 \right),$$

Logarithmic Spectral Distance (LSD)

$$LSD = \frac{2}{KL}\sum_{l}\sum_{k=0}^{K/2-1} \left| G\{X(l,k)\} - G\{\hat{X}(l,k)\} \right|,$$

$$G\{X(l,k)\} = \max\{20\lg(|X(l,k)|), \delta\},$$

$$\delta = \max_{l,k}\{20\lg(|X(l,k)|)\} - 50 ,$$

Bark Spectral Distortion (BSD)

$$BSD = \frac{\sum_{l=1}^{L}\sum_{k=0}^{K/2-1}\left[B\{X(l,k)\} - B\{\hat{X}(l,k)\}\right]^2}{\sum_{l=1}^{L}\sum_{k=0}^{K/2-1}\left[B\{X(l,k)\}\right]^2}$$

are often used. Here $x(l,n)$ and $\hat{x}(l,n)$ are n -th samples of l -th frame of anechoic speech signal $x(t)$ and enhanced signal $\hat{x}(n)$, respectively; $X(l,k)$ and $\hat{X}(l,k)$ are spectrograms of signals $x(n)$ and $\hat{x}(n)$, respectively; $B\{X(l,k)\}$ and $B\{\hat{X}(l,k)\}$ are bark spectrums of l -th frame of signals $x(n)$ and $\hat{x}(n)$, respectively. Furthermore, Perceptual Evaluation of Speech Quality (PESQ) is effective indicator of speech quality. PESQ estimation algorithm is described in [13].

Quantitative evaluation of dereverberation performance can been made also by means of end-to-end quality index "ASR accuracy" [14]:

$$Acc\% = (N - D - S - I)/N \times 100\% .$$

Here N is the total number of labels in the reference transcriptions; D is the number of deletion errors; S is the number of substitution errors; I is the number of insertion errors.

3. Experimental Results

Clean speech signals (single words) were recorded in anechoic room and had been used for ASR system training. Parameters of digitized sounds were: sampling rate 22050 Hz, linear 16 bitquantization. Reverberated signals had been simulated by convolving of clear speech and RIRs for three rooms with reverberation times 0.74 s, 0.89 s and 1.1 s.

Signal frames with 50% overlapping and Hamming window were used for signal processing. Frames duration was 32 ms. Reverberation time was estimated by applying Schroeder's method [15] to a band pass filtered versions of the RIRs. Moreover, it was taken $\eta_y^a(k) = 0{,}5 \cdot \eta_y^d(k)$.

Toolkit HTK [14] had been used for ASR system simulation. Training of ASR system had been made with usage of 269 samples of 27 words of clean speech recorded for two speakers-women. Reverberated discrete speech signal (there were 0.2…0.5 s pauses between single words) was used as test signal, and there were presented, in testing, all 27 words used in training. There were 27 phonemes of Ukrainian language in phoneme vocabulary and there had been used 39 MFCC_0_D_A coefficients when ASR simulating.

Dereverberation procedure was implemented using log MMSE technique [5]. In this case, the power spectrum $\lambda_y(l,k)$ of the reverberated signal was estimated in two ways:

1 in accordance with (4)-(7);

2 in accordance with (4)-(7), but $\eta_z^d(k)$ assumed to be independent from k : $0 \le \eta_z^d < 1$.

Dependences of $Acc\%$, SRR, LSD, BSD and PESQ on the parameters T_l and η_z^d ($\eta_z^d(k)$ assumed to be independent from k) are shown in Fig. 3. These graphics are somewhat different from those given in [9] and the graphs are more correct, because compared signals were normalized by the standard deviation. Note that these graphs are averaged over the parameter T_{60} . $Acc\%$ graphs, which are not average dover parameter T_{60} , can be found in [7].

For further testing of the effectiveness of averaging procedures (3) and (4)-(7), an alternative variant of the power spectrum $\lambda_y(l,k)$ estimation was realised for T_{60} =0.74 s:

$$\hat{\lambda}_y(l,k)=\begin{cases}\left|Y(l,k)\right|^2, & \textit{for vowel};\\ \left|Y(l,k)\right|^2\otimes S_w(k), & \textit{for consonant},\end{cases} \qquad (8)$$

were $S_w(k)$ is Bartlett window (the effective window width was varied in the range of 30-280 Hz). The procedure (8) validation was the assumption that averaging on frequency of period grams of consonants may be useful for reducing the variance of the $\lambda_y(l,k)$ estimate. Decision rule "vowel-consonant" had the form:

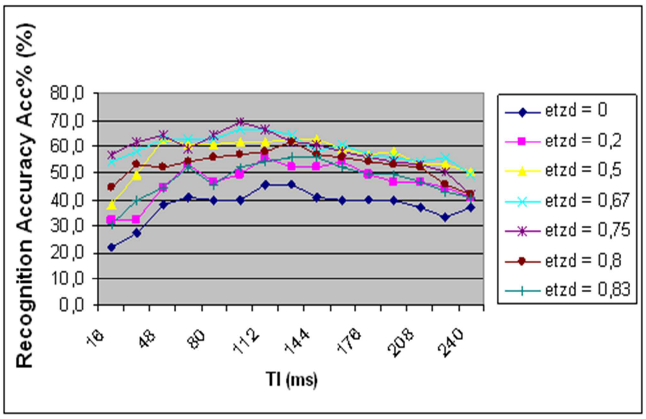
$$phoneme(l)=\begin{cases}vowel, & for\ f_0(l)<1000\,Hz;\\ consonant, & for\ f_0(l)\geq 1000\,Hz,\end{cases}$$

where $f_0(l)$ is the zero-crossings frequency of signal $y(t)$ in the l-th frame.

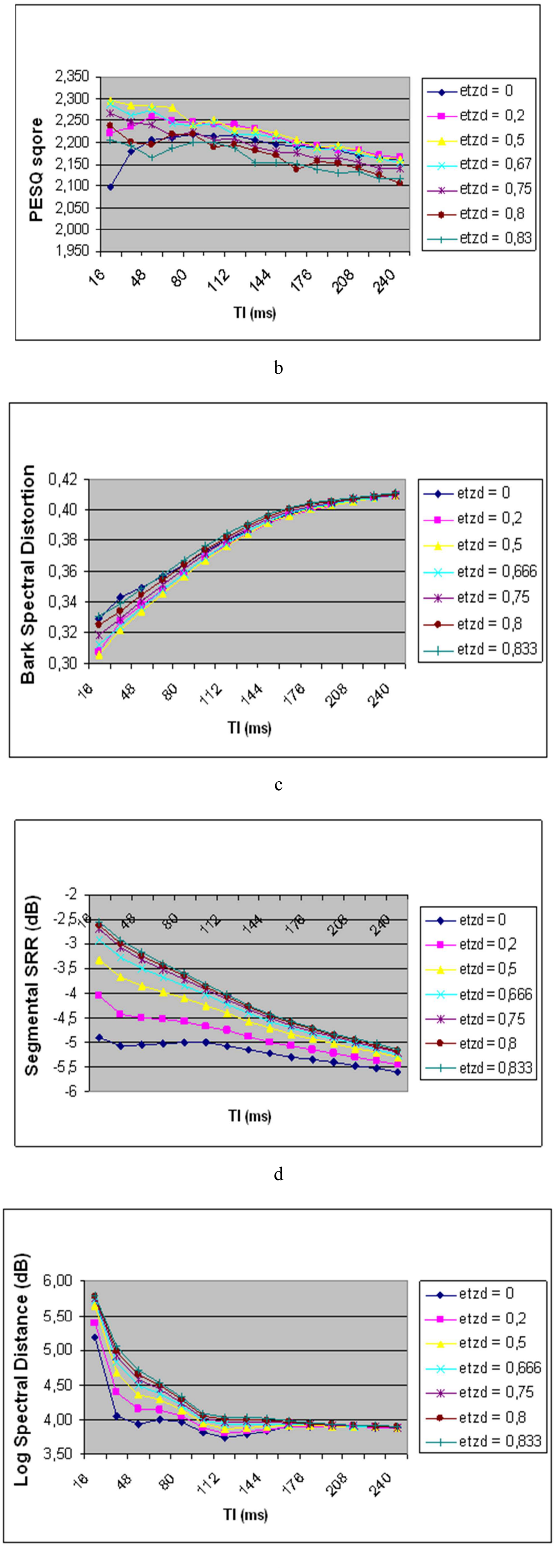
Result of comparing the best curves of Fig. 3,a with similar curve, calculated for $\eta_z^d(k)$, which was chosen as upper-bound of (6) (the curve is denoted as “et zd = old”) is shown in Fig. 4a. It is evident that choice of $\eta_z^d \approx 0.66\ldots0.75$ provides recognition accuracy which is 6% better then one for frequency dependent $\eta_z^d(k)$. As far as speech quality PESQ, results are similar and shown in Fig. 4b. It is evident that choice of $\eta_z^d = 0.5$ provides speech quality, which is aloud better then one for $\eta_z(k)$ chosen as upper-bound of (6).

Experimental studies carried out for the case of T_{60} = 0.74 s, $T_l \approx 100$ ms and $\eta_z^d = 0.5-0.67$, had showed that $Acc\%$ =44% for procedure (8), which is considerably less $Acc\%$ =78% for procedures (3)-(7). Thus, the time averaging in (3)-(7) is much more effective than frequency averaging in (8).

Studying of $Acc\%(T_l)$ and $PESQ(T_l)$ for $\eta_z^d = 0.5$ and $\eta_z^a = m\eta_z^d$ ($m = 0.1-0.9$) shows that $Acc\%$ and PESQ are independent on η_z^a (Fig. 5).



a



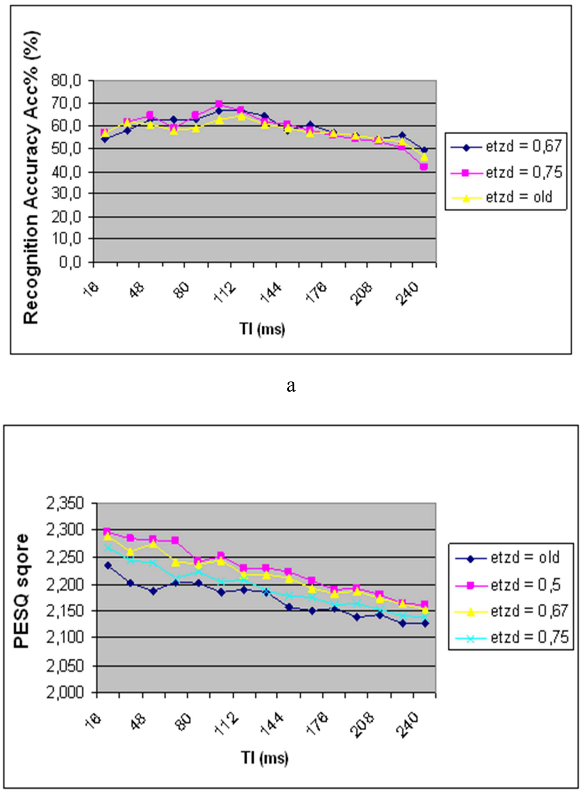
b

c

d

e

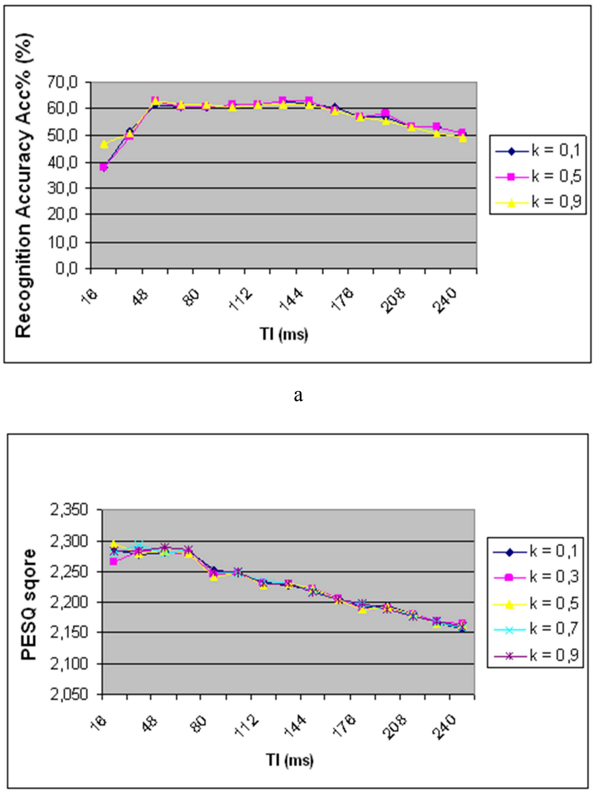
Fig. 3. *Different indicators as functions of* T_l *and* η_z^d



a

b

Fig. 4. *$Acc\%(T_l)$ (a) and $PESQ(T_l)$ (b) for* $\eta_z = const$ *and* $\eta_z(k)$



a

b

Fig. 5. *$Acc\%(T_l)$ (a) and $PESQ(T_l)$ (b) for $\eta_z^a = m\eta_z^d, m = 0.1 - 0.9$,* $\eta_z^d = 0.5$

4. Discussion

Two objectives were formulated in this paper: parameters optimization of late reverberation suppression technique and studying of utility of indexes used for this optimization.

4.1. Quality of Indexes

Presence and severity of extrema in Fig. 3 graphs indicates the ability of the indexes $Acc\%$, PESQ, BSD,SRR and LSD, to inform the user about existence of the parameters T_l and η_z^d optimum values, for which a compromise is reached between the late reverberation suppression and speech distortion. Changing the parameter T_l, as it follows from (2), allows to control the total power of late reverberation spectrum estimator. At the same time, changing the averaging parameter η_z^d, as it follows from (3)-(7), allows to control variance and bias of late reverberation spectrum estimator on different frequencies.

As it can be seen from Fig. 3a, $Acc\%(T_l, \eta_z^d)$ maximum is reached for $T_l \approx 100$ ms and $\eta_z^d \approx 0{,}66 - 0{,}75$, and the maximum is expressed quite clearly. More complicated is the behavior of other indicators. $PESQ(T_l, \eta_z^d)$ (Fig. 3b) has global maximum at $\eta_z^d \approx 0{,}5 - 0{,}66$, $T_l \approx 16$ ms, however, the $PESQ(T_l)$ does not contain a local extremum. At the same time, $PESQ(T_l)$ has local extremum at $T_l \approx 100$ ms and $\eta_z^d \approx 0 - 0{,}2$. Indicator BSD (Fig. 3c) has minimumat $\eta_z^d \approx 0{,}5 - 0{,}66$, however, $BSD(T_l)$ does not contain a local extremum for any η_z^d. Relationships $SRR(T_l)$ and $LSD(T_l)$ (Fig. 3dand 3e) demonstrate existence of weak extrema for $T_l \approx 100$ ms at $\eta_z^d \approx 0 - 0{,}2$. Relationships $SRR(\eta_z)$ and $LSD(\eta_z)$ hasnolocal extreme at any T_l.

Thus, these results indicate disparity between indicators SRR, LSD, BSD and PESQ, as well as their relatively low self-descriptiveness, as compared with indicator $Acc\%$. That is why in the future we prefer using of indicator $Acc\%$, though we will use also indicator PESQ, which is traditional in telecommunications.

4.2. Optimal Parameters Values

Let us compare results reported in this paper with similar results of [3-4] in terms of the choice of parameters T_l and η_z^d. The problem of choosing optimal T_l value was not posed in [3-4], where it was taken $T_l \approx 50$ ms in experimental studies. As follows from Fig. 3 graphs, this T_l choice is somewhat worse (about 8%) compare to $T_l \approx 100$ ms in the sense of indicator $Acc\%$, and is somewhat better (about 0.05) in the sense of indicator PESQ. When listening to the dereverberated signal in case of $T_l \approx 100$ ms, it seems more preferable as compared with the case of $T_l \approx 50$ ms.

Regarding the choice of the parameter η_z^d , analysis of graphs Figs. 4 and 5 shows that simple, in terms of technical implementation, the averaging procedure (3) is more effective compare to more complex procedure (4)-(7). However, parameter $\eta_z = 0{,}9$ value proposed in [3] does not coincide with the value $\eta_z^d \approx 0{,}66 - 0{,}75$ found in our experiments. With regard to ratio (6) proposed in [4], we note two points. Firstly, this ratio had been suggested without any justification. Secondly, it was recognized the approximate nature of (6) in [4]: "…in practice $\eta_z^d(k)$ should be chosen slightly higher than the upper-bound…". Finally, it is surprising that speech constants, such as stationarity duration $T_{stat} \approx$ 30-40 ms [16] and the average phonemes duration $T_{phon} \approx$ 120-150 ms [17], were not considered in [3-4], when choosing averaging parameter η_z^d .

Try to find answers to these questions concerning the choice of η_z^d .

It can be shown that power spectrum $\lambda_y(l,k)$ of reverberated signal $y(t)$ may be regarded as the result of running averaging of clear speech period grams $|X(l,k)|^2$:

$$\lambda_y(l,k) = \frac{\sigma^2}{2\delta}\lambda_{x\,cp}(l,k)\,, \quad (9)$$

$$\lambda_{x\,cp}(l,k) = e^{-2\delta T_{inc}}\,\lambda_{x\,cp}(l-1,k) + \left(1 - e^{-2\delta T_{inc}}\right)|X(l,k)|^2\,,$$

Equation (9) involves averaging parameter

$$\eta_z^d = e^{-2\delta T_{inc}} \approx 1 - 2\delta T_{inc} \approx \frac{1}{1+2\delta T_{inc}}\,, \quad [2\delta T_{inc} << 1]. \quad (10)$$

Matching of (10) and (6) suggests that the ratio (6) had been prepared in the same way. If it is so, then the ratio (6) is not accurate enough since it does not take into account statistical properties of speech signals.

It seems more correct the way of η_z^d choice in which, at first, the effective interval $T_{aver\,eff}$ of adjacent period grams averaging is chosen in accordance with requirement:

$$T_{stat} < T_{aver\,eff} < T_{phon}\,, \quad (11)$$

further, the number of period grams, averaged over the interval $T_{aver\,eff}$ (Fig. 6) is determined:

$$n_{aver\,eff} = (T_{aver\,eff} - T_{frame} + T_{inc})/T_{inc}\,, \quad (12)$$

and, finally, parameter η_z^d value is calculated:

$$\eta_z^d = 1 - \frac{1}{n_{aver\,eff}} = \frac{T_{aver\,eff} - T_{frame}}{T_{aver\,eff} - T_{frame} + T_{inc}}\,. \quad (13)$$

We can now verify the relations (11)-(13) validity. For example, for values T_{frame}=32 ms and T_{inc} =16 ms adopted in the paper, we obtain $T_{aver\,eff}$ = 60…80 ms from (13) for $\eta_z^d \approx 0{,}67\ldots0{,}75$ values, which are optimal in the sense of $Acc\%(\eta_z)$ maximum value. This result is consistent with (11). Another example: there were adopted F_s =8 kHz, T_{frame}=16 ms, T_{inc} =4 ms, η_z^d =0,9 in [5] for $T_{60} = 0.4 - 1.7$ s. Given (13), we can find $T_{aver\,eff}$ = 52 ms, and this value is also in good compliance with (11). Note that it was not taken into account the dependence of η_z^d on T_{60} in (13), though our experiments have shown reality of the dependence.

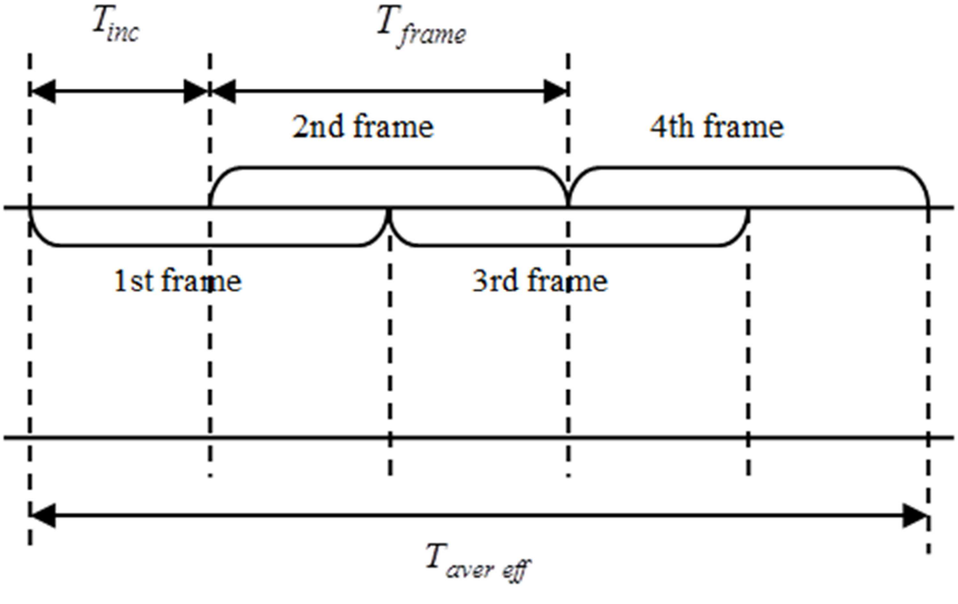


Fig. 6. *Area of effective averaging*

It is possible double interpretation of Fig. 5 graphs behaviour: a relations (5) and (7) do not fulfil their role, or experimental conditions were not correct. It should be recognized that the second variant of the interpretation may be right, because a lot of words in the test speech signal began with the same hissing phoneme "s". Therefore, the question of whether the averaging parameter η_z^d must depend on l , needs of further research.

5. Conclusion

When dereverberator is used as a pre-processor of ASR system, speech recognition accuracy $Acc\%$ has the best ability to inform the user about reached compromise between late reverberation suppression and speech signal distortion. Therefore, it is inexpediently to replace the end-to-end indicator $Acc\%$ by the partial criteria SRR, LSD, BSD and PESQ, commonly used in the examination of telecommunications. Moreover, on the basis of the results obtained, the indicator $Acc\%$ can be recommended as a universal indicator of the speech quality in telecommunication.

The value $T_l \approx 50$ ms of boundary between early reflections and late reverberation, which had been proposed when problems of speech quality and intelligibility were studied, isn't best for ASR systems, for which optimal value is $T_l \approx 100$. It was proposed in the paper to use ratio $\eta_z^d = (T_{aver\ eff} - T_{frame})/(T_{aver\ eff} - T_{frame} + T_{inc})$ for $50\ ms < T_{aver\ eff} < 80\ ms$, when choosing the optimal averaging parameter η_z value. The basis for this proposal is assumption that the averaging time interval should be associated with statistical speech constants, such as the phoneme and stationarity interval durations.

Presented in this paper results had been obtained for case of known RIR, but there are not fundamental obstacles to use them upon blind estimation of $T_{60}(f)$.

References

[1] P. Naylor,N. Gaubitch, Speech Dereverberation, Springer-Verlag: London, 2010.

[2] T. Yoshiokaet al., “Making mashine understand usinreverberantrooms,” IEEE Signal Processing Magazine, Vol. 29, pp. 114-126, November 2012.

[3] K. Lebart, J. Boucher, P. Denbigh, “A new method based on spectral subtraction for speech dereverberation,” Acta Acoustica, Vol. 87, pp. 359-366, April 2001.

[4] E. Habets, Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement, PhD dissertation, Eindhoven, 2007, 257 p.

[5] Y. Ephraim, D. Malah,“Speech enhancement using a minimum mean square error Log-spectral amplitude estimator,” IEEE Trans. Acoust., Speech, Signal Processing,Vol. ASSP-33,pp. 443-445,April 1985.

[6] A. Prodeus, O. Ladoshko,“On existance of optimal boundary value between early reflections and late reverberation,”Proc. of IEEE 34th Int. Sc. Conf. Electronics and Nanotechnology (ELNANO), pp. 442-446, 15-18 April 2014, Kyiv, Ukraine.

[7] A. Prodeus,“Parameters Optimizing of Late Reverberation Spectrum Estimator,” Proc. Xth Int. Conf.“Perspective Technologies and Methods in MEMs Design” (MEMSTECH 2014),pp. 100-103,22-24 June 2014, Lviv, Ukraine.

[8] A. Prodeus, V. Didkovskiy, V. Ovsianyk,“Blind estimation of reverberation time in automatic speech recognition systems,”Information processing systems, No. 7(123), pp. 59-66,Kharkiv, September 2014 (in Russian).

[9] A. Prodeus, O. Ladoshko, “Reverberation suppression systems quality indicators dependency on speech distortion level,” Standartisation, sertification, quality,No. 3(88), pp. 45-49, June 2014 (in Ukrainian).

[10] C.M. Chernick, S. Leigh, K.L. Mills, and R. Toense, “Testing the Ability of Speech Reconizers to Measure the Effectiveness of Encoding Algorithms for Digital Speech Transmission,” IEEE Int. Military Comm. Conf. (MILCOM), 1999.

[11] W. Jiang, H. Schulzrinne,"Speech Recognition Performance as an Effective Perceived Quality Predictor," IEEE Int. Workshop on Quality of Service, pp. 269-275, 2002.

[12] W. Liu, K. Jellyman, J. Mason, N. Evans, "Assessment of Objective Quality Measures for Speech Intelligibility Estimation," IEEE Int. Conf. on Acoustics, Speech and Signal Processing, (ICASSP 2006), Vol. 1, May 14-19, 2006.

[13] J. Beerends, E. Larsen, N. Iyer, J. van Vugt, “Measurement of speech intelligibility based on the PESQ approach,”Proc. Int. Conf. “Measurement of Speech and Audio Quality in Networks” (MESAQIN), Prague, Czech Republic, 2 June 2004.

[14] S. Young, HMMs and Related Speech Recognition Technologies, Springer Handbook of Speech Processing, ed. J.Benesty et al., Berlin Heidelberg: Springer-Verlag, 2008.

[15] R.M. Schroeder, “New method of measuring reverberation time,” J. Acoust. Soc. Am., Vol. 37, pp. 409-412, 1965.

[16] J. R. Deller, J. G. Proakis, J. H. L. Hansen, Discrete-Time Processing of Speech Signals. Macmillan Publishing Company, New York, NY, 1993.

[17] B. Ziolko, M. Ziolko, “Time durations of phonemes in polish language for speech and speaker recognition,” in Human Language Technology, Vol. 6562. Berlin Heidelberg: Springer-Verlag, 2011, pp.105-114.