AASCIT Communications
Volume 3, Issue 1
January 26, 2016 online
ISSN: 2375-3803

AASCIT American Association for Science and Technology

Similarities and Differences of Speech Recognition Accuracy and Speech Quality Measures Behavior

Arkadiy Prodeus Acoustic and Electroacoustic Department, Faculty of Electronics, National Technical University of Ukraine, Kyiv, Ukraine

Keywords

Noise, Late Reverberation, Reduction, Algorithm, Speech Quality Measure, Automatic Speech Recognition Accuracy

Noise and late reverberation reduction algorithms were compared by means of objective speech quality and speech recognition accuracy (Acc%) measures. Negative effects of excessive noise reduction for automatic speech recognition (ASR) had been shown. It was found possibility of improvement the noise suppression algorithms quality, in terms of Acc%, by proper choice of a priori signal-to-noise assessment technique. It was shown that decision-directed technique is the best for speech quality, when "rough" assessment technique is the best for ASR, and the maximum likelihood technique occupies an intermediate position. When studying late reverberation suppression algorithms, it was found existence of optimal, in terms of Acc%, parameters values of the algorithms. It was shown also that these parameters values are different for ASR and for speech enhancement. Thus, late reverberation suppression algorithms behavior is similar to one of noise suppression algorithms. Study of speech quality measures had showed that only few of them were in good agreement with Acc%. But existence of such measures is very important, because it enables use them instead of Acc% and, thus, enables essentially simplify assessment of noise and reverberation robustness in ASR.

Introduction

Noise and late reverberation reduction pre-processors (Fig. 1) are widely used for speech quality improvement in embedded and communication systems, hearing aids, as well as in ASR [1, 2, 3, 4].

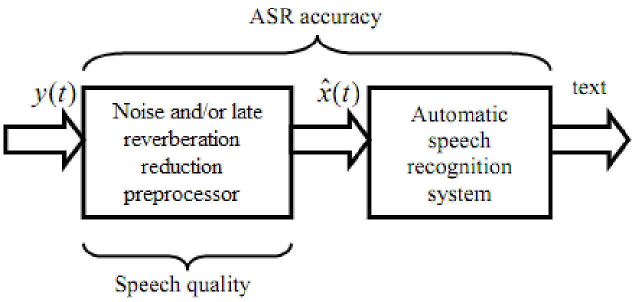

ASR accuracy
$y(t)$
Noise and/or late reverberation reduction preprocessor
$\hat{x}(t)$
Automatic speech recognition system
text
Speech quality

Fig. 1. *Noise and/or late reverberation reduction system as ASR pre-processor.*

Two relatively recent proposed noise reduction algorithms named Wiener-TSNR and Wiener-HRNR [5, 6] were compared in [7] with a set of traditionally used algorithms, such as spectral subtraction (SpecSub), Wiener filtering (Wiener), minimum mean-square error amplitude spectrum estimator (MMSE) and minimum mean-square error log-spectral amplitude estimator (logMMSE) algorithms. It was shown that Wiener-TSNR and Wiener-HRNR algorithms have a serious drawback: when radically suppressing residual noise, they significantly reduce ASR robustness. Since the Wiener-TSNR and Wiener-HRNR algorithms are based on a special correction of a priori SNR evaluation, one would hope to improve the algorithms quality by changing of averaging parameter value of "decision-directed" technique [2] used for a priori SNR estimation. Checking the validity of this assumption had been made in [8] and it was shown that improvement of the algorithms quality in terms of Acc% is possible. Moreover, it was shown also usefulness of other a priori SNR estimation techniques.

Some features of late reverberation reduction algorithms were studied in [9] and it was found that excessive late reverberation suppression is harmful for ASR. At the same time it was shown in [9] that speech quality is much less sensitive to excessive late reverberation reduction.

As it was shown in [7, 8, 9], objective speech quality measures [10, 11] are not in good agreement with Acc% measure, though there can be find a few measures with satisfactory characteristics. Existence of such measures is very important, because it enables use them instead of Acc% and, thus, it enables essentially simplify assessment of noise and reverberation robustness in ASR [12].

In this paper, the results of [7-9] are generalized and equipped with extended comments.

Noise and Late Reverberation Reduction Algorithms

Noise reduction algorithm transforms the mixture $y(t) = x(t) + n(t)$ by means of operator $A\{\cdot\}$ to get restored signal $\hat{x}(t) = A\{y(t)\}$.

Frequency domain speech enhancing technique is the most popular:

$$\left[\hat{\lambda}_{\hat{x}}(f,m)\right]^{1/2} = \hat{G}(f,m)\cdot\left[\hat{\lambda}_y(f,m)\right]^{1/2}$$

where $\hat{\lambda}_y(f,m)$ and $\hat{\lambda}_{\hat{x}}(f,m)$ are power spectrum estimators of the m th frame of signal $y(t)$ and restored signal $\hat{x}(t)$, respectively; $\hat{G}(f,m)$ is estimator of transfer function for each m th frame. Phase spectrum of signal $y(t)$ is usually used when signal $\hat{x}(t)$ is calculated. SpecSub, Wiener, MMSE and logMMSE algorithms [1, 2, 3], and also Wiener-TSNR and Wiener-HRNR algorithms [5, 6] are considered in this paper.

Traditional Noise Reduction Algorithms

Traditional noise reduction algorithms discussed in this paper are SpecSub, Wiener, MMSE and logMMSE algorithms [1, 2, 3, 4], where $\hat{G}(f,m)$ is calculated as follows:

$$\hat{G}_{SpecSub}(f,m) = \left(\frac{\hat{\gamma}(f,m)-1}{\hat{\gamma}(f,m)}\right)^{1/2} \tag{1}$$

$$\hat{G}_{Wiener}(f,m) = \frac{\hat{\xi}(f,m)}{1+\hat{\xi}(f,m)} \tag{2}$$

$$\hat{G}_{MMSE}(f,m) = \Gamma(1{,}5)\sqrt{\frac{\hat{v}(f,m)}{\hat{\gamma}^2(f,m)}}\exp(-\frac{\hat{v}(f,m)}{2})[(1+\hat{v}(f,m))I_0(\frac{\hat{v}(f,m)}{2}) + vI_1(\frac{\hat{v}(f,m)}{2})] \tag{3}$$

$$\hat{G}_{\log MMSE}(f,m) = \frac{\hat{\xi}(f,m)}{1+\hat{\xi}(f,m)}\exp\left\{\frac{1}{2}\int_{\hat{v}(f,m)}^{\infty}\frac{e^{-t}}{t}dt\right\} \tag{4}$$

where $\hat{\xi}(f,m)$ is a priori signal-to-noise ratio (SNR) estimator, $\hat{\gamma}(f,m) = \hat{\lambda}_y(f,m)/\hat{\lambda}_n(f,m)$ is a posteriori SNR, $\hat{v}(f,m) = \hat{\xi}(f,m)\hat{\gamma}(f,m)/[1+\hat{\xi}(f,m)]$, $\Gamma(\cdot)$ is gamma function, $I_0(\cdot)$ and $I_1(\cdot)$ are modified Bessel functions of zero and first order, respectively.

"Decision-directed" method had been proposed in [2] for $\hat{\xi}(f,m)$ calculation:

$$\hat{\xi}_{DD}(f,m) = \alpha\cdot\hat{\lambda}_{\hat{x}}(f,m-1)/\hat{\lambda}_n(f,m-1) + (1-\alpha)\cdot P[\hat{\gamma}(f,m)-1], \quad 0 \le \alpha \le 1, \tag{5}$$

$$P(x) = \begin{cases} x, & x \geq 0; \\ 0, & x < 0. \end{cases}$$

Radical Noise Reduction Algorithms

Wiener-TSNR and Wiener-HRNR algorithms had been proposed relatively recently [5, 6]. Their noise suppression action is much more efficient compared to the aforementioned traditional algorithms. The word «Wiener» in their names means that the transfer functions of the correction filters are formed according to (2). However, this does not mean that the transfer functions are prohibited from forming a different way.

Wiener-TSNR transfer function is formed in two steps.

Step 1:

$$\hat{\xi}_{TSNR}(f,m) = \hat{\xi}_{DD}(f,m+1) \approx \hat{\lambda}_{\hat{x}}(f,m) / \hat{\lambda}_n(f) \tag{6}$$

Step 2:

$$\hat{G}_{TSNR}(f,m) = \frac{\hat{\xi}_{TSNR}(f,m)}{1+\hat{\xi}_{TSNR}(f,m)} \tag{7}$$

When noise suppression is strong as is the case of Wiener-TSNR algorithm, speech signal components are also suppressed intensively. Wiener-HRNR algorithm was proposed for regeneration of the lost signal components. This procedure consists of three steps.

Step 1. Output of TSNR algorithm (or other noise reduction algorithm) is used as input of half-wave rectifier:

$$s_{harm}(t) = \hat{s}(t) \cdot P[\hat{s}(t)] \tag{8}$$

Step 2. Using (8), a priori SNR is calculated:

$$\hat{\xi}_{HRNR}(f,m) = \rho(f,m) \cdot \hat{\lambda}_{\hat{x}}(f,m) / \hat{\lambda}_n(f)[1-\rho(f,m)] \cdot \hat{\lambda}_{harm}(f,m) / \hat{\lambda}_n(f) \tag{9}$$

where $\hat{\lambda}_{harm}(f,m)$ is power spectrum estimator of signal $s_{harm}(t)$; $\rho(f,m)$ ($0 \leq \rho(f,m) \leq 1$) is weight coefficient. Although there is a certain freedom of $\rho(f,m)$ choice, it was proposed designate $\rho(f,m) = \hat{G}_{TSNR}(f,m)$ in [5].

Step 3. Transfer function for HRNR algorithm is formed:

$$\hat{G}_{HRNR}(f,m) = \frac{\hat{\xi}_{HRNR}(f,m)}{1+\hat{\xi}_{HRNR}(f,m)} \tag{10}$$

It is naturally to assume that the ability of Wiener-TSNR and Wiener-HRNR algorithms radically suppress the noise is balanced by unpleasant consequence such as unacceptably high distortion of the speech signal. Therefore it is important to verify the validity of this assumption.

A Priori SNR Estimation Technique

Important component of almost any noise suppression algorithm is a priori SNR $\hat{\xi}(f,m)$ estimation technique [2]. Apart from "decision-directed" (DD) technique (5), other two techniques considered here are maximum likelihood (ML) and "rough" (RO) techniques

$$\hat{\xi}_{ML}(f,m) = P(\overline{\gamma}(f,m) - 1) \tag{11}$$

$$\overline{\gamma}(f,m) = \alpha' \cdot \overline{\gamma}(f,m-1) + (1-\alpha') \cdot \hat{\gamma}(f,m) \,/\, \beta, \qquad 0 \leq \alpha' \leq 1, \beta \geq 1$$

$$\hat{\xi}_{RO}(f,m) = \hat{\gamma}(f,m) - 1 \qquad (12)$$

Late Reverberation Suppression

For reverberated signal $y(t) = x(t) \otimes h(t)$, where $\otimes$ is convolution symbol, room impulse response (RIR) $h(t)$ can be decomposed as follows [13]

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

T_l is time, corresponding to boundary between early reflections and late reverberation. Thus, reverberated signal $y(t)$ can be represented as

$$y(t) = h_i(t) \otimes x(t) + r(t) \qquad (13)$$

where $r(t) = h_l(t) \otimes x(t - T_l)$ is late reverberation component. Terms of (13) are statistically independent, thus late reverberation may be interpreted as kind of noise and its suppression can been made by means of noise reduction algorithm. Late reverberation power spectrum $\lambda_r(l,k) = e^{-2\delta(k)T_l} \cdot \lambda_y(l - N_l, k)$ is used in this case against of noise power spectrum, where $N_l = T_l / T_{inc}$; T_{inc} denotes the frame shift value; $\delta(k) = 2\ln 10 \big/ T_{60}(k)$; $T_{60}(k)$ is reverberation time [13]. Averaging parameter η_z ($0 \le \eta_z < 1$) is used for averaging time interval controlling

$$\hat{\lambda}_y(l,k) = \eta_z^d \cdot \hat{\lambda}_y(l-1,k) + (1 - \eta_z^d)|Y(l,k)|^2 \qquad (14)$$

where $Y(l,k)$ is spectrogram of $y(t)$. Recommendations for choosing optimal, in terms of Acc% maximum, parameters T_l and η_z^d values can be found in [9].

Quality Measures

When speech enhancement algorithm is used as ASR pre-processor, it is naturally use end-to-end quality measure

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

where 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 [12].

For speech enhancement algorithms used in communication systems or in applications for people with hearing loss, it is naturally use speech quality indicators. Nine speech quality measures were explored in [7-9]. They are segmental SNR (SSNR), logarithmic spectral distortion (LSD), weighted spectral slope distance (WSS), log-likelihood ratio (LLR), Itacura-Saito distance (IS), cepstral distance (CEP), composite index [SCI, NCI, OCI], bark spectral distortion (BSD) and perceptual evaluation of speech quality (PESQ).

Measures SSNR, LSD and BSD are described as follows:

$$SSNR = \frac{1}{L}\sum_{l=1}^{L} 10\lg\left[\frac{\sum_{n=Rl}^{Rl+N-1} x^2(l,n)}{\sum_{n=Rl}^{Rl+N-1} [x(l,n) - y(l,n)]^2}\right] \qquad (15)$$

$$LSD = \frac{2}{RL}\sum_{l}\sum_{r=0}^{\frac{R}{2}-1}\left|G\{X(l,r)\} - G\{Y(l,r)\}\right| \qquad (16)$$

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

$$BSD = \frac{\sum_{l=1}^{L}\sum_{k=1}^{K}\left[B_x(l,k) - B_y(l,k)\right]^2}{\sum_{l=1}^{L}\sum_{k=0}^{\frac{K}{2}-1}\left[B_x(l,k)\right]^2} \tag{17}$$

where $x(l,n)$ and $\hat{x}(l,n)$ are n th samples of l th frames of clean speech signal $x(t)$ and enhanced signal $\hat{x}(n)$, respectively, $X(l,k)$ and $\hat{X}(l,k)$ are spectrograms of signals $x(l,n)$ and $\hat{x}(l,n)$, respectively; $B\{X(l,k)\}$ and $B\{\hat{X}(l,k)\}$ are bark spectrums of signals $x(l,n)$ and $\hat{x}(l,n)$, respectively, k is critical bandwidth number.

WSS is described as follows:

$$WSS = \frac{1}{M}\sum_{m}^{M-1}\frac{\sum_{j=1}^{K} W(j,m)(S_c(j,m) - S_p(j,m))^2}{\sum_{j=1}^{K} W(j,m)} \tag{18}$$

where $W(j,m)$ are weighting coefficients, K is the number of bands, M is the total number of frames in the signal, $S_c(j,m)$, $S_p(j,m)$ are the spectral slopes for j th frequency band at m th frame of clean and processed speech signals, respectively.

The LLR, IS and CEP measures are calculated for each frame as follows:

$$LLR_{seg} = \ln\left(\frac{\vec{a}_p \mathrm{R}_c \vec{a}_p^T}{\vec{a}_c \mathrm{R}_c \vec{a}_c^T}\right) \tag{19}$$

$$IS_{seg} = \frac{\sigma_c^2}{\sigma_p^2}\left(\frac{\vec{a}_p \mathrm{R}_c \vec{a}_p^T}{\vec{a}_c \mathrm{R}_c \vec{a}_c^T}\right) + \ln\left(\frac{\sigma_c^2}{\sigma_p^2}\right) - 1 \tag{20}$$

$$CEP_{seg} = \frac{10}{\ln 10}\sqrt{2\sum_{k=1}^{p}[c_c(k) - c_p(k)]^2}\ ,\ c(m) = a_m + \sum_{k=1}^{m-1}\frac{k}{m}c(k)a_{m-k},\quad 1 \le m \le p \tag{21}$$

where $\vec{a}_c$ and $\vec{a}_p$ are LPC vectors of the clean and enhanced speech signals frames, respectively; R_c is the autocorrelation matrix of the original speech signal; σ_c^2 and σ_p^2 are the LPC gains of the clean and enhanced signals, respectively; $c(k)$ are the cepstrum coefficients; p is the order of the LPC analysis.

Composite measure [SCI, NCI, OCI] consists of three "elementary" measures [11]: signal distortion composite index (SCI), noise distortion composite index (NCI), and overall distortion composite index (OCI):

$$\begin{aligned} SCI &= 3{,}093 - 1{,}029 \cdot LLR + 0{,}603 \cdot PESQ - 0{,}009 \cdot WSS, \\ NCI &= 1{,}634 + 0{,}478 \cdot PESQ - 0{,}007 \cdot WSS + 0{,}063 \cdot SSNR, \\ OCI &= 1{,}594 + 0{,}805 \cdot PESQ - 0{,}512 \cdot LLR - 0{,}007 \cdot WSS. \end{aligned} \tag{22}$$

PESQ estimation algorithm description is bulky, its description can be found in [14].

Experimental Results and Discussion

Clean speech signals (single words) were recorded in anechoic room and had been used for ASR system training. Parameters of digitized sounds are as follows: sampling rate 22050 Hz, linear quantization 16 bit, overall SNR about 35 dB.

Noised signals were simulated by adding a discrete white noise to the clean speech signal. This operation was implemented in

Matlab, ensuring a wide range of overall SNR values [–10,+30] dB. 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.

Signals processing was also implemented in Matlab. Signal’s 32 ms frames with 50% overlapping and Hamming window were used for noise and late reverberation suppression.

Toolkit *composite* [15] was used for calculation of SSNR, LLR, WSS, and complex measure [SCI, NCI, OCI]. This toolkit was used also, after some editing, for calculation of IS and CEP distances. Some routines from toolkit *rastamat* [16, 17] were used for BSD calculation. Indicator PESQ was estimated for wideband PESQ version [18, 19].

Toolkit HTK [12] had been used for ASR simulation. Training of ASR system had been made with usage of 269 samples of 27 words of clean speech recorded for two speakers-women. Noised discrete speech signals (with 0.2…0.5 s pauses between single words) were used as test signals, and there were presented in testing all 27 words used in training. There were 27 phonemes of Ukrainian language in phoneme vocabulary and 39 MFCC_0_D_A coefficients were used when ASR simulating.

Comparison of Noise Reduction Algorithms and Quality Measures

Calculation results of Acc% and quality measures (15)-(21) for different SNR values are shown in Figs. 2-5.

As can be seen from Fig. 2, Acc% and perceptual measures PESQ and BSD are not in good matching. For example, when SNR>10 dB, Wiener algorithm is noticeably inferior than SpecSub, MMSE and logMMSE algorithms in terms of Acc%, while in terms of PESQ and BSD measures the algorithm is quite competitive.

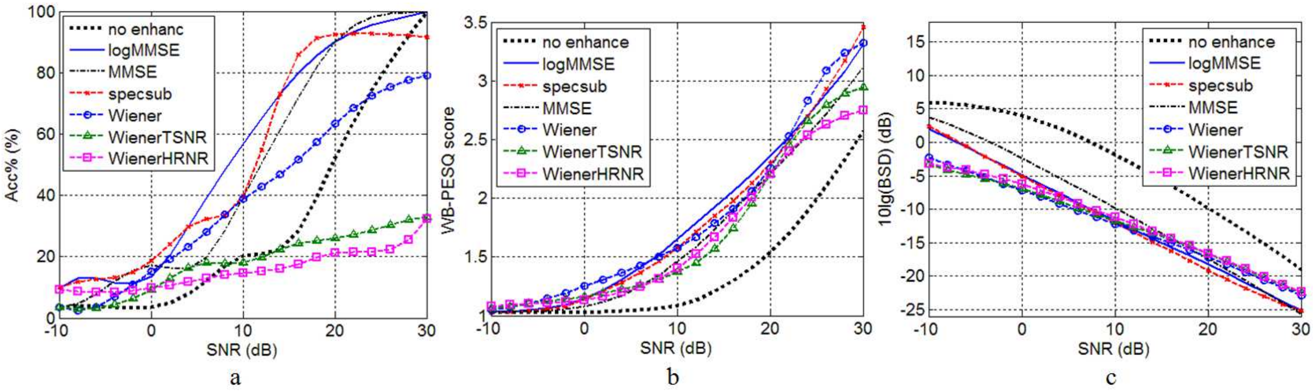


Fig. 2. *Dependency of Acc% (a), WB-PESQ (b) and BSD (c) from SNR.*

Measures SSNR and LSD (Fig. 3) also indicate that the group of "Wiener" algorithms is best for SNR<5 dB, however when SNR>10 dB, SSNR and LSD values are almost the same for all algorithms.

Measures IS, CEP and WSS (Fig. 4) differ from other indicators. They evaluate the action of almost all considered noise reduction algorithms as negative and the group of "Wiener" algorithms is defined as the worst in almost the entire range of SNR values.

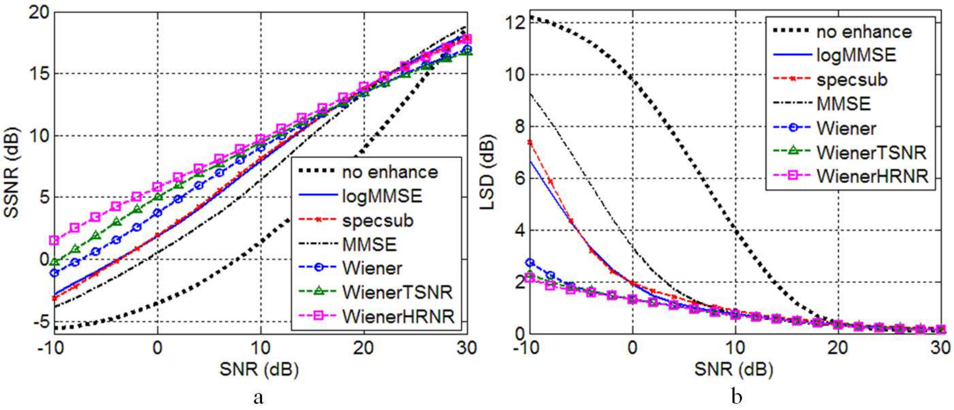


Fig. 3. *Dependency of SSNR (a) and LSD (b) from SNR.*

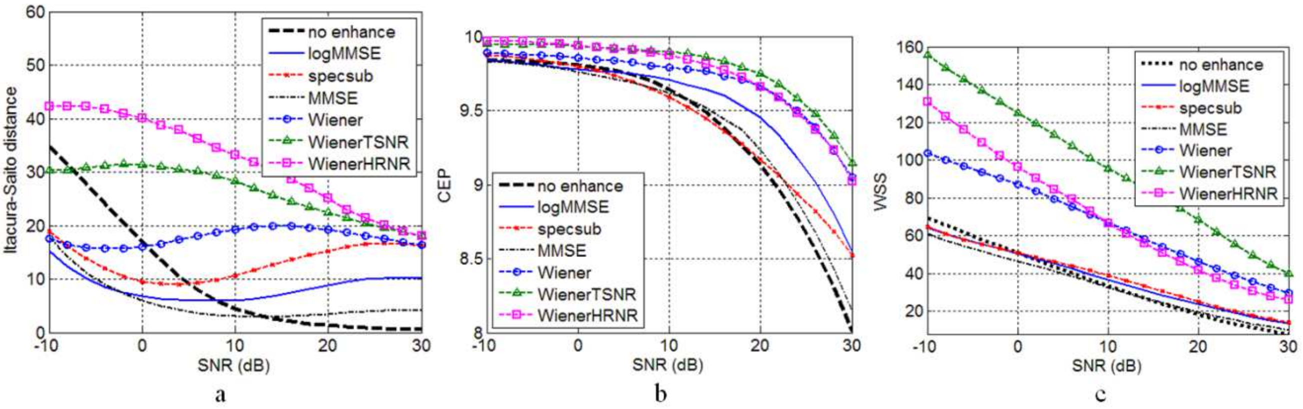


Fig. 4. *Dependency of IS (a), CEP (b) and WSS (c) from SNR.*

Among studied measures, only SCI and LLR were in sufficiently good agreement with the Acc% measure (Fig. 5), indicating that Wiener-TSNR and Wiener-HRNR algorithms are harmful for SNR>10...15 dB. Note that good matching of SCI and LLR could have been predicted, given the relationship (18) [15].

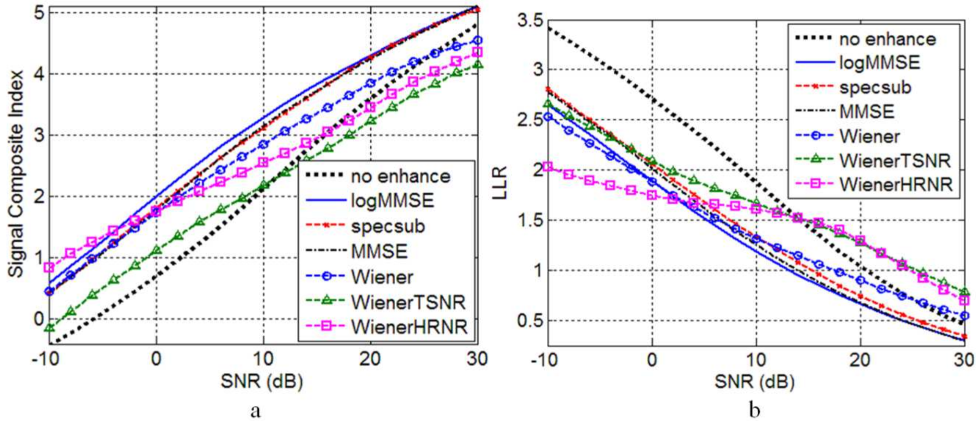


Fig. 5. *Dependency of SCI (a) and LLR (b) from SNR.*

At the same time, the SCI and LLR measures are unable, in contrast to Acc% measure, display a significant efficiency difference between MMSE, logMMSE and SpecSub algorithms as ASR pre-processors.

These results seem some surprising because they do not agree with ones for Wiener-TSNR and Wiener-HRNR algorithms [5, 6]. But the fact that these algorithms strongly suppress noised spectral samples should be alerted anyone, because signal components are also strongly suppressed. It means that consonants which are much more important than vowels for speech recognition and whose power is comparable with the noise power, is strongly distorted too.

Comparison of A Priori SNR Assessment Techniques

Since the Wiener-TSNR and Wiener-HRNR algorithms are based on a special correction of a priori SNR evaluation, one would hope to improve the algorithms quality by changing value of averaging parameter α for "decision-directed" technique. Optimal α value, in terms of subjective speech quality, is $\alpha = 0.98$ for $F_s = 8$ kHz sample rate and $N_{inc} = 64$ frame shift [2]. It can be shown that for arbitrary F_s and N_{inc} optimal is $\alpha_{opt} = \exp(-N_{inc} \big/ (F_s \cdot 0.396))$.

Effectiveness of ML and RO techniques was studied in [2] in terms of subjective speech quality. Speech quality objective measures and Acc% measure were used in [8]. Below is an excerpt of these results.

When varying α values in (5), the results of speech recognition are essentially changed (Fig. 6). For convenience, the values of the time constant τ_{avr} corresponding to different $\alpha = \exp(-N_{inc} \big/ (F_s \cdot \tau_{avr}))$ are shown in Table 1, where τ_{Ephr} values calculated for $F_s = 8000$ Hz and $N_{inc} = 64$ are shown too.

Table 1. *Time constant values.*

α	0.02	0.5	0.9	0.95	0.98
τ_{avr} , s	0.004	0.023	0.152	0.312	0.792
τ_{Ephr} , s	0.002	0.012	0.076	0.156	0.396

As can be seen from Fig. 6, recommended in [2] optimal, in terms of speech quality, value $\alpha = 0.98$ is optimal in terms of Acc% measure only for SNR<10…15 dB, when Acc% values are low. But for SNR>15 dB, Acc% values for α =0.02…0.5 are much higher (10-20% for Wiener and up to 50% for Wiener-TSNR algorithms) than ones for α =0.98. It means that for SNR>15 dB ASR is sufficiently more sensitive to the speech signal distortions than to the residual noise.

The situation is quite different for the assessments of the speech quality. As it follows from Fig. 7 graphs, value $\alpha \approx$ 0.95 is the best in terms of WB-PESQ, and values $\alpha \approx$ 0,02…0,5 are the worst for all considered range of SNR values. This result is in good agreement with optimal value $\alpha \approx$ 0.98 pointed in [2].

When studying the parameter α' values action in ML technique (11), we restrict ourselves to the Wiener algorithm (Fig. 8).

As it can be seen, the value $\alpha' = 0.8$ can be considered as optimal in terms of Acc%. It is interesting that proper values of Acc% are very close to ones obtained by DD technique for $\alpha = 0.02...0.5$.

Because of $\lim\limits_{\alpha\to 0} \hat{\xi}_{DD}(f,m) \approx \hat{\gamma}(f,m) - 1 = \hat{\xi}_{RO}(f,m)$, one would suggest that RO technique (12) can be much more preferable for ASR when SNR>15 dB. It can be expected also that ML technique occupies an intermediate position between RO and DD techniques. Graphs shown in Fig. 9 confirm the validity of these assumptions.

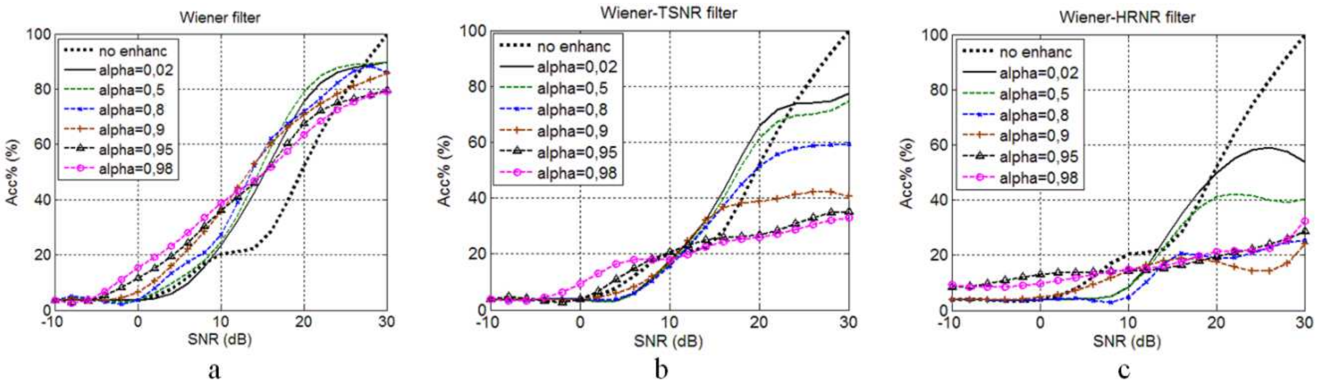


Fig. 6. *Dependency of Acc%(SNR) on α for algorithms: Wiener (a), Wiener-TSNR (b) and Wiener-HRNR (c).*

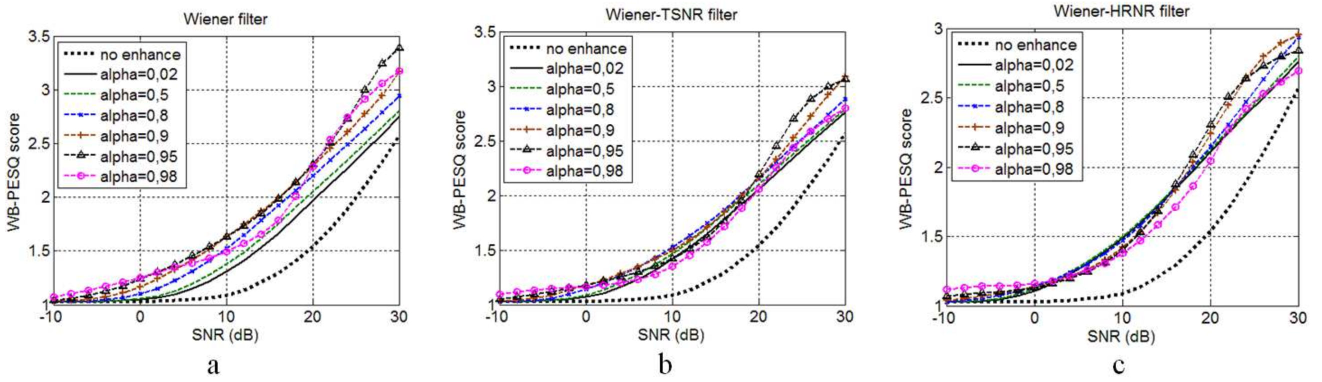
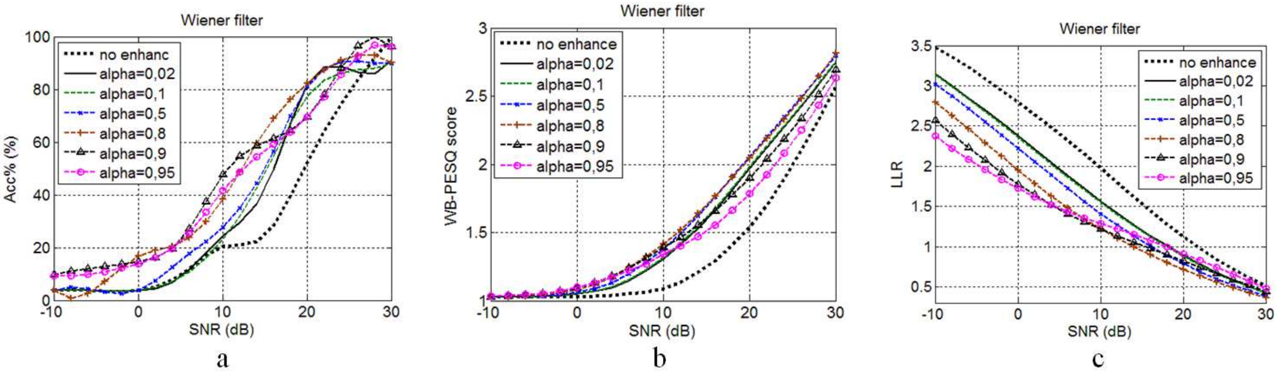


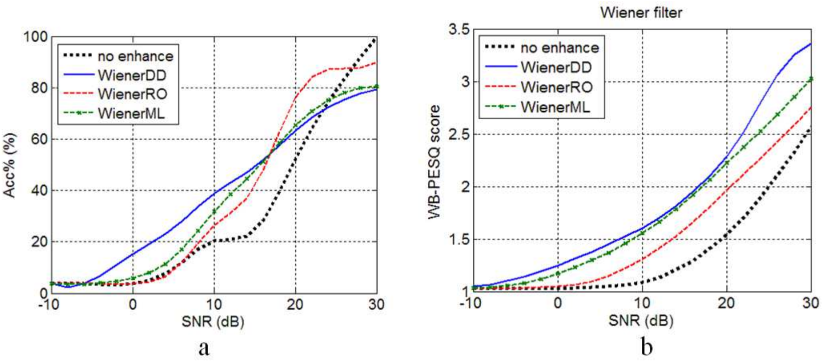
Fig. 7. *Dependency of WB-PESQ ,^$(SNR) on α for algorithms: Wiener (a), Wiener-TSNR (b) and Wiener-HRNR (c).*

Here curve WienerDD corresponds to the DD technique for $\tau_{avr} = 0.396$ which is equivalent to α =0.98 pointed as optimal in [2], and curve WienerRO corresponds to the RO technique. Curve WienerML corresponds to the ML technique where $\alpha' = 0.5273, \quad \beta = 2$ which is equivalent to optimal values $\alpha' = 0.725, \beta = 2$ pointed in [2].



a b c

Fig. 8. *Acc% (a), WB-PESQ (b) and LLR (c) measures for ML technique.*



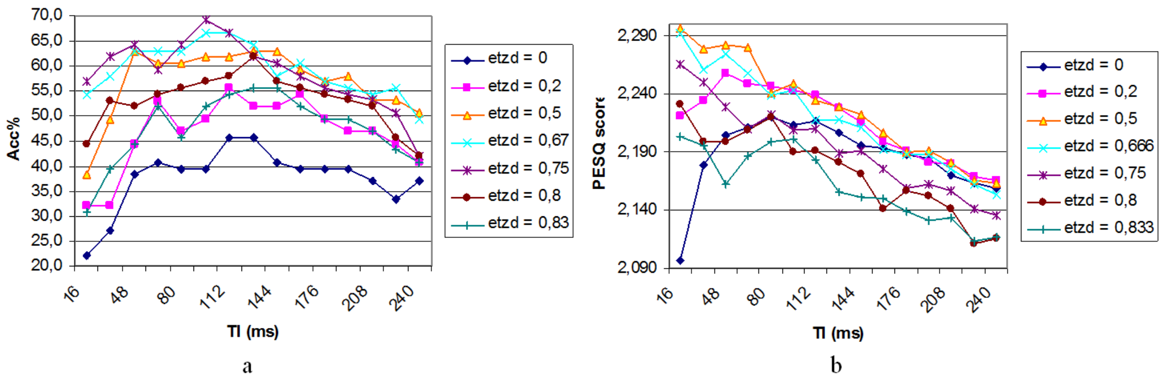
a b

Fig. 9. *Comparison of DD, ML and RO techniques for Acc% (a) and WB-PESQ (b) measures.*

Optimization of Late Reverberation Suppression

It was shown in [9] that when late reverberation is suppressed, boundary value $T_l \approx 100$ ms is the best for ASR systems (Fig. 10a), whereas $T_l \approx 50$ ms value is the best for speech quality [13]. This fact can be explained as sign of high sensitivity of ASR systems to excessive late reverberation reduction, which leads to inappropriate speech signal distortion. Parameter $\eta_z^d \approx 0,66\ldots0,75$ values in (14) are optimal in terms of Acc%.

PESQ measure graphs (Fig. 10b) have no extreme on parameter T_l for $\eta_z^d \approx 0,66\ldots0,75$, but there is extreme when $\eta_z^d \approx 0$ with proper boundary value $T_l \approx 80\ldots100$ ms. It can be shown that signal-to-reverberation ratio SRR (which is identical to SSNR) and LSD measures have similar properties [9], that can be interpreted as low fitness of these speech quality measures to inform us about Acc% reliability when late reverberation suppression is made. Thus, in the future, it is advisable to test the usefulness of LRR and SCI measures for this task.



a b

Fig. 10. *Different indicators as functions of* T_l *and* η_z^d.

Conclusion

Comparison of noise and late reverberation reduction algorithms had showed a risk of excessive reduction of interferences, which leads to strong signal distortions and, as consequence, to essentially reduced robustness of ASR. At the same time speech quality is much less sensitive to excessive reduction of interferences. This result was confirmed after comparing of three techniques of a priori SNR assessment, when it was founded that "rough" assessment technique with its noticeable residual noise is the best for ASR. At the same time decision-directed technique is the best for speech quality, and maximum likelihood technique occupies an intermediate position.

Analysis of speech quality measures reliability showed that when the noise reduction algorithm is used as pre-processor of ASR system, two speech quality measures - LLR and SCI – from nine considered ones are in quite satisfactory agreement with the speech recognition accuracy Acc%. The practical usefulness of the result is the ability to simplify testing of noise reduction algorithms which are used as ASR pre-processors. It is advisable in the future to test the usefulness of LLR and SCI measures for late reverberation suppression task. ■

Arkadiy Prodeus

Arkadiy Prodeus is professor of the Acoustics and Acoustoelectronics Department, National Technical University of Ukraine "Kyiv Polytechnic Institute" (NTUU "KPI"). His research interests are in the areas of digital acoustic signal processing, automatic speech recognition.
Email address: aprodeus@gmail.com

References

[1] J. Benesty, M. Sondhi, Y. Huang (Ed.), Springer Handbook of Speech Processing. Berlin Heidelberg: Springer-Verlag, 2008.

[2] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," IEEE Transactions on Acoustic, Speech, and Signal Processing, vol. ASSP-32, No. 6, Dec. 1984.

[3] Y. Ephraim and 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, 1985.

[4] S. Naida, "Acoustic Theory Problems of Speech Production in the Light of the Discovery of the Formula for the Middle Ear Norm Parameter," Proc. of IEEE 35th Int. Sc. Conf. Electronics and Nanotechnology (ELNANO), pp. 347-350, 21-24 April 2015, Kyiv, Ukraine.

[5] C. Plapous, C. Marro, P. Scalart, "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement," IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, Is. 6, pp. 2098-2108, Nov. 2006.

[6] C. Plapous, C. Marro, P. Scalart, and L. Mauuary, "A Two-Step Noise Reduction Technique," IEEE Int. Conf. on Acoustics, Speech and Signal Proc., Vol. 1, pp. 289–292, 17–21 May, 2004.

[7] A. Prodeus, "Performance measures of noise reduction algorithms in voice control channels of UAVs," Proc. of IEEE 3rd Int. Conf. «Actual Problems of Unmanned Aerial Vehicles Developments», pp. 189-192, October 13-15, 2015, Kyiv, Ukraine.

[8] A. N. Prodeus, V. S. Didkovskyi, "Assessment of a priori signal-to-noise ratio in noise reduction algorithms," Data Processing System, Kharkiv, pp. 29-34, 2015 (in Russian).

[9] A. Prodeus, "Parameter Optimization of the Single Channel Late Reverberation Suppression Technique," Proc. 35th International Conference on Electronics and Nanotechnology (ELNANO-2015), pp. 269-274, 2015, Kyiv, Ukraine.

[10] S. Quackenbush, T. Barnwell, M. Clements, "Objective Measures of Speech Quality," Prentice Hall, Englewood Cliffs, NJ, 1988.

[11] Y. Hu, P. Loizou, "Evaluation of objective quality measures for speech enhancement," IEEE Transactions on Speech and Audio Processing, 16(1), pp. 229-238, 2008.

[12] S. Young et al, "The HTK Book," Cambridge: University Engineering Department, 2009.

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

[14] 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), 2 June, 2004, Prague, Czech Republic.

[15] P. Loizou, "Speech enhancement: Theory and Practice," Boca Raton: CRC Press, 2007.

[16] D. Ellis, “PLP and RASTA (and MFCC, and inversion) in Matlab,” [Online]. Available: http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

[17] M. Brooks, “VOICEBOX: Speech Processing Toolbox for MATLAB,” Imperial College London, Electrical Engineering Department, 2014. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[18] Recommendation P.862 (2001) Amendment 2 (11/05) [Online]. Available: http://www.itu.int/rec/T-REC-P.862-200511-I!Amd2/en

[19] A. Prodeus, “Calculations of speech quality measure PESQ in MATLAB,” Proc. of XIVth Int. Sc. Conf. «The Latest Network Technology in Ukraine», Partenit, proceedings "Vestnik UNIIS", pp. 70-76, 17-19 September, 2012, Kiev, Ukraine (in Russian).