UDC 621.391.7: 004.934.2 (045)

A. M. Prodeus

ASSESSMENT OF NOISE REDUCTION ALGORITHMS QUALITY IN VOICE CONTROL CHANNELS

Acoustics and Electroacoustics Department, National Technical University of Ukraine
"Kyiv Polytechnic Institute", Kyiv, Ukraine
E-mail: aprodeus@gmail.com

Abstract*—*In this paper, traditional noise reduction algorithms such as spectral subtraction, Wiener, MMSE and logMMSE filtering algorithms, and two less known Wiener-TSNR and Wiener-HRNR filtering algorithms had been compared with the use of a set of quality measures. It is found that excessive noise reduction leads to insignificant degradation of the speech signals quality, but significantly reduces the accuracy of the automatic speech recognition (ASR). It is shown the existence of the speech quality measures which satisfactorily are matching with the accuracy of automatic speech recognition. This result is useful for practice because of speech recognition accuracy can be predicted by means of speech quality measures. In addition, it is found that there is no single algorithm among the considered noise reduction algorithms, which is the best in terms of maximum recognition accuracy for a wide range of input signal-to-noise ratio from minus 10 dB to plus 30 dB.

Index Terms—Noise reduction algorithm; speech quality indicator; recognition accuracy; speech signal; noise interference.

I. INTRODUCTION

A number of new aviation systems, and unmanned aerial vehicles (UAVs) are among them, are beginning to utilize speech recognition technology. In particular, it is believed that voice control would enable air battle managers to control their UAVs using voice commands in addition to joystick, mouse, and keyboard inputs [2].

The block diagram shown in Fig. 1 is a schematic diagram of a voice control channel that incorporates natural language processing. A human controller is present to issue directives based on an UAV's current state and the controller's intentions. Once these verbal commands are processed by the ASR system, they are translated into a set of high-level goals and constraints that are then passed on to the UAV's planning algorithms. These planning algorithms then generate a sequence of maneuvers for the UAV.

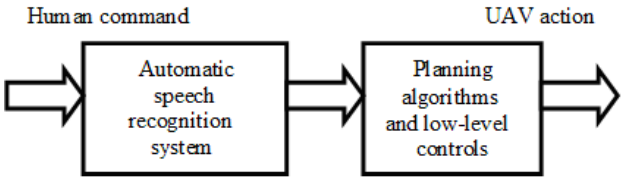


Fig. 1. ASR system incorporation into UAV control channel

Ensuring of acceptable speech quality [4], as well as increasing of automatic speech recognition (ASR) systems robustness [7] to the action of noise interference through the use of noise reduction pre-processors (Fig. 2) is issue of the day. Traditional noise reduction algorithms are spectral subtraction (SpecSub), Winer, minimum mean-square error amplitude spectrum estimator (MMSE) and minimum mean-square error log-spectral amplitude estimator (logMMSE) filtering [3], [4]. Wiener Two-Step Noise Reduction (Wiener-TSNR) and Wiener Harmonic Regeneration Noise Reduction (Wiener-HRNR) algorithms are less known, but they are attractive because of their ability to great noise suppression [5], [6]. Unfortunately, the aforementioned noise reduction algorithms were not compared with each other on speech quality and speech recognition accuracy indicators.

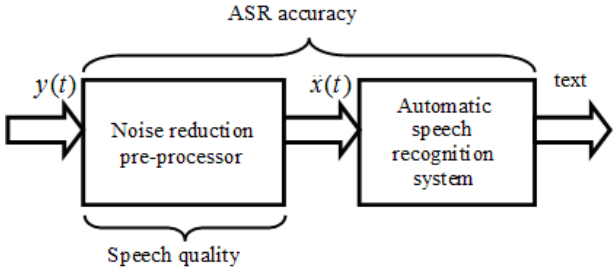


Fig. 2. Noise reduction system as ASR pre-processor

Speech recognition accuracy and different speech quality measures can be used to assess the performance of noise reduction algorithms. While this assessment is fairly typical task, the choice of the best quality measure is largely dependent on the predilections of researchers [1] – [6]. This can be explained by the fact that the choice problem is not enough investigated.

II. PROBLEM STATEMENT

When model $y(t) = x(t) + n(t)$ of distorted speech signal $x(t)$ is considered, noise $n(t)$

reduction algorithm provides recovery of signal $x(t)$ from mixture $y(t)$:

$$\hat{x}(t) = A\{y(t)\},$$

where $\hat{x}(t)$ and $A\{\cdot\}$ are result and operator of speech enhancing, respectively.

Analyzed in this paper noise suppression algorithms implement speech enhancing in frequency domain

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

where $\lambda_y(l,k)$ is power spectrum of signal $y(t)$ l-th frame 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 power spectrum estimator of signal $\hat{x}(t)$ lth frame; $G(l,k)$ is correction filter gain. Usually phase of distorted signal $y(t)$ is used as enhanced signal $\hat{x}(t)$ phase.

The first object of the paper is comparison aforementioned noise reduction algorithms (i.e. different correction filter gains $G(l,k)$) with each other on speech quality and speech recognition accuracy indicators.

When noise reduction algorithm is used as ASR pre-processor, its performance can be evaluated by means of end-to-end quality indicator which is named "ASR accuracy" [8]:

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

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.

The approach drawback is the need for ASR systems simulation. It seems advisable to explore the possibility of replacing Acc% indicator on speech quality measures. Thus, second object of the paper is searching of objective speech quality measures which are matching with speech recognition accuracy Acc%.

III. Noise Reduction Algorithms

SpecSub, Wiener, MMSE and logMMSE traditional noise reduction algorithms [4] are considered in this paper, and proper $\hat{G}(f,m)$ are follows

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

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

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

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

where $\hat{\xi}(f,m) = \hat{\lambda}_x(f,m)/\hat{\lambda}_n(f,m)$ 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 estimator, $\hat{v}(f,m) = \hat{\xi}(f,m)\hat{\gamma}(f,m)/[1+\hat{\xi}(f,m)]$, $\Gamma(\cdot)$ is gamma function, $I_0(\cdot)$ and $I_1(\cdot)$ are modified Bessel functions of zero and first order, respectively.

Decision directed method is usually used for $\hat{\xi}(f,m)$ calculation [4]:

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

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

where α is averaging parameter with $\alpha = 0.98$ optimal value for $F_s = 8$ kHz sample rate and $N_{\text{inc}} = 64$ frame shift. Generalizing this result, it can be shown that for arbitrary values of F_s and N_{inc} optimal value of averaging parameter will be $\alpha_{\text{opt}} = \exp(-N_{\text{inc}}/(0.396\cdot F_s))$.

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 the names of these algorithms means that the transfer functions of the correction filters are formed similar to one of Wiener filter. However, this does not mean that the transfer functions are prohibited from forming otherwise.

Wiener-TSNR transfer function is formed in two steps.

Step 1:

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

Step 2:

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

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:

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

Step 2. A priori SNR is calculated:

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

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

Step 3. Transfer function for HRNR algorithm is formed:

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

It is natural 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 one of the objects of the paper is to verify the validity of this assumption.

IV. Quality Measures

Segmental Signal-to-Noise Ratio (SSNR), Log-Spectral Distortion (LSD), Log-Likelihood Ratio (LLR), Weighted Spectral Slope (WSS), Itakura-Saito distance (IS), cepstral distance (CEP), composite index "Signal Composite Index, Noise Composite Index, Overall Composite Index" (SCI, NCI, OCI), perceptual indicators Bark-Spectral Distortion (BSD) and Perceptual Evaluation of Speech Quality (PESQ) speech quality measures were used in the paper.

Analytically parameters SSNR, LSD and BSD are described as follows

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

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

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

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

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

where $x(l,n)$ and $\hat{x}(l,n)$ are nth samples of lth frame of clear 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 lth frame of signals $x(n)$ and $\hat{x}(n)$, respectively.

Indicators LLR, IS and CEP are computed for each of the frames, and further averaged over all frames:

$$\text{LLR}=\ln\left(\frac{\vec{a}_p\mathbf{R}_c\vec{a}_p^{\,T}}{\vec{a}_c\mathbf{R}_c\vec{a}_c^{\,T}}\right),$$

$$\text{IS}=\frac{\sigma_c^2}{\sigma_p^2}\left(\frac{\vec{a}_p\mathbf{R}_c\vec{a}_p^{\,T}}{\vec{a}_c\mathbf{R}_c\vec{a}_c^{\,T}}\right)+\ln\left(\frac{\sigma_c^2}{\sigma_p^2}\right)-1,$$

$$\text{CEP}=\frac{10}{\ln 10}\sqrt{2\sum_{k=1}^{p}[c_c(k)-c_p(k)]^2},$$

$$c(m)=a_m+\sum_{k=1}^{m-1}\frac{k}{m}c(k)a_{m-k},\quad 1\le m\le p,$$

where $\vec{a}_c$ and $\vec{a}_p$ are linear prediction coefficients of clean and enhanced signals, respectively; $\mathbf{R}_c$ is pure autocorrelation coefficient matrix signal; σ_c^2 and σ_p^2 are variances of clean and enhanced signals, respectively; $c(k)$ are cepstral coefficients; p is filter-predictor order.

The indicator WSS is calculated as follows:

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

where $W(j,m)$ is weight for j th spectral band and m th frame; K is quantity of spectral bands; M is quantity of frames; $S_c(j,m)$ and $S_p(j,m)$ are the spectral slopes of the clean and processed speech signals, respectively. The spectral slope is obtained as the difference between adjacent spectral magnitudes in decibels. In our implementation, the number of bands was set to $K=25$.

PESQ is effective indicator of speech quality, but its analytical description is very cumbersome. Brief description can be found in [4]. We note only that it was used wideband, designed for speech signal analysis over a 7 kHz bandwidth, version of the indicator WB-PESQ in our study.

Composite index was described in [4].

V. Experimental Results

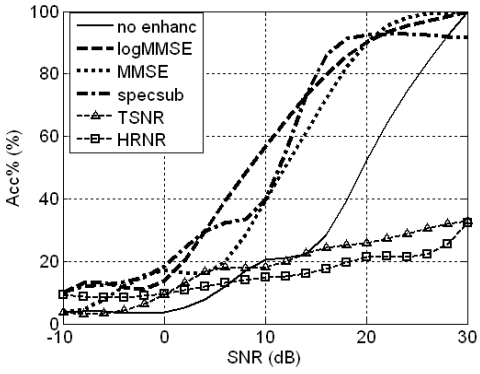
Clean speech signals (single words) were recorded in anechoic room and had been used for ASR system training. Parameters of digitized sounds were: sampling rate 22050 Hz, linear quantization 16 bit. Signal-to-noise ratio (SNR) was near 35 dB for saved clean speech signals.

Signal frames with 50 % overlapping and Hamming window were used for signal processing. Frames duration was 32 ms.

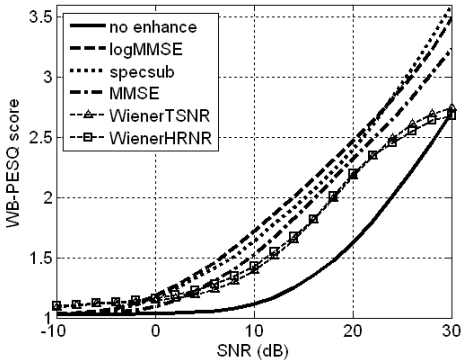
Toolkit HTK [8] had been used for ASR system simulation. Training of ASR system had been made with usage of 269 samples of 27 words of clean speech recorded for two speakers-women. Noised discrete speech signals (with 0.2…0.5 s pauses between single words) were used as test signals, and there were presented, in testing, all 27 words used in training. There were 27 phonemes of Ukrainian language in phoneme vocabulary and there had been used 39 MFCC_0_D_A coefficients when ASR simulating.

The experimental results had showed, first, that the indicators Acc% and PESQ does not agree very well with each other (Fig. 3). Among other indicators had been studied (Figs. 4, 5), only two - LLR and SCI – were in good agreement with the Acc% indicator (Fig. 4). At the same time, the essential disadvantage of LLR and SCI indicators is their inability to display fairly substantial difference of MMSE, logMMSE and spectral subtraction algorithms performance.

Analysis of the Ass% indicator behavior had showed that there is no single noise reduction algorithm, which would be best in terms of maximum Ass% in a broad range of signal-to-noise ratio from minus 10 dB up to plus 30 dB.

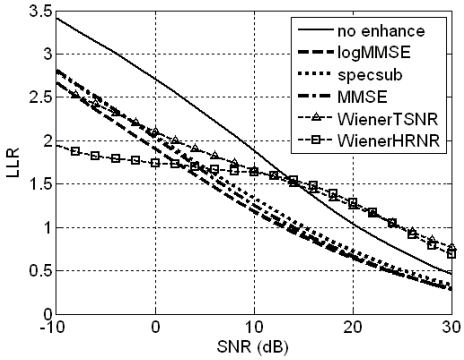


a

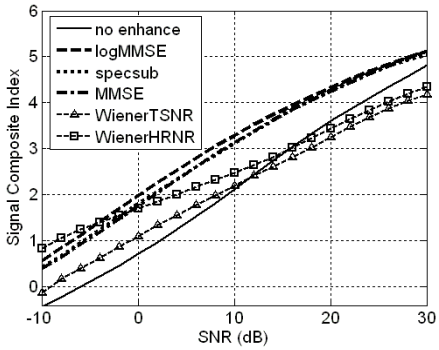


b

Fig. 3. Acc% (SNR) (a) and WB-PESQ (SNR) (b)



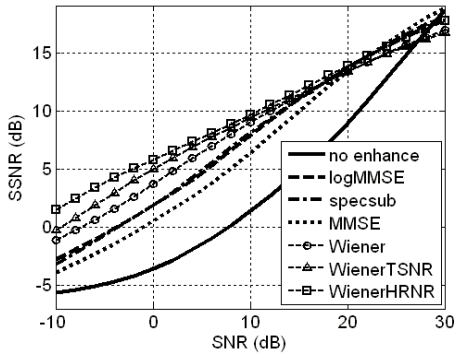
a



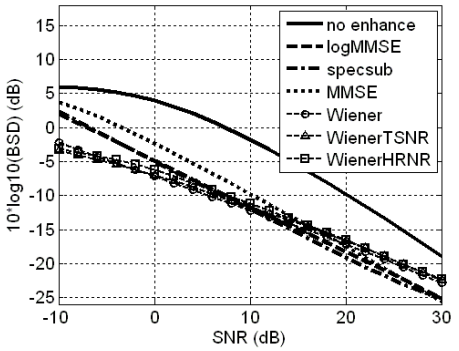
b

Fig. 4. LLR (SNR) (a) and SCI (SNR) (b)

Second, unexpectedly low efficiency of the Wiener-TSNR and Wiener-HRNR algorithms was revealed. Indeed, according to Fig. 3, usage of Wiener-TSNR and Wiener-HRNR algorithms for SNR > 3 dB leads to the lowest Acc% values compared to other algorithms. Moreover, for SNR > 8 dB the situation was even worse than in the case of disabling noise reduction algorithm (curve "no enhance"). LLR and SCI graphs confirm this fact (Fig. 4), although in somewhat "soften" manner: the situation is worse than in the case of disabling noise reduction algorithm only when SNR > 15 dB.



a



b

Fig. 5. Acc % (SNR) (a) and WB-PESQ (SNR) (b)

This result is not consistent with the results of the algorithms authors [5], [6] and can be explained as result of signal distortion. At the same time, these algorithms have shown the best results in all indicators when SNR is below 0 dB.

VI. CONCLUSION

Comparison of six noise reduction algorithms have shown that only two of the nine indicators examined - log-likelihood ratio and signal composite index – are in good matching with speech recognition accuracy Acc% when the noise reduction system is used as pre-processor of automatic speech recognition system.

Unexpectedly low efficiency of the Wiener-TSNR and Wiener-HRNR algorithms had been revealed: when SNR > 8 dB, speech recognition accuracy Acc% is worse than in the case of disabling noise reduction algorithm. This result can be explained as consequence of strong signal distortion. LLR measure and, what is much more important, SCI measure had confirmed this fact, although in somewhat "soften" manner: the situation is worse than in the case of disabling noise reduction algorithm only when SNR > 15 dB.

It was shown that there is no single algorithm among the considered noise reduction algorithms, which is the best in terms of maximum recognition accuracy Acc% for a wide range of input signal-to-noise ratio from minus 10 dB to plus 30 dB. It follows that the choice of noise reduction algorithms for engineering applications should be performed taking into account the value of the signal-to-noise ratio of the distorted signal.

It should be taken into account also that there isn't generally accepted standard ASR system model, so Acc% values will be dependent on the kind of ASR model. However, it is hoped that results obtained in this paper will remain qualitatively correct when using other models of automatic speech recognition system.

REFERENCES

[1] C. M. Chernick, S. Leigh, K. L. Mills and R. Toense, "Testing the Ability of Speech Recognizers to Measure the Effectiveness of Encoding Algorithms for Digital Speech Transmission." *Proceedings of IEEE International Military Communications Conference (MILCOM)*. 1999, vol. 2, pp. 1468–1472.

[2] E. Craparo and E. Feron, "Natural Language Processing in the Control of Unmanned Aerial Vehicles." *Proceeding of AIAA Guidance, Navigation, and Control Conference*, 2004, pp. 1–13.

[3] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, 2008, vol. 16, pp. 229–238.

[4] P. Loizou, *Speech enhancement: Theory and Practice*. Boca Raton: CRC Press, 2007, 632 p.

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

[6] C. Plapous, C. Marro and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement." *IEEE Transactions on Audio, Speech, and Language Processing*, 2006, vol.14, no. 6, pp. 2098–2108.

[7] N. Virtanen; R. Singh and B. Raj, *Techniques for Noise Robustness in Automatic Speech Recognition*. John Wiley, 2013, 501 p.

[8] S. Young, G. Evermann and M. Gales, (ed). *The HTK Book*. Cambridge: University Engineering Department. 2009.

Received August 21, 2015

Prodeus Arkadiy. DrSc. Professor.
Acoustics and Electroacoustics Department, National Technical University of Ukraine "Kyiv Polytechnic Institute", Kyiv, Ukraine.
Education: Kyiv Polytechnic Institute, Kyiv, Ukraine (1972).
Research interests: digital signal processing.
Publications: 167.
E-mail: aprodeus@gmail.com

А. М. Продеус. Оцінювання якості алгоритмів шумозаглушення в каналах голосового управління
Виконано порівняння, із використанням набору показників якості, традиційних алгоритмів шумозаглушення, таких як спектральне віднімання, алгоритми фільтрації Вінера, MMSE і logMMSE, та двох значно менше відомих алгоритмів фільтрації Wiener-TSNR і Wiener-HRNR. Встановлено, що надмірне придушення шуму незначним чином погіршує якість мовлення, проте призводить до суттєвого зниження точності автоматичного розпізнавання мовлення (АРМ). Показано існування показників якості мовлення, які задовільно узгоджуються із точністю автоматичного розпізнавання мовлення. Цей результат є корисним для практики, оскільки дозволяє розраховувати точність розпізнавання мовлення за результатами оцінювання якості мовлення. Крім того, було виявлено, що серед розглянутих алгоритмів шумозаглушення немає єдиного алгоритму, котрий був би найкращим з погляду забезпечення максимальної точності розпізнавання для широкого діапазону співвідношень сигнал-шум від мінус 10 дБ до плюс 30 дБ.
Ключові слова: алгоритм шумозаглушення; показник якості мовлення; точність розпізнавання; мовленнєвий сигнал; шумова завада.

Продеус Аркадій Миколайович. Доктор технічних наук. Професор.
Кафедра акустики та акустоелектроніки, Національний технічний університет України «Київський політехнічний інститут», Київ, Україна.
Освіта: Київський політехнічний інститут, Київ, Україна (1972).
Напрямок наукової діяльності: цифрова обробка сигналів.
Кількість публікацій: 167.
E-mail: aprodeus@gmail.com

А. Н. Продеус. Оценивание качества алгоритмов шумоподавления в каналах голосового управления
Сопоставлены, с использованием набора показателей качества, традиционные алгоритмы шумоподавления, такие как спектральное вычитание, алгоритмы фильтрации Винера, MMSE и logMMSE, и два значительно менее известных алгоритма фильтрации Wiener-TSNR и Wiener-HRNR. Установлено, что чрезмерное подавление шума незначительно ухудшает качество речевых сигналов, однако приводит к существенному снижению точности автоматического распознавания речи (АРР). Показано существование показателей качества речи, которые удовлетворительно согласовываются с точностью автоматического распознавания речи. Этот результат полезен для практики, поскольку позволяет рассчитывать точность распознавания речи по результатам оценивания качества речи. Кроме того, обнаружено, что среди рассмотренных алгоритмов шумоподавления нет единственного алгоритма, который был бы наилучшим с точки зрения обеспечения максимальной точности распознавания для широкого диапазона отношений сигнал-шум от минус 10 дБ до плюс 30 дБ.
Ключевые слова: алгоритм шумоподавления; показатель качества речи; точность распознавания; речевой сигнал; шумовая помеха.

Продеус Аркадий Николаевич. Доктор технических наук. Профессор.
Кафедра акустики и акустоэлектроники, Национальный технический университет Украины «Киевский политехнический институт», Киев, Украина.
Образование: Киевский политехнический институт, Киев, Украина (1972).
Направление научной деятельности: цифровая обработка сигналов.
Количество публикаций: 167.
E-mail: aprodeus@gmail.com