Теория сигналов и систем

UDC 621.391.7: 004.934.2

O. Ladoshko, **A. Prodeus**, Dr.Sc.
National Technical University of Ukraine “Kyiv Polytechnic Institute”,
off. 233, Politekhnichna Str., 16, Kyiv, 03056, Ukraine.

Parameter optimization of late reverberation suppression algorithm

Boundary values between early reflections and late reverberation, optimal in sense of such criteria as speech recognition accuracy and speech quality, had been found. When optimal boundary value is chosen, usage of logMMSE method for late reverberation suppression makes it possible to increase recognition accuracy from 22 ... 30% to 56…74% and speech quality index PESQ from 2.281 to 2.33. Reference 6, figures 4.

Keywords: *late reverberation, speech recognition accuracy, speech quality.*

Introduction

The problem of speech dereverberation in communication and automatic speech recognition (ASR) systems was actively investigated in the last decade due to the rapid development of mobile communications [1-2]. It was found that late reverberation is main detrimental factor which is kind of additive noise. The formula for estimation of late reverberation power spectrum contains parameter T_l, which is time boundary between early reflections and late reverberation. The boundary is blurred: we find $T_l \approx 30...100$ ms in [1-2]. Moreover, these values were experimentally obtained when problems of speech intelligibility and musical clarity were investigated, and it isn't evident that the same values will be good for speech recognition and communication systems. The objective of this paper is searching of parameter T_l optimal values in sense of such criteria as speech recognition accuracy and speech quality.

1. Target setting

The reverberant signal $y(t)$ results from the convolution of the anechoic speech signal $x(t)$ and the causal time-invariant Acoustic Impulse Response (AIR) $h(t)$:

$$y(t) = \int_0^{\infty} h(v)x(t-v)dv = x(t) \otimes h(t).$$

were $\otimes$ is convolution symbol.

When selecting in AIR $h(t)$ (Fig. 1) regions corresponding to early reflections and late reflections

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

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

reverberation action can be described as

$$y(t) = h_i(t) \otimes x(t) + r(t). \quad (1)$$

where $r(t) = h_l(t) \otimes x(t - T_l)$ is component due to late reverberation; T_l is time, corresponding to boundary between early reflections and late reverberation (see Fig. 1).

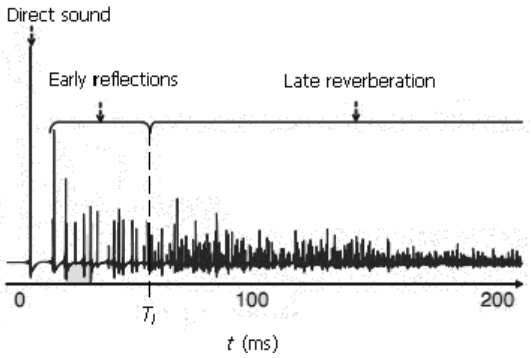


Fig. 1. Room AIR structure

It is clear from (1) that late reverberation may be interpreted as kind of noise. Unfortunately, strong non-stationarity of late reverberation makes ineffective traditional techniques of stationary or slow non-stationary noise suppression [1].

It can be assume that late reverberation suppression may be realized almost by the same remedies which are usually used for noise suppression by estimating of late reverberation spectrum instead of noise spectrum.

Correction in frequency domain is popular noise suppression method [3]:

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

where $\lambda_y(l,k)$ is power spectrum of l-th signal $y(t)$ frame at frequency $f_k = kF_s / N_{fft}$; F_s is sam -

pling frequency; N_{fft} is FFT parameter; k is number of frequency sample; $\hat{\lambda}_x(l,k)$ is power spectrum estimator of l-th frame of signal $x(t)$ for k-th frequency sample; $G(l,k)$ is correction filter gain for l-th signal $y(t)$ frame for k-th frequency sample.

In the paper logMMSE method [3] is considered, for which enhancement filter gain is

$$G(l,k)=\frac{\xi(l,k)}{1+\xi(l,k)}\exp\left(\frac{1}{2}\int\limits_{v(l,k)}^{\infty}\frac{e^{-t}}{t}dt\right)$$

$$v(l,k)=\frac{\xi(l,k)}{1+\xi(l,k)}\gamma(l,k)$$

where $\xi(l,k)=\lambda_x(l,k)/\lambda_n(l,k)$ is prior signal-to-noise ratio (SNR); $\gamma(l,k)=\lambda_y(l,k)/\lambda_n(l,k)$ - posterior SNR; $\lambda_n(l,k)$ - power spectrum of l-th noise $n(t)$ frame at frequency f_k. Fundamentally important and difficult is noise spectrum $\lambda_n(l,k)$ estimation when implementing the logMMSE method for noise suppression. When modifying scheme of noise suppression for late reverberation suppression, we need substitute late reverberation spectrum $\lambda_r(l,k)$ estimator instead of noise spectrum $\lambda_n(l,k)$ estimator.

For distances between speech source and microphone, which are more then critical distance D_c, late reverberation power spectrum $\lambda_r(l,k)$ may be calculated by spectrum $\lambda_y(l,k)$ of signal $y(t)$ [2]:

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

where $N_l=T_lF_s/R$; R denotes the frame rate in samples of the short-time Fourier transform (STFT); $\delta(k)=2\ln 10/T_{60}(k)$; $T_{60}(k)$ is reverberation time.

Smoothing is necessary to enhance the estimation accuracy of the spectrum $\lambda_y(l,k)$ [2]:

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

where $Y(l,k)$ is discrete Fourier transform (DFT) of l-th frame of signal $y(t)$;

$$\eta_y(k)=\begin{cases}\eta_y^d(k), & \left|Y(l,k)\right|^2\le\hat{\lambda}_y(l-1,k);\\ \eta_y^a(k) & \text{otherwise.}\end{cases}$$

Upper-bound of constant $\eta_y^d(k)$ $(0\le\eta_y^d(k)<1)$ is

$$\eta_y^d(k)=\frac{1}{1+2\delta(k)R/F_s}$$

and the constant $\eta_y^a(k)$ is selected from the conditions $0\le\eta_y^a(k)<\eta_y^d(k)$.

2. Experimental organization

There were two groups of experiments: qualitative and quantitative. When realizing qualitative evaluation of dereverberation performance, real speech signal was recorded in room with volume 80 m^3 and time reverberation 1.1 s (sampling frequency 22050 Hz, linear quantization 16 bit). Distance between speaker and microphone was much more of critical distance [1-2].

When realizing quantitative evaluation of dereverberation performance, clear speech signals were convolved with AIRs of three rooms with time reverberation 0.74 s, 0.89 s and 1.1 s for simulation of reverberation action. Sounds of bursting rubber ball were used as AIRs for these rooms. Dereverberation performance had been estimated by means of ASR accuracy:

$$Acc\%=\frac{N-D-S-I}{N}\times 100\%.$$

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. Indicator PESQ had been used for speech quality assessment [4].

Toolkit HTK [5] had been used for ASR system simulation. Training of ASR system had been made with usage of 269 samples of 27 words saved for two speakers-women. Sound file of discrete speech (with 0.2…0.5 s pauses) was used as test signal, there were used all 27 words 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.

VoiceBox [6] routine “ssubmmse.m” designed to reduce the noise was modified in accordance with propositions of previous section. Moreover, it was taken $\eta_y^a(k)=0{,}5\cdot\eta_y^d(k)$.

3. Experimental results

Spectrograms of reverberant and enhanced signals for qualitative experiments are shown in Fig. 2. There is noticeable by ear slight distortion introduced by the dereverberation procedure (it was taken $T_l=48\,ms$ upon the procedure). In-

creasing T_l to 100 ms led to some improvement in sound quality. It demonstrates real problem of true choice of parameter T_l value.

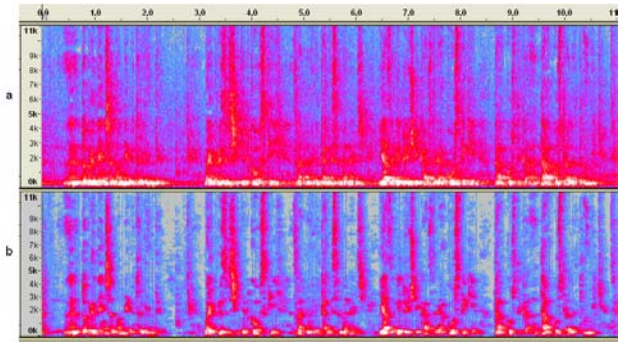
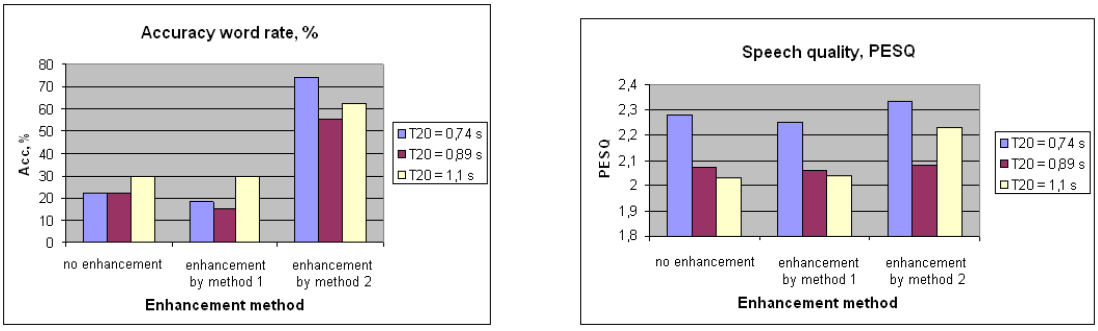


Fig. 2. Reverberant (a) and enhanced (b) spectrograms

It was found for quantitative experiments that reverberation significantly affects both the $Acc\%$ (reduced from 93% to 22 ... 30%) and the PESQ (reduced from 4.5 to 2.03 ... 2.28).

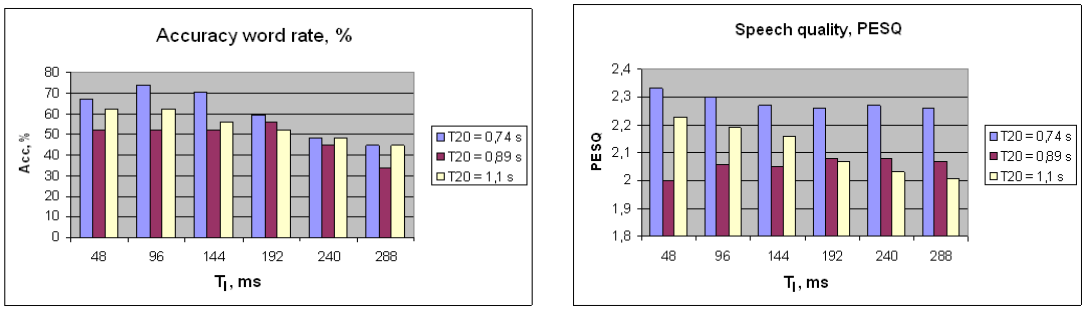
Results of Acc% and PESQ estimation for enhanced speech signals are shown in Fig. 3. As it can be seen, enhancement by method 1 (usage of "classic" logMMSE method) did not lead to positive results. Meanwhile, enhancement by method 2 (usage of modified logMMSE method) had made it possible to significantly increase the Acc% value (raised from 22 ... 30% to 56…74%). It is interesting that PESQ value did not raised so much (increased from 2.281 to 2.33 for $T_{20} = 0.74$ c, and only from 2.073 to 2.08 for $T_{20} = 0.89$ c).

Results of experimental studies of dependencies $Acc\%(T_l)$ and $\mathrm{PESQ}(T_l)$ are shown in Fig. 4. It follows from these results that optimal, in sense of $Acc\%$ maximum, T_l value lies in the interval 100…200 ms. More uncertain is situation with $\mathrm{PESQ}(T_l)$ dependency. Weakly pronounced maximum at $T_l \approx 200...240$ ms was observed only in one from three cases.



a) **b)**

Fig. 3. Recognition accuracy (a) and speech quality (b)



a) **b)**

Fig. 4. $Acc\%(T_l)$ **(a) and** $\mathrm{PESQ}(T_l)$ **(b) dependency**

Conclusions

Experimental studies of dependencies $Acc\%(T_I)$ and $\mathrm{PESQ}(T_I)$ were conducted. It was shown that optimal, in sense of $Acc\%$ maximum, T_I value lies in the interval 100…200 ms. More uncertain is situation with $\mathrm{PESQ}(T_I)$ dependency, where, in two of three cases, the speech quality decreased with increasing T_I values, and only one case was observed with weakly pronounced maximum at $T_I \approx 200...240$ ms.

References

1. *Naylor P., Gaubitch N.* (2010), "Speech Dereverberation". Springer.
2. *Habets E.A.P.* (2007), "Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement". Ph.D Thesis. Eindhoven.
3. *Ephraim Y., Malah D*. (1985), "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator". IEEE Transactions on Acoustic, Speech, and Signal Processing. Vol. ASSP-33. No. 2. Pp. 443-445.
4. *Loizou P.* (2007), "Speech enhancement: Theory and Practice". Boca Raton: CRC Press.
5. *Young S.* (2005), "The HTK Book". Cambridge University Engineering Department. [Online]. Available: http://htk.eng.cam.ac.uk/download.shtml
6. *Brooks M.* (2010), "VOICEBOX: Speech Processing Toolbox for MATLAB". [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/

Поступила в редакцию 20 сентября 2014 г.

УДК 621.391.7: 004.934.2

О.М. Ладошко, **А.М. Продеус**, д.-р.техн.наук
Національний технічний університет України «Київський політехнічний інститут»,
вул. Політехнічна 16, 03056, Київ, Україна.

Оптимізація параметрів алгоритму ослаблення пізньої реверберації

Показано існування оптимальних, в сенсі таких критеріїв як точність розпізнавання мовлення та якість мовлення, значень границі між ранніми відлуннями та пізньою реверберацією. Якщо оптимальне значення границі є обраним, використання методу logMMSE для ослаблення дії пізньої реверберації дозволяє підвищити точність розпізнавання мовлення з 22...30% до 56...74%, а якість мовлення PESQ - з 2.281 до 2,33. Бібл.6, рис. 4.

Ключові слова: *пізня реверберація, точність розпізнавання мовлення, якість мовлення.*

УДК 621.391.7: 004.934.2

О.Н. Ладошко, **А.Н. Продеус**, д.-р.техн.наук
Национальный технический университет Украины «Киевский политехнический институт»,
ул. Политехническая 16, 03056, Киев, Украина.

Оптимизация параметров алгоритма подавления поздней реверберации

Показано существование оптимальных, в смысле таких критериев как точность распознавания речи и качество речи, значений границы между ранними отражениями и поздней реверберацией. Если оптимальное значение границы выбрано, использование метода logMMSE для подавления поздней реверберации позволяет повысить точность распознавания речи с 22 ... 30% до 56 ... 74%, а качество речи PESQ - с 2.281 до 2,33. Библ. 6 , рис. 4.

Ключевые слова: *поздняя реверберация, точность распознавания речи, качество речи.*

Список использованных источников

1. *Naylor P., Gaubitch N*. Speech Dereverberation. - Springer, 2010.
2. *Habets E.A.P.* Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement // PhD Thesis. – Eindhoven, 2007.
3. *Ephraim Y., Malah D.* Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator // IEEE Transactions on Acoustic, Speech, and Signal Processing. – Vol. ASSP-33. – No. 2. – 1985. – Pp. 443-445.
4. *Loizou P*. Speech enhancement: Theory and Practice. – Boca Raton: CRC Press, 2007.
5. *Young S*. The HTK Book. – Cambridge University Engineering Department, 2005. [Online]. Available: http://htk.eng.cam.ac.uk/download.shtml
6. *Brooks M.* VOICEBOX: Speech Processing Toolbox for MATLAB, 2010. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/