Two-Step Noise Reduction Based on Soft Mask for Robust Speaker Identification

Gennadiy Tupitsin, Artem Topnikov, Andrey Priorov
P. G. Demidov Yaroslavl State University
Yaroslavl, Russia
genichyar@genichyar.com, topartgroup@gmail.com, andcat@yandex.ru

***Abstract*—This paper addresses the problem of speaker identification in noisy conditions. A two-step noise reduction algorithm based on soft mask and minimum mean square error short-time spectral amplitude estimator was proposed. It is used in the signal preprocessing stage for more robust speaker identification. The proposed algorithm was tested and compared with the existing noise reduction algorithms in the problem of speaker identification. Testing was carried out with two speech databases and some noise samples from the NOISEX-92 library. The advantage of the new noise reduction algorithm for some noise samples and signal-to-noise ratios was shown.**

I. Introduction

Speaker identification is becoming a high-relevant task in many fields specially in the framework of security remote applications. These systems usually developed under laboratory conditions and severely degrade their performance level when an acoustical mismatch appears among training and testing phases [1]. For example it can occurs in acoustic noise presence. In this case one of the most effective ways to provide more robustness to the recognizer is using noise reduction algorithms for speech signals [1].

The problem of enhancing speech degraded by uncorrelated additive noise, when only the noise speech is available, has widely studied in the past and it is still in active field of research [2]. Some methods in frequency domain using various spectral gain functions depending on *a posteriori* signal-to-noise ratio or/and *a priori* signal-to-noise ratio was proposed in the past. *A priori* signal-to-noise ratio estimation is not required for spectral subtraction gain function [3]. In other cases it can be estimated using decision-directed approach [4], two-step noise reduction technique [5] or other methods [6], [7]. Various gain functions is used for short-time spectral magnitude correction such as Wiener gain function [8], minimum mean square error short-time spectral amplitude estimator [4], minimum mean square error short-time log-spectral amplitude estimator [9], etc. In addition to the frequency domain noise reduction methods there are other approaches [10], [11], [12], [13], [14].

It should be noted that noise reduction algorithms maximizing quality and intelligibility of speech signals are not always effective for signal preprocessing in the problem of speaker identification.

In [15] soft mask noise reduction technique was presented. The soft mask algorithm is similar to other algorithms in the frequency domain, but soft mask's gain function is a probability of speech presence in each point of the time-frequency representation of the speech signal.

In this paper soft masks were generalized. A new approach to soft mask estimation using modified decision-directed approach was proposed. The obtained algorithm was used as first step in the two-step noise reduction algorithm [5]. The two-step noise reduction technique was also modified. The modification related to smoothing *a priori* signal-to-noise ratio for the second stage using exponential moving average with upper limit.

The goal of the research is developing a noise reduction algorithm based on soft mask to improve speaker identification accuracy in noisy conditions.

II. Soft mask for noise reduction

A. Definition

Let $R_{k,w}$ and $A_{k,w}$ denote noisy speech magnitude spectrum and "clean" speech magnitude spectrum respectively (where k is spectral component number, w is analysis frame number). Soft mask's gain function is a probability of speech presence in each point of the time-frequency representation of the speech signal:

$$S_{k,w} = P(H_1) = P(\xi_{k,w}^{local} > 1) \quad (1)$$

where H_1 is speech presence hypothesis; $\xi_{k,w}^{local}$ is local *a priori* signal-to-noise ratio defined as follows:

$$\xi_{k,w}^{local} = \frac{A_{k,w}^2}{D_{k,w}^2} \quad (2)$$

where $D_{k,w}$ is noise magnitude spectrum.

If soft mask is known, speech magnitude spectrum can be estimated as follows:

$$\hat{A}_{k,w} = S_{k,w} R_{k,w} \quad (3)$$

B. Explanation

The ability of using soft mask for noise reduction can be derived from binary mask. It is based on the assumption that additive noise masks some parts of the time-frequency representation of the speech signal and leaves the other parts not strongly affected [16]. It should be noted that binary mask is widely used in the computational auditory scene analysis [17], [18], [19].

Binary mask can be represented as following spectral gain function:

$$B_{k,w} = \begin{cases} 1 & \text{if } H_1 \\ 0 & \text{if } H_0 \end{cases}$$

where H_1 is speech presence hypothesis; H_0 is speech absence hypothesis. Speech magnitude spectrum can be estimated using binary mask as follows:

$$\hat{A}_{k,w} = B_{k,w} R_{k,w} \tag{4}$$

It was proposed to divide points of the time-frequency representation of the speech signal into speech present and speech absent using following rule [19], [20]:

$$\xi_{k,w}^{local} > \tau \tag{5}$$

where τ is threshold that usually equals to one [18], [21].

The gain function $B_{k,w}$ in (5) can be considered to be a random variable as it depends on the $\xi_{k,w}^{local}$. In the context of binary masking, $B_{k,w}$ is a Bernoulli distributed random variable taking the value of 0 or 1, and its parameter p is the hypothesis probability $P(H_1)$. It is difficult to estimate $B_{k,w}$ as it depends on accurate estimates of the local *a priori* signal-to-noise ratio. However, we can obtain $B_{k,w}$ more reliably by taking its expectation [15]. This approach is similar to the soft decision introduced by McAulay and Malpass in [22]. Using it (3) can be transformed to:

$$\hat{A}_{k,w} = E\{B_{k,w}\} R_{k,w} =$$
$$= \left[E\{B_{k,w} \mid H_0\} P(H_0) + E\{B_{k,w} \mid H_1\} P(H_1)\right] R_{k,w}$$

Since $E\{B_{k,w} \mid H_0\} = 0$ and $E\{B_{k,w} \mid H_1\} = 1$ this equation can be transformed to:

$$\hat{A}_{k,w} = P(H_1) R_{k,w} = P(\xi_{k,w}^{local} > \tau) R_{k,w} = S_{k,w} R_{k,w}$$

where $S_{k,w}$ is soft mask.

III. Generalized soft mask

Since $B_{k,w}$ is either 1 or 0 it can be raised to arbitrary power $\theta \geq 0$ in (4). So it can be transformed to:

$$\hat{A}_{k,w} = B_{k,w}^{\theta} R_{k,w}$$

As $\hat{A}_{k,w} \geq 0$ and $R_{k,w} \geq 0$ this equation can be raised to the power $\frac{1}{\theta}$:

$$\hat{A}_{k,w}^{\frac{1}{\theta}} = B_{k,w} R_{k,w}^{\frac{1}{\theta}}$$

We can obtain $\hat{A}_{k,w}^{\frac{1}{\theta}}$ using expectation of $B_{k,w}$. So this equation can be transformed to:

$$\hat{A}_{k,w}^{\frac{1}{\theta}} = E\{B_{k,w}\} R_{k,w}^{\frac{1}{\theta}} =$$
$$= \left[E\{B_{k,w} \mid H_0\} P(H_0) + E\{B_{k,w} \mid H_1\} P(H_1)\right] R_{k,w}^{\frac{1}{\theta}}$$

If we raise this equation to the power θ we obtain following:

$$\hat{A}_{k,w} = S_{k,w}^{\theta} R_{k,w}$$

where $S_{k,w}^{\theta}$ is generalized soft mask.

The power of soft mask θ can be determined based on chosen optimality criterion.

Let us explain the meaning of the parameter θ. In figures below there are normalized histograms for $S_{k,w}^{\theta}$ when $\theta = 1$ (Fig. 1) and $\theta = 2$ (Fig. 2). This histograms were obtained using real speech signal with 15 dB signal-to-noise ratio corrupted by additive white Gaussian noise and noise reduction algorithm described in the next part of the paper.

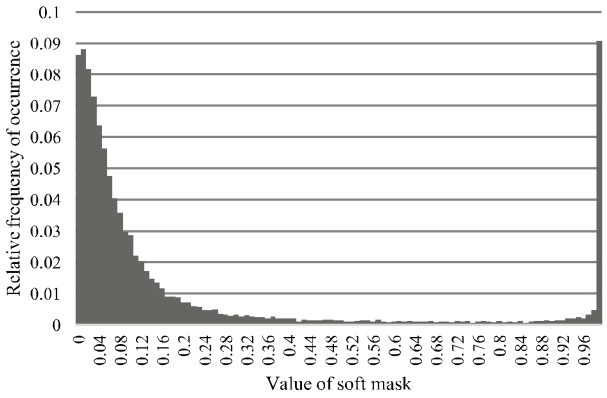

Fig. 1. Normalized histogram for $S_{k,w}^{\theta}$, $\theta = 1$

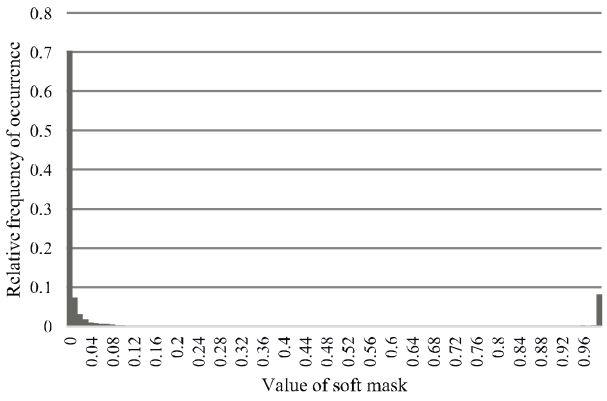

Fig. 2. Normalized histogram for $S_{k,w}^{\theta}$, $\theta = 2$

Conventionally, all points of the time-frequency representation of the signal can be divided into unreliable for which $S_{k,w}^{\theta} \to 0$, reliable for which $S_{k,w}^{\theta} \to 1$, and intermediate that located between them. If we increase θ, quantity of unreliable points increase, but quantity of reliable and intermediate points decrease.

So, by varying the parameter θ, we can set necessary balance between reliable and unreliable points. In other words, higher value of θ provides more noise suppression, and in this case soft mask is more closer to binary mask.

IV. A new approach to soft mask estimation

Let us estimate soft mask using equation (1). To estimate the likelihood of validity of inequality $\xi_{k,w}^{local} > 1$ we express it relatively to $D_{k,w}$ and use assumption that the noise magnitude spectrum is Rayleigh distributed for each spectral component. It can be moved from strict inequality $\xi_{k,w}^{local} > 1$ to non-strict inequality $\xi_{k,w}^{local} \geq 1$ and revealed it using (2):

$$A_{k,w}^2 \geq D_{k,w}^2$$

As $\hat{A}_{k,w} \geq 0$ and $R_{k,w} \geq 0$ square root can be taken:

$$A_{k,w} \geq D_{k,w}$$

"Clean" speech magnitude spectrum $A_{k,w}$ is unknown, but it can be estimated using one of spectral gain functions. In this paper Wiener gain function was using:

$$\frac{\hat{\xi}_{k,w}}{1+\hat{\xi}_{k,w}} R_{k,w} \geq D_{k,w} \tag{6}$$

where $\hat{\xi}$ is estimation of *a priori* signal-to-noise ratio defined as follows:

$$\xi_{k,w} = \frac{\lambda^A_{k,w}}{\lambda^D_k}$$

where $\lambda^A_{k,w} = E\{A_{k,w}\}$ is power spectral density of "clean" speech, $\lambda^D_k = E\{D_k\}$ is power spectral density of noise.

A priori signal-to-noise ratio can be estimated using modified decision-directed approach introduced by Lu and Loizou [6]:

$$\hat{\xi}_{k,w} = \alpha \hat{\xi}^{local}_{k,w-1} + (1-\alpha)\left(\sqrt{\gamma^{local}_{k,w}} - 1\right)^2$$

where α is parameter of the algorithm, $\gamma^{local}_{k,w}$ is *a posteriori* signal-to-noise ratio expressed as follows:

$$\gamma^{local}_{k,w} = \frac{R^2_{k,w}}{D^2_{k,w}}$$

Let us express local *a priori* signal-to-noise ratio through $\hat{A}_{k,w-1}$ and $D_{k,w-1}$, local *a posteriori* signal-to-noise ratio through $R_{k,w}$ and $D_{k,w}$:

$$\hat{\xi}_{k,w} = \alpha \frac{A^2_{k,w-1}}{D^2_{k,w-1}} + (1-\alpha)\frac{R^2_{k,w} + D^2_{k,w} - 2R_{k,w}D_{k,w}}{D^2_{k,w}} \tag{7}$$

Using the assumption of stationarity $D_{k,w}$ it was made the following simplification:

$$D^2_{k,w-1} \to D^2_{k,w} \tag{8}$$

Considered it let us substitute (8) to equation (7) and use it in inequality (6):

$$\frac{\alpha \frac{A^2_{k,w-1}}{D^2_{k,w}} + (1-\alpha)\frac{R^2_{k,w}+D^2_{k,w}-2R_{k,w}D_{k,w}}{D^2_{k,w}}}{1+\alpha \frac{A^2_{k,w-1}}{D^2_{k,w}} + (1-\alpha)\frac{R^2_{k,w}+D^2_{k,w}-2R_{k,w}D_{k,w}}{D^2_{k,w}}} R_{k,w} \geq D_{k,w}$$

This inequality was transformed to inequality relatively to $D_{k,w}$:

$$aD^3_{k,w} + bD^2_{k,w} + cD_{k,w} + d \leq 0 \tag{9}$$

where

$$a = 2 - \alpha$$

$$b = -3(1-\alpha)R_{k,w}$$

$$c = \alpha \hat{A}^2_{k,w-1} + 3(1-\alpha)R^2_{k,w}$$

$$d = -R_{k,w}(\alpha \hat{A}^2_{k,w-1} + (1-\alpha)R^2_{k,w})$$

We exclude a special case when $A_{k,w} = 0 \cap R_{k,w} = 0$ and consider it later. Consider a function

$$f(D_{k,w}) = aD^3_{k,w} + bD^2_{k,w} + cD_{k,w} + d \tag{10}$$

Identify its extremes and equal to zero

$$f'(D_{k,w}) = 3aD^2_{k,w} + 2bD_{k,w} + c = 0 \tag{11}$$

It is easy to show that the discriminant of the equation (11) is always less than zero. So the function (10) has no extremes. Considering this and the fact that $a > 0$, we can conclude that (11) is monotonically increasing function, and the equation $f(D_{k,w}) = 0$ has only one real root.

We need obtain the value of function (10) at the point $D_{k,w} = 0$: $f(0) = d$. Based on the fact that $d < 0$, the function (10) is negative at $D_{k,w} = 0$. So we can conclude that the real root of the equation $f(D_{k,w}) = 0$ lies to the right of zero on the number line. Consequently, the solution of (9) is the following expression:

$$D_{k,w} \leq D_{f(D_{k,w})=0} \tag{12}$$

where $D_{f(D_{k,w})=0}$ is the real root of the equation $f(D_{k,w}) = 0$.

Transforming the equation $f(D_{k,w}) = 0$ to the canonical form we get following equation:

$$y^3 + py + q = 0 \tag{13}$$

where

$$y = D_{k,w} + \frac{b}{3a} \tag{14}$$

$$p = \frac{c}{a} - \frac{b^2}{3a^2}$$

$$q = \frac{2b^3}{27a^3} - \frac{bc}{3a^2} + \frac{c}{a}$$

Let us define Q:

$$Q = \left(\frac{p}{3}\right)^3 + \left(\frac{q}{2}\right)^2$$

It is easy to show that $p > 0$, and hence $Q > 0$. Let us define φ:

$$\varphi = \sqrt[3]{\sqrt{Q} - \frac{q}{2}}$$

In our case there is $\sqrt{Q} > \frac{q}{2}$, so $\varphi > 0$. To find the root of the equation (13) we can use the Cardano's formula:

$$y = \varphi - \frac{p}{3\varphi}$$

Substituting it into (14), we obtain the solution of the equation:

$$D_{f(D_{k,w})=0} = \varphi - \frac{p}{3\varphi} - \frac{b}{3a}$$

Let us consider the special case $A_{k,w} = 0 \cap R_{k,w} = 0$ which was excluded before. It is easy to show that in this case the solution of (9) is the following expression:

$$D_{f(D_{k,w})=0} = 0$$

This solution can be obtained by the algorithm described above, but in formula it should be provided that φ can be equal to zero.

Soft mask can be obtained as follows:

$$S_{k,w} = P(D_{k,w} \leq D_{f(D_{k,w})=0}) = \\ = F_k^{Rayleigh}(D_{f(D_{k,w})=0})$$

where $F_k^{Rayleigh}$ is cumulative distribution function of noise magnitude spectrum for spectral component k.

As we showed in the previous part of the paper, $S_{k,w}$ can be raised to the power θ. A block diagram of proposed approach to soft mask estimation for a frame is presented at Fig. 3. In the figure, $k = 1 : M$ is spectral component iterator (M is quantity of the spectral components).

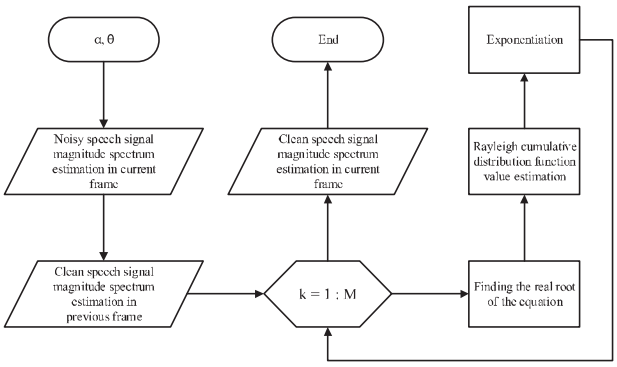

Fig. 3. Soft mask estimation algorithm block diagram for a frame

V. Two-step noise reduction based on soft mask

We can use obtained algorithm as first step in two-step noise reduction algorithm [5]. Minimum mean square error short-time spectral amplitude estimator [4] was chosen for second step but other spectral gain functions can be used such as Wiener gain function [8], minimum mean square error short-time log-spectral amplitude estimator [9], etc.

We also propose to smooth *a priori* signal-to-noise ratio for the second stage using exponential moving average with upper limit [23], [24]:

$$\hat{\xi}_{k,w}^{TSNR} = \epsilon \cdot \min\left(\delta, \frac{\hat{A}_{k,w}^2}{\lambda_k^D}\right) + (1-\epsilon)\hat{\xi}_{k,w-1}^{TSNR}$$

$$0 < \epsilon \leq 1$$

$$\delta >> 1$$

where ϵ is smoothing parameter, δ is limiting parameter preventing *a priori* signal-to-noise ratio overestimate. In this paper $\delta = \infty$.

While listening speech signals processed by the modified algorithm with different values of ϵ, it was observed that decreasing ϵ reduced level of "musical" noise, but speech signals become less intelligibly.

A block diagram of proposed two-step noise reduction algorithm based on soft mask is presented at Fig. 4. In the figure, $w = 1 : W$ is frame iterator (W is quantity of the frames).

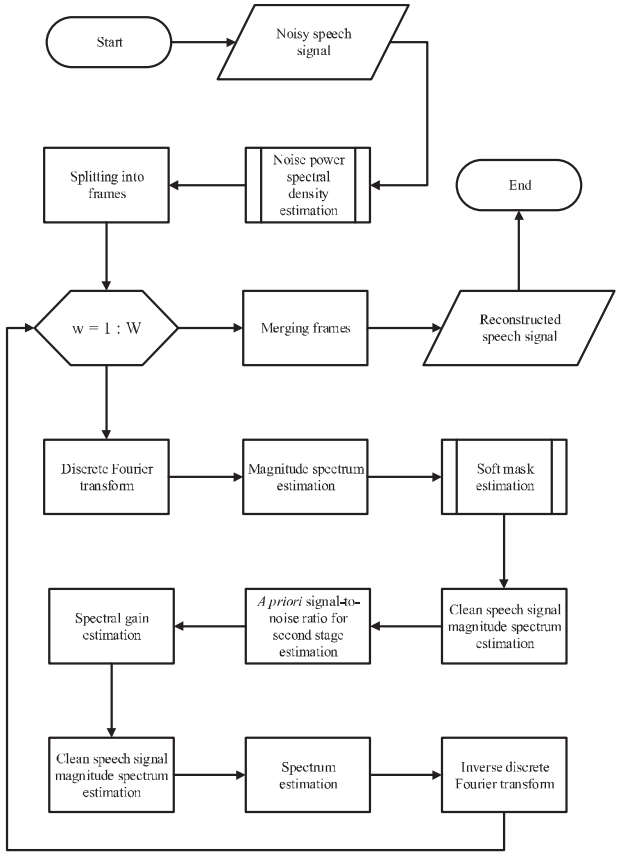

Fig. 4. Two-step noise reduction algorithm based on soft mask block diagram

VI. Results

Speech signals with 8000 Hz sampling rate was separated into 25 ms overlapping frames. The amount of overlap was 60 %. Before discrete Fourier transform performing Hamming window was used.

Mel-frequency cepstral coefficients [25] were used as speech signal features. The triangular filter bank was built for frequency band from 0 to 4000 Hz. 24 coefficients were obtained for each frame, range from 2nd to 16th coefficient was used.

Gaussian mixture models with 128 components were used for speaker modeling. Universal background model [26] was used for Gaussian mixture model pretraining.

Two speech databases were using for experiments. Every speech database used own universal background model to minimize quantity of the Gaussian mixture model components. Specifications of speech databases is presented in Table I. RUS-31-5 database obtained by the authors themselves. ENG-20-5 database is based on the CHAINS Speech Corpus [27] samples.

The proposed algorithms were tested and compared with the existing noise reduction algorithms in the problem of speaker identification. A case of noise reduction algorithm absence was also considered. A list of used algorithms is presented in Table II.

TABLE I. SPECIFICATIONS OF SPEECH DATABASES

Parameter	RUS-31-5	ENG-20-5
Language	Russian	English
Quantity of speakers (men/women)	31 (31/0)	20 (12/8)
Learning signal duration, s	90	72–106
Test signal duration, s	3	2–3
Quantity of test signals	5	5
Quantity of speakers for universal background model (men/women)	132 (132/0)	16 (8/8)
Learning signal duration for universal background model, s	10	35–52

TABLE II. NOISE REDUCTION ALGORITHMS USED IN THE EXPERIMENTS

No.	Algorithm
1	Algorithm based on decision-directed approach and Wiener gain function ($\alpha = 0.98$)
2	Two-step algorithm based on minimum mean square error short-time spectral amplitude estimator ($\alpha = 0.99$)
3	Proposed algorithm based on soft mask ($\alpha = 0.99$)
4	Proposed two-step algorithm based on soft mask and minimum mean square error short-time spectral amplitude estimator ($\alpha = 0.99, \epsilon = 0.75, \delta = \infty$)

Parameters of the algorithms 2, 3, 4 were chosen using fast technique of speaker identification accuracy estimation [24]. The algorithm 1 used standard value of the parameter $\alpha = 0.98$ as it used in some other articles corresponding to the same problem [28], [29], [30].

In our experiments, signals were corrupted by additive white Gaussian noise, Speech babble and Vehicle interior noise from NOISEX-92 library [31]. Three values of signal-to-noise ratio were used. There are 5, 10, 15 dB.

For speaker identification system quality estimation it is used such metrics as speaker identification accuracy (SIA):

$$\text{SIA} = \frac{\text{quantity of correctly identified test signals}}{\text{quantity of test signals}}$$

SIA estimation was performed 10 times, all results were averaged. Results is presented in Table III.

The results shown that the proposed two-step noise reduction algorithm based on soft mask and minimum mean square error short-time spectral amplitude estimator (algorithm 4) is preferable for additive white Gaussian noise, the two-step noise reduction algorithm based on minimum mean square error short-time spectral amplitude estimator (algorithm 2) is preferable for speech babble noise, and the algorithm based on decision-directed approach and Wiener gain function (algorithm 1) is preferable for vehicle interior noise.

VII. CONCLUSION

So, in this paper noise reduction technique based on soft mask introduced by Lu and Loisou was considered. It was generalized: soft mask can be raised to arbitrary power determined based on chosen optimality criterion. Dependence of the power of soft mask was analyzed. Higher value of the power provides more noise suppression, and in this case soft mask is more closer to binary mask.

A new approach of soft mask estimation was introduced. It uses modified decision-directed approach, Wiener gain function, and assumption that the noise amplitude spectrum is Rayleigh distributed in each frequency band.

TABLE III. SPEAKER IDENTIFICATION ACCURACY IN DIFFERENT NOISY CONDITIONS, %

Noise	SNR	Alg.	RUS-31-5	ENG-20-5	Avg.
Additive white Gaussian noise	5 dB	None	26.1	15.6	20.8
		1	60.9	56.3	58.6
		2	74.7	63.1	68.9
		3	73.2	63.7	68.5
		4	77.0	66.8	71.9
	10 dB	None	46.1	25.7	35.9
		1	77.2	71.9	74.5
		2	90.6	82.7	86.7
		3	87.4	84.6	86.0
		4	91.6	86.0	88.8
	15 dB	None	67.5	70.5	69.0
		1	87.2	80.1	83.7
		2	97.9	93.2	95.5
		3	95.5	92.2	93.8
		4	97.8	95.1	96.5
Speech babble	5 dB	None	20.8	26.9	23.9
		1	63.0	59.6	61.3
		2	61.2	63.0	62.1
		3	56.3	63.0	59.7
		4	59.2	63.1	61.1
	10 dB	None	29.4	58.7	44.0
		1	85.4	81.9	83.6
		2	85.8	84.5	85.2
		3	83.4	83.0	83.2
		4	83.6	82.5	83.1
	15 dB	None	45.2	82.1	43.9
		1	92.7	90.5	78.8
		2	95.0	93.1	80.4
		3	93.5	92.6	78.6
		4	93.9	92.5	79.1
Vehicle interior noise	5 dB	None	18.0	47.7	32.8
		1	92.8	96.8	94.8
		2	74.2	98.0	86.1
		3	77.0	98.2	87.6
		4	82.6	98.5	90.5
	10 dB	None	18.1	86.0	52.1
		1	96.3	97.1	96.7
		2	93.7	99.6	96.7
		3	95.3	99.3	97.3
		4	95.7	99.8	97.8
	15 dB	None	20.3	91.8	56.0
		1	98.2	96.5	97.3
		2	99.0	100	99.5
		3	99.4	99.9	99.6
		4	99.2	99.8	99.5

The obtained algorithm was used as first step in two-step noise reduction algorithm. Minimum mean square error short-time spectral amplitude estimator as spectral gain function was chosen for second step.

Smoothing *a priori* signal-to-noise ratio for the second stage using exponential moving average with upper limit was proposed. It can reduce level of "musical" noise, but speech signals become less intelligibly.

In our experiments signals were corrupted by additive white Gaussian noise, Speech babble and Vehicle interior noise from NOISEX-92 library. Three values of signal-to-noise ratio were used. There are 5, 10, 15 dB. Four algorithms were used in our experiments: the algorithm based on decision-directed approach and Wiener gain function, the two-step algorithm based on minimum mean square error short-time spectral amplitude estimator, the proposed algorithm based on soft mask, the proposed two-step algorithm based on soft mask and minimum mean square error short-time spectral amplitude estimator. The proposed two-step algorithm based on soft mask and minimum mean square error short-time spectral amplitude

estimator demonstrates better results than existing methods in additive white Gaussian noise conditions.

ACKNOWLEDGMENT

This work was supported by RFBR grant No. 15-08-99639.

REFERENCES

[1] J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Overview of speech enhancement techniques for automatic speaker recognition," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP 96*, 1996, vol. 2, pp. 929-932.

[2] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, 1996, vol. 2, pp. 629-632.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust.*, vol. 27, no. 2, pp. 113-120, Apr. 1979.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust.*, vol. 32, no. 6, pp. 1109-1121, Dec. 1984.

[5] C. Plapous, C. Marro, L. Mauuary, and P. Scalart, "A two-step noise reduction technique," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, vol. 1, no. 5, pp. 289-292.

[6] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction," *Speech Commun.*, vol. 50, no. 6, pp. 453-466, Jun. 2008.

[7] C. Plapous, C. Marro, and P. Scalart, "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 6, pp. 2098-2108, Nov. 2006.

[8] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586-1604, 1979.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust.*, vol. 33, no. 2, pp. 443-445, Apr. 1985.

[10] K. Khaldi, M. Turki-Hadj Alouane, and A.-O. Boudraa, "A new EMD denoising approach dedicated to voiced speech signals," in *2008 2nd International Conference on Signals, Circuits and Systems*, 2008, pp. 1-5.

[11] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 17-20.

[12] P. Sprechmann, A. Bronstein, M. Bronstein, and G. Sapiro, "Learnable low rank sparse models for speech denoising," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 136-140.

[13] L. Zhou, Y. Hu, S. Xiahou, W. Zhang, C. Zhang, Z. Li, and D. Hao, "Application of denoising algorithm based on LPSO-WNN in speech recognition," in *2013 International Conference on Communications, Circuits and Systems (ICCCAS)*, 2013, pp. 347-349.

[14] S. A. Novoselov, A. I. Topnikov, A. I. Savvatin, and A. L. Priorov, "Speech Denoising by Non-Local Means," *Tsifrovaya obrabotka signalov [Digital signal processing]*, no. 4, pp. 23-28, 2011.

[15] Y. Lu and P. C. Loizou, "Estimators of the Magnitude-Squared Spectrum and Methods for Incorporating SNR Uncertainty," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 5, pp. 1123-1137, Jul. 2011.

[16] P. Renevey and A. Drygajlo, "Detection of reliable features for speech recognition in noisy conditions using a statistical criterion," in *Proceedings of Workshop on CRAC*, 2001, pp. 71-74.

[17] Y. Li and D. Wang, "On the optimality of ideal binary time-frequency masks," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, no. 1, pp. 3501-3504.

[18] J. Jensen and R. C. Hendriks, "Spectral Magnitude Minimum Mean-Square Error Estimation Using Binary and Continuous Gain Functions," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 20, no. 1, pp. 92-102, Jan. 2012.

[19] D. Wang, "On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis," in *Speech Separation by Humans and Machines*, Boston: Kluwer Academic Publishers, 2005, pp. 181-197.

[20] D. Wang, "Time-Frequency Masking for Speech Separation and Its Potential for Hearing Aid Design," *Trends Amplif.*, vol. 12, no. 4, pp. 332-353, Oct. 2008.

[21] Y. Hu and P. Loizou, "Techniques for estimating the ideal binary mask," in *Proc. 11th Int. Workshop Acoust. Echo Noise Control*, 2008, pp. 154-157.

[22] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust.*, vol. 28, no. 2, pp. 137-145, Apr. 1980.

[23] G. S. Tupitsin, A. I. Topnikov, and A. L. Priorov, "Modification of the Two-Step Noise Reduction Techique for Improving the Quality of Speaker Identification in Noisy Conditions," *Informatsionnie systemy i tekhnologii [Information systems and technologies]*, no. 6, pp. 39-47, 2015.

[24] G. S. Tupitsin, "Speech signal preprocessing in automatic speaker identification systems," *Ph.D. dissertation*, Alexander Grigorievich and Nikolay Grigorievich Stoletovs Vladimir State University, Vladimir, Russia, 2015.

[25] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust.*, vol. 28, no. 4, pp. 357-366, Aug. 1980.

[26] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digit. Signal Process.*, vol. 10, no. 13, pp. 19-41, Jan. 2000.

[27] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The CHAINS Speech Corpus: CHAracterizing INdividual Speakers," in *Proc of SPECOM*, 2006, pp. 1-6.

[28] J. Bai, R. Zheng, B. Xu, and S. Zhang, "Robust speaker recognition integrating pitch and Wiener filter," in *SympoTIC 04. Joint 1st Workshop on Mobile Future & Symposium on Trends In Communications (IEEE Cat. No.04EX877)*, 2004, pp. 69-72.

[29] U. Bhattacharjee and P. Das, "Performance Evaluation of Wiener Filter and Kalman Filter Combined with Spectral Subtraction in Speaker Verification System," *Int. J. Innov. Technol. Explor. Eng.*, vol. 2, no. 2, pp. 108-112, 2013.

[30] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, "The effect of noise on modern automatic speaker recognition systems," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4249-4252.

[31] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247-251, Jul. 1993.