An Improved Signal Subspace Algorithm for Speech Enhancement

Xuzheng Dai, Baoxian Yu, Xianhua Dai

An Improved Signal Subspace Algorithm for Speech Enhancement

Xuzheng Dai[1], Baoxian Yu[2], Xianhua Dai[3]✉

[1]School of Information Science and Technology, Sun Yat-sen University, China
463418661@qq.com
[2]School of Information Science and Technology, Sun Yat-sen University, China
yubx@mail2.sysu.edu.cn
[3]School of Information Science and Technology, Sun Yat-sen University, China
issdxh@mail.sysu.edu.cn

Abstract. Most of the algorithms for speech enhancement are designed to improve the speech listening comfort. However the frequency spectrum character is destroyed seriously after the speech enhancement. To achieve better speech listening comfort with less frequency spectral damages, we present an improved signal subspace algorithm for speech enhancement. Compared with the traditional signal space method, the improved algorithm can decrease the Mel-frequency Cepstral Coefficients (MFCC) distance, an evaluation measure which means less frequency spectral damages to the voice and keep the voices' intelligence at the same time. Besides, the method can enlarge the distance of the easily confused voices, which means the improvement of the voice recognition ratio. Thus we get the purpose of the speech enhancement. The improved algorithm is used in a speech recognition program and has a good performance.

Keywords: Speech enhancement, signal subspace method, wiener filtering, prior SNR, Mel-frequency Cepstral Coefficients

1 Introduction

Speech enhancement and voice recognition have been widely used in recent years. In some occasions, the noisy environment will destroy the frequency spectrum character and lead to an erroneous recognition. Thus the speech enhancement algorithm needs to reduce the noise and keep integrality of the frequency spectrum character at the same time. Ephraim (1995) proposed the signal subspace approach to minimize the speech distortion and keep the residual noise below a preset threshold [1]. Hu (2003) proposed a generalized subspace approach for speech enhancement in both white noise and colored noise environment, and derived a time-domain estimator constraint and a spectral domain constraint[2]. The well-known decision-directed technique for speech enhancement limits the musical noise well[3], but the estimated priori signal-to-noise ratio (SNR) is biased since it depends on the speech spectrum estimation of the previous frame which degrades the noise reduction performance. Plapous (2004)

proposed a two-step noise reduction (TSNR) technique to solve this problem while maintaining the effect of the decision-directed approach [4]. Plapous (2005) also proposed a harmonic regeneration noise reduction method (HRNR) for solving the harmonic distortion in enhanced speech by regenerating the degraded harmonics of the distorted signal in an efficient way [5]. Objective and subjective measures prove the improvement than TRNR approach. However, the TSNR method destroyed the frequency spectrum character of the speech more seriously than TRNR. So TSNR is the algorithm we need in this paper.

The key of TRNR is to estimate the SNR, while a deviation of SNR in signal subspace approach significantly affects the speech performance. To solve the problem, we present an improved signal subspace algorithm that combines the signal subspace approach and TRNR that can get a better performance in speech enhancement.

Mel-frequency Cepstral Coefficients (MFCC) is one of the best approaches for voice recognition[6][7]. We propose the distance of the Mel-frequency Cepstral Coefficients as a measure to evaluate the effect of speech enhancement methods. The less the distance means the less damage to the frequency of the voices and better effect of the speech enhancement. The evaluation measure is based on the spectral features so as to evaluate the result of the speech enhancement from the perspective of speech recognition. The experiments verify that the new algorithm has a better performance than the others.

2 Speech enhancment approaches

In this section, we briefly review the signal subspace approach and TSNR algorithm. Then we discuss the shortcoming of each algorithm.

A. Signal Subspcace approach

The signal subspace approach is based on the theory of projecting the signal onto two subspaces: the signal-plus-noise subspace and the noise subspace. Thus we can remove the noise part through the decomposition of the signal. The decomposition can be either the singular value decomposition (EVD) or the eigenvalue decomposition (SVD), and in fact the two-decomposition method can be mutual transformed.

A linear clean signal $\mathbf{x}$ can be described as:

$$\hat{\mathbf{x}} = \mathbf{H}\mathbf{y} \tag{1}$$

Where $\mathbf{y}$ is the noisy signal and $\mathbf{H}$ is a $K \times K$ matrix whose rank is M, and $M < K$. Thus the error of the signal can be obtained by:

$$\begin{aligned} \boldsymbol{\varepsilon} &= \hat{\mathbf{x}} - \mathbf{x} = \mathbf{H}\mathbf{y} - \mathbf{x} = (\mathbf{H} - \mathbf{I})\mathbf{x} + \mathbf{H}\mathbf{d} \\ &= \boldsymbol{\varepsilon}_x + \boldsymbol{\varepsilon}_d \end{aligned} \tag{2}$$

Where $\boldsymbol{\varepsilon}_x$ represents the speech distortion and $\boldsymbol{\varepsilon}_d$ represents the residual noise. So the time-domain constrained optimization is given by making:

$$\min_{H} \overline{\boldsymbol{\varepsilon}}_x^2 \tag{3}$$

Subject to:

$$\frac{1}{K}\overline{\boldsymbol{\varepsilon}}_d^2 \le \alpha\sigma^2 \tag{4}$$

Where σ^2 is a positive constant and $0<\alpha<1$ for scaling. We can get the answer by constructing a Lagrange multiplier, and we can get the optimization of $\mathbf{H}$ in white noise environment:

$$\begin{aligned}\mathbf{H}_{opt} &= \mathbf{R}_d\mathbf{U}\boldsymbol{\Lambda}_\Sigma(\boldsymbol{\Lambda}_\Sigma+\mu\mathbf{I})^{-1}\mathbf{U}^T \\ &= \mathbf{U}^{-T}\boldsymbol{\Lambda}_\Sigma(\boldsymbol{\Lambda}_\Sigma+\mu\mathbf{I})^{-1}\mathbf{U}^T \\ &= \mathbf{U}^{-T}\mathbf{G}\mathbf{U}^T\end{aligned} \tag{5}$$

Where $\mathbf{G}=\boldsymbol{\Lambda}_\Sigma(\boldsymbol{\Lambda}_\Sigma+\mu\mathbf{I})^{-1}$. μ is the multiplier factor, $\mathbf{R}_x$ is the covariance matrix of clean signal, and $\mathbf{R}_d$ is the covariance matrix of noise. $\boldsymbol{\Lambda}_\Sigma$ and $\mathbf{U}$ are the eigenvalue matrix and eigenvector matrix of $\boldsymbol{\Sigma}$,where $\boldsymbol{\Sigma}=\mathbf{R}_d^{-1}\mathbf{R}_x$.

Searle in his book tells the theory that a matrix $\mathbf{U}$ which can simultaneously diagonalize $\mathbf{R}_\mathbf{x}$ and $\mathbf{R}_d$ [8]. Thus we can get:

$$\begin{aligned}\mathbf{U}^T\mathbf{R}_x\mathbf{U} &= \boldsymbol{\Lambda}_\Sigma \\ \mathbf{U}^T\mathbf{R}_d\mathbf{U} &= \mathbf{I}\end{aligned} \tag{6}$$

The matrix $\mathbf{G}$ is a diagonal matrix, thus the k th diagonal element g_{kk} is given by:

$$g_{kk}=\begin{cases}\dfrac{\lambda_x^k}{\lambda_x^k+\mu}, & k=1,2,\cdots,M \\ 0, & k=M+1,\cdots K\end{cases} \tag{7}$$

The value of μ affects the quality of the enhanced speech directly. According to Dendrinos's theory, μ depends on the short time of SNR [9]:

$$\mu=\begin{cases}\mu_0-(SNR_{dB})/s, & -5\le SNR_{dB}\le 20 \\ \mu_{\min}, & SNR_{dB}>20 \\ \mu_{\max}, & SNR_{dB}<5\end{cases} \tag{8}$$

Where：

$$
\begin{aligned}
SNR_{dB} &= 10\log_{10} SNR \\
\mu_0 &= \mu_{\min} + 20 * s \\
s &= \frac{(\mu_{\max} - \mu_{\min})}{25}
\end{aligned} \tag{9}
$$

The value of $\mu_{\max}$ and $\mu_{\min}$ represent the maximum and minimum of μ , and they are chosen experimentally. SNR can be given by:

$$
SNR = \frac{tr(\mathbf{V}^T \mathbf{R}_x \mathbf{V})}{tr(\mathbf{V}^T \mathbf{R}_d \mathbf{V})} = \frac{\sum_{k=1}^{M} \lambda_x^{(k)}}{K} \tag{10}
$$

Thus we can get the time-domain constrained optimization of the signal subspace approach. Form (10) we can find that the SNR in each frame is a value rather than a vector, which means that the transmission function we got based on SNR is not that accurate, so we need to find some other ways to get more accurate SNR.

B. Two Step Noise Reduction approach

In some speech enhancement algorithms based on the SNR, two parameters are needed: the posteriori SNR and the priori SNR, which are computed by:

$$
\mathrm{SNR}_{post}^{local}(p,k) = \frac{\left|Y(p,k)\right|^2}{\left|N(p,k)\right|^2} \tag{11}
$$

and

$$
\mathrm{SNR}_{pri}^{local}(p,k) = \frac{\left|X(p,k)\right|^2}{\left|N(p,k)\right|^2} \tag{12}
$$

Where $Y(p,k)$, $X(p,k)$ and $N(p,k)$ represent the frequency spectral of the noisy speech, the clean speech and the noise. The directed-decision algorithm says that the posterior SNR can be given by[10]:

$$
\hat{SNR}_{pri}^{DD}(\mathrm{p,k}) = \beta \frac{|\hat{X}(p-1,k)|^2}{\hat{\gamma}_n(p,k)} + (1-\beta) P[\hat{SNR}_{post}(p,k) - 1] \tag{13}
$$

Where $\hat{SNR}_{pri}^{DD}(\mathrm{p,k})$ represents the priori SNR got from the directed-decision algorithm, $\hat{\gamma}_n(p,k)$ represents the estimated noise as we can't know the exact noise, and β is a constant to balance the result, usually, β=0.97 .

According to the Weiner filtering theory, the transmission function $H_{DDopt}(p,k)$ is given by:

$$H_{DDopt}(p,k)=\frac{\hat{SNR}_{pri}^{DD}(\mathrm{p,k})}{1+\hat{SNR}_{pri}^{DD}(\mathrm{p,k})} \tag{14}$$

The experiment demonstrates that the directed-decision algorithm can reduce the "music noise" well, however the SNR got from (13) has a frame delay compares with the speech, especially at the speech onset and offset moment, which will limit the noise reduction performance and bring in some new reverberation effect.

The TSNR approach computes the SNR of the next frame using the transmission function got by directed-decision approach as:

$$\begin{aligned}&\hat{SNR}_{pri}^{\mathrm{TSNR}}(p,k)=\hat{SNR}_{pri}^{DD}(p+1,k)\\&=\beta^{'}\frac{\left|H_{DDopt}(p,k)X(p,k)\right|^2}{\hat{\gamma}_n(p,k)}+(1-\beta^{'})P[\hat{SNR}_{post}(p+1,k)-1]\end{aligned} \tag{15}$$

Where $\beta^{'}$ has the same effect as β, and we make $\beta^{'}=1$ because we can't know the information in the $p+1$th frame. The experiment suggests that the TSNR approach has a good performance on estimating the SNR. However, the TSNR has a large attenuation of the signal energy, which leads to a low sound of the voice signal. The low energy of the voice is disadvantage for the speech recognition. So even though the TSNR has a good performance in SNR and listening intelligence, it is not a good algorithm for speech recognition.

3 The improved signal subspace algorithm

In signal subspace approach, the SNR calculated in (10) is a value rather than a vector, which means that the SNR is not accurate. Thus we instead the (10) by (15), which is much more accurate than (10). Thus the whole process of the new algorithm is shown in Figure 1.

The proposed algorithm can be formulated in the following ten steps. For each frame of the voice signal:

Step 1: Compute the covariance matrix $\mathbf{R}_y$ of the noisy signal, and compute $\mathbf{\Sigma}=\mathbf{R}_d^{-1}\mathbf{R}_y-\mathbf{I}$. Then update the matrix of the noise $\mathbf{R}_d$.

Step 2: Compute the decomposition of matrix $\mathbf{\Sigma}$ by $\mathbf{\Sigma U}=\mathbf{U\Lambda}_{\Sigma}$.

Step 3: Sorting the eigenvalue of the matrix $\mathbf{\Sigma}$ as $\lambda_{\Sigma,1}\geq\lambda_{\Sigma,2}\geq\cdots\geq\lambda_{\Sigma,P}$, and we can estimate the rank of the speech signal subspace as $M=\max_{1\leq k\leq P}\arg\{\lambda_{\Sigma,k}>0\}$.

Step 4: Get the frequency spectrum of the noisy speech signal by FFT, and estimate the frequency spectrum of the noise signal at the same time.

Step 5: Compute the posteriori SNR by (11), and then compute the priori SNR by (13).

Step 6: Compute the TSNR SNR by (15).

Step 7: Compute the multiplier factor μ by (8) and (9) using the TSNR SNR we get in the previous step.

Step 8: Compute the diagonal elements g_{kk} of the matrix $\mathbf{G}$, and get the matrix $\mathbf{G}$ by:

$$G = diag\{g_{11}, g_{22}, ..., g_{MM}\} \tag{16}$$

Step 9: Compute the optimization $\mathbf{H}$ by (5).

Step 10: Estimate the enhanced speech signal by $\hat{\mathbf{x}} = \mathbf{Hy}$.

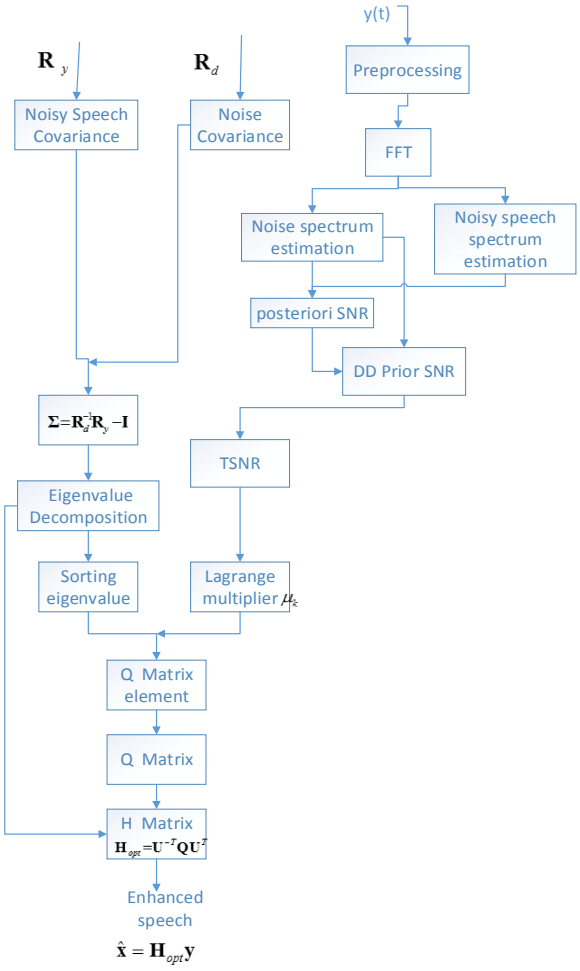

Figure 1. The process of the new speech enhancement algorithm

To illustrate the better performance of the improved algorithm, we take the speech enhancement result of a 10s length voice which is polluted by the factory noise from database NOISEX.92 as the example. The SNR after speech enhancement improves about 2.5dB in heavy noise environment. Figure 2 shows the enhancement result in time and frequency field.

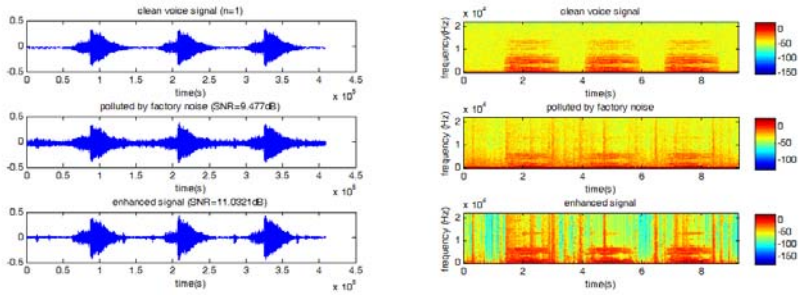

Figure 2. Speech enhancement result shown in time and frequency field.

Figure 2 shows that after the enhancement, the noise has been reduced pretty well. We can also get a comfort listening feeling after the enhancement. However the listening comfort and SNR cannot evaluate the performance of the algorithm in all directions. Thus we need some new evaluation measures to evaluate the algorithm.

4 Performance evaluation of the improved signal subspece algorithm by MFCC distance

After speech enhancement, we need to evaluate the performance of the algorithm. We propose MFCC distance as a new measure to evaluate the algorithm. The key of MFCC is transforming the speech signal from frequency into mel-frequency by:

$$mel(f) = 295 * \log_{10}(1 + f / 700) \tag{17}$$

Where f represents the frequency, and the specific calculation process of MFCC is in Figure 3:

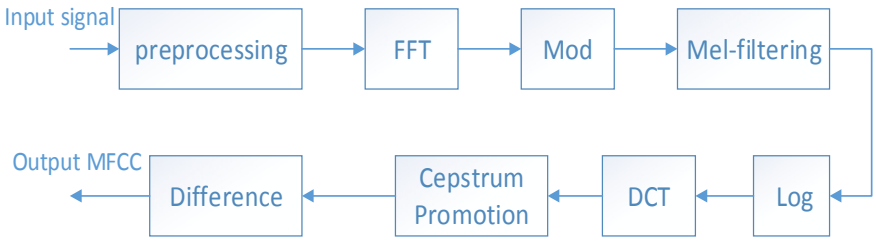

Figure 3. Process of MFCC

Figure 3 shows the process of getting the MFCC of the speech. The preprocessing includes Frame Blocking, Pre-emphasis and Hamming-windowing. Then get the frequency spectrum by FFT (Fast Fourier Transform Algorithm), and get the mode of the spectrum. Mel-filtering is as shown as below:

$$H_m(k)=\begin{cases}0 & ,k<f(m-1)\\ \dfrac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1)\le k\le f(m)\\ \dfrac{f(m+1)-k}{f(m)-f(m-1)}, & f(m)\le k\le f(m+1)\\ 0 & ,k>f(m-1)\end{cases} \tag{18}$$

Where $H_m(k)$ represents the coefficients of the Mel-filtering, k represents the frequency, and $f(m)$ represents the center frequency of the Mel-filtering, where $m=1,2,...,20$.

And we can get the logarithm of the frequency spectrum by:

$$s(p,m)=\ln(\sum_{k=0}^{N-1}\left|X_a(p,k)\right|^2 H_m(k)),\quad 0\le m\le M \tag{19}$$

Where $X_a(p,k)$ is the frequency spectrum through FFT of the signal, and $s(p,m)$ is the logarithm of the frequency spectrum in p th frame, the m th order of the Mel-filtering.

The DCT means Discrete Cosine Transformation:

$$C(p,n)=\sum_{m=0}^{N-1}s(p,m)\cos(\frac{\pi n(m-0.5)}{M}),\quad n=1,2,...,L \tag{20}$$

Where $C(p,n)$ is the MFCC coefficient, and L is the MFCC order number, usually we make $L=12$.

Usually the MFCC coefficients value in the low frequency is more easily interfered by the channel than in the high part, while the high part has a too high influence on speech recognition, thus the center part of the coefficients is the most useful and important. Thus we need a Cepstrum Promotion for the signal to promote the center part of the coefficients by:

$$\begin{aligned}&W(n)=1+\frac{L}{2}\sin\frac{\pi n}{L}\\ &mf(p,n)=C(p,n)W(n),\quad n=1,2,...,L\end{aligned} \tag{21}$$

Where $W(n)$ is the Cepstrum promotion transmission function, and $mf(p,n)$ is the MFCC coefficients after Cepstrum Promotion. However the MFCC coefficients now can only reflect the voice parameters of the current frame without the change of

the front and rear frame. So we take the difference of the MFCC coefficients in the front and rear frame into account by:

$$dmf(p,n) = \frac{2(mf(n+2) - mf(n-2))}{3} + \frac{+(mf(n+1) - mf(n-1))}{3} \tag{22}$$

$dmf(p,n)$ represents the difference coefficients. Then we combine the MFCC coefficients $mf(p,n)$ and difference coefficients $dmf(p,n)$ as the whole MFCC coefficients of the speech. We calculate the distance of the coefficients between the noisy speech and the clean speech. The less the distance is, the better performance of the enhancement algorithm is. And if the distance between the easily confused speech signals gets larger after enhancement, it means the enhancement algorithm has a better performance.

Table 1. Coefficients distance

Noise kind / algorithm	White Noise	Factory Noise	Average of both Noise
DD approach	2.803	2.759	2.713
TSNR	3.125	3.012	2.930
Signal Subspace	2.860	2.658	2.780
New algorithm	**2.000**	**2.236**	**2.144**

Table I shows the MFCC distance between the noisy speech and the clean speech. The speech signal is a toy voice of 10s length, and the speech signal is chosen from NOISEX.92 database. We can see that the distance of the new algorithm is less than TSNR and signal subspace approach, which concludes that the new algorithm has a better performance than TSNR and signal subspace approach.

Almost every speech enhancement algorithm will damage the frequency feature of the speech, then why do we still need enhancement algorithm? The reason is that some speeches' feature polluted by noise are easily got confused. Thus after the speech enhancement, the easily confused speech should be judged correctly. There are two German alphabets 'e' and 'i' which sound very similar, and are easily confused in severe noisy environment. After the enhancement, the misjudged signal can be judged correctly as seen in below:

Table 2. The new algorithm for confused voice acting

Clean speech / speech	i	e	Distance difference
Polluted i	1.8651	1.9104	0.0453

Polluted e	2.1306	2.1709	**-0.0403**
Enhanced i	3.2937	3.3849	0.0912
Enhanced e	3.9398	3.7787	**0.1611**

From the table Ⅱ we can get the conclusion that after the enhancement, the distance difference of alphabet ‘i’ is enlarged from 0.0453 to 0.0912 which means that the new algorithm makes the signal easier to be identified. Also, the distance difference of alphabet ‘e’ is -0.0403, which means that the alphabet ‘e’ can’t be identified correctly because of the noisy pollution. But after the enhancement, the distance becomes 0.1611, means that ‘e’ can be identified correctly, which demonstrates the role of the new method for speech recognition.

5 The application of the algorithm

In cooperation of our laboratory with a toy company, we are asked to design software which is used at the production line in the factory to recognize if the voices of the toys are right or not. The toys can sing many kinds of voices, and many of them are easily confused, such as the German alphabets ‘e’ and ‘i’ we test in the fourth part. The alphabets are more difficult to recognize in the noisy factory because of the loud noise. So we need some algorithms to remove the noise. We used MFCC algorithm for the voice recognition, but most of the speech enhancement algorithms would enlarge the MFCC distance, which will lead voice recognition to a failure. The algorithm can decrease the back noise, improve the feeling of our hearing and do not destroy the spectral features of the voice at the same time as far as possible, which is useful for the voice recognition. The use of our algorithm is shown in Figure 4:

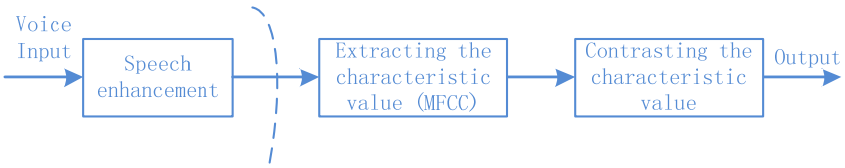

Figure 4. The use of the algorithm

Figure 4 shows the use of the improved algorithm. The left pat of the dotted line is the place where the algorithm is used in this paper. Because of the loud noise in the factory, the voices of the toys are polluted seriously, which will influence the recognition rate. We tried many algorithms include the traditional signal subspace algorithm, but none of them can meet their demands of the recognition rate. So we tried many improvements until we found the algorithm we propose in this paper. The improved algorithm meets the demand they want. As is shown in the fourth part, the improved algorithm can decrease the MFCC distance, keep the frequency correlation and voice intelligence at the same time. Also, the method can make the distance of the easily confused words and letters larger, which make it not that easy to be erroneously

judged. After we use the improved algorithm the average recognition rate of the toys' voices in the factory production line is from 73% to 91%, thus we get the purpose of the speech enhancement, and that is the value of the improved algorithm.

However, the complexity of the algorithm in this paper becomes larger, which is a big drawback of the algorithm. In the following research, the main focus is to reduce the complexity of the algorithm, and continue to improve the speech recognition rate toys.

6 Summary and conclusions

In this paper we propose a new algorithm for speech enhancement that combine the TSNR and signal subspace approach. And we propose MFCC coefficients distance to evaluate the performance of the speech enhancement algorithms. The experiments verify that the new algorithm has a better performance than the single.

References

1. Ephraim, Y. and Van Trees, H. L. A signal subspace approach for speech enhancement. Speech and Audio Processing, IEEE Transactions, 3(4): 251-266. (1995)
2. Hu, Y. and Loizou, P. C. A generalized subspace approach for enhancing speech corrupted by colored noise. Speech and Audio Processing, IEEE Transactions, 11(4): 334-341. (2003)
3. Ephraim, Y. and Malah, D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. Acoustics, Speech and Signal Processing, IEEE Transactions on, 32(6): 1109-1121. (1984)
4. Plapous, C., Marro, C., Mauuary, L. et al. A two-step noise reduction technique[C].Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on. IEEE, 2004, 1: I-289-92 vol. 1. (2004)
5. Plapous, C., Marro, C., Scalart, P. Speech enhancement using harmonic regeneration[C].IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2005). (2005)
6. Samal, A., Parida, D., Satapathy, M. R. et al. On the Use of MFCC Feature Vector Clustering for Efficient Text Dependent Speaker Recognition[C]//Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2013. Springer International Publishing: 305-312. (2014)
7. Sahidullah, M. and Saha, G. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech Communication. 54 (4): 543–565. doi:10.1016/j.specom.2011.11.004 (May 2012).
8. Searle, S. R. Matrix algebra useful for statistics. New York, (1982)
9. Dendrinos, M. Bakamidis, S., and Carayannis, G. Speech enhancement from noise: A regenerative approach. Speech Communication, 10(1): 45-57. (1991)
10. Cohen, I. On the decision-directed estimation approach of Ephraim and Malah. Acoustics, Speech, and Signal Processing. Proceedings. (ICASSP'04). IEEE International Conference on. IEEE, 2004, 1: I-293-6 vol. 1. (2004)