NONNEGATIVE MATRIX FACTORIZATION BASED NOISE ROBUST SPEAKER VERIFICATION

S.H. Liu, Y.X. Zou, H.K. Ning*

ADSPLAB/ELIP, School of Electronic Computer Engineering
Peking University, Shenzhen 518055, China
*zouyx@pkusz.edu.cn

ABSTRACT

The performance of speaker verification system (SVS) declines dramatically in noisy environments. To suppress the adverse impact of the noise on SVS, this paper investigates employing the nonnegative matrix factorization (NMF) technique to reconstruct the speech based on the pre-trained speech basis matrix (SBM) and noise basis matrix (NBM). The contribution of this research lies in utilizing the time correlation of the speech signal to obtain a more appropriate SBM. An enhanced NMF-based speech enhancement algorithm (ENMF-SE) is derived. Accordingly, the robust SVS based on ENMF-SE (ENMF-SE-SVS) is constructed and evaluated by intensive experiments with a public speech database. Experimental results show that the proposed ENMF-SE-SVS provides up relative improvement EER compared with the traditional NMF-SE based SVS algorithm under different SNR noise conditions.

Index Terms—speech enhancement; nonnegative matrix factorization; time correlation; speaker verification;

1. INTRODUCTION

Voiceprint is one of the unique biometric and has found wide applications including access control, providing forensic evidence, and user authentication in telephone banking, etc. Speaker verification (SV) aims at using voiceprint to verify the identity of the claimed speaker [1]. The state-of-art SV techniques perform perfectly in the laboratory environment, while in the real-world applications the performance degrades dramatically, which is mainly due to the mismatch between training data and testing data. Obviously, the mismatch is always present since there are various interfering sources, such as additive noise and channel distortion. The robust technology of the SV is still one of the challenging and most demanding techniques.

The mainstream technology of robust SV includes the pre-processing such as speech enhancement, robust feature such as RPCC [2],GFCC [3], robust model such as JFA [4], ivector [5]. This research will focus on single channel speech enhancement (SCSE) in the pre-processing module, as shown in Fig.2.. SCSE has a lot of research outcomes, such as famous spectral subtraction (SS) [6], short time spectral amplitude estimation based on minimum mean-squared error criterion (MMSE-STSA) [7], Wiener filtering [8], Kalman filtering [9] and magnitude spectrum enhancement (MSE) [10]. Research results showed that the methods discussed above achieved good performance under stationary or quasi-stationary noise condition, but in practical applications, there are also non-stationary noise.

To tackle this problem, some new methods are proposed, such as nonnegative matrix factorization (NMF) [11] and deep neural network (DNN) [12]. NMF was first applied to solve the speech separation in [13][14] and then adopted in the speech enhancement by [15][16]. Research results show NMF has great capacity of modeling the non-stationary noise [15].

NMF is a powerful mathematic tool, which can factorize a nonnegative matrix into two nonnegative sub-matrices known as building blocks termed as basis matrix and time-varying activation levels of building block termed as encoding matrix. This provides a way of decomposing a signal into a convex combination of nonnegative building blocks. NMF based speech enhancement (NMF-SE) technique is to project the mixture signal onto speech basis matrix (SBM) and noise basis matrix (NBM) separately. Then the speech signal will be reconstructed through the corresponding SBM. Generally, a supervised training is involved to train a robust and expressive SBM and NBM from the prior knowledge.

Our research utilizes the time correlation of the speech signal to obtain a more expressive SBM in the training stage. Then an enhanced NMF based speech enhancement algorithm (ENMF-SE) is proposed. Accordingly, a robust SVS based on ENMF-SE (ENMF-SE-SVS) is constructed.

The rest of this paper is organized as follows: Sect.2 provides the problem formulation, including the UBM-MAP-GMM based SVS, standard NMF algorithm and NMF-SE algorithm. The proposed ENMF-SE is introduced in Sect.3. Experimental results are presented in Sect.4. We conclude our work at last.

978-1-4799-1948-2/15/$31.00 ©2015 IEEE

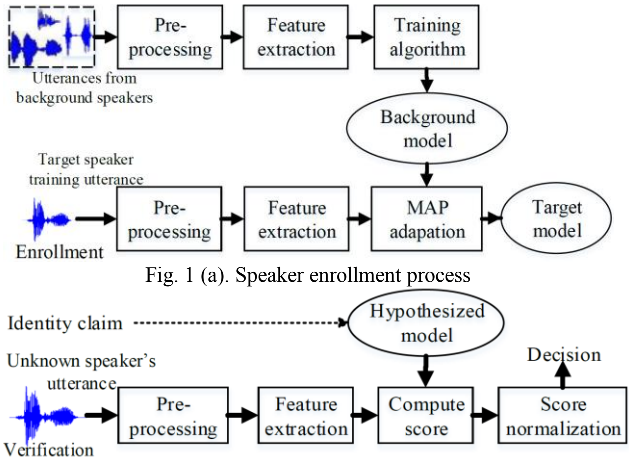


Fig. 1 (a). Speaker enrollment process

Fig. 1 (b). Speaker verification process

Fig. 1. Block diagram of the UBM-GMM-MAP based speaker verification system

2. PROBLEM FORMULATION

2.1. Speaker verification system

The UBM-MAP-GMM based speaker verification (SV) is one of the famous SV techniques. Its system diagram is shown in Fig.1. Fig. 1(a) shows the speaker enrollment process, where the GMM model of each speaker is computed and stored. Specifically, the feature extraction module computes the mel-frequency cepstral coefficients (MFCCs) by the method proposed in [17]. The background model is firstly trained by expectation maximization (EM) training algorithm [18], taking MFCCs as the training data. And MAP adaption is used to derive the target speaker's GMM model by taking the MFCCs of the target speaker's training utterances. Fig. 1(b) presents the SV process. We can compute the score by taking the MFCCs of the tested utterance and the claimed speaker's GMM model. Moreover, to achieve the better decision, a normalization module is added after the scoring module.

The performance of SVS dramatically degrades under noise condition. As we have mentioned above, there are many methods to develop a robust SVS, but our research aims at using the NMF algorithm to enhance the speech in pre-processing module of SVS, which is shown in Fig.2..

2.2. Nonnegative matrix factorization (NMF) algorithm

To make the presentation clear, in this section, the basic NMF algorithm [11] will be briefly described. Given a nonnegative matrix $\boldsymbol{V}$ ($\boldsymbol{V}\in\boldsymbol{R}^{m\times n}$) and a constant r, NMF factorizes $\boldsymbol{V}$ into two nonnegative matrices $\boldsymbol{W}$ ($\boldsymbol{W}\in$ $\boldsymbol{R}^{m\times r}$) and $\boldsymbol{H}$ ($\boldsymbol{H}\in\boldsymbol{R}^{r\times n}$)

$$\boldsymbol{V}\approx\hat{\boldsymbol{V}}=\boldsymbol{W}\boldsymbol{H},\ \ (\boldsymbol{W},\boldsymbol{H}\geq 0) \tag{1}$$

Typically, r is chosen by $(m+n)r<=mn$, giving a lower-rank approximation of $\boldsymbol{V}$. The matrix $\boldsymbol{W}$, which is always termed as basis matrix. While the matrix $\boldsymbol{H}$ is termed as encoding matrix. NMF performs the decomposition by minimizing a cost function, which is usually a distance matric between $\boldsymbol{V}$ and $\boldsymbol{WH}$. The Kullback-Leibler (KL) is used as the cost function, which is denoted as follows [15].

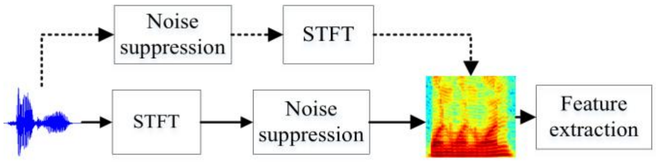


Fig. 2. Block diagram of the pre-processing module in SVS

$$D\left(\boldsymbol{V}\,\|\,\hat{\boldsymbol{V}}\right)=\sum_{ij}\left(\boldsymbol{V}_{ij}\log(\boldsymbol{V}_{ij}/\hat{\boldsymbol{V}}_{ij})+\boldsymbol{V}_{ij}-\hat{\boldsymbol{V}}_{ij}\right)\quad \hat{\boldsymbol{V}}_{ij}=(\sum_{j=1}^{r}\boldsymbol{H}_{j.}\boldsymbol{W}_{.j})_{ij} \tag{2}$$

where the matrix $\boldsymbol{V}$-hat is the estimation of $\boldsymbol{V}$. The update rule for $\boldsymbol{W}$, $\boldsymbol{H}$ is given as [19]:

$$\boldsymbol{W}_{ik}\leftarrow\boldsymbol{W}_{ik}\frac{\sum_j \boldsymbol{H}_{kj}\boldsymbol{V}_{ij}/(\boldsymbol{WH})_{ij}}{\sum_j \boldsymbol{H}_{kj}}\quad \boldsymbol{H}_{kj}\leftarrow\boldsymbol{H}_{kj}\frac{\sum_i \boldsymbol{W}_{ik}\boldsymbol{V}_{ij}/(\boldsymbol{WH})_{ij}}{\sum_i \boldsymbol{W}_{ik}} \tag{3}$$

2.3. NMF based speech enhancement (NMF-SE)

In this section the basic principle of the NMF based speech enhancement (NMF-SE) will be introduced following Wilson's work [15]. First of all, we consider the noisy speech $x(t)$ which is generated by the clean speech $s(t)$ and noise $n(t)$ in an additive manner and denoted as

$$x(t)=s(t)+n(t) \tag{4}$$

Take short-time Fourier Transform (STFT) on (4) and assume $\boldsymbol{V}_{speech}$, $\boldsymbol{V}_{noise}$, and $\boldsymbol{V}_{noisy}$ is the spectra magnitude matrix of clean speech signal, noise, and noisy speech signal, respectively. Approximately, we can get (5)

$$\boldsymbol{V}_{noisy}=\boldsymbol{V}_{speech}+\boldsymbol{V}_{noise} \tag{5}$$

According to the NMF principle described in (1), $\boldsymbol{V}_{speech}$ $\boldsymbol{V}_{noise}$ and $\boldsymbol{V}_{noisy}$ can be expressed as follows, respectively.

$$\boldsymbol{V}_{speech}\approx\boldsymbol{W}_{speech}\boldsymbol{H}_{speech} \tag{6}$$

$$\boldsymbol{V}_{noise}\approx\boldsymbol{W}_{noise}\boldsymbol{H}_{noise} \tag{7}$$

$$\boldsymbol{V}_{noisy}\approx\boldsymbol{W}_{noisy}\boldsymbol{H}_{noisy} \tag{8}$$

Substituting (6)-(8) into (5)，we can get the following:

$$\begin{aligned}\boldsymbol{V}_{noisy}&\approx\boldsymbol{W}_{noisy}\boldsymbol{H}_{noisy}\approx\boldsymbol{W}_{speech}\boldsymbol{H}_{speech}+\boldsymbol{W}_{noise}\boldsymbol{H}_{noise}\\&\approx\left[\boldsymbol{W}_{speech}\boldsymbol{W}_{noise}\right]\begin{bmatrix}\boldsymbol{H}_{speech}\\ \boldsymbol{H}_{noise}\end{bmatrix}\end{aligned} \tag{9}$$

In the second equation of (9), since the noisy data is available, $\boldsymbol{W}_{noisy}$ and $\boldsymbol{H}_{noisy}$ can be obtained by NMF decomposition From the last equation of (9), if $\boldsymbol{W}_{speech}$ and $\boldsymbol{W}_{noise}$ are obtained in advance, it is able to get the $\boldsymbol{H}_{speech}$ from $\boldsymbol{H}_{noisy}$.. To make the presentation clear, a diagram of NMF-SE is given in Fig.3. It is clear that NMF-SE has two stages, in the training stage, $\boldsymbol{W}_{speech}$ and $\boldsymbol{W}_{noise}$ are trained separately with training speech data and training noise data. The procedure of training $\boldsymbol{W}_{speech}$ and $\boldsymbol{W}_{noise}$ is same. In the enhancement stage, the matrix $\boldsymbol{W}$ is constructed as $\boldsymbol{W}$ = [$\boldsymbol{W}_{speech}$ $\boldsymbol{W}_{noise}$]. The coefficient matrix $\boldsymbol{H}_{noisy}$ is determined by the NMF and the corresponding $\boldsymbol{H}_{speech}$ can be derived. As a result, the enhanced speech data can be reconstructed as follows.

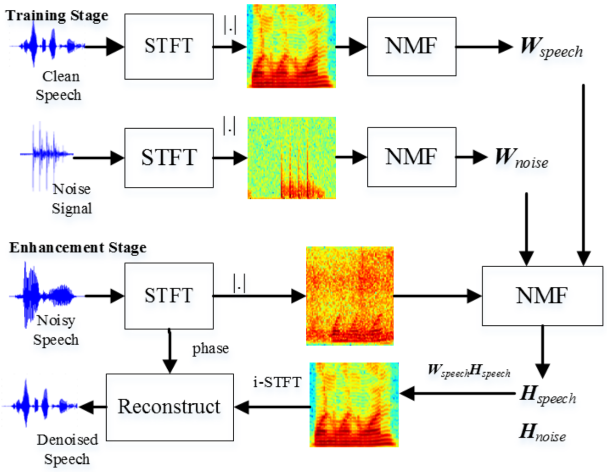


Fig. 3. Block diagram of NMF based speech enhancement

$$\hat{V}_{speech} = \boldsymbol{W}_{speech}\boldsymbol{H}_{(1:r_s,:)} \tag{10}$$

where r_s represents the number of SBM vectors, which is chosen by empirical values.

3. THE PROPOSED NMF BASED SPEECH ENHANCEMENT ALGORITHM

It is noted that, in NMF-SE algorithm, the spectrogram of the speech and noise are uniformly divided as the data matrix in time domain. However, speech and noise have different time-spectral characteristics. For example, speech has harmonic structure in time frequency (TF) spectrogram, but noise doesn't [20]. So in this section, we will take the unique speech TF property into account to train a more expressive SBM by using the clean speech data. More specifically, time correlation will be measured to develop an enhanced NMF-based speech enhancement (ENMF-SE) algorithm. The block diagram of the ENMF-SE algorithm is shown in Fig.4.

As shown in Fig.4., $\boldsymbol{X}$={$\boldsymbol{X}_1$, $\boldsymbol{X}_2$,…, $\boldsymbol{X}_T$} is computed by taking the STFT of the clean training speech $\boldsymbol{x}(t)$, where T is the number of speech frames. Then in the frame dividing module, $\boldsymbol{X}$ is automatically divided into Q segments in sequence, which is expressed as $\boldsymbol{X}$={$\boldsymbol{b}_1$, $\boldsymbol{b}_2$,…,$\boldsymbol{b}_Q$}. For each segment b_i=[$\boldsymbol{b}_{i,1}$, $b_{i,2}$,…,$\boldsymbol{b}_{i,m}$], the distance $d(i)$ of the frames in b_i is shown in (11), where m_i is the number of frames in b_i and μ_i indicates the magnitude mean value of b_i. And $dist(\boldsymbol{b}_{i,j},\mu_i)$ represents the distance between $\boldsymbol{b}_{i,j}$ and μ_i, the Euclidean Distance is employed to measure $dist(\cdot)$. Then according to (12), the Q segment is derived from $\boldsymbol{X}$ by m.

$$D(i,m_i) = \sum_{j=1}^{m_i} dist(b_{i,j},\mu_i) \tag{11}$$

$$\arg\min_{m_i} D(i,m_i) \tag{12}$$

It is noted the acoustic characteristics of different phoneme during a median time span are captured through this way. After the frame dividing module, Q segments are derived. To deal with the different time duration of acoustic characteristics, the segment normalization module is added to normalize the length of the Q segments by using the resample method. Then an expressive SBM can be trained by taking the output of the segment normalization module. Besides, the other steps in ENMF-SE are the same as those in NMF-SE.

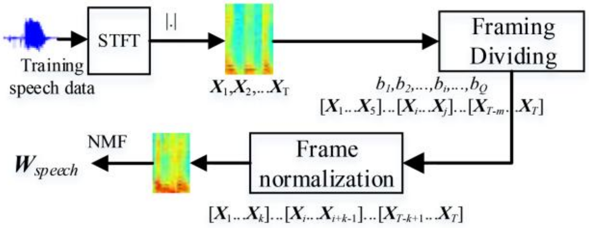

Training speech data
STFT
|.|
$X_1,X_2,...X_T$
Framing Dividing
$b_1,b_2,...,b_i,...,b_Q$
$[X_1...X_5]...[X_i...X_j]...[X_{T-m}...X_T]$
Frame normalization
$[X_1...X_k]...[X_i...X_{i+k-1}]...[X_{T-k+1}...X_T]$
NMF
W_{speech}

Fig. 4. Block diagram of proposed ENMF-SE training stage

From Fig.3., it is clear that the computational complexity of NMF-SE in training stage is $O(n)$ ($O(n)$=$O(mnlogn)$+$O(kmn)$), where m and n are the number of frames and frequency bins, respectively. And k is the number of basis vectors of the NMF used for training SBM and NBM. While in the proposed ENMF-SE algorithm, the complexity is $O'(n)$ ($O'(n)$=$O(mnlogn)$+$O(kmn)$+$O(k_1m)$), where k_1 is a constant and indicate the max number of the relevant frames.

To make the presentation clear, the proposed ENMF-SE algorithm is summarized in Table 1.

Table 1: The proposed ENMF-SE algorithm

Training stage:

1. Convert the training clean speech data and noise data into time-frequency (TF) domain by STFT, in which the magnitude spectra of the speech frames formed $\boldsymbol{X}$ and the noise frames formed $\boldsymbol{V}_{noise}$.
2. Dividing $\boldsymbol{X}$ into Q segments according to (11) and (12) to form $\boldsymbol{V}_{speech}$.
3. Compute $\boldsymbol{W}_{speech}$ and $\boldsymbol{W}_{noise}$ described in Section 2.2

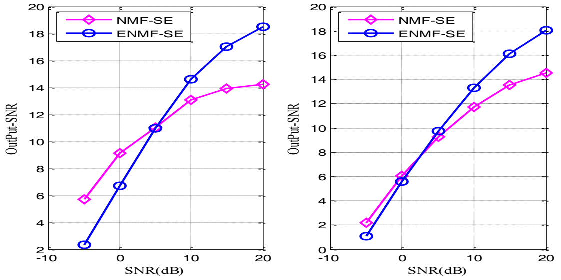
Enhancement stage:

1. Convert the noisy speech data into TF domain by STFT
2. Construct $\boldsymbol{W}$ ($\boldsymbol{W}$= [$\boldsymbol{W}_{speech}$ $\boldsymbol{W}_{noise}$])
3. Compute $\boldsymbol{H}$ by (9) and the updating rules are shown in Section 2.2
4. Reconstruct the speech signal according to (10) and the corresponding phase

4. EXPERIMENTS

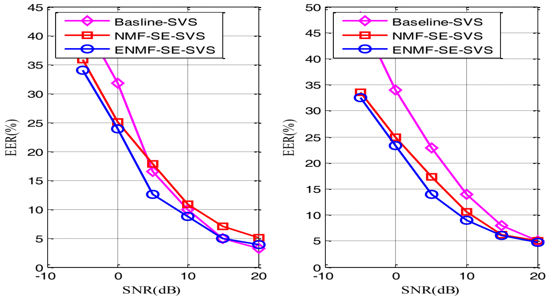
4.1. Experiment setup

TIMIT database [21] is employed in our experiment. 530 utterances from the 630 speakers are randomly chosen, which gives 25 minutes training speech. The training speech is down-sampled to 8kHz with the frame length of 256 samples (32 msec) and a frame shift of 128 samples. Then it is transferred to 513 dimensions spectra magnitude by STFT to form the training feature data set, which is used to train the speech basis matrix (SBM). The max number of the update epoch for (3) is set as 500. The factorization factor (r) for $\boldsymbol{W}_{speech}$ in (6) and $\boldsymbol{W}_{noise}$ in (7) is set to 40 and 20, respectively.



(a) Machine-Gun noise (b) Car noise

Fig.5. Output SNR versus Input SNR levels



(a) Babble noise (b) Car noise

Fig.6. EER versus SNR levels

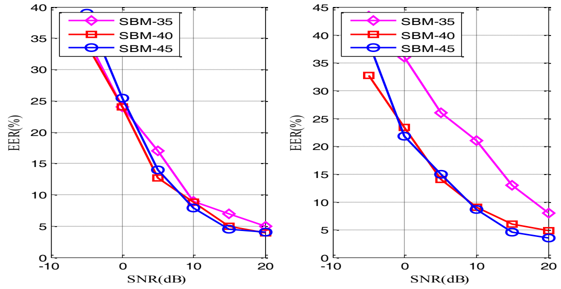
MSR Identity Toolbox [22] is employed to achieve the SVS. The parameters are chosen followed the setting given in [23]. 20 dimension MFCCs are extracted and used in SVS. Besides, compared to [23], 22 types of noise are used to mix additively with the clean training speech to generate the noisy speech dataset to train the UBM model. The number of GMM for UBM model is set to 1024 and the epoch for EM algorithm is set to 20.

We take the commonly used equal error rate (EER) as the measure for evaluating the performance of the speaker verification systems (SVS).

4.2. Experimental results and analysis

Experiment 1: As discussed before, the proposed ENMF-SE algorithm essentially works as the speech enhancement (SE), which aims at eliminating the adverse impact of the noise on the SVS. Therefore, this experiment is conducted to evaluate the SE ability of ENMF-SE, where the most commonly used Signal-to-Noise Ratio (SNR) will be taken as the performance measure. The results with machine-gun noise and car noise are shown in Fig.5. From Fig.5 (a) and (b), although the curves are slightly different, we can see that the ENMF-SE outperforms the NMF-SE when SNR is about higher than 5dB. This results explain the advantages by utilizing the time correlation of the speech signal to obtain a more appropriate SBM and then better SE is achieved when SNR is about higher than 5dB.

Experiment 2: Since we are targeting on the robust speaker verification problem, the EER performance under different SNR levels is evaluated in this experiment. The baseline-SVS, NMF-SE-SVS and ENMF-SE-SVS are considered, in which the parameter settings for the baseline-SVS and NMF-SE-SVS are the same as those for ENMF-SE-SVS. The results are plotted in Fig. 6. From Fig.6., it is clear to see that the proposed ENMF-SE-SVS outperforms the NMF-SE-SVS under about all SNR levels under babble noise and car noise. It is encouraged to see that with the decrease of the SNR, the performance of ENMF-SE-SVS increase more compared to that of the baseline SVS. For example, when SNR level is 0dB, the EER of ENMF-SE-SVS is about 6% higher and 10% higher than that of the baseline under babble noise and car noise, respectively.



(a) Babble noise (b) Car noise

Fig.7. EER versus SNR levels used ENMF-SE in SVS

Experiment 3: It is noted that, for ENMF-SE-SVS, the factorization factor (*r*) is an important parameter. In this experiment, we aim to evaluating the impact of the factor *r* on EER performance. The experimental settings are the same as those in Experiment 2 except we vary *r* from 35 to 45 with step-size of 5. Experimental results are shown in Fig. 7. From Fig.7.(a), we can see that for babble noise with different SNR levels, the impact of *r* is not too big. The largest variation is about 5%. However, from Fig.7.(b), we can see that, for car noise, *r* does have a big impact on EER. For example, for *r*=35, the EER degraded about 20% for each SNR level compared with the EER when *r* is set to 40 or 45. These results may tell us that using small *r* may cause poor EER performance. From the results, it suggests to set *r* to 40 considering the tradeoff between the computational complexity and EER performance for ENMF-SE-SVS.

5. CONCLUTION

As the performance of speaker verification system (SVS) declines dramatically in noisy environment, a NMF-SE is employed as the pre-processing method for a robust SVS. Considering the time correlation of the speech signal, an appropriate SBM is trained and then an ENMF-SE algorithm is derived to develop a robust ENMF-SE-SVS. Under the experimental settings, it is encouraged to see the proposed ENMF-SE-SVS compared to the baseline GMM-UBM-SVS is able to obtain about 5%-10% EER improvement when SNR lower than 10dB.

6. ACKNOWLEGEMENT

This work is partially supported by National Natural Science Foundation of China (No: 61271309) and Shenzhen Science Research Program (No. CXZZ20140509093608290).

7. REFERENCES

[1]. Reynolds, D.A., Speaker identification and verification using Gaussian mixture speaker models. Speech communication, 17(1): p. 91-108. 1995.

[2]. Wang, J. and M.T. Johnson. Residual Phase Cepstrum Coefficients with Application to Cross-lingual Speaker Verification. in Thirteenth Annual Conference of the International Speech Communication Association. 2012.

[3]. P. Clemins, M. Trawicki, K. Adi, J. Tao, and M. Johnson, \Generalized perceptual features for vocalization analysis across multiple species," in 2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006), vol. 1, 2006.

[4].] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," IEEE Trans. Audio, Speech, Lang.Process.,vol. 15, no. 4, pp. 1435–1447,May2007.

[5]. Dehak, N., et al., Front-end factor analysis for speaker verification. Audio, Speech, and Language Processing, IEEE Transactions on, 19(4): p. 788-798. 2011.

[6]. S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 27, no. 2, pp. 113-120, 1979.

[7]. Y. Ephraim, and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 33, no. 2, pp. 443-445, 1985.

[8]. Lim, J. S. and A. V. Oppenheim. "Enhancement and bandwidth compression of noisy speech." Proceedings of the IEEE 67(12): 1586-1604. 1979.

[9]. V. Grancharov, J. Samuelsson, and B. Kleijn, "On causal algorithms for speech enhancement," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 14, pp. 764-773, 2006.

[10]. Hung, J.-w., et al. "Enhancing the magnitude spectrum of speech features for robust speech recognition." EURASIP Journal on Advances in Signal Processing 2012(1): 1-20. 2012.

[11]. Lee, D. D., & Seung, H. S.. Algorithms for non-negative matrix factorization. In Advances in neural information processing systems (pp. 556-562). 2001.

[12]. Yong, X., et al.. "An Experimental Study on Speech Enhancement Based on Deep Neural Networks." Signal Processing Letters, IEEE 21(1): 65-68. 2014.

[13]. Schmidt, M. and R. Olsson. "Single-channel speech separation using sparse non-negative matrix factorization." 2006.

[14]. Smaragdis, P.. "Convolutive Speech Bases and Their Application to Supervised Speech Separation." Audio, Speech, and Language Processing, IEEE Transactions on 15(1): 1-12. 2007.

[15]. Wilson, K. W., et al.. Speech denoising using nonnegative matrix factorization with priors. ICASSP, Citeseer. 2008.

[16]. Wilson, K. W., et al.. Regularized non-negative matrix factorization with temporal dependencies for speech denoising. Interspeech. 2008.

[17]. Davis, S. and P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing, 28(4): p. 357-366. 1980.

[18]. Reynolds, D.A., T.F. Quatieri, and R.B. Dunn, Speaker verification using adapted Gaussian mixture models. Digital signal processing, 10(1): p. 19-41. 2000.

[19]. Lee, D. D. and H. S. Seung. "Learning the parts of objects by non-negative matrix factorization." Nature 401(6755): 788-791. 1999.

[20]. Guo, Yifan, Y. X. Zou, and Yongqing Wang. "A robust high resolution speaker DOA estimation under reverberant environment." Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on. IEEE, 2014.

[21]. J. S. Garofolo, Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database NIST Tech Report, 1988.

[22]. http://www.mathworks.com/products/compiler/mcr/

[23]. Lee H, Pham P, Largman Y, et al. Unsupervised feature learning for audio classification using convolutional deep belief networks[C]//Advances in neural information processing systems. 1096-1104. 2009.