Speech Enhancement Based on Sparse Representation Using Joint Dictionary

Ming Wei[1], Zheng Liu[1,2], Xueqin Chen[1,*] and Heming Zhao[1]
[1]School of Electronic and Information Engineering, Soochow University, China
[2]Suzhou Industrial Park Institute of Services Outsourcing, China
[*]Corresponding author

***Abstract*—A new speech denoising method that aims for processing corrupted speech signal which is based on the sparse representation theory of speech signal. In this paper, we train a composite dictionary consisting of the concatenation of the speech dictionary and the noise dictionary by using the K-SVD algorithm. Noise is divided into structured and unstructured noise in this paper. For structured noise, we train speech and noise dictionary firstly, and then according to the different coherence between speech and noise, we use LARC algorithm with a suitably chosen residual coherence threshold to realize the separation of the speech and the noise. For unstructured noise, we only need speech dictionary to extract the clean speech from corrupted speech. Experiments indicate that the proposed method gives better enhancement results in terms of quality measures of speech. The proposed method outperforms the universal dictionary speech enhancement algorithm.**

Keywords—speech denoising; composite dictionary; K-SVD algorithm; speech enhancement; LARC algorithm

I. Introduction

Speech is important information carrier of language communication. It narrows the communication gap between people and speeds up the exchange of information [1]. However, there are many different kinds of noise in the real world, which tends to reduce speech clarity and intelligibility [1].Therefore, the research of speech enhancement technology is particularly important. Speech enhancement technology has been applied in many areas, such as mobile communications, hearing aids and pre-treatment of speech recognition system [2].The purpose of speech enhancement is to extract the clean speech signal from the noisy speech signal as much as possible, so as to improve the clarity and intelligibility of the speech signal. At present, there are many ways to implement speech enhancement. The traditional methods including spectral subtraction, wiener filtering and subspace methods [1].

In recent years, with the development of sparse representation theory, sparse representation and dictionary learning techniques have been widely used in the field of signal processing [3].The signal sparse representation refers to the over-complete atomic dictionary select a few atoms for linear combination to approximate the signal [4].The purpose of dictionary learning is to find the optimal set of atoms which can capture the characteristics of these signals well [4]. Sparse representation in speech signal processing applications are mainly focus on Voice Activity Detection (VAD), pitch estimation, speaker recognition and speech recognition [4]. At present, speech enhancement algorithms based on sparse representation are also discussed in some literature. For example, in the literature [2], a speech enhancement algorithm for sparse representation in the time domain is proposed. In this algorithm, a dictionary is obtained by using the K-singular value decomposition (K-SVD) algorithm and the clean speech is reconstructed by the Orthogonal Matching Pursuit (OMP) [2].

The traditional speech enhancement algorithm based on sparse representation is a universal dictionary denoising algorithm [1]. It trains a universal dictionary and separates the speech and the noise by controlling the sparsity of the signal decomposition on the dictionary according to the sparsity of speech and noise [2]. This method has some obvious effects on denoising of unstructured noise (such as Gaussian white noise). However, in real life, there are many structured noises, which have the same sparsity as the speech [5]. Therefore, it is very difficult to achieve the separation of speech and noise by this method. This paper uses a speech enhancement algorithm based on the joint dictionary learning, which is implemented in the short-time Fourier transform (STFT) magnitude domain [5]. First, assuming that the magnitude of the noisy speech is a linear additivity of the noise magnitude and the speech magnitude [6], ignoring the influence of the phase [7]. Then, we use the training samples of speech signal and noise signal to learn the speech dictionary and the noise dictionary respectively, then concatenate these two dictionaries. In the enhancement part, we use the least angle regression (LARS) with a coherence criterion (LARC) algorithm [2], the noisy signal is sparsely represented on this composite dictionary. According to the coherence between the signal and the dictionary, the speech and the noise can be respectively represented by the corresponding dictionary. Finially, we extract the clean speech from the corrupted speech signal.

II. Method

Consider the following model for noisy speech:

$$y = s + i \tag{1}$$

where y, s and i denote noisy speech, clean speech and noise, respectively. The purpose of speech enhancement is to estimate the speech components $\hat{s}$ from the noisy signal and reconstruct the original clean speech. The proposed method in this paper is divided into two parts: dictionary learning part and speech enhancement part [8]. The specific process is as follows.

Copyright © 2018, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

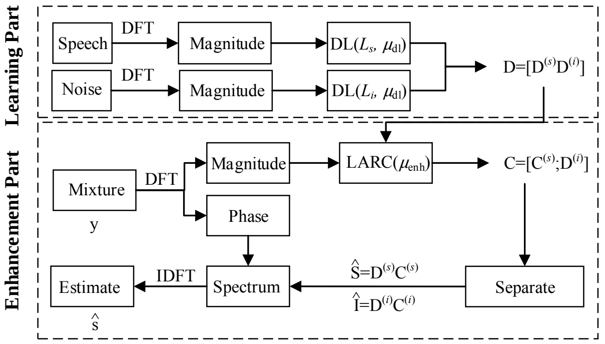


FIGURE I. SPEECH ENHANCEMENT BLOCK DIAGRAM BASED ON JOINT DICTIONARY

A. Dictionary Learning Part

We consider a signal x and an over-complete dictionary $\text{D} = [\text{d}_{(1)}\text{d}_{(2)}\text{d}_{(3)}\cdots\text{d}_{(L)}]\in\mathbf{R}^{D\times L}$ consisting of L unit-norm atoms. We slide a window to divide the signal into N frames and make short-time Fourier transform with this signal, and then we obtain a signal matrix $\text{X}=\left[\text{x}_{(1)}\text{x}_{(2)}\cdots\text{x}_{(N)}\right]\in\mathbf{R}^{D\times N}$. Signal matrix X can be decomposed into a dictionary D and a coding $\text{C}=\left[\text{c}_{(1)}\text{c}_{(2)}\ldots\text{c}_{(N)}\right]\in\mathbf{R}^{L\times N}$. The target function of dictionary learning is given by

$$\arg\min_{\text{D,C}} \|\text{X} - \text{DC}\|_F^2 \tag{2}$$

subject to a sparsity constraint on C. $\|\cdot\|_F$ denotes the Frobenius norm.

In this paper, we choose the K-SVD algorithm for dictionary learning [9]. The redundant dictionary training based on K-SVD learning algorithm includes three steps: dictionary initializing, sparse coding and dictionary atoms updating [9]. When considering this available scheme, sparse coding and dictionary updating should be alternately performed, thus the redundant dictionary and the sparse matrix can be updated synchronously [3]. General steps are as follows:

1) Dictionary initializing

The selection of the initial dictionary is random [8], but the proper initial dictionary can effectively reduce the number of the dictionary learning iterations and increase the rate of learning [11]. In this paper, we select the over-complete cosine basis as the initial dictionary [8].

2) Sparse coding

The initial dictionary must be assumed as fixed firstly in this step. Aligning with this, the signal matrix X should be decomposed over this initial dictionary by using the least angle regression (LARS) with a coherence criterion algorithm (LARC). Since the iterative termination condition is decided by the coherence between decomposition residual and the current atom of dictionary [10]. In a word, the sparse matrix can be got by solving the following formula.

$$\min_{\text{D,C}}\|\text{c}_i\|_0 \quad \text{s.t. } \|\text{X} - \text{DC}\|_F^2 \le \varepsilon. \tag{3}$$

3) Dictionary updating

This step is aiming to update the redundant dictionary [11]. It is important to fix the sparse matrix which has been trained at the last step. It is updated atom-by-atom and this is an iterative process. The residual norm that separates the contribution of k, $k \in [1, L]$ th atom which is being updates can be formulated as

$$\begin{aligned}\|\text{X} - \text{DC}\|_F^2 &= \left\|\text{X} - \sum_{j=1}^{L} \text{d}_{(j)}\text{c}^{[j]}\right\|_F^2 \\ &= \left\|\left(\text{X} - \sum_{j\neq k} \text{d}_{(j)}\text{c}^{[j]}\right) - \text{d}_{(k)}\text{c}^{[k]}\right\|_F^2 \\ &= \left\|\text{E}_k - \text{d}_{(k)}\text{c}^{[k]}\right\|_F^2 \end{aligned} \tag{4}$$

The residual norm is minimized by seeking for a rank-one approximation [12]. The approximation is based on computing the singular value decomposition (SVD) [13].

Let ω be the set of indices of columns that corresponding to the signals that use the atoms.

$$\omega_k = \left\{\, i \mid 1 \le i \le N, \text{c}^{[k]}(i) \neq 0 \,\right\} \tag{5}$$

Define E_R^k as the set of columns in E_k indexed by ω_k. Compute the SVD of E_R^k,

$$\text{E}_R^k = \text{U}\Delta\text{V}^T \tag{6}$$

We update $\text{d}_{(k)}$ as the first column of U, and $\text{c}^{[k]}$ as $\Delta_{1,1}$ times the first row of V^T [14].

B. Speech Enhancement Part

In this part, our purpose is to obtain an estimate $\hat{\text{s}}$ of clean speech and an estimate $\hat{\text{i}}$ of the noise, given noisy speech y, a speech dictionary $\text{D}^{(s)} \in \mathbf{R}^{D\times L_s}$ and a noise dictionary $\text{D}^{(i)} \in \mathbf{R}^{D\times L_i}$. For the formal analysis, noise can be divided into unstructured and structured noise [5]. Due to the structure of the noise magnitude, structured noise can be sparsely coded in a suitable dictionary and unstructured noise cannot be sparsely coded in any fixed dictionary [5], in particular not in speech dictionary. Therefore, we can make use of the characteristic of the structural noise and enhance noisy speech in different ways.

1) Unstructured noise

Due to the non-sparsity of unstructured noise, we cannot train dictionary for noise. For the suppression of this noise, in the enhancement step, the noisy speech y is only sparsely coded in the speech dictionary using LARC with a suitably chosen residual coherence threshold μ_{enh}. LARC coding captures the structured speech signal components which have a coherence to the speech dictionary that is above the threshold, while discarding the noise components, as they fall below the residual coherence threshold [2]. The vector of coding coefficient matrix $\text{C}^{(s)} \in \mathbf{R}^{L_s}$ is obtained as follows.

$$C^{(s)} \leftarrow \mathrm{LARC}\left(D^{(s)}, y, \mu_{\mathrm{enh}}\right) \tag{7}$$

An estimate of the clean speech is obtained as $\hat{S} = D^{(s)}C^{(s)}$.

2) Structured noise

A structured noise can be sparsely represented with low approximation error in a suitably trained noise dictionary [2]. In order to enhance speech degraded by structured noise which are partially coherent to the speech dictionary, the noisy speech is sparsely coded in the composite dictionary consisting of the concatenation of the speech and the noise dictionary [5]. Noisy speech y is sparsely coded in the composite dictionary $D = [D^{(s)}D^{(i)}]$ using LARC with a suitably chosen residual coherence threshold μ_{enh}. The vector of coding coefficient matrix C is obtained as follows.

$$C \leftarrow \mathrm{LARC}\left([D^{(s)}D^{(i)}], y, \mu_{\mathrm{enh}}\right) \tag{8}$$

The matrix $C = [C^{(s)};C^{(i)}]$ consists of weights $C^{(s)}$ corresponding to the speech dictionary $D^{(s)}$, as well as weights $C^{(i)} \in \mathbf{R}^{L_i}$ corresponding to the noise dictionary $D^{(i)}$. An estimate of the clean speech is obtained as $\hat{S} = D^{(s)}C^{(s)}$.

III. Experiment and Performance Evaluation

A. Experimental Settings

In this paper, speech data is obtained from the GRID audio-visual corpus [2]. This large multi-talker corpus has been provided for studies in speech perception with a recording of 1000 sentences. The corpus involves speech recording of 34 speakers of both genders with sentences of about 1.1-2.2s. We have used 30 sentences of different male speakers as speech dictionary learning samples in our experiments. We chose 4 different kinds of noise (white, babble, volvo and leopard) from the NOISEX-92 corpus as noise dictionary learning samples [14]. The samples is resampled at 16kHz. The time-domain signals are transformed into STFT domain and the frame length is set to 512 point with 50% overlap. Noisy speech is synthetically generated by adding clean speech and noise at various SNRs, since objective measures require access to the clean speech signal.

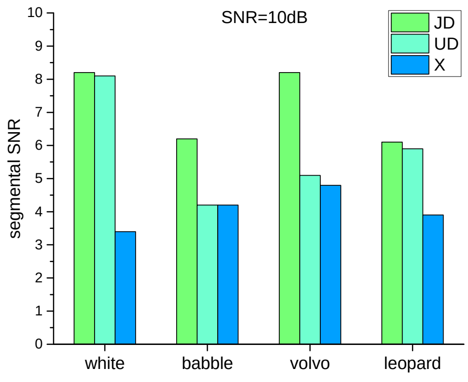
Dictionaries are trained using K-SVD algorithm, initialized with cosine basis. Each dictionary contains 800 atoms and the number of iterations are 30. For dictionary learning, the residual coherence threshold μ_{dl} was set to 0.2 for all dictionaries. μ_{enh} during enhancement was set to 0.15.

B. Performance Evaluation

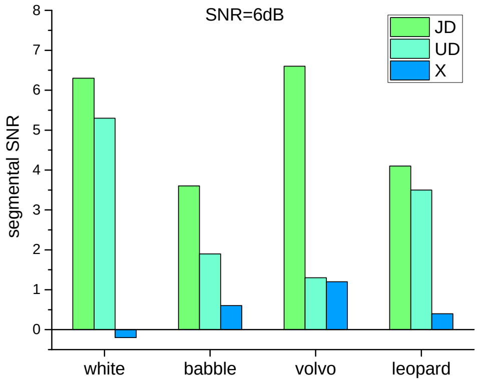
The performance of a speech enhancement algorithm can be measured both subjectively and objectively [2]. Subjective measurements are based on the judgment of human listeners, and are important because in many applications (such as hearing aids) the output of the enhancement algorithm has to appeal to the human ear. However, subjective evaluation takes time, is expensive, and usually requires trained listeners. As an alternative, objective measures provide mathematical models of some perceptual aspects of the human auditory system [2].

We evaluate our proposed method in the segmental SNR (SegSNR). SegSNR is a simple objective measure, computed on individual signal frames, and the per-frame SNRs are averaged over time. The larger the value of the SegSNR , the better the denoising performance.

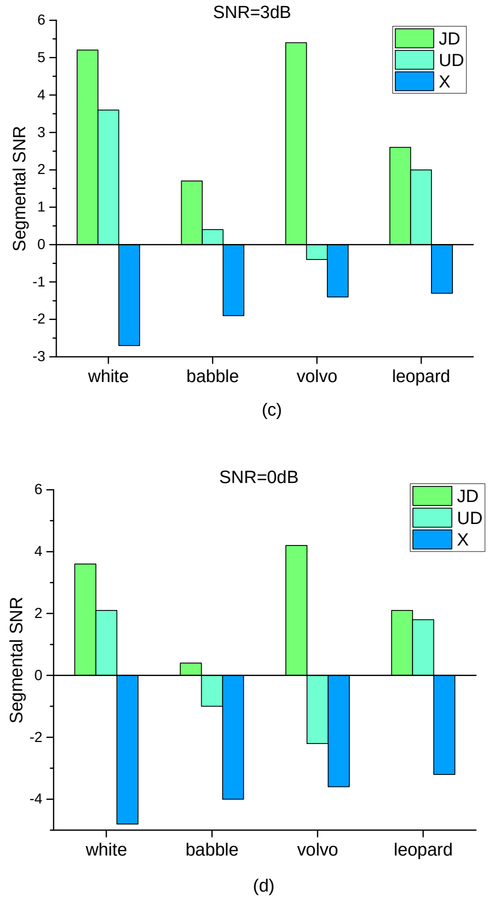
In order to evaluate the performance, we compare our proposed method with universal dictionary learning algorithm[1]. The universal dictionary learning algorithm is also based on sparse representation, which uses K-SVD algorithm to train a universal dictionary and reconstruct the clean speech with OMP algorithm.



(a)



(b)



(c)

(d)

FIGURE II. (a), (b), (c) AND (d) DENOTE THE DENOISING PERFORMANCE OF "JD" AND "UD" AT DIFFERENT SNR.

FIGURE II shows the results of segmental SNR for the proposed and universal dictionary learning method. "JD" denotes joint dictionary learning method (proposed method). "UD" denotes universal dictionary learning method. "X" denotes the objective measurement before any enhancement. The results consistently indicate that the proposed method works better than the "UD" and "X", which means proposed method outperforms better than universal dictionary learning method.

IV. Conclusion

In this paper, we adopt a novel method based on sparse representation theory to enhance the degraded speech. The method has trained composite dictionary consisting of the concatenation of the speech dictionary and the noise dictionary by using the KSVD algorithm, then takes advantage of the different coherence between the speech and the noise and uses LARC with a suitably chosen residual coherence threshold to realize the separation of the speech and the noise. In experimental results, we have showed that our method is outperform than the universal dictionary speech enhancement algorithm. Although this method has effectively improved the speech signal quality to a certain extent, but there is a large research space that do not be explored in this work. For example, the research of the optimization algorithm for greed algorithm when training the dictionary, the research of a redundant dictionary and more. In a word, all of these are the research direction.

Acknowledgment

This work is supported by the national natural science foundation of China under grant No.61340004 and by Jiangsu higher vocational college teachers professional leader advanced research project under grant No.2017GRGDYX058.

References

[1] L. Huang, L. Li and S. He, "Speech Enhancement Based on Sparse Representation Using Universal Dictionary", International Conference on Anti-Counterfeiting Security and Identification, pp. 1-4, 2013.

[2] C. D. Sigg, T. Dikk and J. M. Buhmann, "Speech enhancement using generative dictionary learning", IEEE Trans. on Audio Speech Lang. Process., vol. 20, no. 6, pp. 1698-1712, 2012.

[3] Y. Zhou, H. Zhao, X. Chen, T. Liu, D. Wu and L. Shang, "Speech denoising based on sparse representation algorithm", 12th International Conference on Intelligent Computing Theories and Application, pp. 202-211, 2016.

[4] Tak W. Shen and Daniel P. K. Lun, "A speech enhancement method based on sparse reconstruction on log-spectra", HKIE Transactions Hong Kong Institution of Engineers, vol. 224, no. 1, pp. 24-34, 2017.

[5] C. Sigg, T. Dikk, and J. Buhmann, "Speech enhancement with sparse coding in learned dictionaries", Proc. IEEE Int. Conf. Acoust. Speech Signal Process., pp. 4758-4761, 2010.

[6] G. Bao, Y. Xu, and Z. Ye, "Learning a discriminative dictionary for single-channel speech separation", IEEE Trans. on Audio Speech and Lang. Process., vol. 22, no. 7, pp. 1130-1138, 2014.

[7] Y. Zhao, X. Zhao and B. Wang, "A speech enhancement method employing sparse representation of power spectral density", Journal of Information and Computational Science, vol. 10, no. 6, pp. 1705-1714, April 2013.

[8] N. Zhao, X. Xin and Y. Yang, "Sparse representations for speech enhancement", Chinese Journal of Electronics, vol. 20, no. 2, pp. 268-272, April 2011.

[9] M. Aharon, M. Elad, A. Bruckstein and Y. Katz, "K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation", IEEE Trans. Signal Process., vol. 54, no. 11, pp. 4311-4322, Nov. 2006.

[10] M. Yang, L. Zhang, J. Yang and D. Zhang, "Metaface learning for sparse representation based face recognition", IEEE International Conference on Image Processing, pp. 1601-1604, 2010.

[11] Y. Hao and C. Bao, "An Improved Dictionary Learning Method for Speech Enhancement", Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 144-147, 2015.

[12] S. Mavaddaty, S. M. Ahadi and S. Seyedin, "Modified coherence-based dictionary learning method for speech enhancement", IET Signal Processing, vol. 9, no. 7, pp. 537-545, 2015.

[13] M. Sun; Y. Li, J. F. Gemmeke and X. Zhang, "Speech Enhancement Under Low SNR Conditions Via Noise Estimation Using Sparse and Low-Rank NMF with Kullback–Leibler Divergence", IEEE Trans. on Audio Speech and Lang. Process., vol. 23, no. 7, pp. 1233 - 1242, 2015.

[14] Y. He, J. Han, S. Deng, T. Zheng and G. Zheng, "A solution to residual noise in speech denoising with sparse representation", Proc. IEEE Int. Conf. Acoust. Speech Signal Process., pp. 4653-4656, 2012.