

Single Channel Speech Separation Using Deep Neural Network

Linlin Chen^{1,2}, Xiaohong Ma^{1(✉)}, and Shuxue Ding²

¹ School of Information and Communication Engineering,
Dalian University of Technology, Dalian, China
maxh@dlut.edu.cn

² School of Computer Science and Engineering,
University of Aizu Fukushima, Aizuwakamatsu, Japan
sding@u-aizu.ac.jp

Abstract. Single channel speech separation (SCSS) is an important and challenging research problem and has received considerable interests in recent years. A supervised single channel speech separation method based on deep neural network (DNN) is proposed in this paper. We explore a new training strategy based on curriculum learning to enhance the robustness of DNN. In the training processing, the training samples firstly are sorted by the separation difficulties and then gradually introduced into DNN for training from easy to complex cases, which is similar to the learning principle of human brain. In addition, a strong discriminative objective function for reducing the source interference is designed by adding in the correlation coefficients and negentropy. The efficiency of the proposed method is substantiated by a monaural speech separation task using TIMIT corpus.

Keywords: Single channel speech separation · Deep neural network · Discriminative object function

1 Introduction

Single channel speech separation (SCSS) is to extract speech from one signal that is a mixture of multiple sources. It is a vital issue of speech separation and may play an important role in many applications. Researchers have devoted to solving SCSS problems from various perspectives, which can be categorized into computational auditory scene analysis (CASA) based and model based.

Approaches based on CASA have proven to be effective to attack the SCSS problem in an unsupervised mode. In [15], Wang et al. proposed to utilize temporal continuity and cross-channel correlation for speech segregation. The method can segregate speech from interfering sounds but it does not perform well for high frequency part of speech. Hu et al. improve it by segregating resolved and unresolved harmonics differently [5]. A common problem occurred in many CASA-based methods is that the recovered speech usually miss some parts. In order

to solve this problem, shape analysis techniques in image processing such as labeling and distance function are applied to speech separation in [10].

In model based approaches, non-negative matrix factorization (NMF) is one of the most popular techniques for SCSS in recent years. Conventionally, the basis of each source is trained separately first and then the magnitude spectra of mixed signal is decomposed into a linear combination of the trained basics. Finally, the separated signals can be obtained from the corresponding parts of decomposed mixture [4, 12]. However, the separation become difficult when sources are overlap in subspaces. Various attempts have been made to solve this problem [13, 16].

Deep neural networks (DNN) has achieved state-of-art results in many applications such as object detection [3], speech recognition [17] owing to its strong mapping ability. Kang et al. use DNN to learn the mapping between the mixture and the corresponding encoding vectors of NMF [8]. In [6], a simple discriminative training criterion which takes into account the squared error of prediction and other sources is proposed.

In this paper, we focus on training strategy and objective function. This paper considers DNN as a kind of system learning rules from a mess of training samples. These training samples are sorted by a ranking function and fed to DNN in ascending order of learning difficulty. Furthermore, we use correlation coefficients and negentropy, rather than the criterion in [6], to model the similarity of recovered signals, which aim at reducing the interference of other sources. Experimental results demonstrated that the proposed method outperformed NMF and approach in [6].

The organization of this paper is as follows: Sect. 2 introduces the proposed methods, including the learning strategy and discriminative objective function, Sect. 3 presents the experimental setting and results based on the TIMIT corpus and conclusion is given in Sect. 4.

2 Proposed Method

2.1 Problem Formulation

In this paper, we assume the observed signal is a mixture of source signals of two speakers. Ignored the attenuations of the path, the problem can be formulated as

$$\mathbf{x}(t) = \mathbf{s}^T(t) + \mathbf{s}^I(t) \quad (1)$$

where $\mathbf{s}^T(t)$ and $\mathbf{s}^I(t)$ represent the target speech and interfering speech respectively. Denoting $\mathbf{x}(t, f)$ as the short time Fourier transform (STFT) of $\mathbf{x}(t)$, the formula in the STFT domain can be represented as

$$\mathbf{x}(t, f) = \mathbf{s}^T(t, f) + \mathbf{s}^I(t, f) \quad (2)$$

Phase recovery is ignored in this paper since the human is not so sensitive to phase distortion. The magnitude spectrum in Eq. (2) can be written in matrix form as follows

$$\mathbf{X} \approx \mathbf{S}^T + \mathbf{S}^I \quad (3)$$

where \mathbf{S}^T and \mathbf{S}^I are the unknown magnitude spectrums need to be estimated by DNN.

2.2 System Framework

The overall framework of the proposed method is showed in Fig. 1. Firstly, pairs of mixed signal and the sources are transformed to time-frequency domain by STFT and frames of magnitude spectrum can be obtained. Then, these training samples are sorted by a ranking function from “easy” to “hard” and fed to the DNN gradually. After the model is mature enough, it is applied to process the magnitude spectrum of test data and predict the spectrums of sources. Finally, an overlap add method is used to synthesize the waveform of the estimated signals [7].

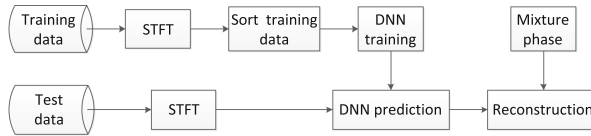


Fig. 1. System framework

2.3 DNN Learning Strategy

For all we know, one starts with small and easy curriculums, and then gradually increases the difficulty level. Inspired by this principle, Bengio et al. proposed a new learning paradigm in machine learning framework called curriculum learning (CL) [1]. The main idea is sorting the training samples by a difficulty measurement and then introducing them from easy to complex to the model. This strategy is proved to be effective to alleviate the bad local optimum problem in non-convex optimization and improve generalization ability [9, 11]. The key of CL methodology is to find a ranking function that assigns learning importance to training samples. So the key to us is to find an appropriate ranking function for source separation problem.

In source separation, the target speech is corrupted by interfering speech to varying degrees over time. Empirically, we define the ranking function as follows:

$$f(\mathbf{x}_i, \mathbf{s}_i^T) = 10 \log \frac{P_{\mathbf{s}_i^T}}{P_{\mathbf{x}_i}} \quad (4)$$

where \mathbf{x}_i denotes i th frame of mixed magnitude spectrum which is fed to DNN, \mathbf{s}_i^T represents i th frame of target speech, $P_{\mathbf{s}_i^T}$ and $P_{\mathbf{x}_i}$ are the energy of \mathbf{s}_i^T and \mathbf{x}_i respectively. It is easy to see that the bigger the value of f is the “easier” a sample is, since it means a larger proportion of energy the source accounts for. According to the function, the training samples can be sorted and the system will learn from easy to hard.

Formally, let $J(g(\mathbf{x}_i, \mathbf{w}), \mathbf{y}_i)$ denote the objective function of neural network which calculates the cost between the target output $\mathbf{y} = [\mathbf{s}_i^T; \mathbf{s}_i^T]$ and the estimated magnitude spectrums $g(\mathbf{x}_i, \mathbf{w})$. Here \mathbf{w} represents the model parameters inside DNN. Then DNN can be optimized by minimizing:

$$\min_{\mathbf{w}, \mathbf{v}} \sum_{i=1}^m v_i J(g(\mathbf{x}_i, \mathbf{w}), \mathbf{y}_i) \tag{5}$$

where v_i is determined by:

$$v_i = \begin{cases} 1 & f(\mathbf{x}_i, \mathbf{s}_i^T) > \lambda \text{ or } f(\mathbf{x}_i, \mathbf{s}_i^I) > \lambda \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Equation (6) indicates that a sample is considered as easy one as long as either one of two sources occupies most energy of mixture, since it implies this sample could be separated more easily and should be introduced to the model earlier. The parameter λ controls the pace at which the model learns new samples. It decreases over time. When λ is large, only easy samples will be fed to DNN. As time goes on, λ decreases and more samples with more severe corrosion will be gradually appended to train a more mature model.

2.4 Discriminative Objective Function

In the training stage, the magnitude spectrums from pairs of mixed signal and the sources are utilized to train the DNN. Given \mathbf{x}_i as input of DNN, the output $g(\mathbf{x}_i, \mathbf{w}) = [\tilde{\mathbf{s}}_i^T; \tilde{\mathbf{s}}_i^I]$ are expected to have small error with the target output \mathbf{y} , where $\tilde{\mathbf{s}}_i^T$ and $\tilde{\mathbf{s}}_i^I$ represent DNN estimates of target output \mathbf{s}_i^T and \mathbf{s}_i^I , so conventionally one can optimize the neural network parameters by minimizing the squared error:

$$J_{mse}(g(\mathbf{x}_i, \mathbf{w}), \mathbf{y}_i) = \frac{1}{2} \sum_{i=1}^m (\|\mathbf{s}_i^T - \tilde{\mathbf{s}}_i^T\|^2 + \|\mathbf{s}_i^I - \tilde{\mathbf{s}}_i^I\|^2) \tag{7}$$

Equation (7) enables DNN to separate two sources after training a set of samples. In order to further improve the separation quality, here we propose a different criteria to enhance the discrimination of the two predicted sources. An important fact is that Eq. (7) does not take into account source interference. If the two sources are similar, DNN may be confused and mistakes the target speech for the interfering speech. Correlation coefficient is a metric that measures the correlation between two signals and we expect to minimize the correlation coefficients of the sources to reduce the interference. Moreover, starting from an information theoretic viewpoint, the discrimination problem can be formulated as reducing the mutual information. Mutual information is a natural measure of the dependence between random variables. The mutual information can be approximated by the negentropy. Minimizing the mutual information is roughly equivalent to finding directions in which the negentropy is maximized. Taking into account these two measures, we add the following two parts to the original objective function in Eq. (7):

$$\begin{aligned} J_{cor}(g(\mathbf{x}_i, \mathbf{w}), \mathbf{y}_i) &= \sum_{i=1}^m \text{corr}(\tilde{\mathbf{s}}_i^T, \mathbf{s}_i^I) + \text{corr}(\tilde{\mathbf{s}}_i^I, \mathbf{s}_i^T) \\ &= \sum_{i=1}^m \frac{\text{cov}(\tilde{\mathbf{s}}_i^T, \mathbf{s}_i^I)}{\sqrt{D(\tilde{\mathbf{s}}_i^T)}\sqrt{D(\mathbf{s}_i^I)}} + \frac{\text{cov}(\tilde{\mathbf{s}}_i^I, \mathbf{s}_i^T)}{\sqrt{D(\tilde{\mathbf{s}}_i^I)}\sqrt{D(\mathbf{s}_i^T)}} \end{aligned} \tag{8}$$

$$\begin{aligned}
J_{neg}(g(\mathbf{x}_i, \mathbf{w}), \mathbf{y}_i) &= \sum_{i=1}^m (\text{H}_G(\tilde{\mathbf{s}}_i^T) + \text{H}_G(\tilde{\mathbf{s}}_i^I)) \\
&= \sum_{i=1}^m \left(\int p(\tilde{\mathbf{s}}_i^T) \log \frac{p(\tilde{\mathbf{s}}_i^T)}{p_G(\tilde{\mathbf{s}}_i^T)} d\tilde{\mathbf{s}}_i^T + \int p(\tilde{\mathbf{s}}_i^I) \log \frac{p(\tilde{\mathbf{s}}_i^I)}{p_G(\tilde{\mathbf{s}}_i^I)} d\tilde{\mathbf{s}}_i^I \right)
\end{aligned} \tag{9}$$

where Eqs. (8) and (9) represent the correlation coefficients and negentropy of the sources, respectively. In Eq. (8), $\text{cov}(\cdot)$ denotes covariance and $D(\cdot)$ denotes variance. In Eq. (9), $p_G(\theta)$ is the density of a Gaussian random variable with the same covariance matrix as θ . To simplify the calculations, here we use nonlinear correlation coefficients [2] and negentropy approximate formula [7] instead,

$$J_{cor}(g(\mathbf{x}_i, \mathbf{w}), \mathbf{y}_i) = \sum_{i=1}^m \left(\sum_{j=1}^n \tilde{s}_{i,j}^T g(s_{i,j}^I) + \sum_{k=1}^n \tilde{s}_{i,k}^I g(s_{i,k}^T) \right) \tag{10}$$

$$J_{neg}(g(\mathbf{x}_i, \mathbf{w}), \mathbf{y}_i) = -\frac{1}{a} \sum_{i=1}^m E \left(e^{-\frac{a(\tilde{\mathbf{s}}_i^T)^2}{2}} + e^{-\frac{a(\tilde{\mathbf{s}}_i^I)^2}{2}} \right) \tag{11}$$

In Eq. (10), $r(\cdot)$ denotes nonlinear function and $2n$ denotes the dimension of DNN output. $E(\cdot)$ in Eq. (11) represents the statistical expectation of variables and the parameter a is usually chosen as 1. We expect to minimize J_{cor} and maximize J_{neg} for enhancing the discrimination. In order to estimate the unknowns in the model, we solve the following problem:

$$\text{argmin} \quad J(g(\mathbf{x}_i, \mathbf{w}), \mathbf{y}_i) \tag{12}$$

where

$$J(g(\mathbf{x}_i, \mathbf{w}), \mathbf{y}_i) = J_{mse} + \eta_1 J_{cor} - \eta_2 J_{neg} \tag{13}$$

is the joint discriminative function which we seek to minimize. η_1 and η_2 are regularization parameters which are chosen experimentally.

3 Experiments

3.1 Experimental Setup

In order to evaluate the performance of the proposed method, we conduct experiments on speech separation with TIMIT corpus. Two speakers, one male and one female, are chosen from database. To each speaker, 80% of the sentences are for training and 20% for testing. Mixed speech utterances were generated by mixing the sentences randomly from the two speakers at 0 dB signal-to-noise ratio (SNR). For increasing the number of training samples, we circularly shift points of the signal from male speaker and mix it with the female source. The time frequency representations are computed by the 512 point short time Fourier transform using a 32 ms window with a step size of 16 ms. Then 257-dimensional magnitude spectrums are used as input features to train DNN.

The separation result obtained from the proposed algorithm is compared with that of the standard NMF and DNN-based separation method [6]. In the NMF experiment, The number of basis vectors is set to 40 for each source. As for DNN, the architecture which jointly optimizes time-frequency masking functions as a layer with DNN in [6] is applied here. The neural network has 2 hidden layers with 160 nodes each, 2 hidden layers with 300 nodes each and 3 hidden layers with 160 nodes each. Pre-training is not adopted here benefits from the activation function Rectified Linear Unit (ReLU), which can reach the same performance without requiring any unsupervised pre-training on purely supervised tasks with large labeled datasets [6]. Empirically, the nonlinear function in Eq. (8) is chosen as $\tanh(\cdot)$ and the value of parameters η_1 and η_2 is in the range of 0.1 and 0.3.

The separation performance is evaluated in terms of three metrics, signal to distortion ratio (SDR), signal to interference ratio (SIR), and signal to artifacts ratio (SAR) [14].

3.2 Experimental Results

The separation results with 2 hidden layers and 160 nodes are reported in Table 1. The DNNori in Table 1 means the basic DNN model with the objective function in Eq. (7). It is obvious from the results that all DNN-based methods outperform NMF, which confirms that neural network has better generalization and separation capacity. Compared the results between DNNori and DNNori-cl, which denotes the DNNori using the learning strategy proposed in Sect. 2, the SDR, SIR, and SAR all have been improved. It confirms the need for curriculum learning. Sorting the training samples by the ranking function and making the neural network learn like human brain from easy to complex does help the system to be more robust. As for DNNori-dis, we use the discriminative cost function in (13) instead of the function in Eq. (7). It is interesting to find that the improvement in SDR is so slight that can be ignored, but the SIR achieves around 1.2 dB gain compared to DNNori. The results match our expectation when we design the objective function which aims at enhancing the discrimination and reducing the interfering of the other source. The disadvantage is that some artifacts are introduced into the separation and the SAR is lower than DNNori.

Table 1. Speech separation results of various separation algorithms.

Method	Measurement (dB)		
	SDR	SIR	SAR
NMF	6.008	8.722	7.624
DNN[13]	7.70	11.53	8.07
DNNori	7.58	10.81	8.31
DNNori-cl	7.73	10.88	8.78
DNNori-dis	7.62	12.06	7.86
DNNori-cl-dis	7.87	12.12	8.15

Finally, in DNNori-cl-dis, the curriculum learning strategy and the discriminative objective function are both added in. We can see from the results that the two techniques both play an effective role in single channel source separation. To strongly demonstrate the jointly function of the curriculum learning strategy and the discriminative objective function proposed, we do some experiments on different layers and different notes respectively. And the results are shown in Table 2. According to Table 2, we can see that the case having 3 hidden layers and 160 nodes for each layer and the case having 2 hidden layers and 200 nodes for each layer have achieved different level improvement, which is fit for the conclusion we make above. The model achieves best performance in SDR and SIR.

Table 2. Speech separation results of various network structures.

Method	Measurement (dB)		
	SDR	SIR	SAR
DNNori-cl-dis (3*160)	7.883	11.693	8.522
DNN[13] (3*160)	7.486	10.962	8.257
DNNori-cl-dis (2*300)	7.79	11.477	8.459
DNN[13] (2*300)	7.763	11.343	8.442

4 Conclusions

In this paper, the DNN is used to model each source signal and trained to separate the mixed signal. Two novel improvements have been proposed to further enhance the separation performance: a learning strategy based on curriculum learning and a discriminative objective function that reduces the interference from the other source. We have proved that the proposed algorithm achieves better results through a series of experiments on speech separation. The future work will focus on improving the proposed method by combining the phase separation with DNN training.

References

1. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 41–48. ACM (2009)
2. Cichocki, A., Unbehauen, R., Rummert, E.: Robust learning algorithm for blind separation of signals. *Electron. Lett.* **30**(17), 1386–1387 (1994)
3. Erhan, D., Szegedy, C., Toshev, A., Anguelov, D.: Scalable object detection using deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2147–2154 (2014)

4. Grais, E.M., Erdogan, H.: Single channel speech music separation using nonnegative matrix factorization and spectral masks. In: 2011 17th International Conference on Digital Signal Processing (DSP), pp. 1–6. IEEE (2011)
5. Hu, G., Wang, D.: Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Trans. Neural Netw.* **15**(5), 1135–1150 (2004)
6. Huang, P.S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P.: Deep learning for monaural speech separation. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1562–1566. IEEE (2014)
7. Hyvarinen, A.: Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.* **10**(3), 626–634 (1999)
8. Kang, T.G., Kwon, K., Shin, J.W., Kim, N.S.: Nmf-based target source separation using deep neural network. *IEEE Sig. Process. Lett.* **22**(2), 229–233 (2015)
9. Khan, F., Mutlu, B., Zhu, X.: How do humans teach: on curriculum learning and teaching dimension. In: *Advances in Neural Information Processing Systems*, pp. 1449–1457 (2011)
10. Lee, Y.K., Kwon, O.W.: Application of shape analysis techniques for improved casa-based speech separation. *IEEE Trans. Consum. Electron.* **55**(1), 146–149 (2009)
11. Ni, E.A., Ling, C.X.: Supervised learning with minimal effort. In: Zaki, M.J., Yu, J.X., Ravindran, B., Pudi, V. (eds.) *PAKDD 2010. LNCS*, vol. 6119, pp. 476–487. Springer, Heidelberg (2010). doi:[10.1007/978-3-642-13672-6_45](https://doi.org/10.1007/978-3-642-13672-6_45)
12. Smaragdis, P., Raj, B., Shashanka, M.: Supervised and semi-supervised separation of sounds from single-channel mixtures. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) *ICA 2007. LNCS*, vol. 4666, pp. 414–421. Springer, Heidelberg (2007). doi:[10.1007/978-3-540-74494-8_52](https://doi.org/10.1007/978-3-540-74494-8_52)
13. Sun, D.L., Mysore, G.J.: Universal speech models for speaker independent single channel source separation. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 141–145. IEEE (2013)
14. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. *IEEE Trans. Audio Speech Lang. Process.* **14**(4), 1462–1469 (2006)
15. Wang, D.L., Brown, G.J.: Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Trans. Neural Netw.* **10**(3), 684–697 (1999)
16. Wang, Z., Sha, F.: Discriminative non-negative matrix factorization for single-channel speech separation. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3749–3753. IEEE (2014)
17. Xue, S., Abdel-Hamid, O., Jiang, H., Dai, L., Liu, Q.: Fast adaptation of deep neural network based on discriminant codes for speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **22**(12), 1713–1725 (2014)