DEEP STACKING NETWORKS WITH TIME SERIES FOR SPEECH SEPARATION

Shuai Nie[1,2] *Hui Zhang*[2] *XueLiang Zhang*[2] *WenJu Liu*[1]

[1] National Laboratory of Patten Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[2] College of Computer Science, Inner Mongolia University, Huhhot 010021, China
nss90221@gmail.com alzhu.san@163.com cszxl@imu.edu.cn lwj@nlpr.ia.ac.cn

ABSTRACT

In many present speech separation approaches, the separation task is formulated as a binary classification problem. Several classification-based approaches have been proposed and performed satisfactorily. However, they do not explicitly model the correlation in time and each time-frequency (T-F) unit is still classified individually. As we know, the speech signal has a very rich time series and temporal dynamic information that can be exploited for speech separation. In this study, we incorporate the correlation in time into classification. Compared with the previous approaches, the proposed approach achieves better separation and generalization performance by using deep stacking networks (DSN) with time series and re-threshold method.

Index Terms— Speech Separation, Deep Stacking Networks, Binary Classification, Computational Auditory Scene Analysis (CASA)

1. INTRODUCTION

A good speech separation system has many important real world applications, such as hearing aids design and robust automatic speech recognition (ASR). However, the performance of present separation systems in general acoustic environments is far from being satisfactory. In particular, for the monaural speech separation, the separation is more difficult due to the lack of sound directions and spatial information. In this paper, we focus on the monaural speech separation from non-speech interference.

Inspired by human auditory perception, computational auditory scene analysis (CASA) attempts to mimic human auditory processing to solve speech separation problem and has shown considerable promise [1, 2]. Compared to other speech separation approaches, such as spectral subtraction [3], Weiner filtering [4] and model based approaches [5, 6], CASA doesn't need to make strong assumptions about interference and have more potential to deal with the more general interference by using grouping and segmentation principles [1, 2, 7, 8]. However, CASA heavily relies on the reliable detection of pitch and onset/offset segments in noise that are both formidable tasks, and has limited capacity in dealing with unvoiced speech due to the lack of harmonic structure.

The ideal binary mask (IBM) has been proposed as a primary goal of CASA and proven to be able to produce substantial improvements of human speech intelligibility in noise [9, 10, 11]. The IBM is a binary matrix of the time-frequency (T-F) units, where the matrix element is 1 if the signal-to-noise ratio (SNR) within the corresponding T-F unit is greater than a local SNR criterion (LC=0) and is 0 otherwise. Adopting IBM as the goal of CASA, we naturally cast IBM estimation as a binary classification problem, which has been proven to be an effective formulation [10, 12, 13, 14].

Influenced by the speech production mechanism and linguistic constraints, speech signal contains rich time series and temporal dynamic information that could be exploited for speech separation. However, so far, there have been few classification-based methods explicitly incorporating these information into classification. Existing classification-based methods, such as GMM [10] and SVM [12], mainly attempt to model the correlation with delta features and context information. Although they can capture some temporal variations on the feature level, each T-F unit is still classified individually. Furthermore, modeling the correlation on the feature level is extremely difficult and greatly enlarges the input vector and increases the complexity of the model. To address this deficiency, we propose a new variant of DSN [15] that is capable of effectively capturing interactions between the neighboring T-F units by taking time series into account, which is named "DSN with the time series" (DSN-TS). To further improve speech separation quality, we also use re-threshold method to convert the DSN-TS's output into binary IBM, the threshold is estimated by maximizing hit minus false alarm rates (HIT-FA), which is shown to be highly correlated with human speech intelligibility [10]. Here, the HIT rate is the percent of correctly classified target-dominant T-F unit(1s) in the IBM and the FA rate is the percent of the wrongly classified interference-dominant(0s) T-F units in the IBM.

This paper is organized as follows. In section 2, we present an overview of the proposed approach. Section 3 gives the evaluation. We conclude the paper in Section 4.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2. APPROACH DESCRIPTION

2.1. DSN-TS

The proposed DSN-TS is a variant of DSN [15]. DSN is composed of a variable number of basic modules, and those basic modules are stacked up one by one to be a DSN (Fig. 1(b)). The basic module is a one hidden layer perceptron, which includes one input layer, one linear output layer, one sigmoidal nonlinear hidden layer and two sets of trainable weights (Fig. 1(a)). We denote the upper-layer weight matrix by $\mathbf{U}$ and the lower-layer weight matrix by $\mathbf{W}$. The linearity in the output units permits the weight matrix $\mathbf{U}$ to be able to be determined through a closed-form solution given the weight matrix $\mathbf{W}$. And the weight matrix $\mathbf{W}$ can also be trained efficiently with stochastic gradient descent algorithm by descending the gradient of the loss function with respect to $\mathbf{W}$. See [15] for the detailed description on training algorithm.

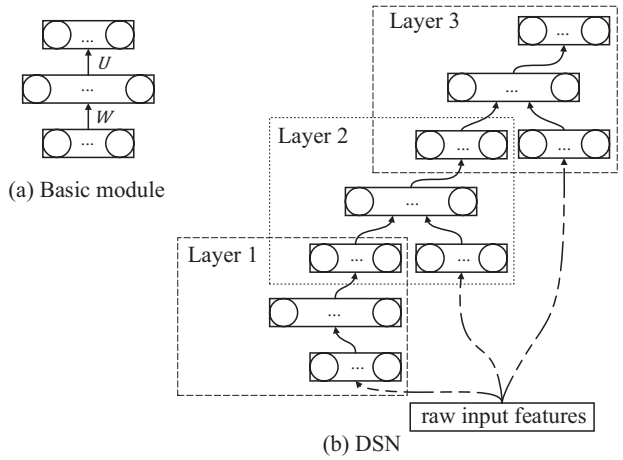

Fig. 1. Architecture of DSN. (only 3 layers are shown)

DSN exploits supervision information by stacking up all basic modules. The input units of all basic modules take the same raw features. But a key difference is that the input units of the higher module also take the outputs of the adjacent lower module. It means the outputs of higher layer module rely on not only the original input features but also the prediction of lower layer module. In other words, the lower layer module will help the higher one for classification.

We can regard DSN as a finite expansion of recurrent neural network, as shown in Fig. 2. This recurrent neural network is used to model the joint probability distribution $p(in, out)$, and note that the input units are clamped with the same original input after each cycle, so DSN can not model the correlation in time for the neighboring T-F units. To address this deficiency, we expand the DSN with finite time series and get a variant of DSN, which is denoted as DSN-TS. Assume $T = ..., (in_{k-1}, out_{k-1}), (in_k, out_k), (in_{k+1}, out_{k+1}), ...$ is the time series, where (in_k, out_k) are the corresponding input and output of a basic module. We built a model to capture the relationship between the adjacent two items in the time series T with the joint probability distribution $p(out_{k-1}, in_k)$ like a Markov model. This probability distribution can also be modeled with DSN-TS as the recurrent neural network is shown in Fig. 3. But, compared to the DSN, the only difference is that we clamp the model with the next input item in the time series T after each cycle instead of the original input.

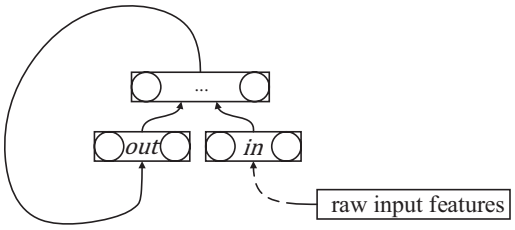

Fig. 2. DSN as a recurrent neural network.

In the paper, we take DSN-TS as the speech separation model, and train one DSN-TS for each frequency channel. The time series item (in_k, out_k) in this application are the features extracted from the mixture and the corresponding IBM estimation in the k-th frame T-F unit, respectively for in_k and out_k. We illustrate the DSN-TS in Fig. 3.

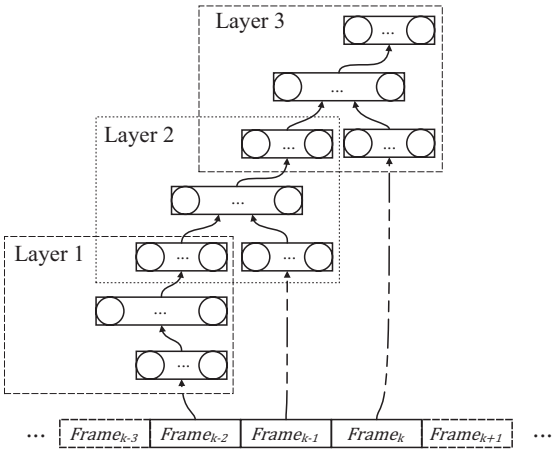

Fig. 3. Architecture of DSN-TS. (only 3 layers are shown)

The initialization of $\mathbf{W}$ is important [16]. Recently, restricted Boltzmann machines (RBM) are widely used to pretrain a deep neural networks (DNN) and have achieved great success. In this paper, we initialize the $\mathbf{W}$ of the lowest layer module with RBM trained with the raw features and the $\mathbf{W}$ of the higher layer modules with RBM trained with the current T-F unit's features and the label of the previous T-F unit.

2.2. Re-threshold

Since the outputs of DSN-TS modules are linear and for one frequency channel, 1 and 0 elements in the IBM are unbalance, the distribution of the DSN-TS modules' outputs would

look like something shown in Fig. 4. So we need to find a threshold based on some criterions to convert the outputs into binary results. To improve separation quality, we take HIT-FA as the criterion to select threshold. HIT-FA is highly correlative with human speech intelligibility [10]. Instead of a hard threshold, we set a soft threshold to convert the DSN-TS's outputs into a number between 0 and 1 that can be interpreted as the probability of the T-F unit being target-dominant. The converting function used here is a logistic function [17]:

$$P(Y=1|\mathbf{x}) = \frac{1}{1+\exp\left(\alpha f(\mathbf{x})+\phi\right)}$$

where $f(\mathbf{x})$ denotes the output of the DSN-TS, and the parameters α and ϕ decide the shape of the logistic function, which are fixed by maximizing the HIT-FA in the training phase.

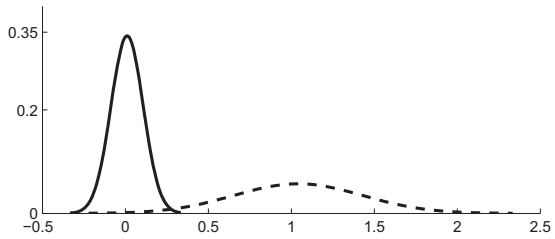

Fig. 4. The distribution of the DSN-TS's example outputs.

Due to speech variety and immanent pattern, the local thresholds varying with time may be more suitable for improving the HIT-FA rate than the overall threshold. To find the local thresholds, we divide the full output of a module into a set of slices with 256 frames length and 128 frames shift. For each of the slices, we estimate a corresponding threshold by maximizing the HIT-FA. Firstly, we estimate the distribution of each slice by using Gaussian kernel function. And then we use a one hidden layer DNN to model the correspondence between the thresholds and the distributions of the slices for predicting the local thresholds in testing phase. Note that we respectively evaluate the local thresholds for each frequency channel and apply the proposed re-threshold method to the output of each module for providing a higher classification starting point for the next module.

2.3. Feature Extraction

We extract the features to estimate IBM as follows. We first apply the 64-channel Gammatone filterbank with their center frequencies evenly distributed over the frequency space from 50 to 8000 Hz in the log axis [1] to the input mixture signal. Then, the output of each channel is windowed into 20-ms time frames with 10-ms frame shift, and this processing decomposes the one-dimensional input signal into a two-dimensional T-F representation.

From the viewpoint of classification, we need to extract acoustic features from the subband signal slice for each T-F unit. This kind of unit-level feature is more suitable than the conventional frame-level one for separation task [18]. A set of T-F unit level complementary features has been shown very effective for separation in the previous researches, such as [18]. In this paper, we also use this set of features, which consists of amplitude modulation spectrogram (AMS), relative spectral transform and perceptual linear prediction (RASTA-PLP), Mel-frequency cepstral coefficients (MFCC) and pitch-based features. For pitch-based features, we use ideal pitches for training [19] but estimated ones for testing [20].

3. EXPERIMENT AND RESULTS

3.1. Dataset and Evaluation Metrics

We systematically evaluate the proposed approach on Chinese National Hi-Tech Project 863 corpus, which contain 100 female speakers and 100 male speakers, with 500 utterances for each speaker. For training, 50 utterances from one female speaker are randomly chosen to mix with 6 non-speech noises (babble, cocktail party, factory, siren, white and speech-shaped noise) at 0dB SNR. For testing, we randomly choose 20 different utterances from the same speaker to mix with 14 noises at -10,-5, 0, 5 and 10 dB SNR respectively, in which 6 noises come from training set and the remaining 8 noises are unseen in training set for testing the generalization of the proposed approach, including bird chirp, crow noise, crowd, machine operation, engine start noise, alarm, traffic, and wind noise. These noises mainly cover a variety of daily noises and most of them are highly non-stationary.

In this paper, we take classification accuracy, HIT-FA rate and SNR as the evaluation metrics.

3.2. Related Models

In order to systematically evaluate the separation and generalization performance of the proposed approach (denoted as DSN-TS), we compare with the GMM-based [10](denoted as GMM) and the DNN-SVM-based [10] (denoted as DNN-SVM) speech separation systems in both matched and unmatched conditions. We also present the result of the DNN-based separation system (denoted as DNN).

The DNN-based and the DNN-SVM-based systems are implemented in a similar way but with a key difference in the last classification step. Rather than taking the outputs from the sigmoidal nonlinear output layer of DNN as the estimated IBM, the DNN-SVM-based system take linear SVMs as classifiers trained with the concatenated features of the raw data and the last hidden layer activations of the DNN with a linear output layer. The DNNs used in the two systems both have two hidden layers with 200 units for each one. We use 400 epochs of mini-batch gradient descent for RBM pre-training and 500 epochs of L-BFGS for network fine-tuning.

For the DSN-TS-based system, the DSN-TS consists of 5 stacked modules, and the number of the hidden layer units for

each module is also 200. We use similar RBM pre-training configuration. But owing to the efficient training algorithm of DSN [15], we do only 20 epochs network fine-tuning for each module and no global fitting over the entire architecture and thus reducing over-fitting that is good for generalization.

All systems use the same feature sets. But because of the DSN-TS-based system involving the previous four frame T-F units, other systems use the context features with 5 frames length.

3.3. Evaluation

We first verify our beliefs that modeling the correlation in time between neighboring T-F units can bring about a better separation performance. As the result shown in Fig. 5, the HIT-FA metric is getting better with the increase of the number of stacked modules.

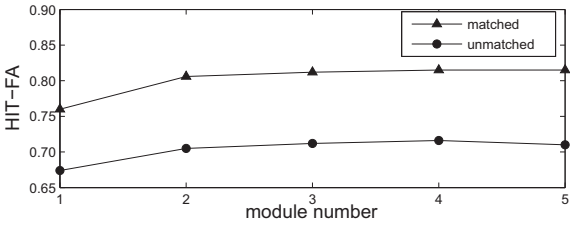

Fig. 5. Performance improves when more modules are stacked. (at 0dB SNR)

Table. 1 shows the classification performance of the different approaches on both matched-noise and unmatched-noise test set at 0dB SNR. The GMM performs the worst, which suggests that deep architectures are likely more suitable for the speech separation problem than shallow ones. We also note that the DNN-SVM outperforms the DNN, which mainly is due to the stronger generalization ability of linear SVM than sigmoid. Clearly, the proposed method remarkably outperforms the other comparisons on both matched and unmatched noise, which indicates modeling the correlation in time can improve the classification and generalization performance of the classifier.

Moreover, we also compare the generalization to unmatched SNRs with other approaches, shown in Fig. 6. As we see, The GMM-based approach fails to generalize and the proposed approach similarly performs the best mainly due to the re-threshold and the stability of the DSN-TS model.

4. CONCLUSION AND RELATION TO PRIOR WORK

The work presented here mainly focuses on classification-based monaural speech separation, which formulates the separation task as a binary classification problem. Under the subject of CASA, Kim and Wang et al proposed several GMM-based and SVM-based separation methods, such as [10], [12], [13], [18] etc. The key differences of this study are that we propose DSN-TS to model the correlation in time and each T-F unit is classified with considering the previous T-F units' classification information. The proposed DSN-TS is a variant of DSN proposed by Deng et al [15]. To improve separation quality, we also employ re-threshold method to maximize the HIT-FA rate. In our experiments, the proposed approach outperforms the previous ones, which is mainly benefit from the classification information of previous T-F units. Furthermore, as a result of the modular and layered stacking structure, DSN's training can be independently performed module by module through using convex optimization, which is much better than DNN [16] in terms of the time complexity.

Table 1. Performance of different systems at 0dB SNR.

	System	HIT	FA	HIT-FA	Accuracy	SNR(dB)
Matched	GMM	79.99	34.92	45.07	69.15	5.00
	DNN	73.18	8.36	64.82	88.63	9.34
	DNN-SVM	77.06	6.03	71.02	91.19	9.10
	DSN-TS	**80.04**	**4.28**	**75.76**	**92.99**	**9.84**
Unmatched	GMM	80.40	36.58	43.82	66.92	5.11
	DNN	68.09	11.08	57.01	84.13	7.82
	DNN-SVM	69.89	8.05	61.84	87.14	7.30
	DSN-TS	**75.40**	**8.79**	**66.61**	**87.69**	**7.44**

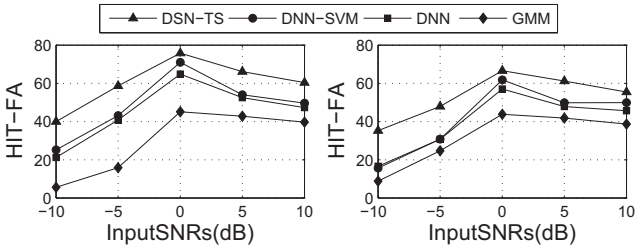

Fig. 6. Performance of different systems at different SNRs. (the left figure is matched-noise case, and the right figure is unmatched-noise case)

5. ACKNOWLEDGEMENTS

This research was supported in part by the China National Nature Science Foundation (No.61263037, No.61365006, No.91120303, No.61273267, No.90820011 and No.90820303) and inspired by the 2013 Dragon Star courses, hold in Tianjin University and Shanghai Jiao Tong University. We would like to thank Professor *Deliang Wang* in The Ohio State University, Researcher *Li Deng* in Microsoft Research, Redmond and Dr. *Shan Liang* in Institute of Automation for their helps.

6. REFERENCES

[1] D.L. Wang, G.J. Brown, et al., *Computational auditory scene analysis: Principles, algorithms, and applications*, vol. 147, Wiley interscience, 2006.

[2] A.S. Bregman, *Auditory scene analysis: The perceptual organization of sound*, MIT press, 1994.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 2, pp. 113–120, 1979.

[4] J.D. Chen, J. Benesty, Y.T. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1218–1234, 2006.

[5] S.T. Roweis, "One microphone source separation," in *Proc. NIPS*, 2000, pp. 793–799.

[6] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1118–1133, 2012.

[7] G.N Hu and D.L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions. Neural Networks.*, vol. 15, no. 5, pp. 1135–1150, 2004.

[8] G.N. Hu and D.L. Wang, "Segregation of unvoiced speech from nonspeech interference," *The Journal of the Acoustical Society of America.*, vol. 124, pp. 728–739, 2008.

[9] M.C. Anzalone, L. Calandruccio, K.A. Doherty, and L.H. Carney, "Determination of the potential benefit of time-frequency gain manipulation," *Ear and hearing.*, vol. 27, no. 5, pp. 480–492, 2006.

[10] G. Kim, Y. Lu, Y. Hu, and P.C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *The Journal of the Acoustical Society of America.*, vol. 126, pp. 1486–1494, 2009.

[11] D.L. Wang, U. Kjems, M.S Pedersen, J.B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *The Journal of the Acoustical Society of America.*, vol. 125, pp. 2336–2347, 2009.

[12] K. Han and D.L. Wang, "An svm based classification approach to speech separation," in *Proc. ICASSP*, 2011, pp. 4632–4635.

[13] Y.X. Wang and D.L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, pp. 1381–1390, 2013.

[14] Y.X. Wang and D.L. Wang, "Cocktail party processing via structured prediction," in *Proc. NIPS*, 2012, pp. 224–232.

[15] L. Deng, D. Yu, and J. Platt, "Scalable stacking and learning for building deep architectures," in *Proc. ICASSP*, 2012, pp. 2133–2136.

[16] G.E. Hinton and R.R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[17] J. Platt et al., "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[18] Y.X. Wang, K. Han, and D.L. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, pp. 270–279, 2012.

[19] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot international.*, vol. 5, no. 9/10, pp. 341–345, 2002.

[20] Z.Z. Jin and D.L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 625–638, 2009.