LSTM Network Supported Linear Filtering For The CHiME 2016 Challenge

Xiaofei Wang, Ziteng Wang, Xu Li, Yueyue Na, Qiang Fu, Yonghong Yan

Insitute of Acoustics, Chinese Academy of Sciences

xiaofei.wang1987@gmail.com, wangziteng@hccl.ioa.ac.cn

Abstract

This paper explores the combination of the emerging long short-term memory (LSTM) and the well established linear filtering techniques, parametric multi-channel Wiener filtering (PMWF) as well as single-channel minimum variance distortionless response (MVDR), for robust front-end signal processing in a speech recognition system. LSTM is employed for the estimation of speech and noise statistics, which are then used to compute the filter coefficients. PMWF is utilized in a novel way that the residual noise power remains constant along the frequency axis, while single-channel MVDR exploits inter-frame correlation coefficient vector, taking advantage of LSTM network based mask prediction, for linear filter estimation. With the baseline recognition system, our proposed methods reach a final word error rates (WER) of 5.69% on the 6ch real evaluation set of CHiME-4 challenge.

Keywords: CHiME 2016 Challenge, Supervised Time-frequency Masking, Parametric Multichannel Wiener Filtering, Single-channel MVDR

1. Introduction

The technique of neural network has greatly promoted speech recognition in everyday environments. It also quickly expands its scope to the signal processing area. Articles apply deep neural network (DNN) for spectral mask estimation [1] or predicting the clean spectrum [2]. Both tasks report promising results. However, most neural network based approaches only deal with problems in the signal channel case.

While multi-channel algorithms are more capable of extracting the desired source and suppressing undesired components at the same time, microphone arrays are becoming commonplace in modern human-machine interaction systems. The well established minimum variance distortionless response (MVDR) and multi-channel Wiener filter (MWF), which have solid theoretical foundations, arose new interests.

MVDR filter is also proposed for single-microphone noise reduction [3]. This filter takes the speech correlations of consecutive time frames into account. Under the assumption that noise spectrum is known previously, the MVDR filter could achieve promising performance in terms of speech distortion which is a key factor that affects speech recognition accuracy rate.

For the task of robust speech recognition of CHiME-4 [4], one practical front-end signal processing solution is the combination of the above two techniques [5][6]. DNN deals well with the noisy data and makes no extra assumptions as in conventional methods. Meanwhile, the multi-channel algorithms and single-channel MVDR provide optimized solutions.

Specifically, long short-term memory (LSTM) is employed for the estimation of speech and noise masks as originally suggested in [6][7]. With short-time Fourier transform performed in 1024 points, the network input is of 513 nodes. We have the following one bi-directional LSTM layer of 256 nodes and two feed-forward layers of 513 nodes. The training targets are ideal binary masks of both speech and noise, which are calculated by weighting the local signal-to-noise ratio (SNR) and the local threshold (LC)

$$\mathcal{M} = \begin{cases} 1, & \mathrm{SNR} > LC \\ 0, & else \end{cases} \tag{1}$$

The Adam optimization algorithm [8] is used for tuning the network. Dropout and batch normalization techniques are also employed for improving the generalization performance.

In the testing phase, the predicted masks $\mathcal{M}'_c$ ($c = speech$ and $noise$) are used to calculate the power spectral density (PSD) matrixes that are needed by our proposed PMWF and MVDR.

$$\mathbf{\Phi}_{cc} = \sum \mathcal{M}'_c \, \mathbf{y}\mathbf{y}^H \tag{2}$$

where $\mathbf{y}$ is the observation vector and superscript H denotes Hermitian transpose.

2. Parametric multi-channel Wiener filter

In the 2ch and 6ch tasks, the multi-channel processing problem is formulated in the frequency domain. With an array of M microphones, we have

$$Y_p(j\omega) = X_p(j\omega) + N_p(j\omega), \quad p = 1, 2, ..., P \tag{3}$$

In order to extract the desired source $X(j\omega)$ from the noisy observations, we apply an optimal filter $\mathbf{h}(j\omega)$

$$X(j\omega) = \mathbf{h}^H(j\omega)\mathbf{y}(j\omega) \tag{4}$$

where $\mathbf{y}(j\omega) = [Y_1(j\omega)...Y_p(j\omega)...Y_P(j\omega)]^T$.

The solution of PMWF [9][10] is known as

$$\mathbf{h}(j\omega) = \frac{\mathbf{\Phi}_{nn}^{-1}(j\omega)\mathbf{\Phi}_{xx}(j\omega)}{\mu + \lambda(\omega)}\mathbf{u}_{ref} \tag{5}$$

where Φ_{nn}, Φ_{xx} are respectively the noise and speech PSD matrixes which can be derived by (2), $\mathbf{u}_{ref}$ is one zero vector except for the index of reference channel being one (The first channel was used as reference in CHiME-4). $\lambda(\omega) = \mathrm{tr}\{\mathbf{\Phi}_{nn}^{-1}(j\omega)\mathbf{\Phi}_{xx}(j\omega)\}$, μ is the hyper-parameter that controls the tradeoff between speech distortion and noise reduction. With a higher value, we get more noise reduction at the expense of more distortion.

In speech recognition applications, it is still unclear how the speech distortion and noise reduction factors will influence the final recognition performance. Here, we propose a novel parameter control strategy that proves quite effective. Particularly, the residual noise power (RNP) in the filter output is constrained to

be constant along the frequency axis. From Eq.(5), the output RNP is

$$\mathbf{h}^H(j\omega)\mathbf{\Phi}_{nn}(j\omega)\mathbf{h}(j\omega) = \frac{\phi_{x_{ref}x_{ref}}\lambda(\omega)}{[\mu+\lambda(\omega)]^2} \quad (6)$$

We denoted the desired RNP as r_{nn}. Hence, we have

$$\mu(\omega) = \sqrt{\phi_{x_{ref}x_{ref}}(\omega)\lambda(\omega)/r_{nn}} - \lambda(\omega) \quad (7)$$

It should be noted that the value of r_{nn} only scales the output rather than changes the spectral shape of speech. It is set to 1.0. By regulating the RNP, a bin-wise controller $\mu(\omega)$ is derived. The reason why r_{nn} is constant across frequencies is that the spectrums of filtered signals would be preferred flat, avoiding transient changes between adjacent bins.

3. Single-channel MVDR

In the 1ch task, the single-channel problem is formulated as follows in the frequency domain. The complex spectral noisy observation $Y(k,m)$ is thus given by

$$Y(k,m) = X(k,m) + N(k,m) \quad (8)$$

where k is the frequency bin number and m is the frame index. The estimate of the clean speech spectral component $X(k,m)$ is obtained by applying an FIR filter

$$\hat{X}(k,m) = \mathbf{h}^H(k,m)\mathbf{y}(k,m) \quad (9)$$

where L is the order of the filter (set 20), and

$$\mathbf{h}(k,m) = [H(k,m,0)...H(k,m,L-1)]^T \quad (10)$$

$$\mathbf{y}(k,m) = [Y(k,m)...Y(k,m-L+1)]^T \quad (11)$$

By introducing the speech inter-frame correlation (IFC) coefficient vector $\gamma_x(k,m)$, which is defined by (The operator $E[\cdot]$ denotes the expectation),

$$\gamma_x(k,m) = \frac{E[\mathbf{x}(k,m)X(k,m))]}{E[\|X(k,m)\|^2]} \quad (12)$$

Therefore, from [3] the single-channel MVDR filter is

$$\mathbf{h}_{mvdr}(k,m) = \frac{{\mathbf{\Phi}_\mathbf{y}}^{-1}(k,m)\gamma_x^*(k,m)}{\gamma_x^T(k,m){\mathbf{\Phi}_\mathbf{y}}^{-1}(k,m)\gamma_x^*(k,m)} \quad (13)$$

$$\mathbf{\Phi}_\mathbf{y}(k,m) = \lambda_y\mathbf{\Phi}_\mathbf{y}(k,m) + (1-\lambda_y)\mathbf{y}(k,m)\mathbf{y}^H(k,m) \quad (14)$$

where λ_y is the forgetting factor(set 0.95). Also, to calculate ${\mathbf{\Phi}_\mathbf{y}}^{-1}(k,m)$, the regularization is used,

$${\mathbf{\Phi}_\mathbf{y}}^{-1}(k,m) = \{\mathbf{\Phi}_\mathbf{y}(k,m) + \frac{\delta\cdot tr[\|\mathbf{\Phi}_\mathbf{y}(k,m)\|]}{L}\mathbf{I}_{L\times L}\}^{-1} \quad (15)$$

where $\delta > 0$ is the regularization parameter (set 0.04).

Specifically, IFC $\gamma_x(k,m)$ can be estimated as follows,

$$\begin{aligned}\gamma_x(k,m) &= \frac{\Phi_Y(k,m)}{\Phi_Y(k,m)-\Phi_N(k,m)}\gamma_y(k,m) \\ &- \frac{\Phi_N(k,m)}{\Phi_Y(k,m)-\Phi_N(k,m)}\gamma_n(k,m) \quad (16)\end{aligned}$$

$\Phi_Y(k,m)$ and $\Phi_N(k,m)$ represent the second-order statistics of observed signal $Y(k,m)$ and noise $N(k,m)$, respectively.

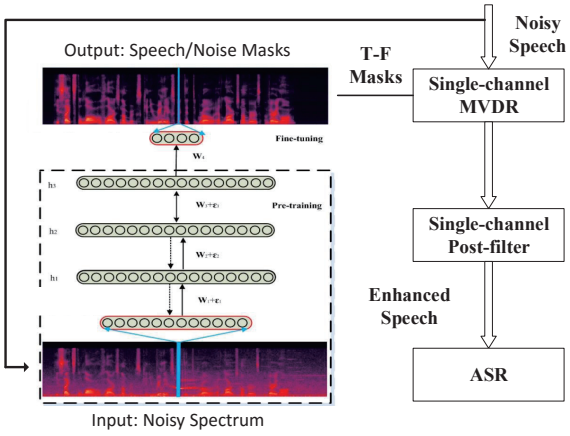

Figure 1: Diagram for the 1ch recognition task.

We use the speech soft mask $\mathcal{M}_c'$ to get the estimated noise component $\hat{N}(k,m)$ as follows,

$$\hat{N}(k,m) = (1-\mathbf{max}(\epsilon, 1-\mathbf{max}(\sqrt{\mathcal{M}_c'}, \epsilon)))Y(k,m) \quad (17)$$

where ϵ is an extremely small number to avoid sudden changes between frames.

Following the single-channel MVDR filtering, a stationary noise reduction algorithm [11][12] is applied to the filtered signal as a post-filter shown in Fig.1.

4. Experimental evaluation

4.1. 2ch and 6ch results

For all the recognition tasks, we always apply matched training. In the case of 2ch track, we randomly select two channels from all six channels to compose the training set. The channels selected for development and evaluation are kept unmodified. In the back-end (2ch and 6ch tasks), only one modification is made to the standard scripts. We make use of the fact that we have all six channels data available. Besides the enhanced data, we also use all six channel real and one channel simulated recordings in the training stage.

In the front-end, LSTM is trained with all the six channel simulated data [7]. The mask estimation is actually single-channel based, so we get separate outputs for each channel. For 2ch and 6ch tasks, the masks are then taken median between specific channels for robustness to outliers.

The results of 2ch and 6ch tasks using sequence training and RNN language model rescoring are given in Table 1, 2. The WERs of real test data in the 2ch and 6ch tasks are 9.64% and 5.69%, respectively.

4.2. 1ch results

In the 1ch task, we use 6 channels' data for matched training. The results are given by Table 3. A relative 15.59% WER decrease on real test data using GMM acoustic model is achieved compared to the official baseline, in which both training and testing data are noisy signals. Single-channel MVDR and post-filtering achieve the best performance since MVDR could filter the non-stationary noise without speech distortion, meanwhile, post-filtering is good at suppressing the stationary noise. Single-channel MVDR and post-filtering benefit each other.

Table 1: Average WER (%) for the multi-channel tested systems.

Track	System	Dev		Test	
		real	simu	real	simu
2ch	Baseline	8.23	9.50	16.58	15.33
	GMM	12.95	16.06	21.08	20.53
	DNN+sMBR	8.34	9.54	12.16	13.27
	DNN+RNNLM	**5.58**	**7.18**	**9.64**	**8.77**
6ch	Baseline	5.76	6.77	11.51	10.90
	GMM	9.25	9.24	12.70	10.49
	DNN+sMBR	5.43	5.19	8.25	6.51
	DNN+RNNLM	**3.65**	**3.71**	**5.69**	**4.38**

Table 2: WER (%) per environment for the current multi-channel best system.

Track	Envir.	Dev		Test	
		real	simu	real	simu
2ch	BUS	6.73	5.68	12.82	6.16
	CAF	5.97	10.10	10.59	10.29
	PED	4.34	6.03	8.56	9.15
	STR	5.26	6.92	6.57	9.47
6ch	BUS	4.81	3.33	7.35	3.46
	CAF	3.20	4.69	5.27	4.76
	PED	2.99	3.07	5.66	4.28
	STR	3.58	3.75	4.50	5.01

Besides GMM acoustic model, results of DNN acoustic model are also given in Table 4. The best results of test set are achieved using single-channel MVDR or single-channel MVDR + postfiltering, however, the best results of development set are achieved using unprocessed data. This phenomenon is different from the consistent improvements using GMM acoustic model. It is still expected to achieve a better tradeoff between noise reduction and distortion.

5. Conclusion

The main contributions of the submitted systems were two proposed front-end processing methods, which were multi-channel and single-channel noise reductions for specific recognition tasks, respectively. With a fine-tuning parametric multi-channel Wiener filter, WERs on 2ch and 6ch Real Test sets of CHiME-4 were reduced to 9.64% and 5.69%. Meanwhile, supervised time-frequency masking based single-channel MVDR filter with a post-filter performed well in the 1ch task. The results showed that WER of Real Test set decreased much on GMM acoustic model but slightly on DNN model. Experimental results also showed that enlarging the training data could bring benefits for CHiME-4 tasks.

6. Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 61601453) and the China Scholarship Council (No. 201604910007).

Table 3: Average WER (%) for the 1ch tested systems using the GMM-HMM acoustic model. Baseline(official) used only CH5 for training. Baseline(all channels) used all the 6 channels' data for training. sMVDR/sMVDR + Postfilter means all the 6 channels' data and test data are processed with single-channel MVDR/MVDR + Postfilter.

System	Dev		Test	
	real	simu	real	simu
Baseline(official)	22.15	24.49	37.54	33.3
Baseline(all channels)	20.87	23.07	35.01	31.65
sMVDR	19.71	**20.76**	34.64	28.96
sMVDR + Postfilter	**19.27**	20.92	**31.69**	**27.70**

Table 4: Average WER (%) for the 1ch tested systems using the DNN acoustic model (without sMBR and RNNLM rescore).

System	Dev		Test	
	real	simu	real	simu
Baseline(official)	14.86	15.47	27.27	24.09
Baseline(all channels)	**14.10**	**15.25**	25.74	22.79
sMVDR	15.24	16.10	25.21	**21.73**
sMVDR + Postfilter	15.90	16.92	**25.05**	22.34

7. References

[1] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[2] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[3] J. Benesty and Y. Huang, "A single-channel noise reduction mvdr filter," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 273–276.

[4] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language, to appear.*

[5] S. Sivasankaran, A. A. Nugraha, E. Vincent, J. A. Morales-Cordovilla, S. Dalmia, I. Illina, and A. Liutkus, "Robust asr using neural network based speech enhancement and feature simulation," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 482–489.

[6] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "Blstm supported gev beamformer front-end for the 3rd chime challenge," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 444–451.

[7] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 196–200.

[8] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[9] J. Benesty, J. Chen, Y. Huang, and B. Rafaely, "Microphone array signal processing," *Journal of the Acoustical Society of America*, vol. 125, no. 6, pp. 4097–4098, 2009.

[10] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 260–276, 2010.

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

[12] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on speech and audio processing*, vol. 9, no. 5, pp. 504–512, 2001.