DNN-BASED AR-WIENER FILTERING FOR SPEECH ENHANCEMENT

Yan Yang, Changchun Bao

Speech and Audio Signal Processing Laboratory, Beijing University of Technology, Beijing, China

1. A BRIEF OVERVIEW

The key ideas in our work contain:

- This paper presents a novel approach for estimating auto-regressive (AR) model parameters using deep neural network (DNN)
- The problem of residual noise between harmonics is overcome by speech-presence probability (SPP)
- The proposed approach is found to be significantly better than reference approaches on SSNR, PESQ and STOI

2. AR model parameters estimation

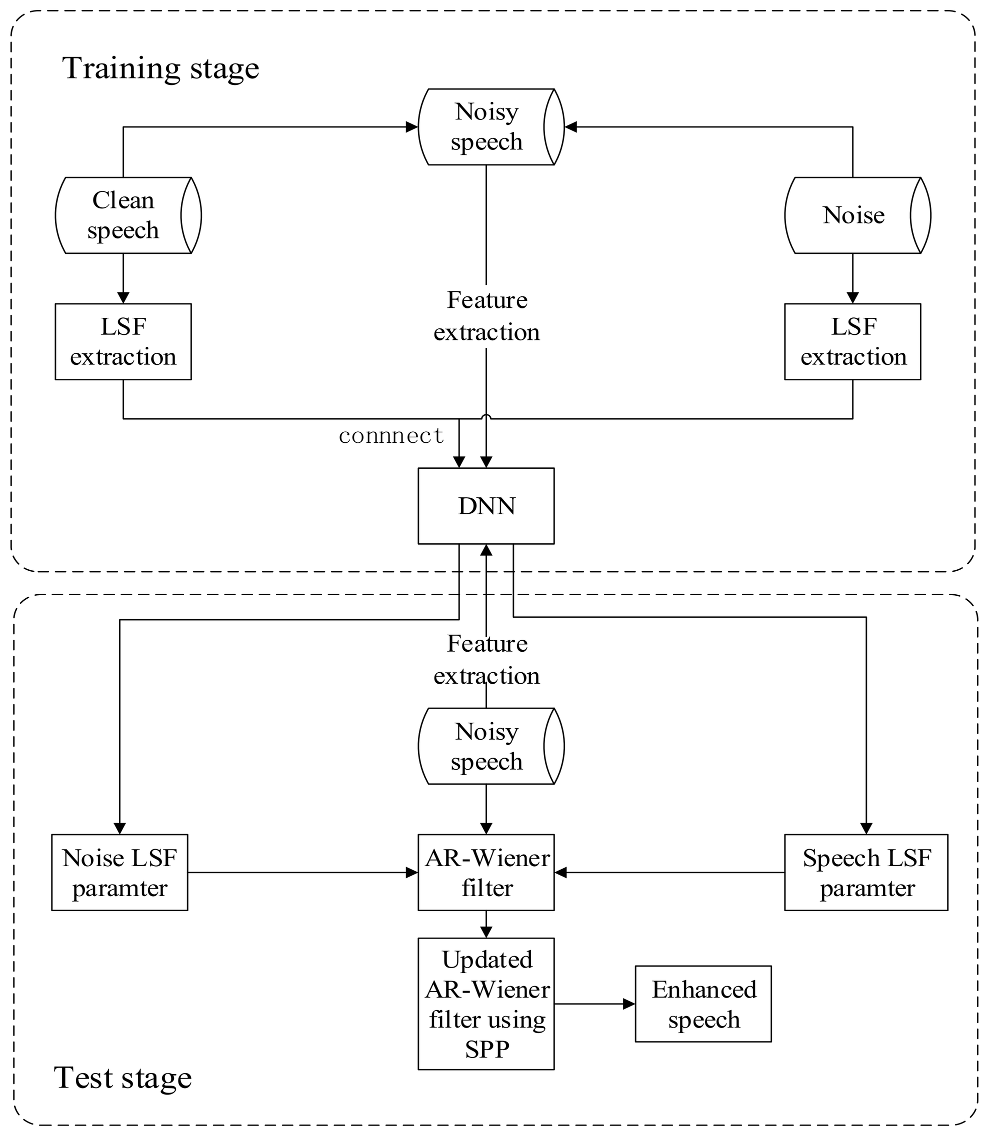

Fig. 1. Block diagram of the proposed method

● The training stage

→ The training feature is the log power spectrum（LPS）of noisy speech

$$\boldsymbol{x}=[x_1^{l-5},x_2^{l-5},\ldots x_m^{l-5}\ldots,x_1^{l},\ldots,x_m^{l},\ldots,x_1^{l+5},\ldots x_m^{l+5}]$$

→ The training target is the connected vector that combines the linear spectrum frequency（LSF）parameters of speech and noise

$$\boldsymbol{d}=[p_1^x,p_2^x,\ldots p_n^x,p_1^w,\ldots,p_n^w]$$

1)The cost function of the DNN is constructed based on the Euclidean distance minimum criterion

$$J=\frac{1}{M}\sum_{m=1}^{M}\left\|\boldsymbol{d}^{(m)}-f_{\boldsymbol{w},\boldsymbol{b}}(\boldsymbol{x}^{(m)})\right\|^2$$

2)We calculate the gradient of all weights and bias

$$\nabla w_{ji}=\frac{\partial J}{\partial w_{ji}} \qquad \nabla b_j=\frac{\partial J}{\partial b_j}$$

3)Gradient descent method is used to optimize the DNN

$$w_{ji}^{k+1}\leftarrow w_{ji}^{k}-\eta[(1-\alpha)\nabla w_{ji}^{k}+\alpha\nabla w_{ji}^{k-1}]$$

Repeat the step (2) and (3) until reaching the training epochs

Input dimension	129×11
Output dimension	20
Hidden layers number	3
Units number of each hidden layer	512
Activation function of hidden unit	Rectified Linear Unit
Activation function of output unit	tanh
Training epoch (K)	50
Learning rate (μ)	0.03×0.98^k
Momentum (α)	0.2 (first 20 epochs) →0.5

Tab. 1. Parameters setup of DNN

●The test stage

1)We extract the LPS of test speech as the input feature of the DNN and the output is the estimated LSF parameters of speech and noise.

$$\hat{\boldsymbol{a}}_{lsf}=f_{w,b}(\boldsymbol{x})$$

2)We transform the LSF to LPC parameters and calculate the spectral shape of speech and noise

$$A_x(k)=\sum_{m=0}^{p}\hat{a}_{x,m}\exp(-\frac{2\pi m}{K}k) \qquad A_w(k)=\sum_{m=0}^{p}\hat{a}_{w,m}\exp(-\frac{2\pi m}{K}k)$$

3)The AR gain is calculate by multiplicative update rule

$$\hat{g}_x\bullet\frac{(H_x)^T[(H_yW_y)^{-2}\bullet P_y]}{(H_x)^T(H_yW_y)^{-1}}\to\hat{g}_x \qquad \hat{g}_w\bullet\frac{(H_w)^T[(H_yW_y)^{-2}\bullet P_y]}{(H_w)^T(H_yW_y)^{-1}}\to\hat{g}_w$$

where

$$\hat{\boldsymbol{P}}_y=[\hat{P}_y(0),\ldots,\hat{P}_y(K-1)]^T \qquad \boldsymbol{H}_x=[\frac{1}{|A_x(0)|^2},\ldots,\frac{1}{|A_x(K-1)|^2}]^T \qquad \boldsymbol{H}_w=[\frac{1}{|A_w(0)|^2},\ldots,\frac{1}{|A_w(K-1)|^2}]^T \qquad \boldsymbol{W}_y=[g_y]$$

4)The AR-Wiener filter is constructed by the AR model parameters and the AR gains

$$WF_{AR}(k)=\frac{\hat{g}_x}{|A_x(k)|^2}\Big/\frac{\hat{g}_x}{|A_x(k)|^2}+\frac{\hat{g}_w}{|A_w(k)|^2}$$

3. SPEECH-PRESENCE PROBABILITY（SPP）

$H_0^k\to$ the state that speech is absent in frequency bin k

$H_1^k\to$ the state that speech is present in frequency bin k

Under the Gauss distribution of speech and noise

$$P(Y_k\mid H_0^k)=\frac{1}{\pi\lambda_d(k)}\exp(-\frac{Y_k^2}{\lambda_d(k)}) \qquad P(Y_k\mid H_1^k)=\frac{1}{\pi(\lambda_x(k)+\lambda_d(k))}\exp[-\frac{Y_k^2}{(\lambda_x(k)+\lambda_d(k))}]$$

So, the SPP is calculated

$$P(H_1^k\mid Y_k)=\frac{P(Y_k\mid H_1^k)P(H_1^k)}{P(Y_k)}=\frac{P(Y_k\mid H_1^k)P(H_1^k)}{P(Y_k\mid H_1^k)P(H_1^k)+P(Y_k\mid H_0^k)P(H_0^k)}=\frac{1-q_k}{1-q_k+q_k(1+\xi_k')\exp(-v_k')}$$

where

$$\xi_k'=\frac{\xi_k}{1-q_k} \quad v_k'=\frac{\xi_k'}{\xi'+1}\gamma_k \quad \xi_k=\frac{X_k^2}{D_k^2}\approx\frac{P_x(k)}{P_w(k)} \quad \gamma_k=\frac{Y_k^2}{D_k^2}\approx\frac{Y_k^2}{P_w(k)} \quad q_k=P(H_0^k) \quad P_x(k)=\frac{\hat{g}_x}{|A_x(k)|^2} \quad P_w(k)=\frac{\hat{g}_w}{|A_w(k)|^2}$$

Finally, the SPP is used to update the AR-Wiener filter

$$WF_{updated}(k)=P(H_1^k\mid Y_k)WF_{AR}(k)$$

4. PERFORMANCE EVALUATION

● Experimental setup

Speech dataset	TIMIT	FFT size	256
Training hours	8	Window	hamming
Fs	8khz	Noise dataset	Noisex-92
Frame size	256	Noise type	babble f16 factory buccaneer
Frame shift	128	Input SNR	-5dB 0dB 5dB 10dB

●Reference Methods

Ref. A	DNN-based amplitude recovering [11]
Ref. B	Codebook-Based Sparse Hidden Markov Models method [8]
Ref. C	DNN-based ideal ratio mask (IRM) method [12]
Pro. A	AR-Wiener filtering without SPP
Pro. B	AR-Wiener filtering with SPP

●SSNR test results

	-5dB	0dB	5dB	10dB
Ref. A	12.0243	9.1580	6.3286	3.3413
Ref. B	11.5093	9.1606	6.1674	5.8473
Ref. C	10.7430	9.9106	8.7555	7.2213
Pro. A	13.8090	12.3177	10.1973	7.4086
Pro. B	**14.1283**	**13.1526**	**11.6552**	**9.4908**

●PESQ test results

	-5dB	0dB	5dB	10dB
Noisy	1.4180	1.6824	2.0107	2.3455
Ref. A	1.3209	1.6338	2.0067	2.3140
Ref. B	1.4120	1.8497	2.3050	2.6476
Ref. C	1.5932	1.9532	2.3352	**2.7484**
Pro. A	1.5819	1.9854	2.3414	2.6585
Pro. B	**1.6666**	**2.0452**	**2.3942**	2.7318

●STOI test results

	-5dB	0dB	5dB	10dB
Noisy	0.5148	0.6300	0.7453	0.8433
Ref. A	0.5253	0.6381	0.7456	0.8158
Ref. B	0.4981	0.6118	0.7211	0.8093
Ref. C	0.5989	0.7119	0.8114	0.8876
Pro. A	0.6114	0.7218	0.8298	0. 8829
Pro. B	**0.6312**	**0.7490**	**0.8351**	**0.8945**

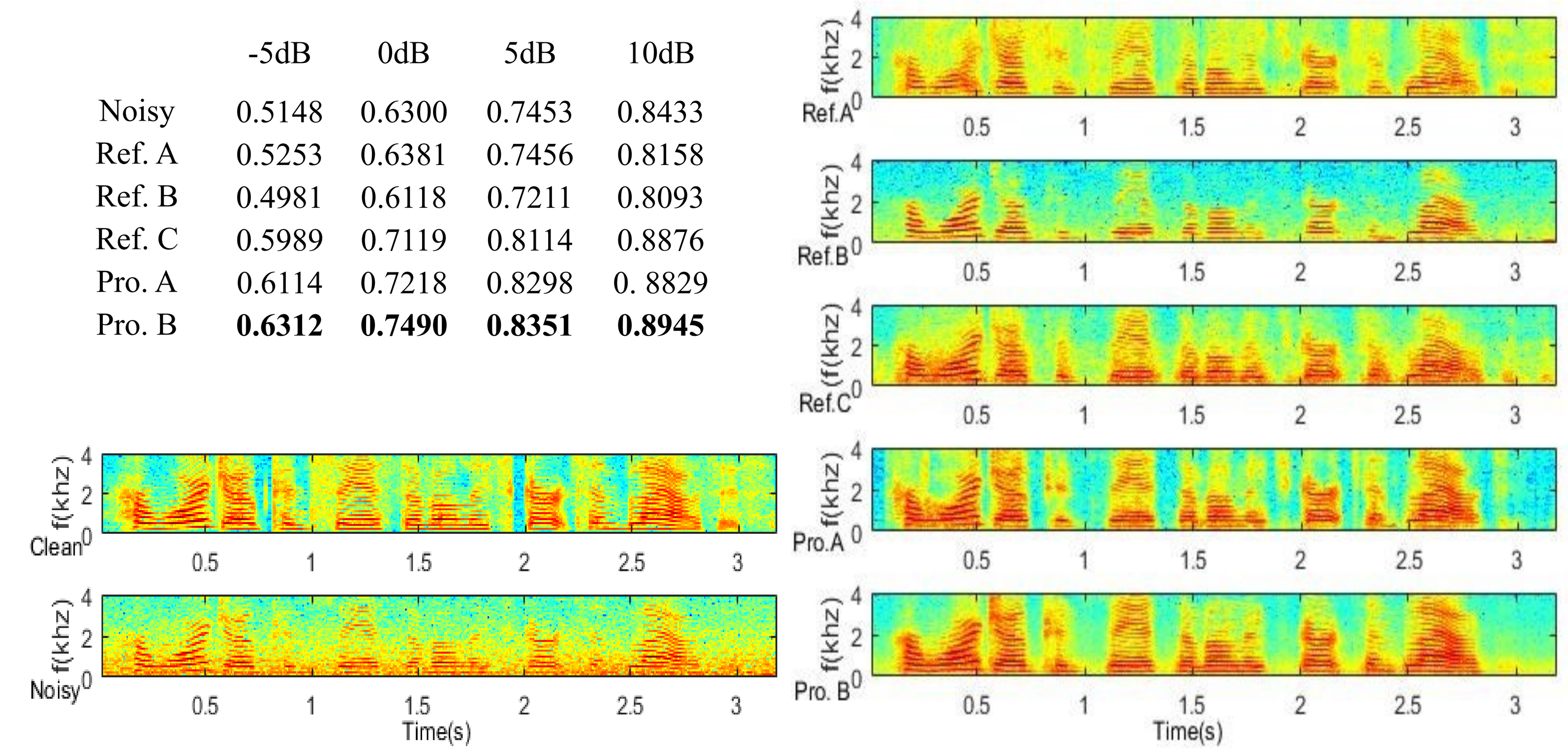

Fig. 2. Spectrum comparison

5. CONCLUSIONS AND FUTURE WORK

● Conclusions

＊ The DNN is used to estimate the AR model parameters of speech and noise simultaneously

＊ The AR-Wiener filter is constructed by the AR model parameters of speech and noise

＊ In order to remove the noise between harmonics, we use the speech-presence probability to update the AR-Wiener filter.

● Future Work

＊ More robust features should be explored

＊ We can try other network structure which takes into account the temporal correlations

E-mails: yangyan00800@emails.bjut.edu.cn, chchbao@bjut.edu.cn