LINEAR PREDICTION-BASED PART-DEFINED AUTO-ENCODER USED FOR SPEECH ENHANCEMENT

Zihao Cui, Changchun Bao

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

1. The Main Contributions of This Paper

- This paper presents a novel linear prediction-based part-defined auto-encoder (PAE) network
- The parallel Network is used to estimate the modification factor of AR-wiener filter mask
- The PESQ and STOI results of the LP-based PAE are better than baseline method at lower signal noise ratio (SNR) levels

2. PAE-based Speech Enhancement

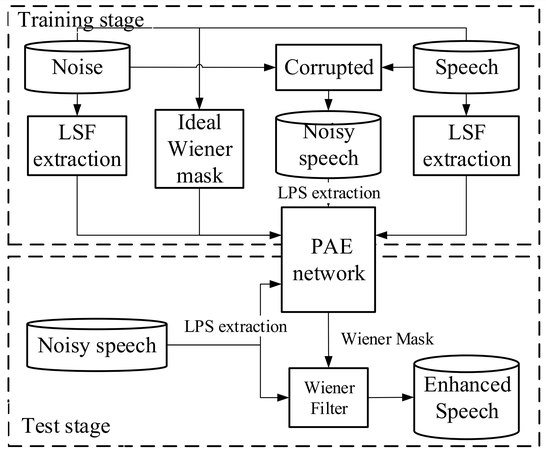

Fig. 1. The block diagram of the proposed PAE

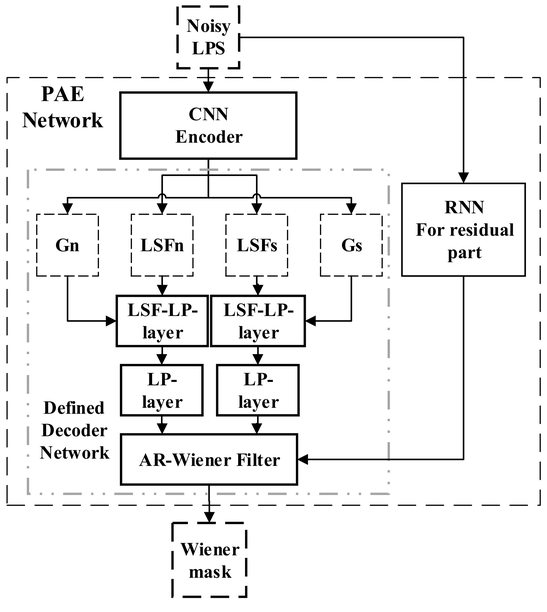

Fig. 2. The proposed Part-defined AutoEncoder

3. The Loss Function of The PAE

● PAE based on the AR filter

The Decoder as synthesizer is based on the AR Wiener filter

$$H_{AR-WF}(k)=\frac{\dfrac{g^{(s)}}{\left|A^{(s)}(k)\right|^2}}{\dfrac{g^{(s)}}{\left|A^{(s)}(k)\right|^2}+\dfrac{g^{(n)}}{\left|A^{(n)}(k)\right|^2}} \quad (1)$$

The Loss function based on the Eq. (1)

$$\begin{aligned}E_r=&\frac{1}{K}\left\|\widehat{\mathbf{WM}}\left(\mathbf{Y}_{l-t}^{l+t},\mathbf{W},\mathbf{b}\right)-\mathbf{WM}_l\right\|_2^2\\&+\frac{1}{N}\left\|\widehat{\mathbf{LSF}^{(s)}}\left(\mathbf{Y}_{l-t}^{l+t},\mathbf{W},\mathbf{b}\right)-\mathbf{LSF}_l^{(s)}\right\|_2^2\\&+\frac{1}{N}\left\|\widehat{\mathbf{LSF}^{(n)}}\left(\mathbf{Y}_{l-t}^{l+t},\mathbf{W},\mathbf{b}\right)-\mathbf{LSF}_l^{(n)}\right\|_2^2\end{aligned} \quad (2)$$

The LSF coefficients are more suitable for the target of Coding layer, not only for the boundary [0,2π] of it, but concentrate on the frequency of the formant.

● Modified AR-Wiener filter with the residual

Linear predicition residual with AR model an infinite impulse response (IIR) filter

$$g\bullet r(n)=x(n)*\left[\delta(n)-\delta(n)*h_{IIR}(p)\right]$$

In power spectral density (PSD):

$$X(k)=g\frac{R(k)}{1-H_{IIR}(k)}$$

The modified AR-Wiener filter and its approach estimation

$$H_{M-AR-WF}(k)=\frac{\dfrac{g^{(s)}R^{(s)}(k)}{\left|A^{(s)}(k)\right|^2}}{\dfrac{g^{(s)}R^{(s)}(k)}{\left|A^{(s)}(k)\right|^2}+\dfrac{g^{(n)}R^{(n)}(k)}{\left|A^{(n)}(k)\right|^2}} \quad (3) \qquad H_{M-AR-WF}(k)=H_{AR-WF}(k)H_{r-WF}(k) \quad (4)$$

● Details of the proposed PAE

Input: 11 frames noisy speech with 129 frequency bins				
Layer index	Type of layer	Number of filter	Output shape (nodes)	Previous output layer
RNN structure				
9	RU(2FC)		1419	input
10	FC		2048	9
11	FC		129	9
Output of RNN: Modified factors in 129 frequency bins				

Input: 11 frames noisy speech with 129 frequency bins				
Layer index	Type of layer	Number of filter	Output shape (nodes)	Previous output layer
CNN structure				
1	Conv2D(2,2) Maxp2D(2,2)	6	(10,128) (5,64)	input
2	Conv2D(2,2) Maxp2D(2,2)	16	(4,63) (2,31)	1
3	FC		2048	2
4	FC		2048	3
5	FC		12	4 (first 1500)
6	FC		12	4 (last 1500)
7	FC/Maxout		20/1	3 (first 100)
8	FC/Maxout		20/1	3 (last 100)
Output of CNN: LSFs and gains of speech and noise				

4. Performance Evaluation

● Experimental setup

Tab. 1. *Parameters setup of DNN*

Speech dataset	TIMIT
AR order	12
Fs	8khz
Frame size	256
Frame shift	128

FFT size	256
Window	hamming
Noise type	babble f16 factory(training and test) street and office (test)
Input SNR	-5dB 0dB 5dB 10dB

● Reference Methods

Tab. 2. *Reference Methods*

W-DNN	DNN with the target of wiener filter mask [1]
Pro. A	PAE without residual estimation
Pro. B	PAE with residual estimation as the modified method

● Test results

Tab. 3. PESQ details

SNR(dB)	Noise Type	Noisy	W-DNN	Pro A	Pro B
-5	Babble	1.719	1.941	1.793	**1.986**
	F16	1.595	2.125	2.037	**2.139**
	Factory	1.761	2.298	2.207	**2.333**
	Office	1.860	2.067	1.975	**2.078**
	Street	1.990	2.570	2.502	**2.614**
	Average	1.785	2.200	2.103	**2.230**
0	Babble	1.931	2.338	2.180	**2.377**
	F16	1.832	2.520	2.401	**2.523**
	Factory	2.073	2.699	2.577	**2.734**
	Office	2.175	2.489	2.403	**2.511**
	Street	2.320	2.946	2.868	**2.983**
	Average	2.066	2.598	2.486	**2.625**
5	Babble	2.221	2.733	2.567	**2.749**
	F16	2.136	2.898	2.771	**2.908**
	Factory	2.391	3.064	2.932	**3.086**
	Office	2.492	2.889	2.811	**2.906**
	Street	2.650	3.287	3.198	**3.297**
	Average	2.378	2.974	2.856	**2.989**
10	Babble	2.518	**3.098**	2.930	**3.098**
	F16	2.444	3.225	3.105	**3.239**
	Factory	2.705	**3.374**	3.244	3.37
	Office	2.803	**3.256**	3.171	3.253
	Street	2.973	**3.577**	3.477	3.565
	Average	2.689	**3.306**	3.185	3.305

Tab. 4. The average results of the PESQ and STOI (%)

SNR	Noisy		W-DNN		Pro. B	
	PESQ	STOI	PESQ	STOI	PESQ	STOI
-5	1.785	57.142	2.200	68.392	**2.230**	**69.043**
0	2.066	68.242	2.598	78.674	**2.625**	**78.802**
5	2.378	78.008	2.974	**85.773**	**2.989**	85.559
10	2.689	85.704	**3.306**	**90.453**	3.305	90.06
Avg.	2.230	72.274	2.770	80.823	**2.787**	**80.866**

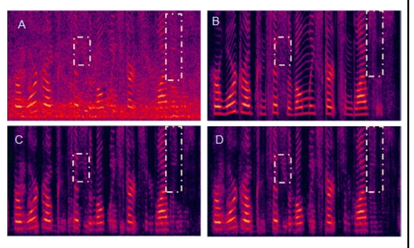

Fig. 3. Spectrogram comparison. A. Speech corrupted by Babble at 0 dB; B. Clean speech; C. Enhanced speech by the W-DNN; D. Enhanced speech by Pro. B

✲ The PAE can estimate the wiener mask based on the AR model constriant

5. Conclusions and Future Work

● Conclusions

✲ the PAE is used to estimate the AR model parameters of speech and noise and the mask of wiener-filter simultaneously

✲ the RNN is given to modify the estimated wiener mask ratios of PAE

✲ the proposed neural network concentrate more spectrum structure based on the AR model

● Future Work

✲ Learning the perception of the AR model and the mask design based on PAE

✲ the structure design or the voiced speech model for better residual estimation

Main Reference:

[1]Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 22, no. 12, pp. 1849-1858, 2014.

[2]G. Kang, and L. Fransen, "Application of line-spectrum pairs to low-bit-rate speech encoders," in Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85., 1985, pp. 244-247.

[3]Y. Yang, and C. Bao, "Dnn-Based Ar-Wiener Filtering for Speech Enhancement," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 2901-2905.

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