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

![Diagram of proposed method]

**Fig. 1. Block diagram of the proposed method**

- **The training stage**
  - The training feature is the log power spectrum (LPS) of noisy speech
  - The training target is the connected vector that combines the linear spectrum frequency (LSF) parameters of speech and noise

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

\[ J = \frac{1}{2} \int \left[ f_d(x) - f_m(x) \right]^2 dx \]

2) We calculate the gradient of all weights and bias

\[ \nabla V_s = \frac{\partial J}{\partial V_s}, \quad \nabla b_s = \frac{\partial J}{\partial b_s} \]

3) Gradient descent method is used to optimize the DNN

\[ w_{s,t+1}^{n+1} = w_{s,t}^{n+1} - \eta (1-\alpha) \nabla V_s \]

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

3. SPEECH-PRESENCE PROBABILITY (SPP)

\[ R_c \rightarrow \text{the state that speech is absent in frequency bin k} \]

\[ R_s \rightarrow \text{the state that speech is present in frequency bin k} \]

Under the Gausssian distribution of speech and noise

\[ P(V|R_s) = \frac{1}{\sqrt{2\pi} \sigma} \exp \left( -\frac{(V-\mu)^2}{2\sigma^2} \right) \]

So, the SPP is calculated

\[ R_{s,m} = \frac{P(V|R_s) P(R_s)}{P(V|R_s) P(R_s) + P(V|R_m) P(R_m)} \]

where

\[ \sigma_s^2 = \frac{\lambda_s \sigma_s^2}{\lambda_s + \sigma_s^2}, \quad \lambda_s = \frac{\lambda_s}{\lambda_s + \sigma_s^2}, \quad \sigma_m^2 = \frac{1-\lambda_s}{\lambda_s + \sigma_s^2} \]

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

\[ W_{s,m} = \frac{P(V|R_s) P(R_s)}{P(V|R_s) P(R_s) + P(V|R_m) P(R_m)} \]

4. PERFORMANCE EVALUATION

- **Experimental setup**
  - Speech dataset: TIMIT
  - FFT size: 256
  - Training hours: 8
  - Fs: 8kHz
  - Noise dataset: Noise-92
  - Frame size: 256
  - Noise type: bubble & factory bucket
  - Frame shift: 128

- **Reference Methods**
  - 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. B: 11.5093 11.4666 5.4981 4.6976

- **PESQ test results**
  - -5dB 0dB 5dB 10dB
  - Ref. A: 1.2080 1.6027 2.0107 2.6855
  - Ref. B: 1.1260 1.6358 2.0667 2.3143
  - Pro. A: 1.2477 1.7407 2.2071 2.4976
  - Pro. B: 1.3392 1.9352 2.3352 2.7454

**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