REINFORCEMENT LEARNING BASED SPEECH ENHANCEMENT FOR ROBUST SPEECH RECOGNITION

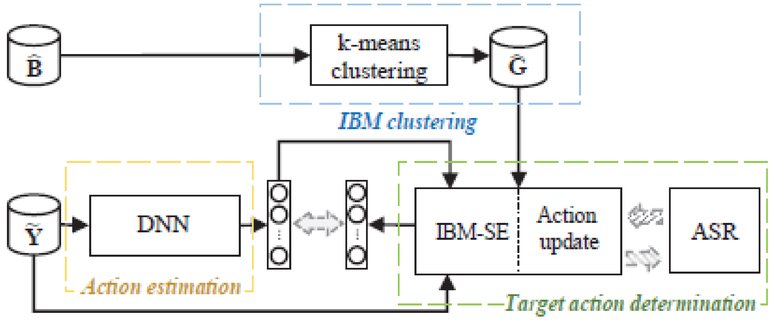
Yih-Liang Shen[1], Chao-Yuan Huang[1], Syu-Siang Wang[2], Yu Tsao[3], Hsin-Min Wang[2,4], and Tai-Shih Chi[1]

[1]Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, R.O.C
[2] Joint Research Center for AI Technology and All Vista Healthcare, MOST, Taipei, R.O.C
[3] Research Center for Information Technology Innovation, Academia Sinica, Taipei, R.O.C
[4] Institute of Information Science, Academia Sinica, Taipei, R.O.C

Introduction

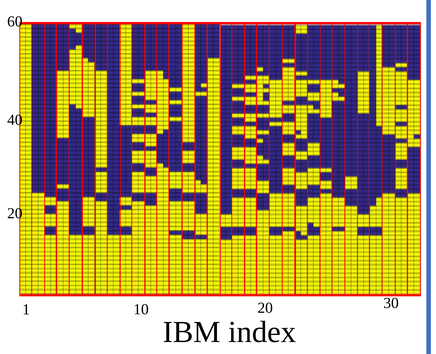
- The mean square error (MSE) optimized model may not directly improve the performance of an automatic speech recognition (ASR) system. If the target is to minimize the recognition error, the recognition results should be used to design the objective function for optimizing the speech enhancement (SE) model.
- we propose to adopt the reinforcement learning (RL) algorithm to optimize the SE model based on the recognition results. We evaluated the proposed RL-based SE system on the Mandarin Chinese broadcast news corpus (MATBN). Experimental results demonstrate that the proposed SE system can effectively improve the ASR results.

Proposed SE system

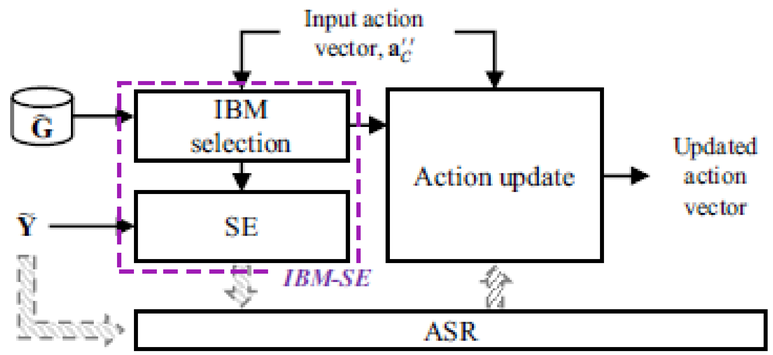


IBM clustering

- The 64-dimensional MPS features were extracted from all utterances. An IBM filter is computed for each feature vector. The IBM clustering module groups the entire set of IBM vectors collected from the training data to A clusters based on the K-means algorithm.



Target action determination module



1. Reward calculation

$$R = \tanh\{\alpha(Z_y - Z_{s'})\}$$

$$\hat{E}_c = (\log(\hat{S}_c) - \log(\hat{S'}_c))^T(\log(\hat{S}_c) - \log(\hat{S'}_c))$$

$$\tilde{E}_c = \frac{\hat{E}_c}{\max_{0\le c\le C-1}\hat{E}_c}$$

$$r_c = \begin{cases}(1-\tilde{E}_c)R, & R>0\\ \tilde{E}_c R, & R\le 0\end{cases}$$

2. Action update

$$[\mathbf{a}_c]_{a_c} = \begin{cases} r_c + \max_{a_c\in A}[\mathbf{a}''_c]_{a_c}, & R>0\\ [\mathbf{a}''_c]_{a_c}, & R=0\end{cases}$$

$$[\mathbf{a}_c]_{a_{\hat{B}_c}} = [\mathbf{a}''_c]_{a_{\hat{B}_c}} - r_c, \qquad R<0.$$

Experiments and Results (1)

- To train the RL–SE system, 460 utterances with signal-to-noise ratio (SNR) level at 5 dB. The overall RL–SE and ASR systems were evaluated using another 30 utterances at 0 and 5 dB SNR levels, where the utterance lengths were around 1 to 6 seconds. we used the baby-cry noise as the background noise.
- we established two RL-based SE models, with two parameters p for the number of frame in an analysis chunk: the systems with p = 1 and p = 2 are termed $RLSE_1$ and $RLSE_2$.

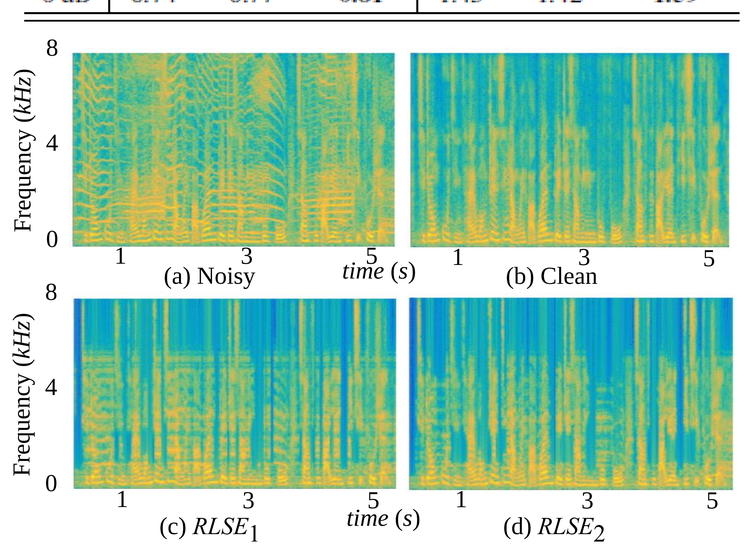
Experiments and Results (2)

Table 1. The average CERs of Noisy (the baseline), $1nnSE$, $DNNSE$, $RLSE_1$, and $RLSE_2$ at 0 and 5 dB SNR conditions.

SNR	*Noisy*	$1nnSE$	$DNNSE$	$RLSE_1$	$RLSE_2$
5 dB	56.14	73.09	60.76	55.60	**49.18**
0 dB	81.40	85.79	73.88	77.20	**65.75**

Table 2. The STOI and PESQ scores of $RLSE_1$, $RLSE_2$, and Noisy at 0 and 5 dB SNR conditions.

SNR	STOI			PESQ		
	Noisy	$RLSE_1$	$RLSE_2$	*Noisy*	$RLSE_1$	$RLSE_2$
5 dB	0.82	0.82	**0.86**	1.85	1.67	**1.96**
0 dB	0.74	0.77	**0.81**	1.45	1.42	**1.59**



Conclusion

- We present an RL-based SE for robust speech recognition without retraining the ASR system in this study. By using the recognition errors as the objective function, the RL-based SE can effectively reduce CERs by 12.40% and 19.23% at 5 and 0 dB SNR conditions, respectively.
- In the future, We will try to implement the whole system with only noisy speech without the paired clean speech.