A NOVEL LSTM-BASED SPEECH PREPROCESSOR FOR SPEAKER DIARIZATION IN REALISTIC MISMATCH CONDITIONS

Lei Sun[1], Jun Du[1], Tian Gao[1], Yu-Ding Lu[2], Yu Tsao[2], Chin-Hui Lee[3], Neville Ryant[4]

[1]University of Science and Technology of China, Hefei, Anhui, China

[2]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan

[3]Georgia Institute of Technology, Atlanta, Georgia, USA

[4]Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA

Background

- The performance of speaker diarization suffers huge degradation in quite challenging realistic environments
- Previous researches have mostly focused on multi-channel speech preprocessing, few on single-channel scenes
- Deep learning techniques become mainstream methods in speech enhancement

What's Important in This Study

- Investigate on the effects of different speech enhancement methods as a preprocessor to speaker diarization
- Propose a novel LSTM-based architecture for speech enhancement
- Explore the generalization capability of the preprocessor in highly mismatched conditions

Baseline Diarization System

Information bottleneck framework:

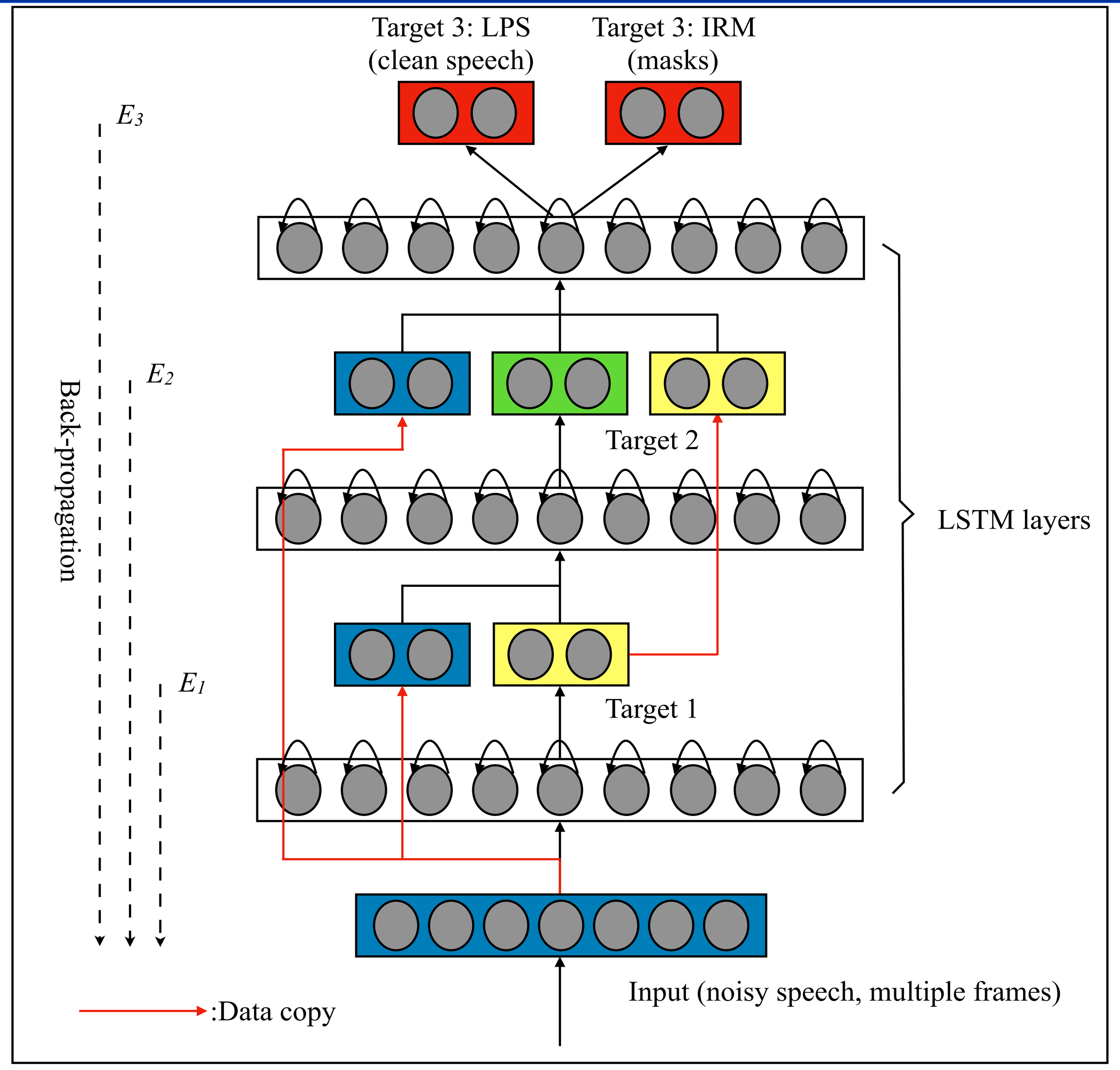
Suppose we have the speech segment $\boldsymbol{X} = \{x_1, x_2, ..., x_m\}$ to be clustered, and the set of relevance variables are $\boldsymbol{Y} = \{y_1, y_2, ...y_n\}$, the desired cluetering outputs are $\boldsymbol{C} = \{c_1, c_2, ...c_p\}$.

The optimization function is:

$$F = I(C, Y) - \frac{1}{\beta} I(C, X)$$

$I(\cdot,\cdot)$ denotes the mutual information between two sets of random variables.

The Novel Architecture For Preprocessor



The corresponding objective function is:

$$E = \sum_{k=1}^{K} \alpha_k E_k + E_{\text{IRM}}$$

$$E_k = \frac{1}{N} \sum_{n=1}^{N} \| \mathcal{F}_k(\hat{\mathbf{x}}_n^0, \hat{\mathbf{x}}_n^1, ..., \hat{\mathbf{x}}_n^{k-1}, \mathbf{\Lambda}_k) - \mathbf{x}_n^k \|_2^2$$

$$E_{\text{IRM}} = \frac{1}{N} \sum_{n=1}^{N} \| \mathcal{F}_{\text{IRM}}(\hat{\mathbf{x}}_n^0, \hat{\mathbf{x}}_n^1, ..., \hat{\mathbf{x}}_n^{K-1}, \mathbf{\Lambda}_{\text{IRM}}) - \mathbf{x}_n^{\text{IRM}} \|_2^2$$

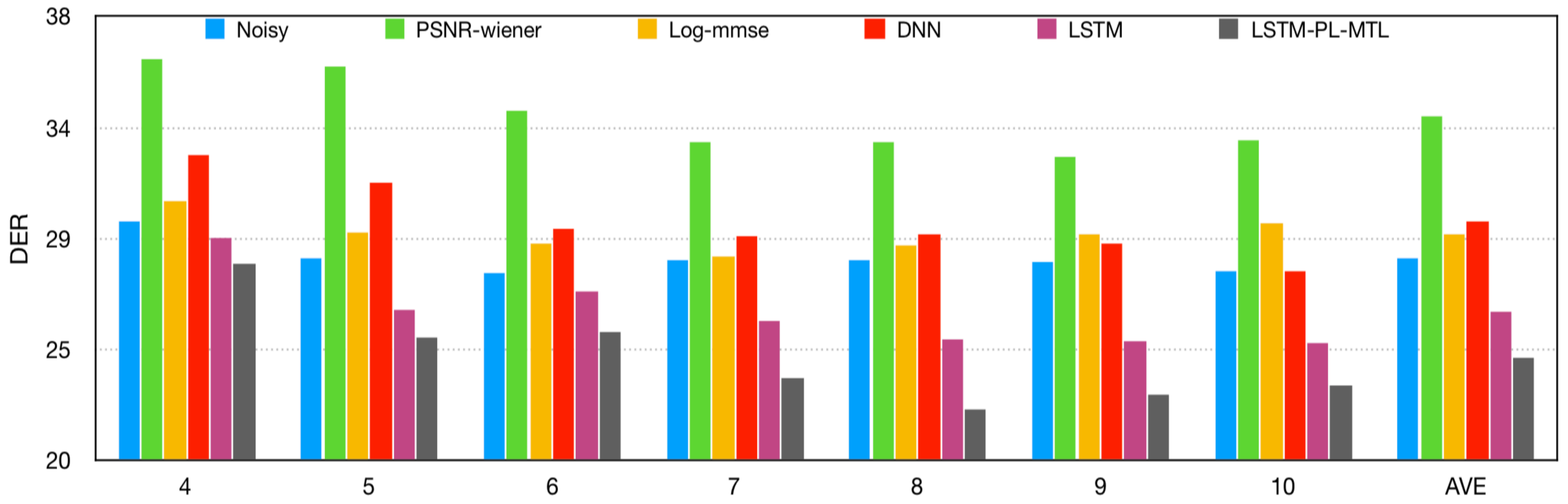
where $\hat{\mathbf{x}}_n^k$ and $\mathbf{x}_n^k$ are the n^{th} *D*-dimensional vectors of estimated and target LPS feature for k^{th} target layer. $\mathcal{F}_k(\hat{\mathbf{x}}_n^0, \hat{\mathbf{x}}_n^1, ..., \hat{\mathbf{x}}_n^{k-1}, \mathbf{\Lambda}_k)$ is the layer function with the dense structure using the learned intermediate targets from $\hat{\mathbf{x}}_n^0$ to $\hat{\mathbf{x}}_n^{k-1}$, and $\mathbf{\Lambda}_k$ represents the parameter set before k^{th} target layer. E_k *and* E_{IRM} are MSE for multi-target learning in the final output layer.

Experiments

High mismatches between training and testing data:

	Training	Testing		
Corpus	WSJ0	ADOS	SeedLings	AMI
Distance	Near	Far	Near	Far
Style	Reading	Conversation		
Interferences	Additive noise	Background noises, reverberations		
Interaction	Simulation	Unknown, real noisy speech		
Child?	None	Kids	6-month baby	None

Proposed architecture performs better than all other previous speech enhancement methods in terms of DER on AMI's SDM data:



For MDM data (after beamforming algorithm) in AMI:

Noisy	DNN	LSTM	LSTM-PL-MTL
25.9	26.4	22.5	21.6

For data which involves child's speech:

	Noisy	Log-mmse	PSNR-wiener	LSTM-PL-MTL
ADOS	36.3	40.0	36.0	29.2
SeedLings	45.3	47.0	46.7	39.2