A Good Title is Worth a Thousand Words

Three Zhang[1,2] Four Li[1] Five Wang[2]

[1]Organization 1
[2]Organization 2 has such a looooooooooooong name that it cannot properly fit in a single line

Abstract

Don't cite references in abstract. Usually we don't write multiple paragraphs in abstract. Mind the word count limit in the paper template (can differ across conferences/journals)! **Basic structure:** (1) 1 sentence describing the background; (2) 1 "however" sentence stating current problems; (3) 2–3 sentences introducing your goal and contributions; (4) 1–2 sentences highlighting the experimental results or key findings.

1 Introduction

Define the **full names** of abbreviations (initials) when they appear for the first time! Even if they have already been defined in *Abstract*, they should be defined again in the main body (starting from *Introduction*). For example, convolutional neural networks (CNNs) and Transformer [1] can be combined to form new architectures.

2 Related Works

In `AISHELL-3` [2], it is reported that ...

3 Proposed Method

We propose a new model called HahahahaNet. It takes as input encoded features extracted from the DNSMOS P.835 [3] model[1].

[1]`https://github.com/microsoft/DNS-Challenge/blob/master/DNSMOS/DNSMOS/sig_bak_ovr.onnx`

3.1 Problem formulation

Let $\mathbf{A} \in \mathbb{C}A^{F \times F}$ be the xxx matrix, where F is the frequency dimension. We have ...

$$\mathbf{X} = \mathbf{\Phi}^{\mathsf{H}}\mathbf{U}^{\mathsf{T}}\mathbf{A}^{-1} \quad \in \mathbb{C}^{T \times F}, \tag{1}$$

$$M_f = \frac{1}{T}\left(\sum_{t=0}^{T-1} w_t |X|_{t,f}^2\right)^{0.5}, \tag{2}$$

where M_f represents the estimated mask of the f-th frequency bin.

Jepsen *et al.* [4] found that training speech enhancement models with noisy signal labels based on scale-invariant signal-to-noise ratio (SI-SNR) [5] can lead to inferior performance.

3.2 First method

CHiME-6 [6]

3.3 Second method

We use the mask M_f to refine the final prediction.

4 Experiments

We evaluate the model performance using the SI-SNR metric.

Figure 1: Overview of the proposed model.

Table 1: Tag occurrences in the dataset.

Tag		Note	Fullset	Subset
Data Domain	real_recording	recorded directly		0
	enhanced	(suspiciously) processed by SE systems	0	
	synthetic	(suspiciously) synthetic speech	0	
Speech Type	non_speech	w/o any intelligible speech[2]		
	read_speech			

4.1 Experimental setup

4.2 Results and analysis

5 Conclusion

Acknowledgment

This work was supported by xxx.

References

[1] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[2] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A multi-speaker Mandarin TTS corpus," in *Proc. ISCA Interspeech*, 2021, pp. 2756–2760.

[3] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP*, 2022, pp. 886–890.

[4] S. D. Jepsen, M. G. Christensen, and J. R. Jensen, "A study of the scale invariant signal to distortion ratio in speech separation with noisy references," in *Accepted by ASRU*, 2025.

[5] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *ICASSP*, 2019, pp. 626–630.

[6] S. Watanabe *et al.*, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, 2020, pp. 1–7.

[2] This means ...