Text-Annotated Noisy Speech as Supervision
A Dual-Learning Framework for Target-Domain Clean-Free Speech Enhancement

Xin Liu | Shulin He | Xueliang Zhang
Inner Mongolia University, China | Southern University of Science and Technology, China

Fine-tune a speech enhancement model in a real target domain with noisy speech plus transcripts, without paired target clean speech.

Core Contributions

01 Clean-free target adaptation

Real CHiME-3 noisy speech with transcripts supplies target-domain supervision when aligned clean speech is unavailable.

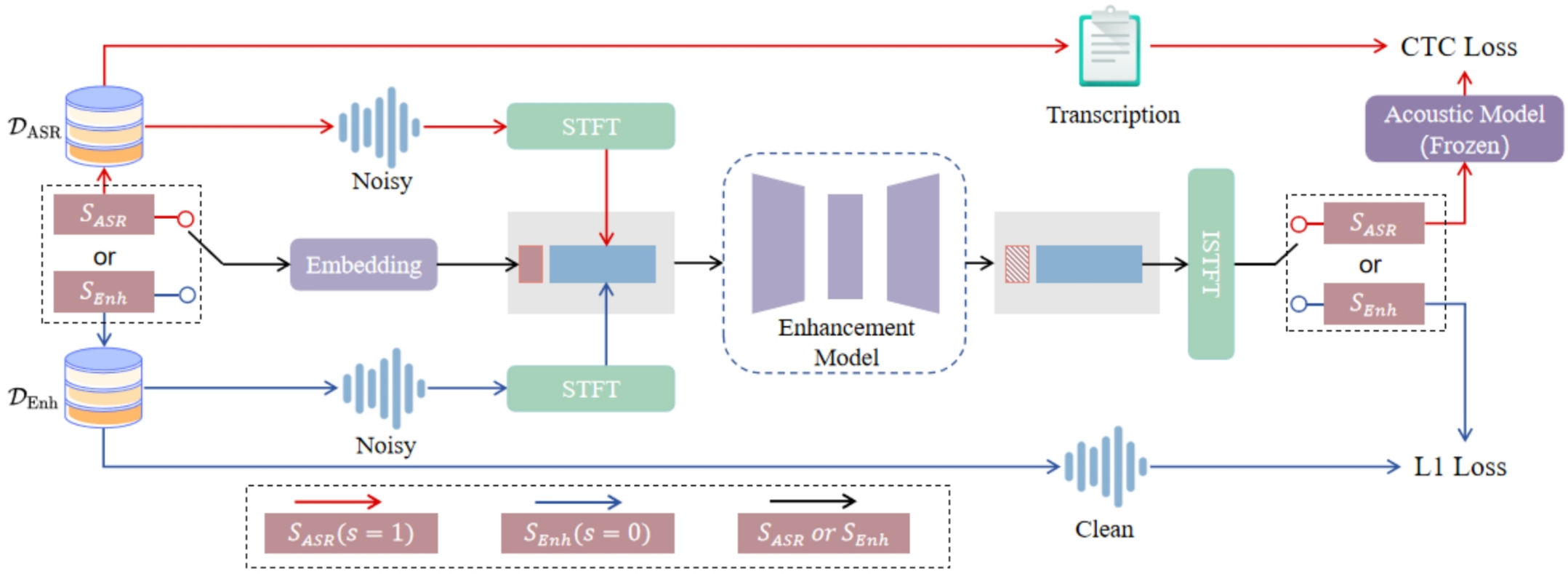
02 Switch-conditioned dual learning

A learned embedding controls two modes: enhancement trains on noisy-clean pairs with L1 loss, while ASR trains on noisy-transcript pairs with CTC loss.

03 Controllable operating modes

The same model exposes quality-oriented S_{Enh} and ASR-oriented S_{ASR} inference modes instead of fixing a single quality-recognition trade-off.

SwitchSE Fine-Tuning Framework



A switch embedding prepends a mode token to the noisy spectrogram. Enhancement batches use L1 reconstruction; noisy-transcript batches use CTC through a frozen U2++ acoustic model.

Experimental Setting **16 kHz. VCTK clean-noisy pairs plus CHiME-3 noisy-transcript pairs. CHiME-3 target training: 1,600 utterances (2.9 h).**

Cross-Dataset Results and Mode Trade-Off

Method	Switch	Train Enh	Train ASR	VCTK				CHiME	
				WER	P.808 DNSMOS	PESQ	STOI	WER	P.808 DNSMOS
Noisy	-	-	-	37.51	2.33	1.19	0.58	18.38	2.34
GCRN	-	VCTK	-	36.93	3.09	1.68	0.66	63.66	2.77
GCRN CTC	-	-	CHiME	35.33	2.33	1.25	0.61	**18.35**	2.36
SwitchSE	S_{Enh}	**VCTK**	**CHiME**	35.53	**3.14**	**1.70**	**0.66**	23.63	**3.23**
	S_{ASR}	**VCTK**	**CHiME**	**35.24**	3.13	1.68	**0.66**	20.27	3.19

VCTK evaluates source-domain preservation; CHiME-3 evaluates target-domain adaptation. Higher is better except WER.

- Perceptual speech quality first. On CHiME-3, S_{ASR} reaches the best perceptual score (P.808 = 3.23) using noisy speech and transcripts without paired target clean speech.
- Controlled ASR trade-off. Switching to S_{Enh} reduces CHiME WER from 23.63 to 20.27 while preserving P.808 at 3.19.

Switch Ablation

Method	Switch	VCTK WER	VCTK PESQ	CHiME WER	CHiME P.808 DNSMOS
SwitchSE	S_{Enh}	35.53	**1.70**	23.63	**3.23**
	S_{ASR}	**35.24**	1.68	20.27	3.19
Switch SE-4	S_{Enh}	36.48	1.68	27.53	3.17
	S_{ASR}	35.78	1.42	**18.09**	2.69
w/o z_s	-	35.93	**1.70**	22.44	3.20

A meaningful operating boundary. A longer switch further lowers WER to 18.09 (lower than GCRN CTC), but reduces P.808 to 2.69; the default system therefore prioritizes perceptual quality.

DEMO WEBSITE

SwitchSE turns real noisy speech plus text into target-domain supervision and lets one enhancement model trade perceptual quality for ASR-friendliness at inference.