Empowering Whisper as a Joint Multi-Talker and Target-Talker Speech Recognition System

Lingwei Meng, Jiawen Kang, Yuejiao Wang, Zengrui Jin, Xixin Wu, Xunying Liu, Helen Men

The Chinese University of Hong Kong, Hong Kong SAR, China

2025.11.02
HyorinJung

Index

Background

Prior adaptation methods: Multi-talker speech recognition.

Cascade systems

- 1)Separate → 2)recognize
- suffer from objective mismatch and limited joint optimization

End-to-end (E2E)

- ASR methods (PIT, HEAT, SOT) show promise
- but usually need training from scratch or full fine-tuning

Background

Prior adaptation methods: Target-Talker ASR.

Goal : Efficiently recognize one target speaker's speech in multi-talker conditions

- End-to-end approaches have shown remarkable progress

Limitation

- However, they generally require:
  - External modules (e.g., x-vector, ECAPA-TDNN)
  - Internal embedding extractors
- Usually, only the target speaker's transcript is produced, while non-target speech is ignored.
- Speaker-attributed ASR can handle multiple speakers, but requires embeddings for all speakers
- Only one prior work addressed joint multi- and target-talker ASR, but still depends on an external extractor

Introduction

Main Contributions:

- Novel joint ASR framework — transcribes **multi-talker speech** while highlighting the **target speaker** without any speaker embedding extractor.
- Parameter-efficient & loosely coupled — only a small number of trainable parameters.
- State-of-the-art performance on LibriMix / LibriSpeechMix (English) and strong zero-shot generalization on AishellMix (Mandarin)

Introduction

Method

- **Whisper (Foundation Model)** : Serves as the base encoder-decoder model, providing robust pre-trained speech representations
- **Sidecar Separator** : Separates mixed embeddings from multiple speakers within Whisper’s encoder layers
- **Target Talker Identifier (TTI)** : Identifies the target speaker’s embedding flow dynamically using a 3-second enrollment cue
- **Soft Prompt Embedding** : Adapts the Whisper decoder to multi-/target-talker ASR tasks through lightweight soft prompt tuning

Background

Whisper as the Speech Foundation Model

- Whisper is an attention-based encoder–decoder speech recognition model.
- Trained on massive web-scale labeled speech data
- Widely used beyond ASR as a speech foundation model
- Inspired by its versatility, this study extends Whisper to handle joint multi-talker and target-talker ASR tasks efficiently

Background

Empowering Whisper as a Multi-Talker ASR System

- SS + Whisper >> converts a well-trained single-talker ASR model into a multi-talker ASR model
- The SS is a temporal convolutional network inserted between the early encoder layers
- It consists of stacked 1-D dilated convolutional blocks inspired by Conv-TasNet.
- Since the shallow encoder layers of ASR capture more acoustic than linguistic information, the SS separates mixed embeddings into speaker-specific representations by generating talker-dependent masks.

Background

Placement in Whisper:

- The Sidecar Separator, along with two additional 1-D Conv layers, is placed after the 2nd encoder block.
- It generates masks corresponding to each speaker,
  which are element-wise multiplied with the mixed embeddings to obtain separated speaker embeddings.
- The remaining encoder blocks and decoder then process each branch to produce the final transcriptions for each speaker.

Background

Log-Mel Spectrogram

$$X \in \mathbb{R}^{D \times T}$$

CNN Down-sampling

Encoder

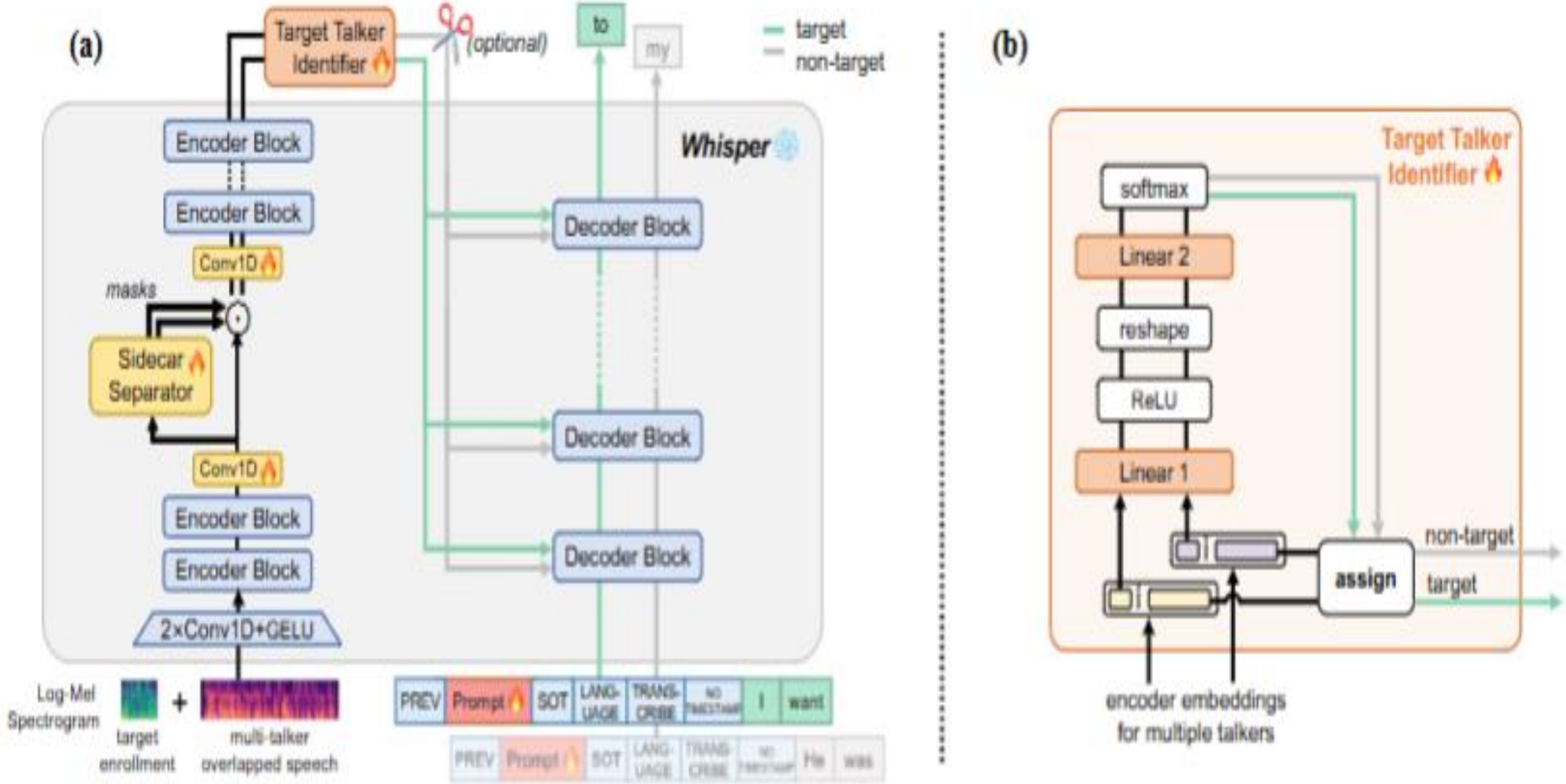
$$H_e = \text{Encoder}(\text{Conv}(X))$$

Decoder

$$\hat{y}_m = \text{Decoder}(s, \hat{y}_{1:m-1}, H_e)$$

- s : special token sequence
  ex<|PREV|>, decoder prompt,
<|SOT|>, <|LANGUAGE|>,
<|TRANSCRIBE|>, <|NO-TIMESTAMP|>

Text


(a)
Target Talker Identifier
(optional)
to
my
target
non-target
Whisper
Encoder Block
Encoder Block
Conv1D
masks
Sidecar Separator
Conv1D
Encoder Block
Encoder Block
2×Conv1D+GELU
Decoder Block
Decoder Block
Decoder Block
Log-Mel Spectrogram
target enrollment
multi-talker overlapped speech
PREV
Prompt
SOT
LANG-UAGE
TRANS-CRIBE
I
want
He
was
(b)
Target Talker Identifier
softmax
Linear 2
reshape
ReLU
Linear 1
assign
non-target
target
encoder embeddings for multiple talkers

Background

Target Talker Identifier (TTI)

- Introduces target-talker ASR capability into the system
- During the forward process :
  - Encoder outputs (from multiple speakers) are divided into:
    - Prefix segment — corresponds to the *3-second enrollment speech*
    - Main segment — corresponds to the *multi-talker speech*
- The prefix segment is sent to the TTI module,
  which determines which branch belongs to the target talker.
- Only the main segment of the target branch is passed to the Whisper decoder for transcription.

Background

- B: 배치 크기 (Batch Size)
- S: 화자 수 (Number of Talkers)
- 150: 시간 프레임 수 (Time Frames), 각 프레임은 20ms이므로 150프레임은 3초에 해당합니다.
- C: 채널 수 (Number of Channels)

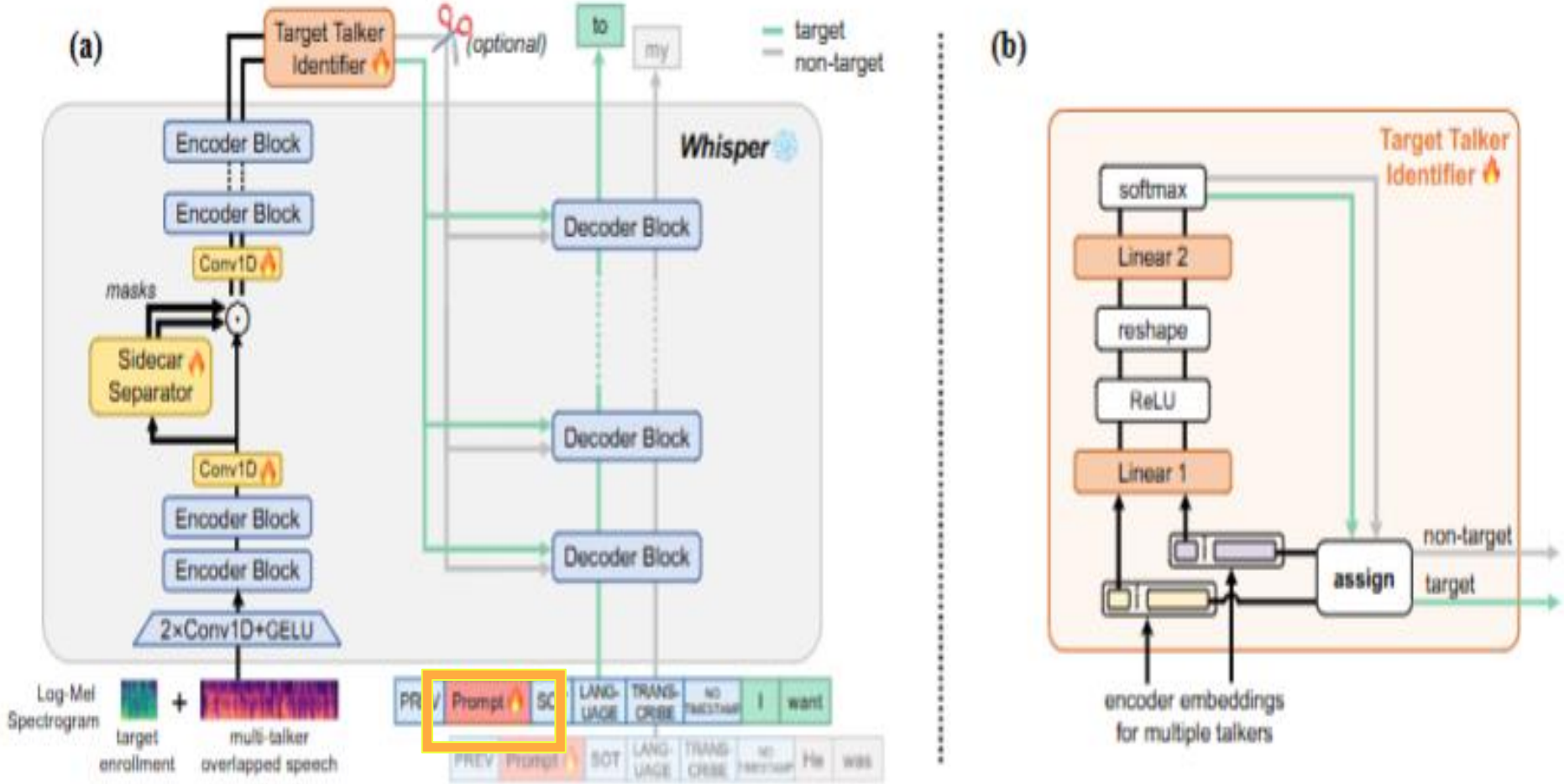
How TTI Works

Step	Operation	Output shape
1	Prefix segment input	(B × S, 150, C)
2(reshape)	Linear layer + ReLU	(B × S, 150, 1)
3(softmax)	Squeeze(크기1차원 제거) + reshape → Linear + Softmax	Probability (B, S) of each branch being the target speaker
	Select target branch (highest probability)	Minimal computational overhead

Background

Soft Prompt Tuning

- Whisper model text prompt tokens >> prefixes in the decoder input sequence to specify tasks and conditions
- Combine this inherent property of Whisper with
  Soft Prompt Tuning >> To adapt the model efficiently
  for multi-talker and target-talker ASR


(a)
Target Talker Identifier
(optional)
to
my
target
non-target
Whisper
Encoder Block
Encoder Block
Conv1D
masks
Sidecar Separator
Conv1D
Encoder Block
Encoder Block
2×Conv1D+GELU
Decoder Block
Decoder Block
Decoder Block
Log-Mel Spectrogram
target enrollment
multi-talker overlapped speech
Prompt
LANG-UAGE
TRANS-CRIBE
I
want
PREV
Prompt
SOT
LANG-UAGE
TRANS-CRIBE
He
was
(b)
Target Talker Identifier
softmax
Linear 2
reshape
ReLU
Linear 1
assign
non-target
target
encoder embeddings for multiple talkers

Background

How It Works

- A learnable embedding (soft prompt) is inserted between <|PREV|> and <|SOT|> tokens — where hard prompts are usually placed
- The position of the soft prompt is masked when calculating loss, since the model doesn't need to generate it
- The soft prompt embedding is updated as the model learns to transcribe multi-talker speech

Experiments : Setting

Datasets

- **English datasets:**
  a. **LibriMix** : 2- and 3-speaker fully overlapped mixtures (Libri2Mix, Libri3Mix)
  b. **LibriSpeechMix** : partially overlapped mixtures (random delay)

- **Mandarin dataset:**
  a. **Aishell1Mix** : generated from Aishell-1 corpus
- Each enrollment speech = **3-second clip** randomly sampled from LibriSpeech.

- Audio longer than **30 s** is time-stretched to fit Whisper’s input limit.

Experiments : Setting

Implementation Details

- At each training step:
  - 80% probability: multi-talker ASR training
  - 20% probability: joint multi-talker + target-talker ASR training
- Two losses are optimized:
  - ASR loss
  - TTI cross-entropy loss
- Both require speaker-order permutation assignment (to resolve label ambiguity)

Experiments : Setting

Model Settings and Evaluation Metrics

- Base models: Whisper-small, Whisper-medium,Whisper-large-v3
- Frozen weights; only train: Sidecar Separator, TTI module, Soft Prompt embeddings
- Optimizer: AdamW (lr: 2e-4 → 1e-4)
- Training: up to 200k steps on 8×NVIDIA V100 (batch=16)
- λ = 0.01 for TTI loss weight.
- Evaluation metrics:
  - Multi-talker ASR → WER/CER (permuta
  - Target-talker ASR → standard WER

Foundation Model	2-speaker	3-speaker
Whisper-small	8.69 M (3.47%)	8.79 M (3.51%)
Whisper-medium	13.16 M (1.69%)	13.29 M (1.71%)
Whisper-large	18.41 M (1.18%)	18.58 M (1.19%)

Experiments : Setting

Implementation Details

- The assignment is determined by Permutation Invariant Training (PIT): matching

$$\hat{\pi} = \arg\min_{\pi \in \mathcal{P}} \sum_{s=1}^{S} \text{Loss}_{\text{ASR}}(Y^s, R^{\pi(s)}) \Rightarrow L_{\text{ASR}} = \sum_{s-1}^{S} \text{Loss}_{\text{ASR}}(Y^s, R^{\hat{\pi}(s)})$$

- Final objective: $L = L_{ASR} + \lambda L_{TTI}$
- λ = 0.01
- No CTC loss is used, since Whisper was not trained with CTC.

Ablation Study

Ablation:

- Tested prompt lengths {0, 2, 4, 8, 16}
- Best performance achieved at **soft prompt length = 4.**
- Too long → harder optimization (frozen model → longer sequences hurt)

Table 5: *Ablation study on soft prompt, evaluated by WER (%).*

	Soft Prompt Length				
System	0	2	4	8	16
Whisper-medium-SS-TTI	7.21	6.82	**6.56**	6.84	7.5
Whisper-large-SS-TTI	5.27	4.98	**4.66**	4.74	5.43

Table 2: *Multi-talker ASR on the test sets of LibriMix and LibriSpeechMix. Evaluated by WER (%). "SS" denotes "Sidecar Separator", "TTI" denotes "Target Talker Identifier".*

System	LibriMix		LibriSpeechMix	
	2spk	3spk	2spk	3spk
(a) WavLM Base+ PIT [9]	18.45	-	-	-
(b) C-HuBERT-Large [35]	7.80	-	-	-
(c) SURT [11]	-	-	7.20	-
(d) SOT-Conformer [27]	-	-	4.90†	**6.20†**
(e) D2V-Sidecar-DB [21]	9.69	33.91	7.49	11.94
(f) Whisper-small-SS	10.04	29.20	5.27	9.85
(g) Whisper-small-SS-TTI	9.39	26.76	5.18	8.61
(h) Whisper-medium-SS	6.95	22.58	4.32	7.80
(i) Whisper-medium-SS-TTI	6.56	21.47	4.01	7.50
(j) Whisper-large-SS	4.98	17.55	3.81	7.13
(k) Whisper-large-SS-TTI	**4.66**	**16.79**	**3.43**	6.80

† with extremely heavier training efforts.

Table 3: *Target-talker ASR on LibriMix and LibriSpeechMix. Evaluated by WER (%). "-limited" denotes using the same training data as in [24].*

System	LibriMix		LibriSpeechMix	
	2spk	3spk	2spk	3spk
WavLM-Base$^+$-TSE [9]	12.32	-	-	-
Whisper-TS-ASR [24]	11.98	-	-	-
Whisper-small-SS-TTI-limited	15.75	-	-	-
Whisper-medium-SS-TTI-limited	11.39	-	-	-
Whisper-large-SS-TTI-limited	10.79	-	-	-
Whisper-small-SS-TTI	11.81	30.52	8.89	15.85
Whisper-medium-SS-TTI	9.14	25.75	7.58	12.4
Whisper-large-SS-TTI	**7.97**	**21.97**	**6.99**	**11.4**

Table 4: *Zero-shot and one-batch-tuning multi-talker ASR on Aishell1Mix Mandarin dataset. Evaluated by CER (%).*

System	zero-shot	one-batch-tuning
Whisper-small-SS-TTI	55.87	28.95
Whisper-medium-SS-TTI	36.28	19.83
Whisper-large-SS-TTI	**28.94**	**17.81**

Experiments : Result

- Evaluated whether multilingual capability is retained after fine-tuning on English.
- Tested on AishellMix (Mandarin) – first use for multi-talker ASR.
- Two evaluation modes:
  - Zero-shot: directly tested without retraining
  - One-batch tuning: fine-tuned for one epoch on AishellMix
- Medium & large models show acceptable CER in zero-shot; further improved with minimal tuning.
- Confirms that Whisper's multilingual ability is preserved

Limitations

Limitations:

- Current method relies on **PIT**, requiring pre-defined max speaker count.
- Future work: explore **SOT / HEAT** to handle variable speakers.
- Large target-speaker delay can reduce accuracy.
- Plan to enhance **TTI** with longer contextual awareness (beyond 3 seconds).

Conclusion

- We proposed a novel methodology that harnesses Whisper,
  a speech foundation model, to jointly transcribe multi-talker speech
  while highlighting the target talker's speech,
  without any external speaker embedding extractor.
- Key components:
  1. **Frozen Whisper backbone**
  2. **Sidecar Separator** for embedding separation
  3. **Target Talker Identifier (TTI)** for on-the-fly target detection
  4. **Soft Prompt Tuning** for efficient task adaptation
- Extensive experiments show that our approach:
- Outperforms previous methods on LibriMix and LibriSpeechMix
- Achieves **strong zero-shot performance** on **AishellMix (Mandarin)**