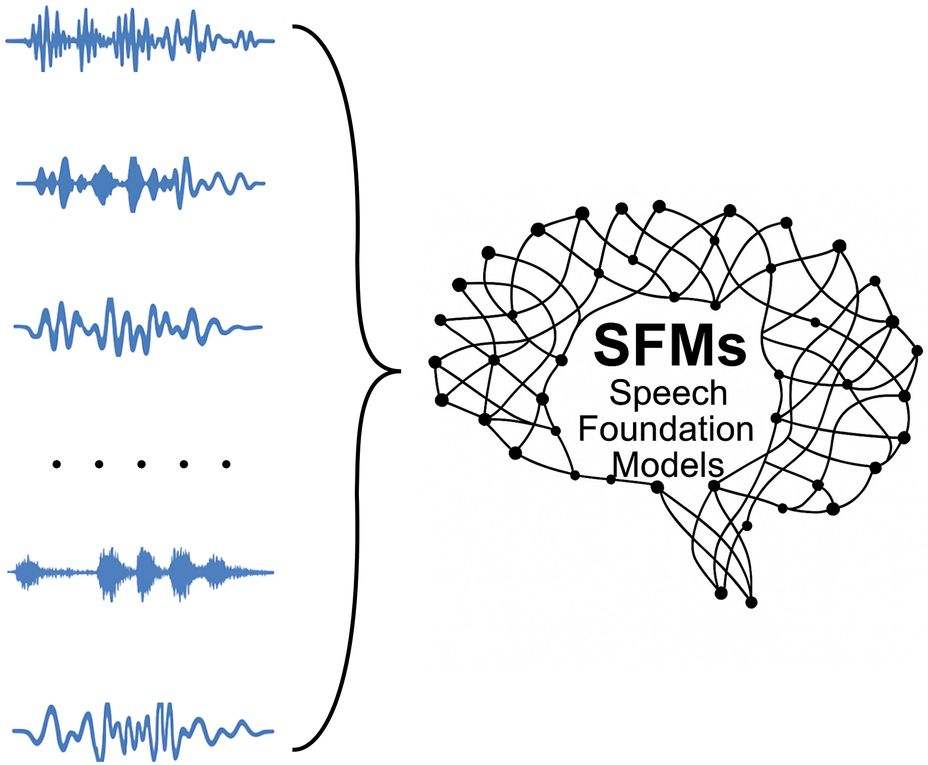
E-BATS: Efficient Backpropagation-Free Test-Time Adaptation for Speech Foundation Models

Jiaheng Dong, Hong Jia, Soumyamjit Chatterjee, Abhirup Ghosh, James Bailey, Ting Dang

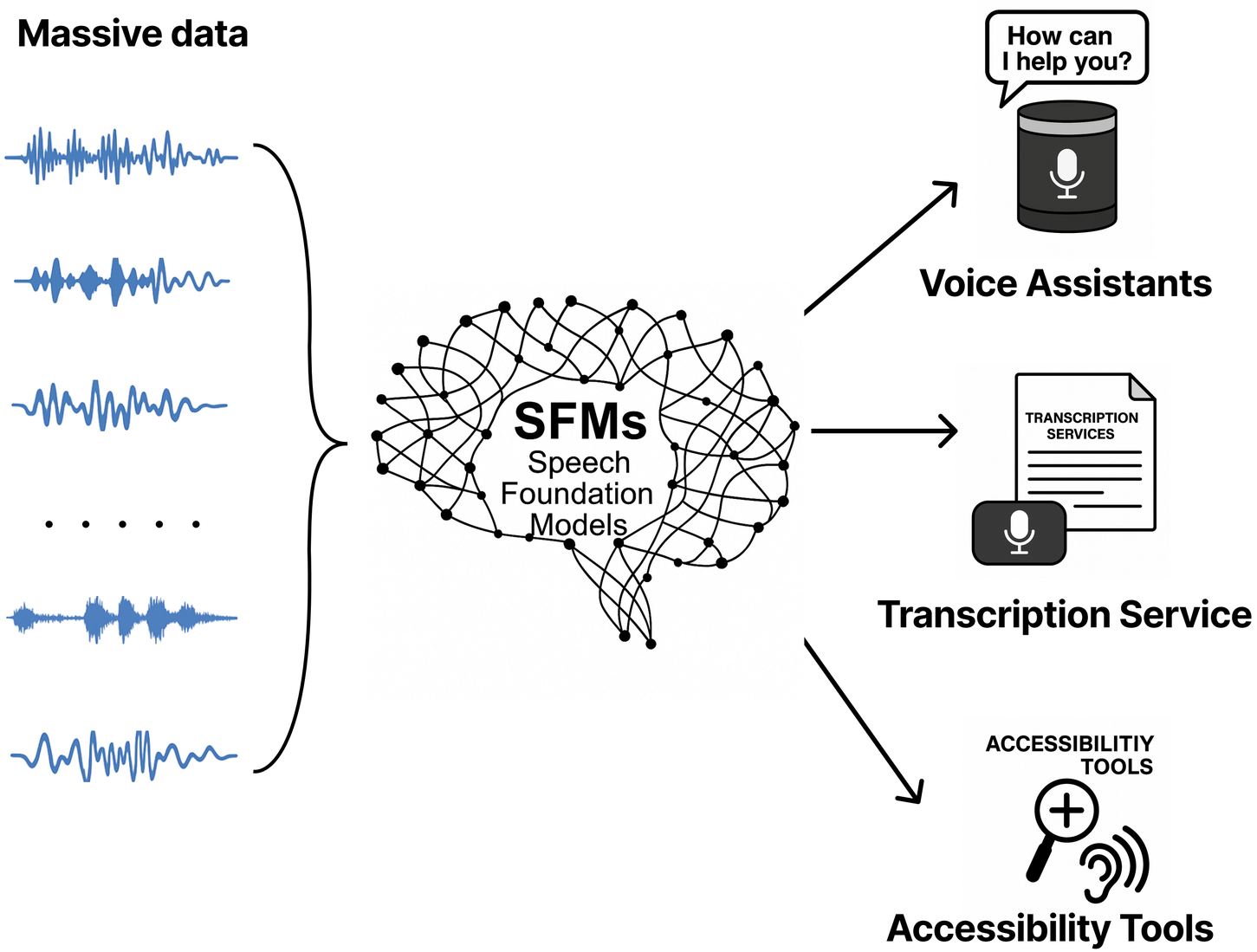
NeurIPS 2025

SFMs in the Wild

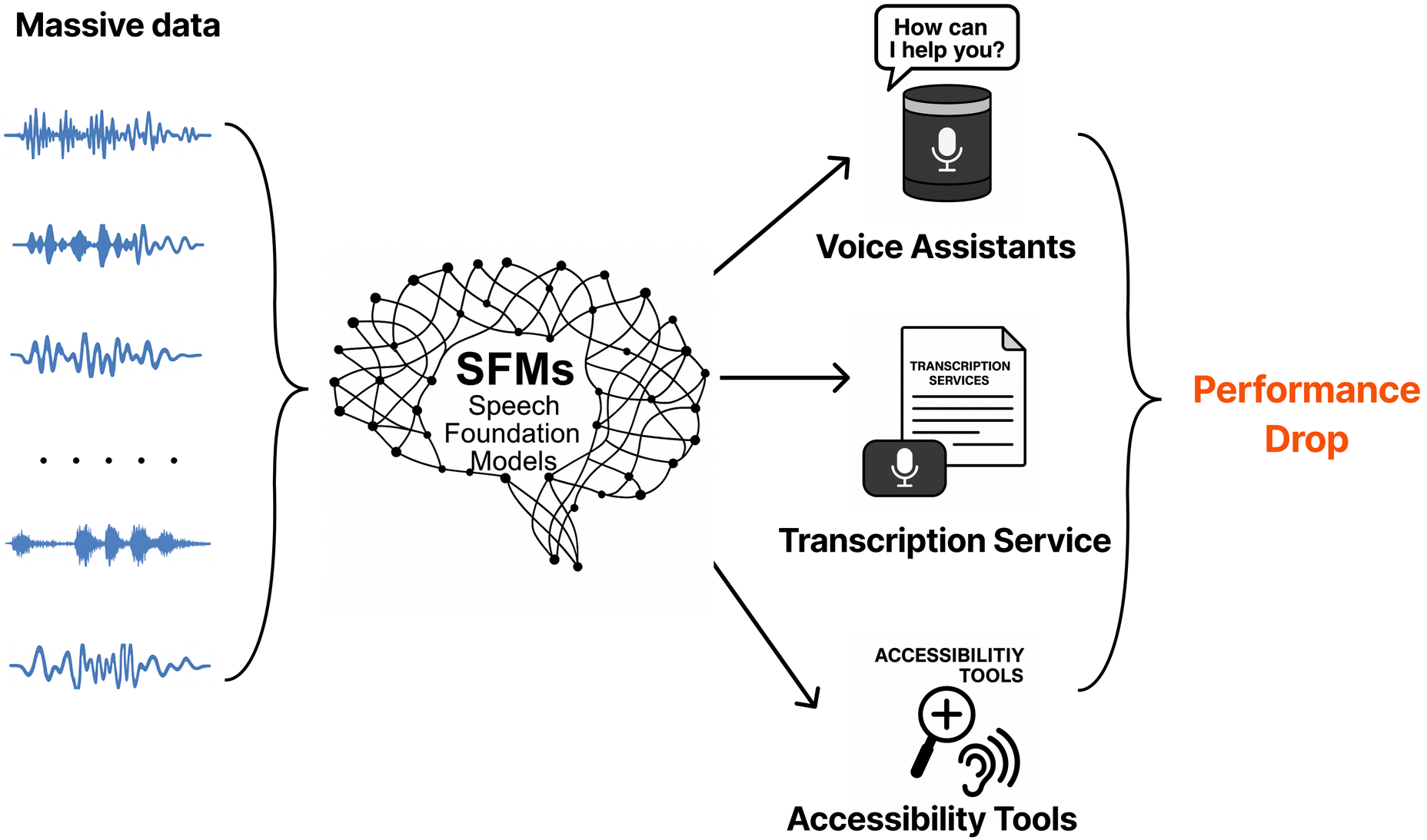
Massive data



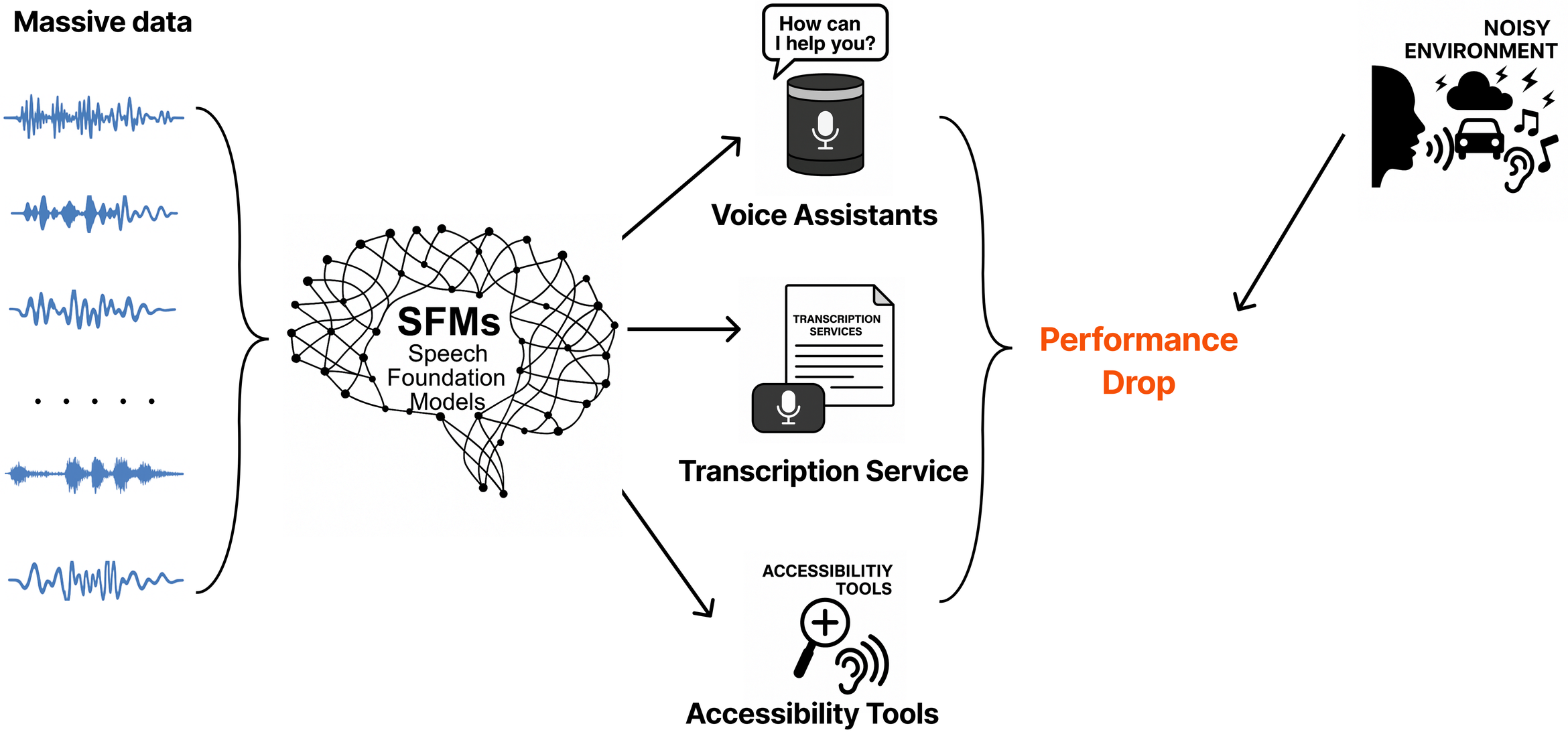
SFMs in the Wild



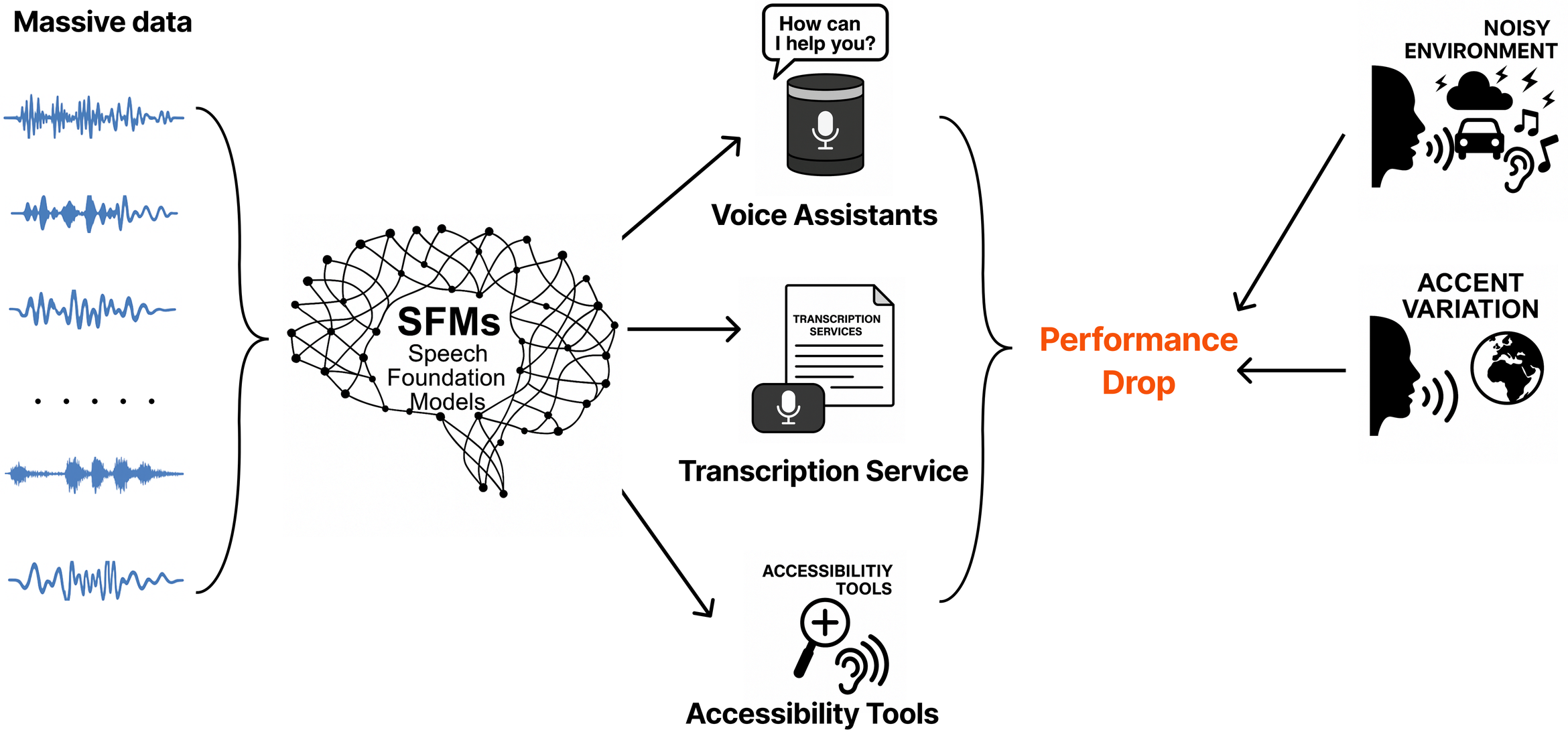
SFMs in the Wild



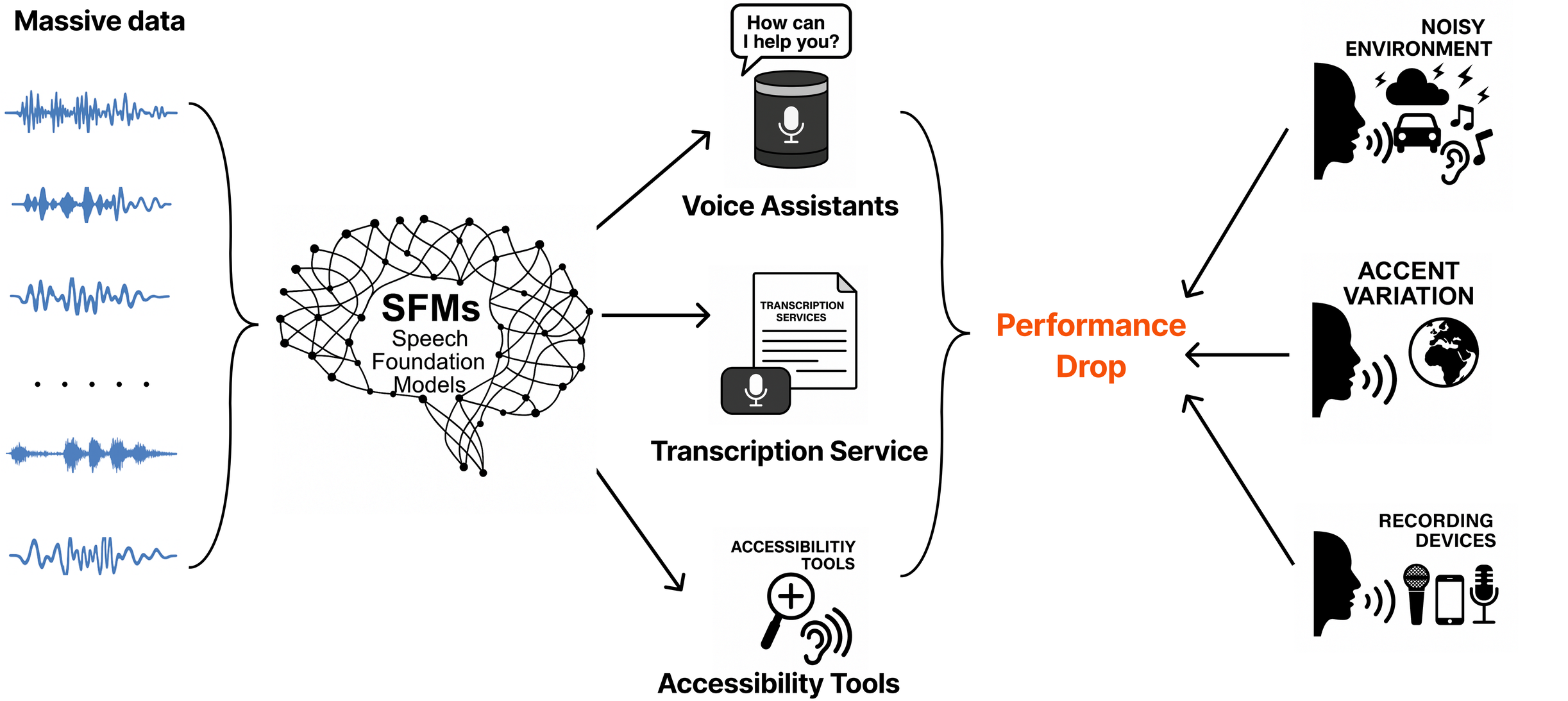
SFMs in the Wild



SFMs in the Wild



SFMs in the Wild

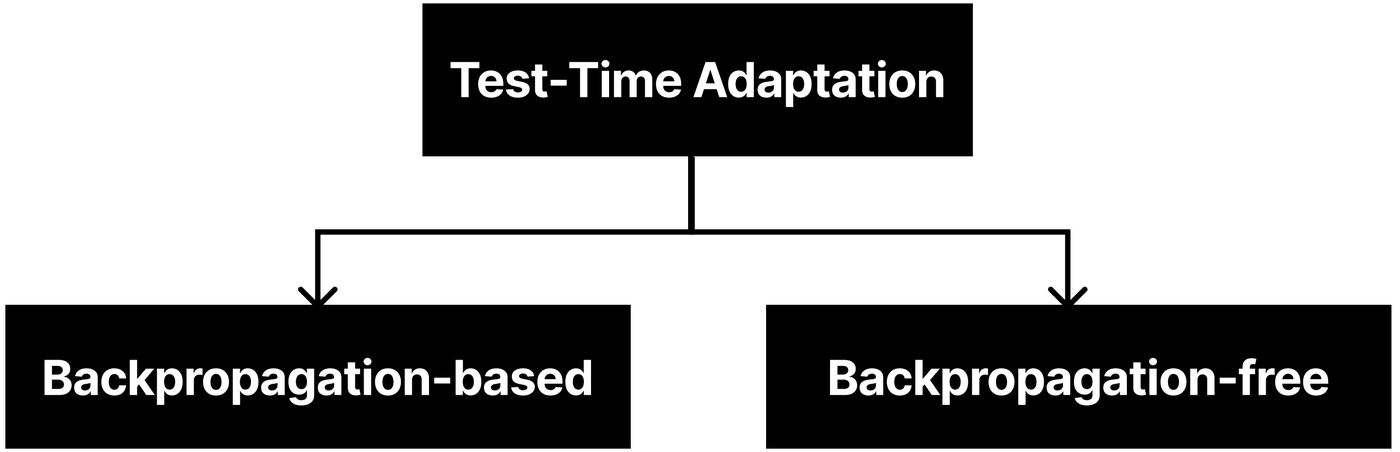


Test-Time Adaptation

- SFMs need to be updated after deployment
- Only unlabelled real world data are available
- Test-Time Adaptation (TTA) is an attractive solution

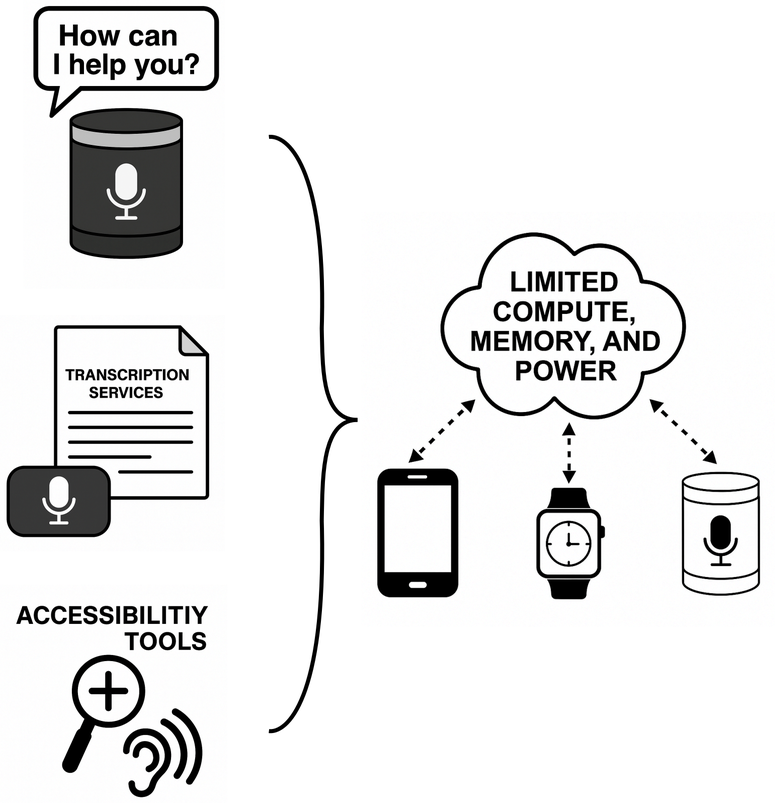
Test-Time Adaptation

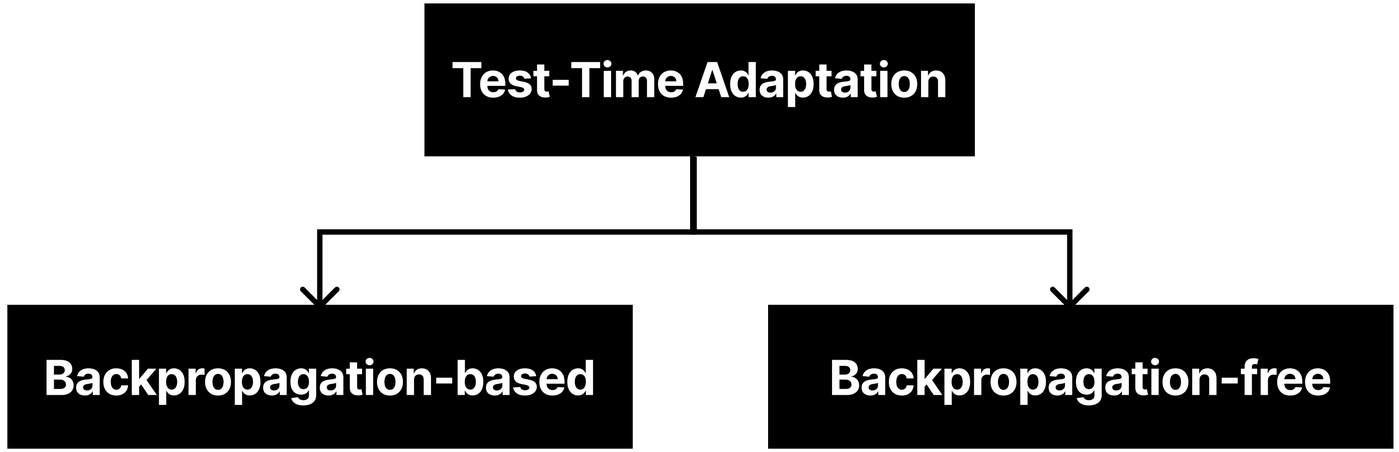
- SFMs need to be updated after deployment
- Only unlabelled real world data are available
- Test-Time Adaptation (TTA) is an attractive solution



Test-Time Adaptation

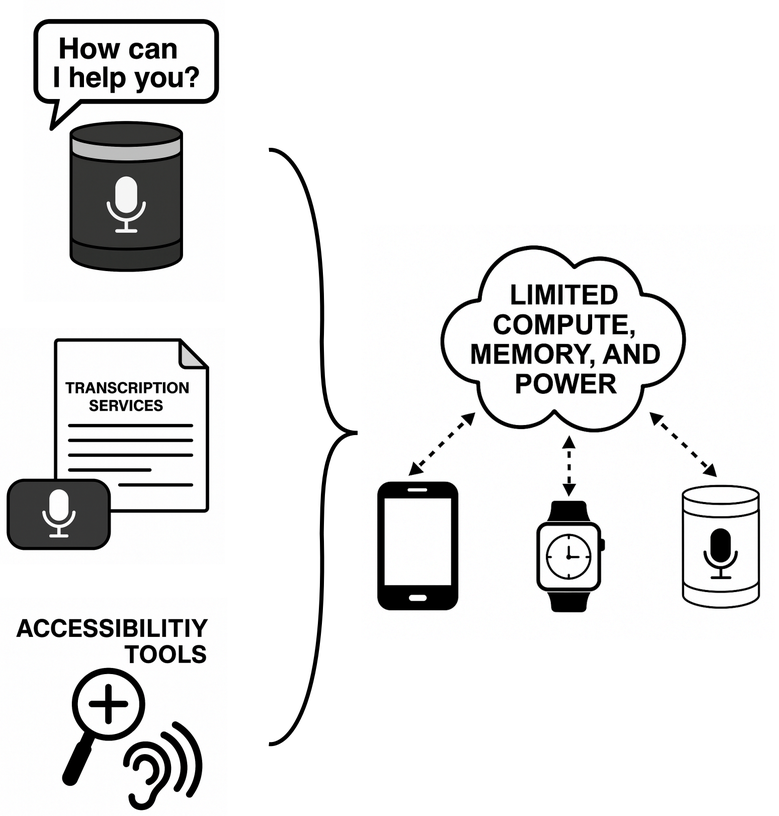
- SFMs need to be updated after deployment
- Only unlabelled real world data are available
- Test-Time Adaptation (TTA) is an attractive solution

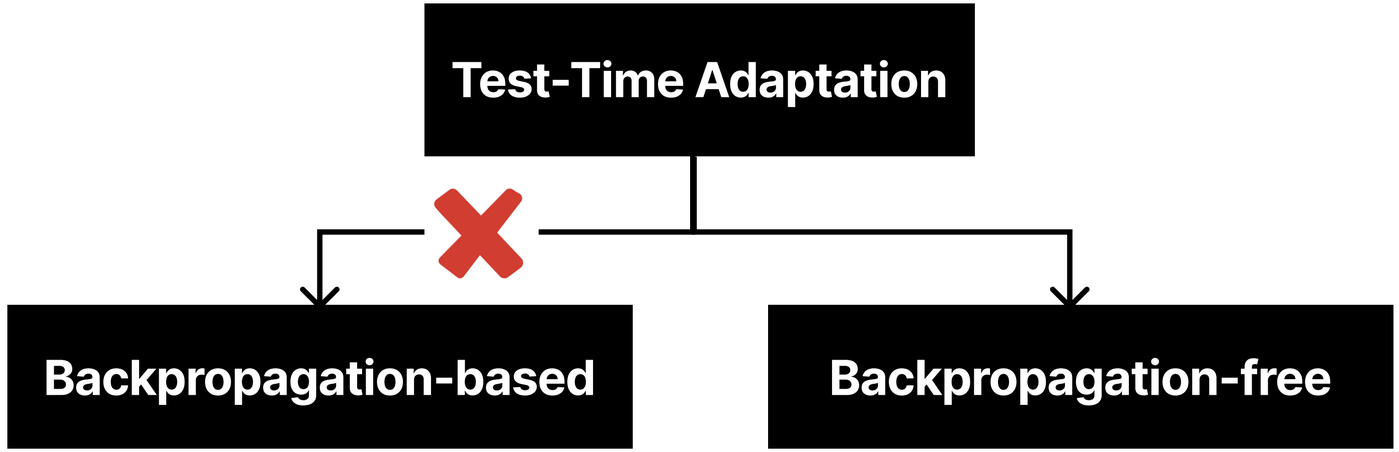




Test-Time Adaptation

- SFMs need to be updated after deployment
- Only unlabelled real world data are available
- Test-Time Adaptation (TTA) is an attractive solution

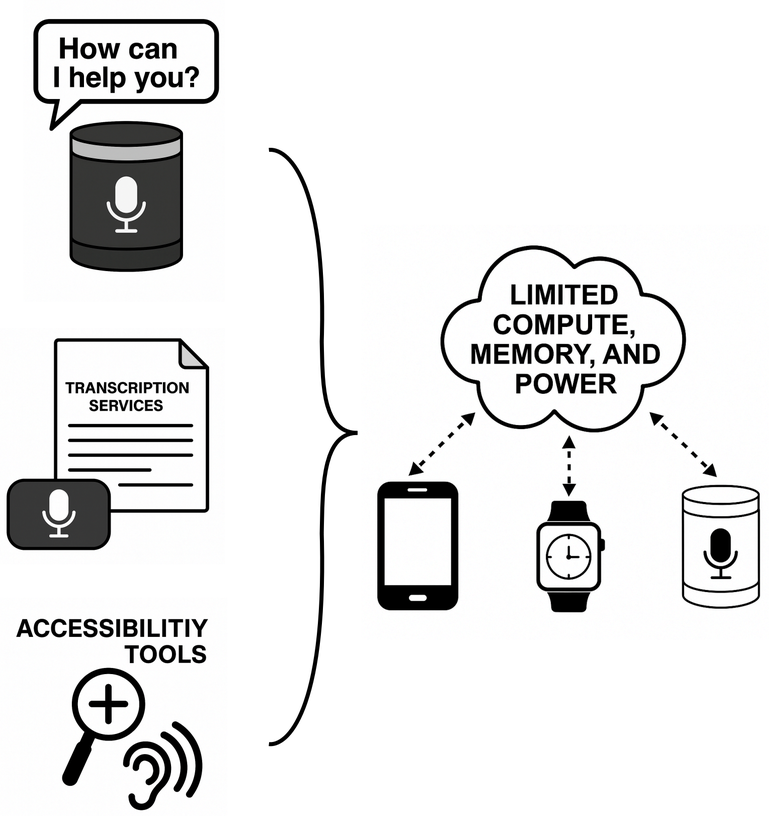


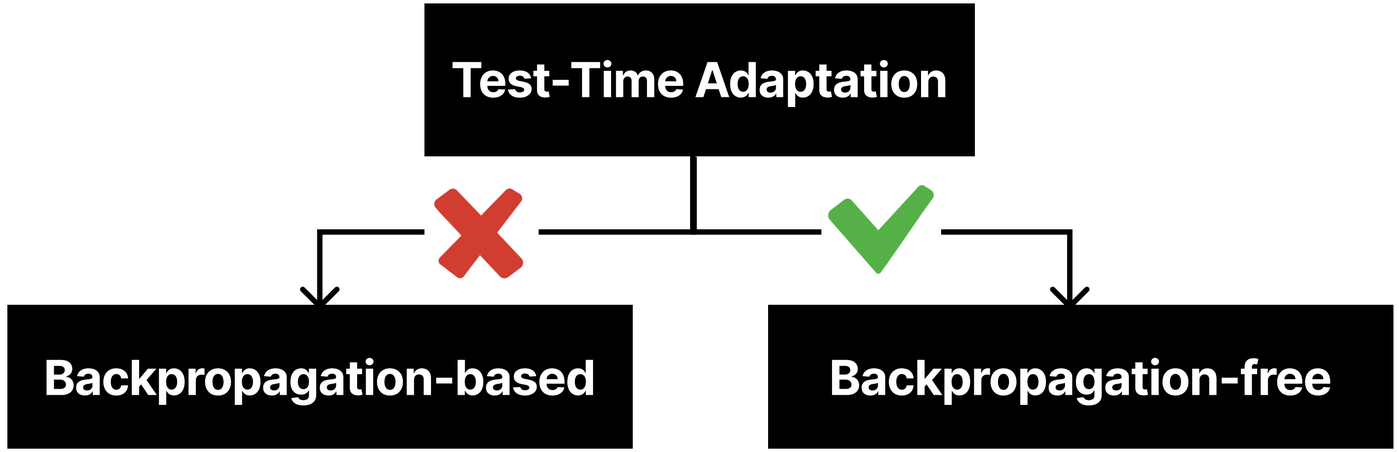


- Large memory overhead
- Not practical for SFMs after deployment

Test-Time Adaptation

- SFMs need to be updated after deployment
- Only unlabelled real world data are available
- Test-Time Adaptation (TTA) is an attractive solution



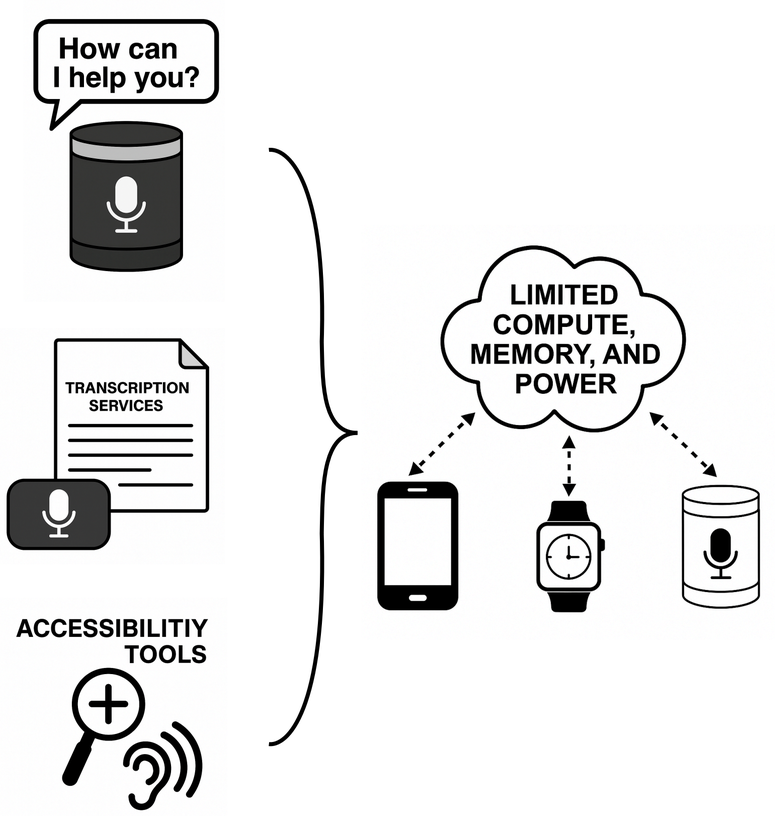


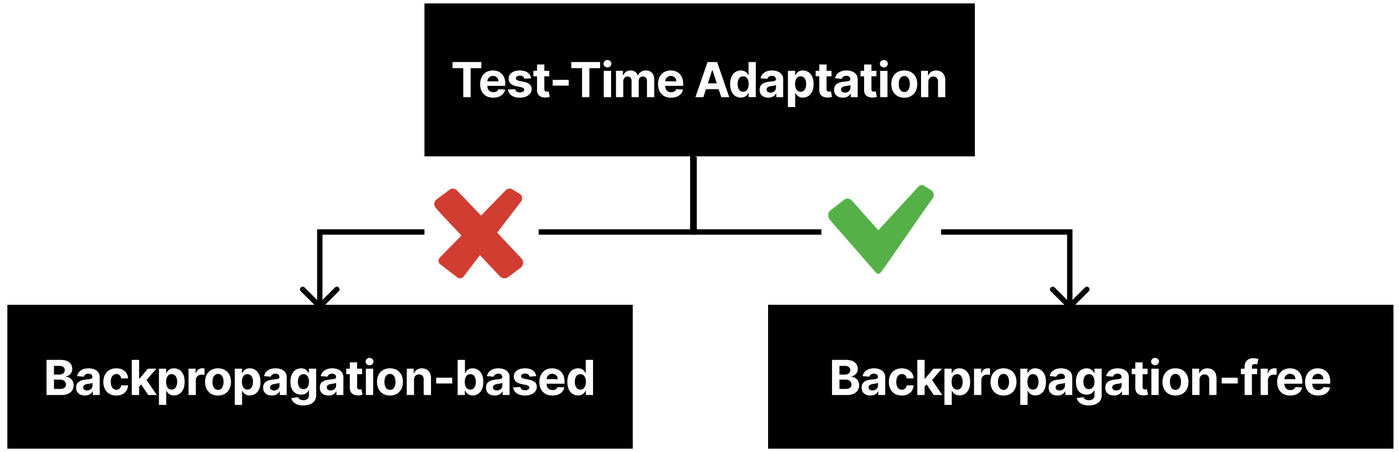
- Large memory overhead
- Not practical for SFMs after deployment

- Memory efficient

Test-Time Adaptation

- SFMs need to be updated after deployment
- Only unlabelled real world data are available
- Test-Time Adaptation (TTA) is an attractive solution





- Large memory overhead
- Not practical for SFMs after deployment

- Memory efficient
- Model architecture & task differences pose unique challenges for SFMs

Challenges of TTA on SFMs

Model architecture differences:

- **No batch normalization layers in SFMs**
- **SFMs: CNN layers + Transformer layers;**
  VFMs: CNN layers or Transformer layers

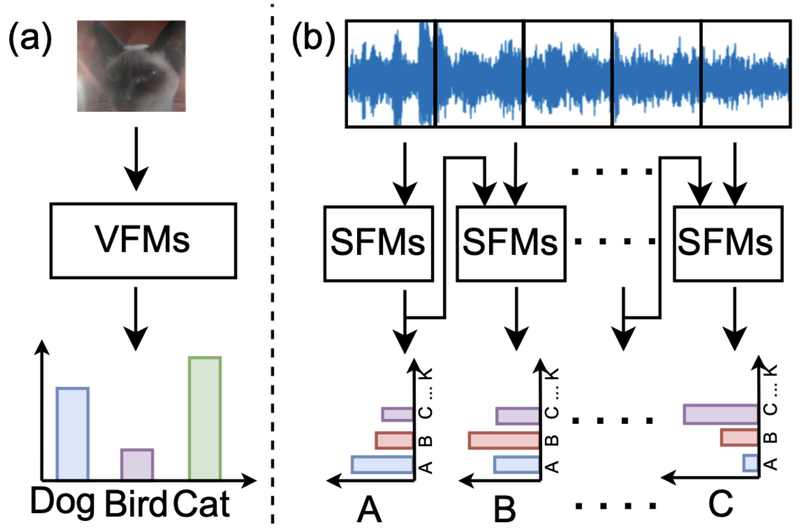
Challenges of TTA on SFMs

Model architecture differences:

- **No batch normalization layers in SFMs**
- **SFMs: CNN layers + Transformer layers;**
  VFMs: CNN layers or Transformer layers

Downstream tasks and noise characteristic differences:

- **Image classification**
  - **One-to-one mapping**
  - **Spatial perturbations of pixels**
- **Speech recognition**
  - **Sequence-to-sequence mapping**
  - **Dynamic, temporally varying across frames**



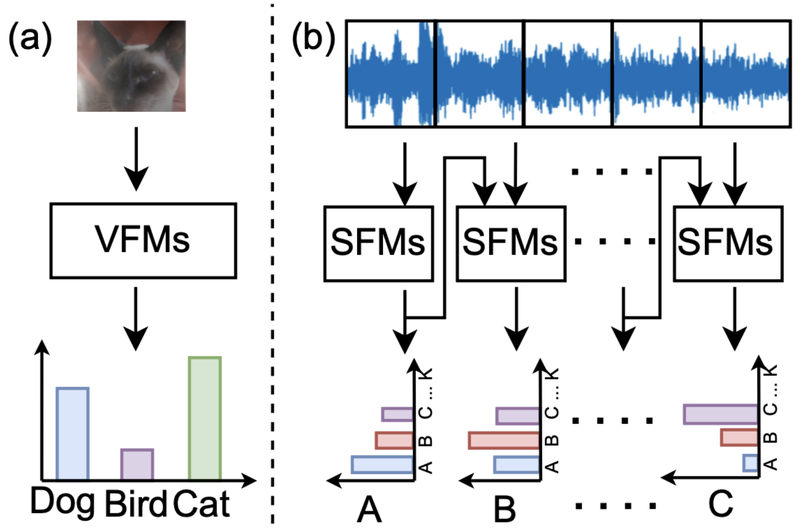
Challenges of TTA on SFMs

Model architecture differences:

- **No batch normalization layers in SFMs**
- **SFMs: CNN layers + Transformer layers;**
  VFMs: CNN layers or Transformer layers

Downstream tasks and noise characteristic differences:

- **Image classification**
  - **One-to-one mapping**
  - **Spatial perturbations of pixels**
- **Speech recognition**
  - **Sequence-to-sequence mapping**
  - **Dynamic, temporally varying across frames**



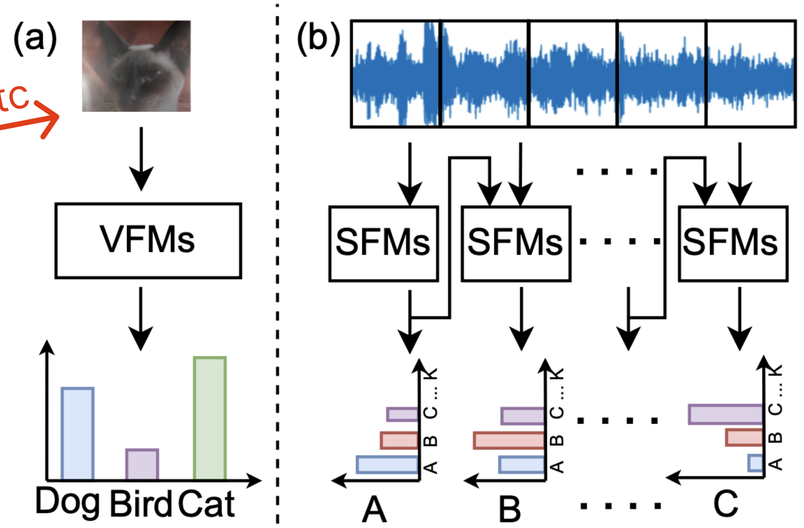
Challenges of TTA on SFMs

Model architecture differences:

- **No batch normalization layers in SFMs**
- **SFMs: CNN layers + Transformer layers;**
  VFMs: CNN layers or Transformer layers

Downstream tasks and noise characteristic differences:

- **Image classification**
  - **One-to-one mapping**
  - **Spatial perturbations of pixels**
- **Speech recognition**
  - **Sequence-to-sequence mapping**
  - **Dynamic, temporally varying across frames**



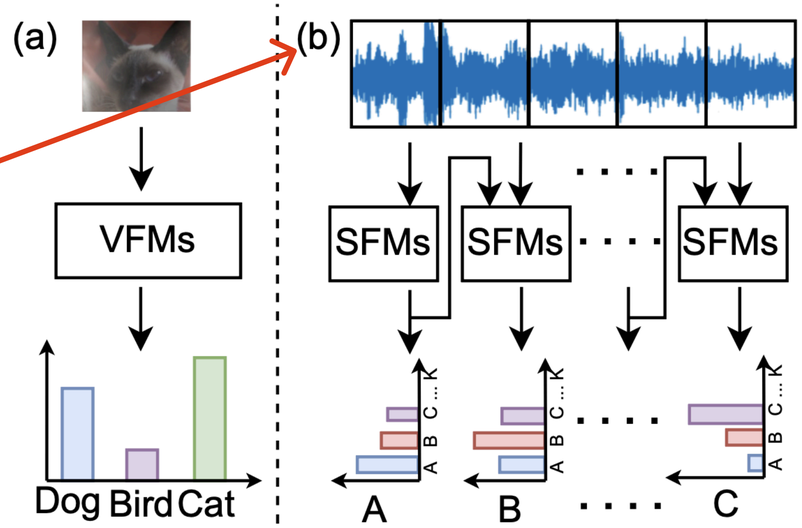
Challenges of TTA on SFMs

Model architecture differences:

- **No batch normalization layers in SFMs**
- **SFMs: CNN layers + Transformer layers;**
  VFMs: CNN layers or Transformer layers

Downstream tasks and noise characteristic differences:

- **Image classification**
  - **One-to-one mapping**
  - **Spatial perturbations of pixels**
- **Speech recognition**
  - **Sequence-to-sequence mapping**
  - **Dynamic, temporally varying across frames**



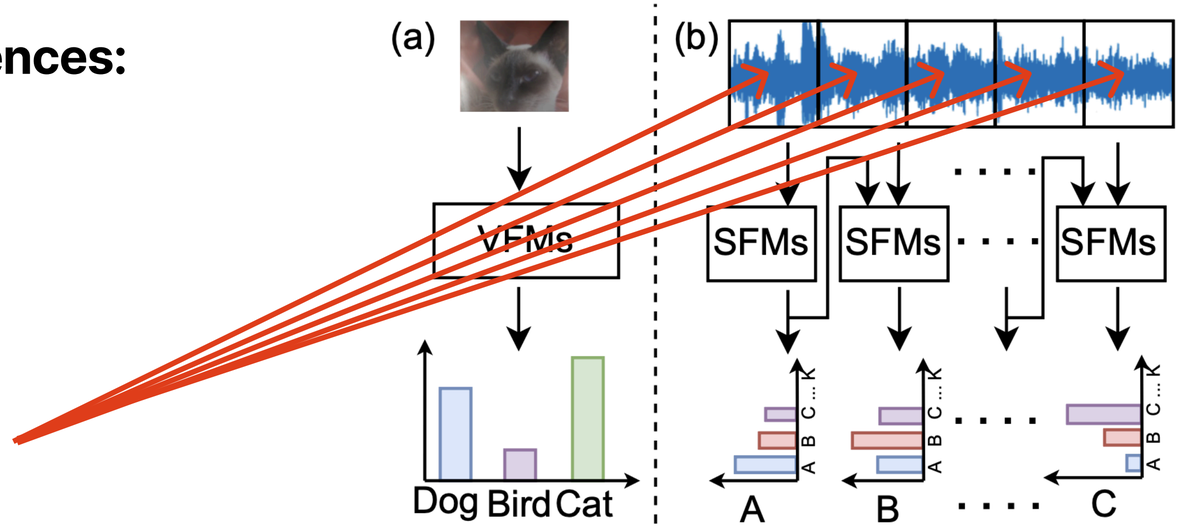
Challenges of TTA on SFMs

Model architecture differences:

- **No batch normalization layers in SFMs**
- **SFMs: CNN layers + Transformer layers;**
  VFMs: CNN layers or Transformer layers

Downstream tasks and noise characteristic differences:

- **Image classification**
  - **One-to-one mapping**
  - **Spatial perturbations of pixels**
- **Speech recognition**
  - **Sequence-to-sequence mapping**
  - **Dynamic, temporally varying across frames**



Challenges of TTA on SFMs

Model architecture differences:

- **No batch normalization layers in SFMs**
- **SFMs: CNN layers + Transformer layers;**
  VFMs: CNN layers or Transformer layers

Downstream tasks and noise characteristic differences:

- **Image classification**
  - **One-to-one mapping**
  - **Spatial perturbations of pixels**
- **Speech recognition**
  - **Sequence-to-sequence mapping**
  - **Dynamic, temporally varying across frames**

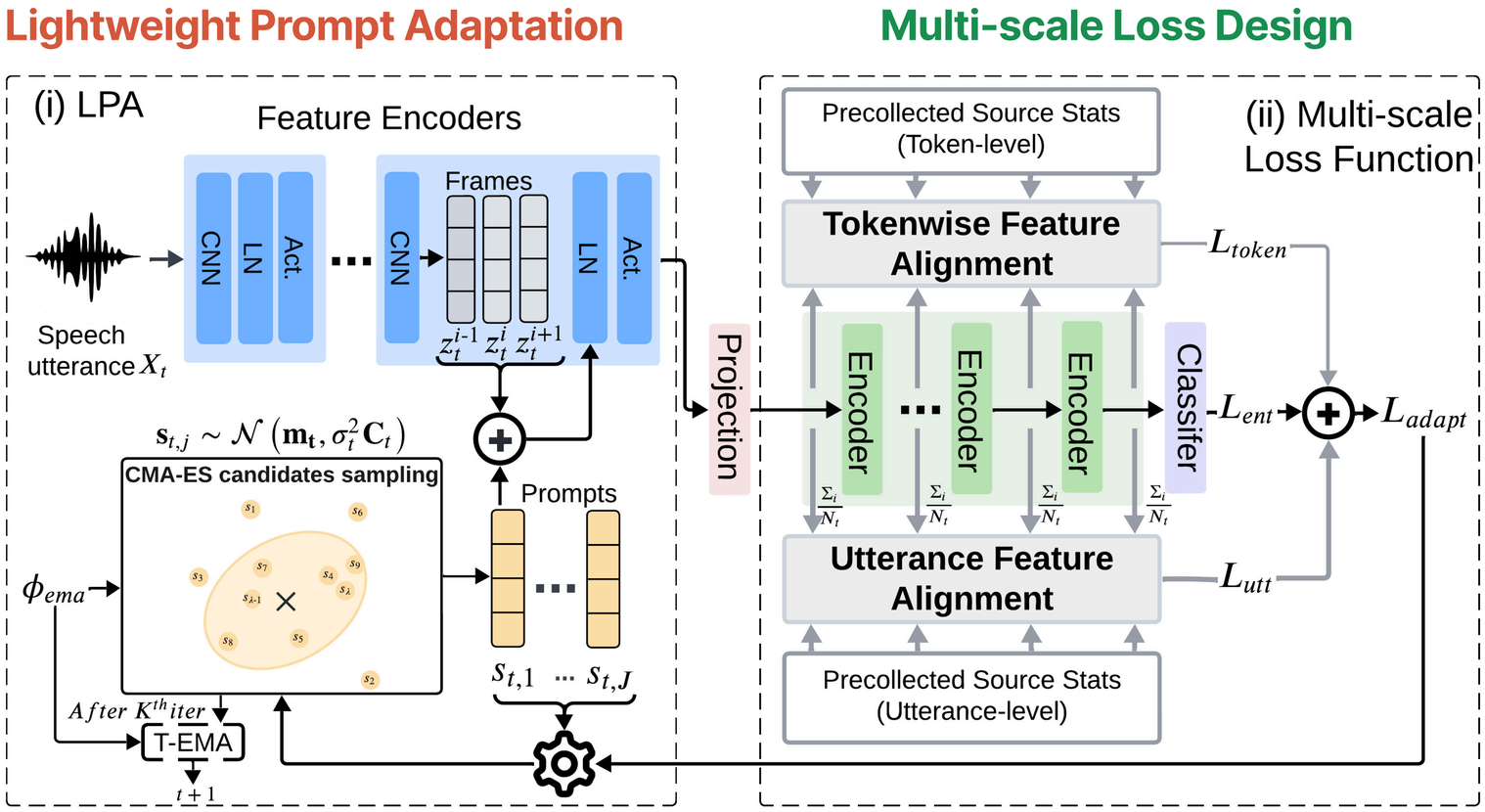
Adaptation batch size differences:

- **Current TTA depend on large batch size for reliable adaptation**
- **TTA in speech task need to process one utterance at a time**

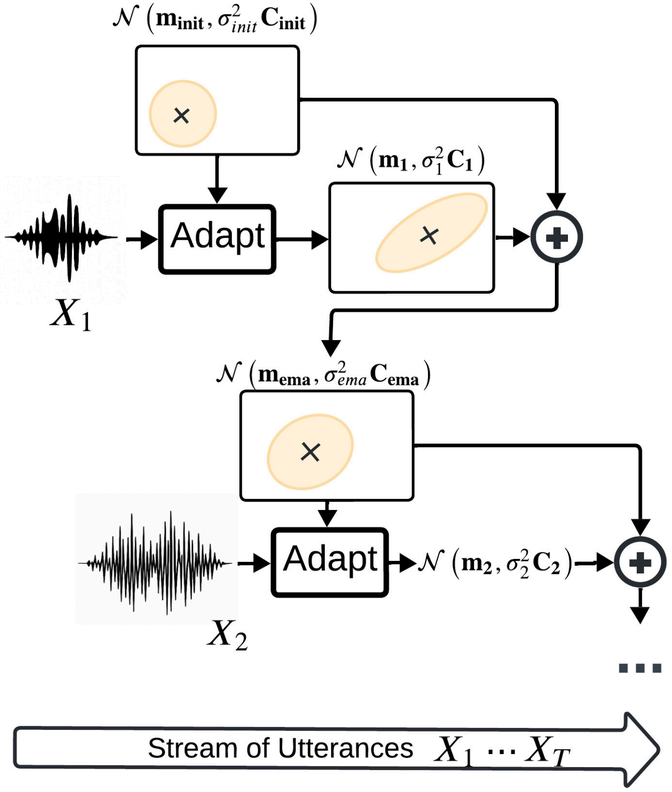
E-BATS

First efficient, backpropagation-free, and robust TTA for SFMs:

- Lightweight Prompt Adaptation (LPA)
- Multi-scale Loss Function
- Test-time Exponential Moving Average (T-EMA) across utterances

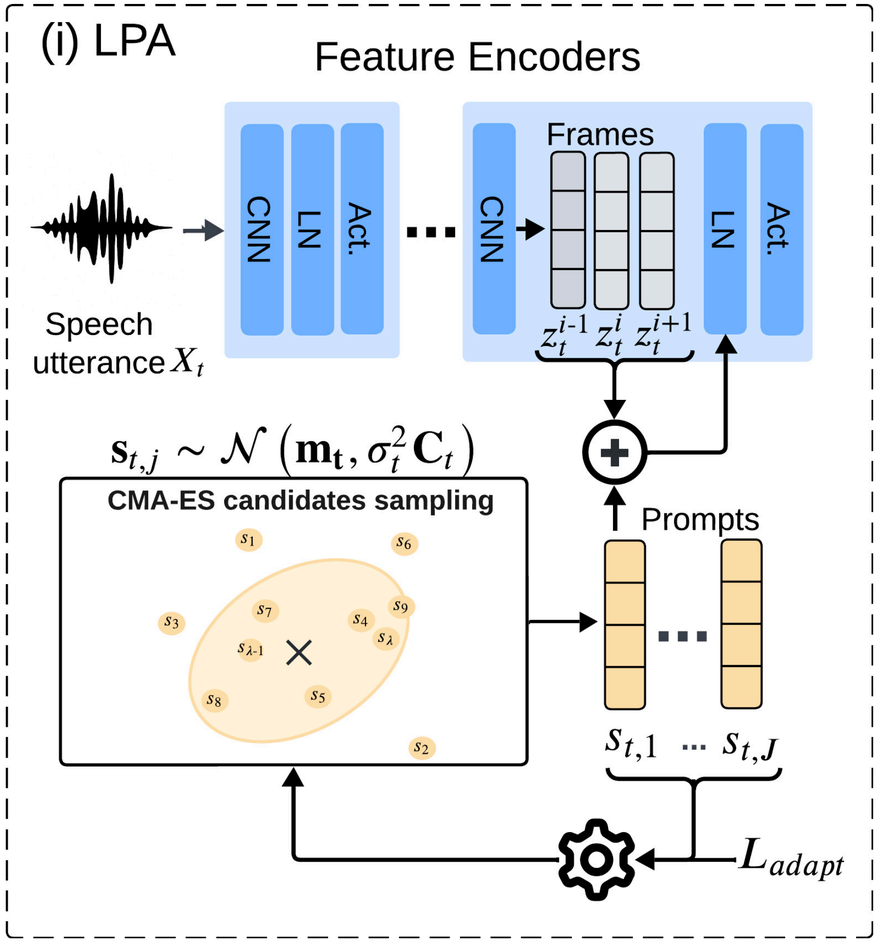


(a) Single-utterance Backpropagation-free Adaptation



(b) T-EMA across utterances

Lightweight Prompt Adaptation (LPA)

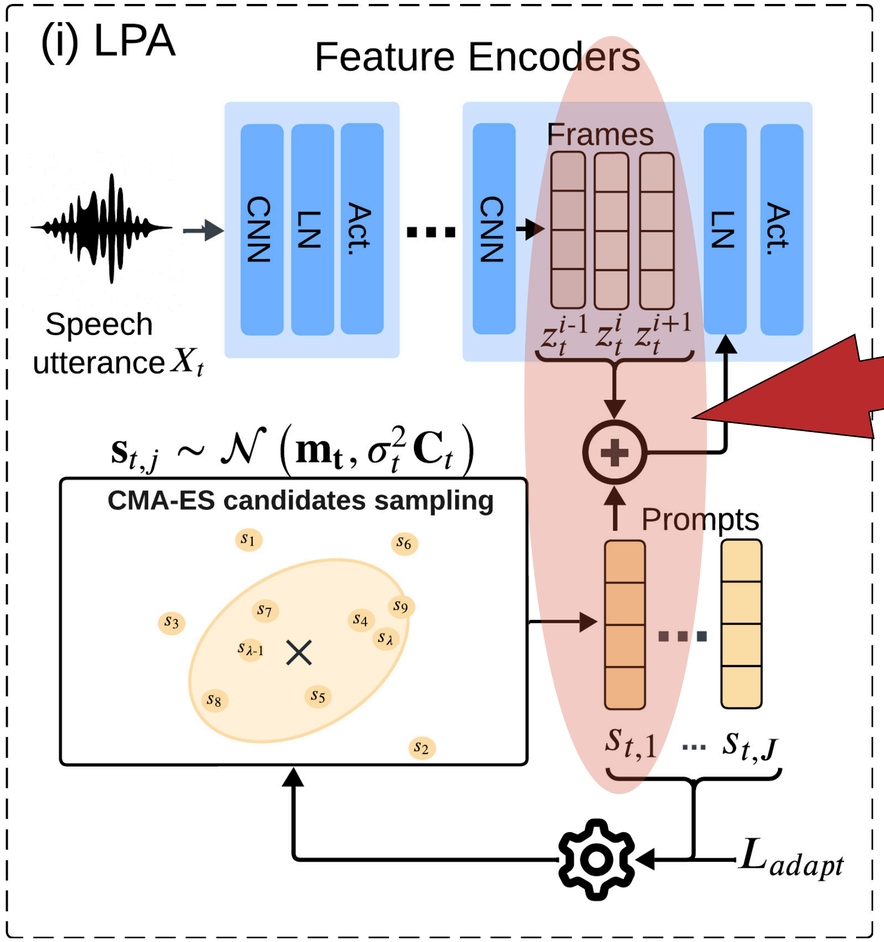


- **Learning Prompts to Adapt (address mean shift)**
  - **Focus on Feature encoders (CNN-based part)**
  - **Lightweight**

$$\hat{\boldsymbol{Z}}_t = \boldsymbol{Z}_t + \boldsymbol{s}_t \cdot \mathbf{1}_N^\top$$

$$\boldsymbol{Z}_t = \left[\boldsymbol{z}_t^1, \ldots, \boldsymbol{z}_t^{N_t}\right]$$

Lightweight Prompt Adaptation (LPA)

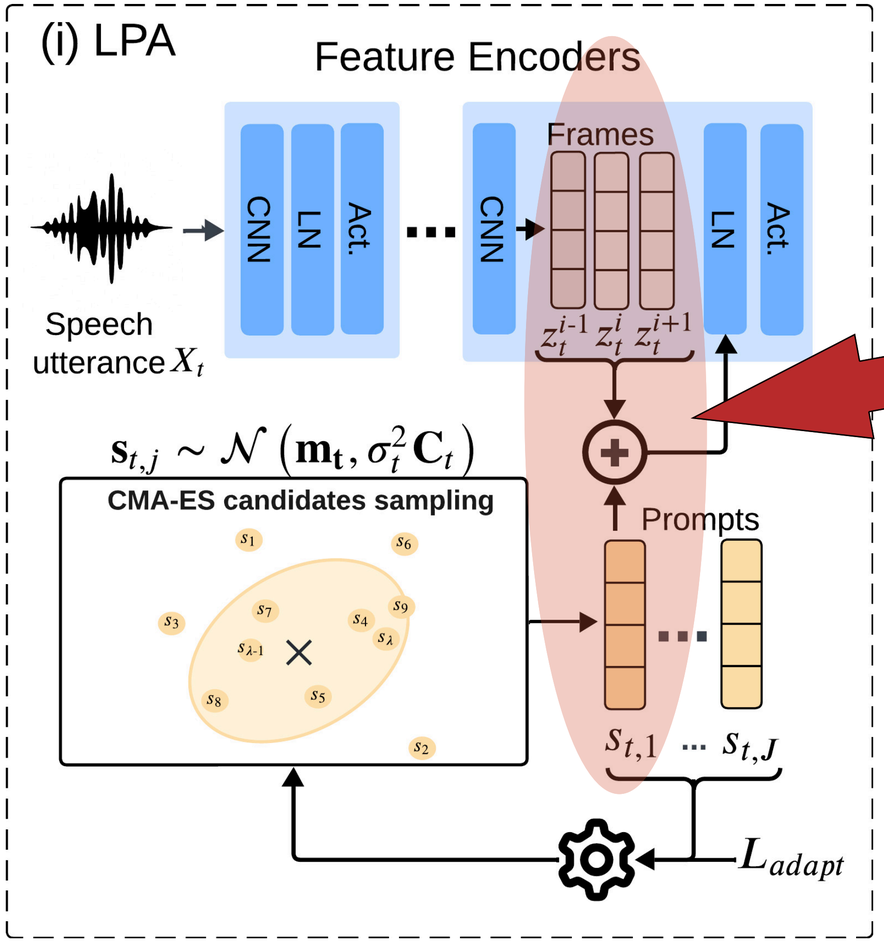


- **Learning Prompts to Adapt (address mean shift)**
  - **Focus on Feature encoders (CNN-based part)**
  - **Lightweight**

$$\hat{\boldsymbol{Z}}_t = \boldsymbol{Z}_t + \boldsymbol{s}_t \cdot \mathbf{1}_N^\top$$

$$\boldsymbol{Z}_t = \left[\boldsymbol{z}_t^1, \ldots, \boldsymbol{z}_t^{N_t}\right]$$

Lightweight Prompt Adaptation (LPA)



- **Learning Prompts to Adapt (address mean shift)**
  - **Focus on Feature encoders (CNN-based part)**
  - **Lightweight**

$$\hat{\boldsymbol{Z}}_t = \boldsymbol{Z}_t + \boldsymbol{s}_t \cdot \mathbf{1}_N^\top$$

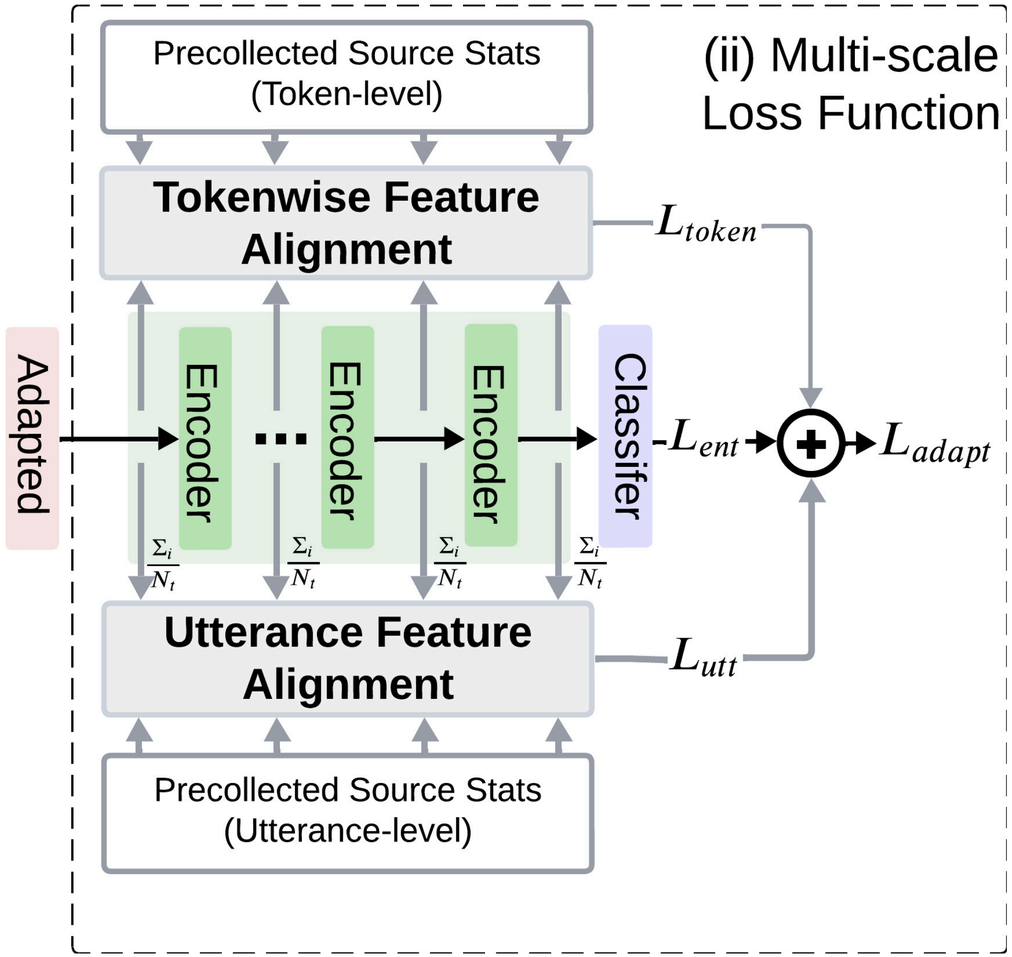
$$\boldsymbol{Z}_t = \left[\boldsymbol{z}_t^1, \ldots, \boldsymbol{z}_t^{N_t}\right]$$

- **Prompt Optimization with CMA-ES**
  - **Sample candidate prompt vectors**
  - **Update parameters with candidate prompts + Loss**

$$\boldsymbol{s}_t = \arg \min_{\boldsymbol{s}_{t,j} \in \mathbb{R}^d} L_{adapt}(\boldsymbol{X}_t, \boldsymbol{s}_{t,j})$$

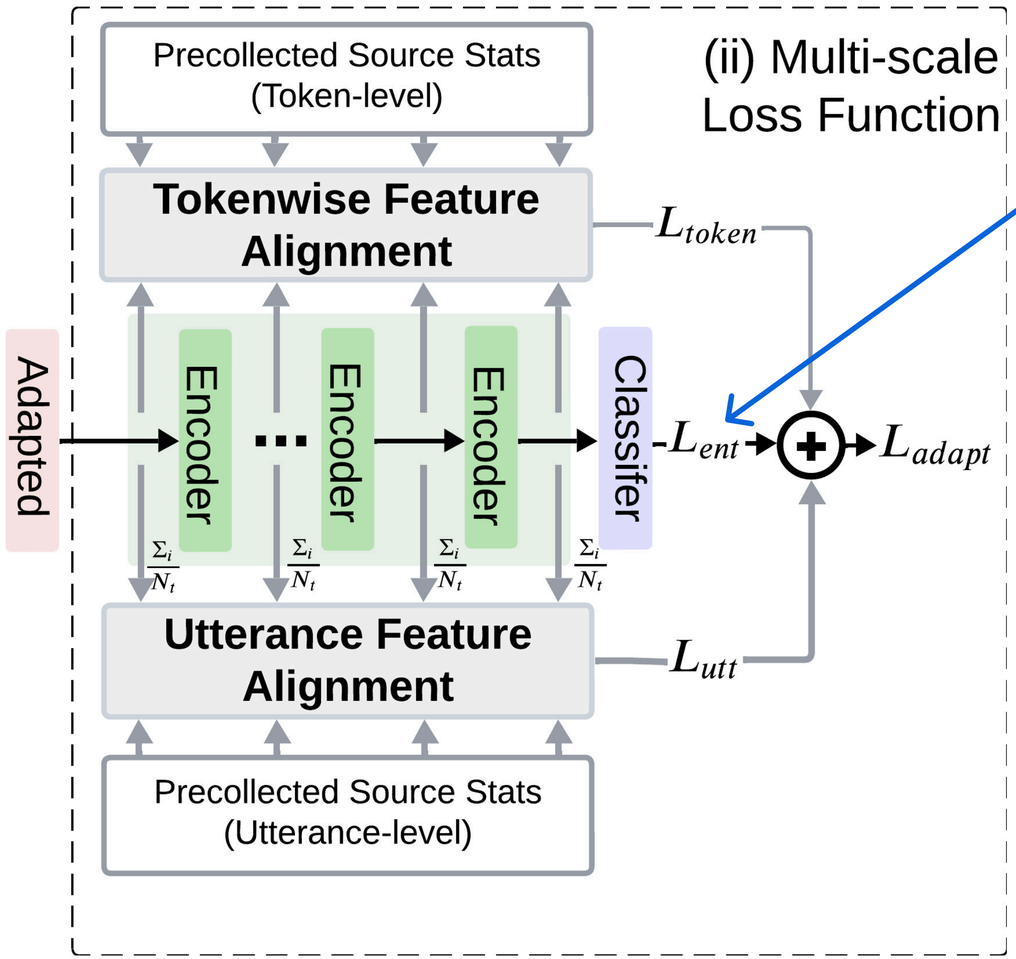
Multi-scale Loss Function

$$L_{adapt} = \alpha L_{ent} + \beta L_{utt} + cL_{token}$$



Multi-scale Loss Function

$$L_{adapt} = \alpha L_{ent} + \beta L_{utt} + cL_{token}$$

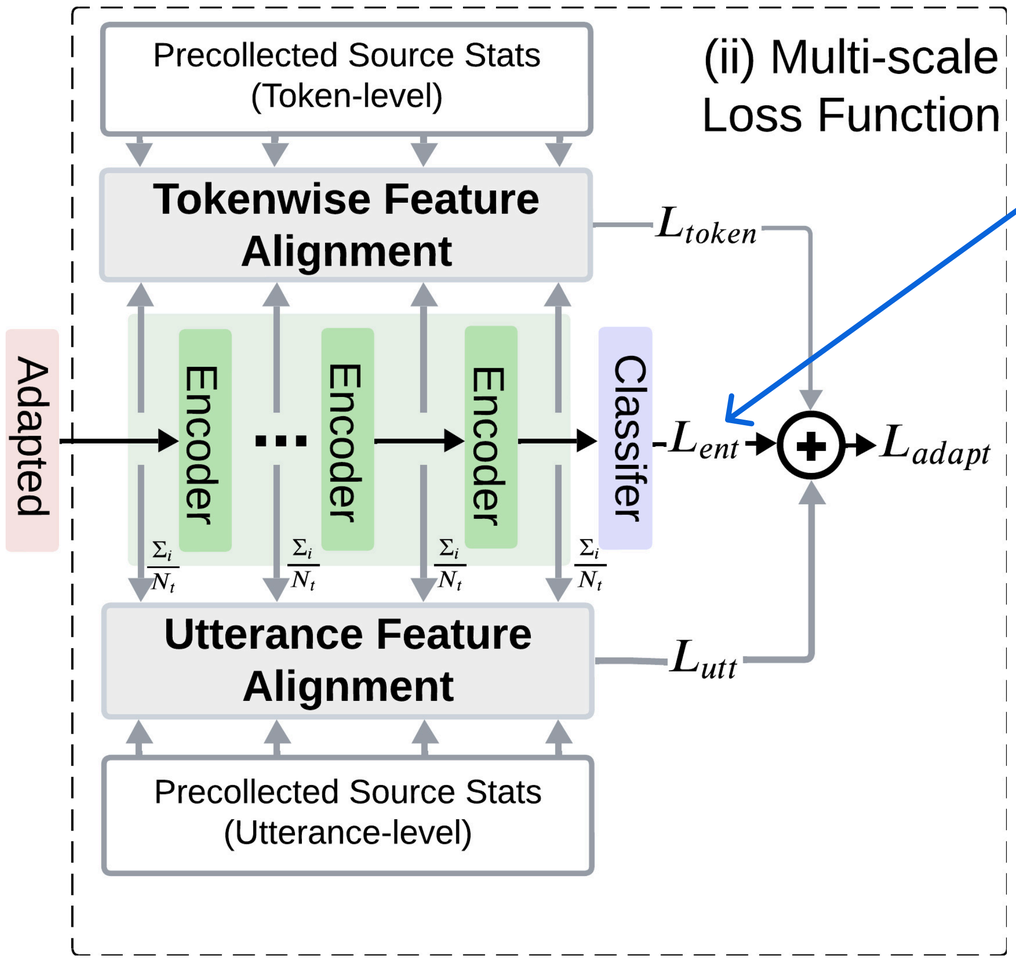


- **Entropy Loss**

$$L_{ent} = -\frac{1}{|\tilde{X}_t|} \sum_{x_t^i \in \tilde{X}_t} \mathcal{H}(\Theta(x_t^i))$$

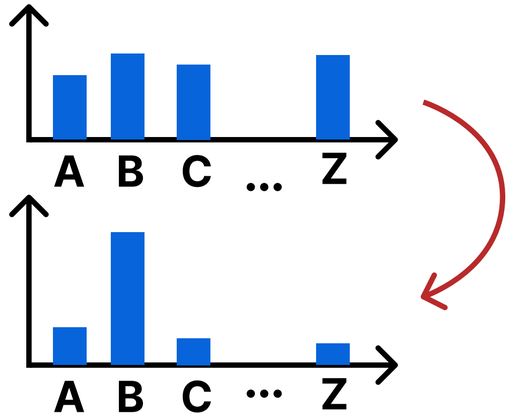
Multi-scale Loss Function

$$L_{adapt} = \alpha L_{ent} + \beta L_{utt} + cL_{token}$$



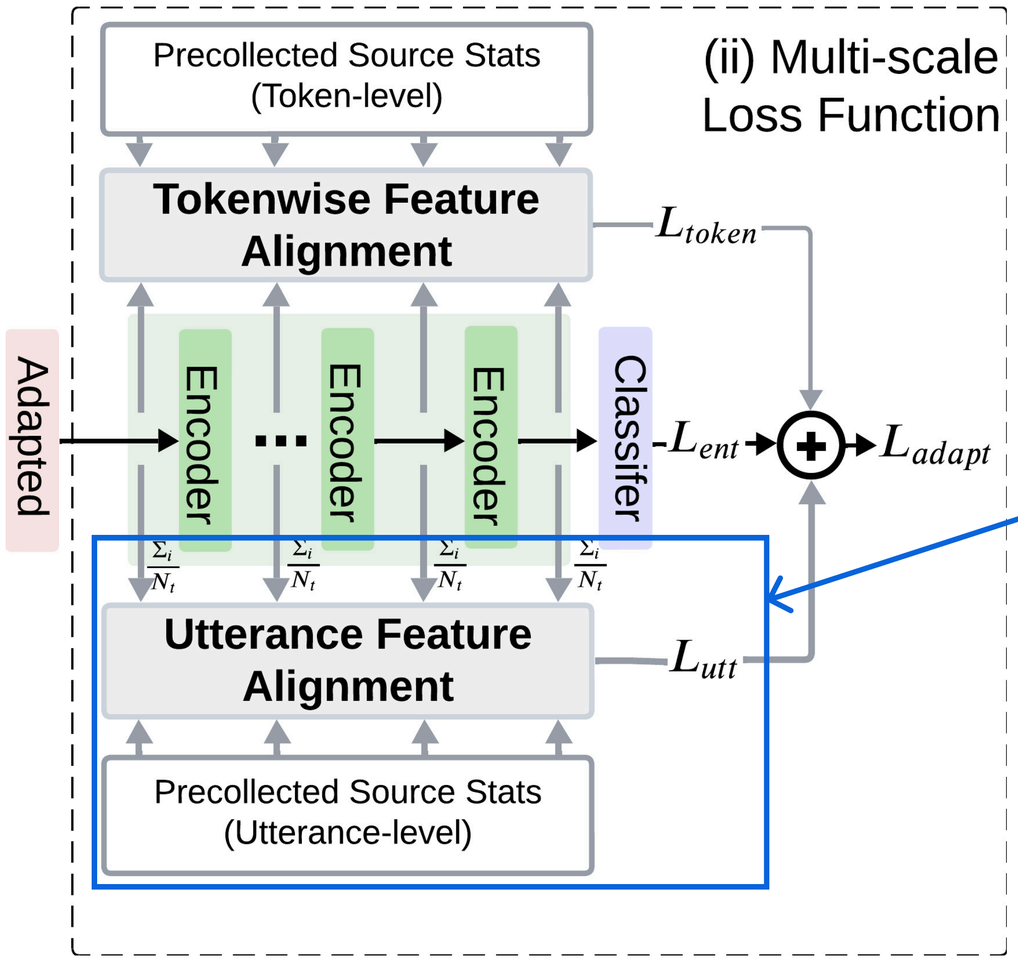
- **Entropy Loss**

$$L_{ent} = -\frac{1}{|\tilde{X}_t|} \sum_{x_t^i \in \tilde{X}_t} \mathcal{H}(\Theta(x_t^i))$$



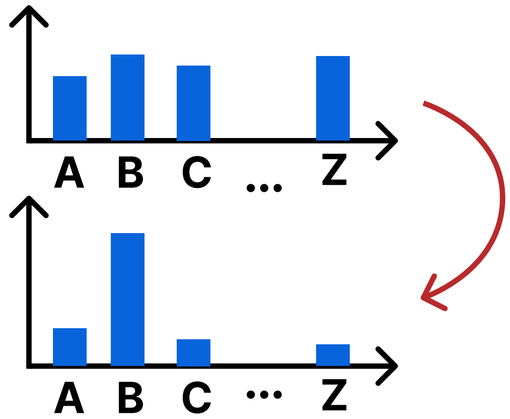
Multi-scale Loss Function

$$L_{adapt} = \alpha L_{ent} + \beta L_{utt} + cL_{token}$$



- **Entropy Loss**

$$L_{ent} = -\frac{1}{|\tilde{X}_t|} \sum_{x_t^i \in \tilde{X}_t} \mathcal{H}(\Theta(x_t^i))$$

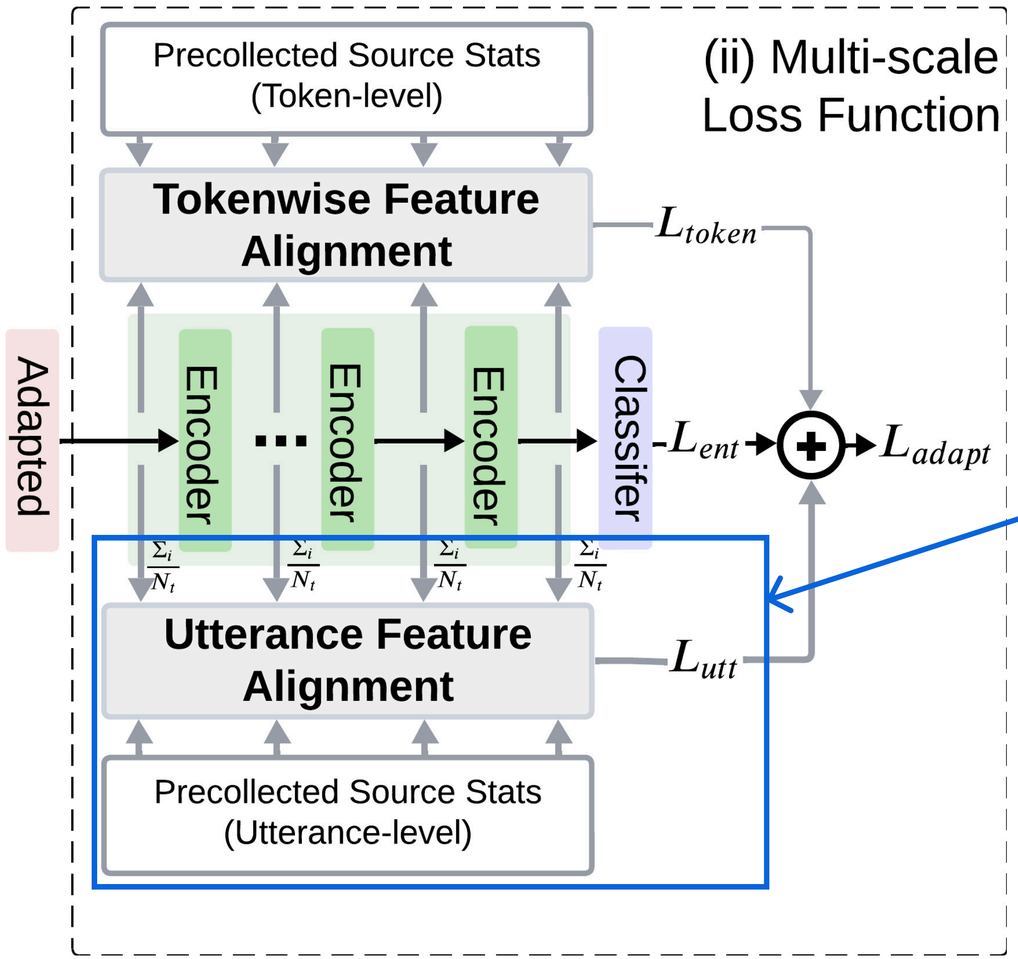


- **Utterance-level Latent Embedding Alignment Loss**

$$L_{\text{utt}} = \frac{1}{L} \sum_{l=0}^{L} \|\mu_{\text{tgt}}^l - \mu_{\text{src}}^l\|_2^2$$

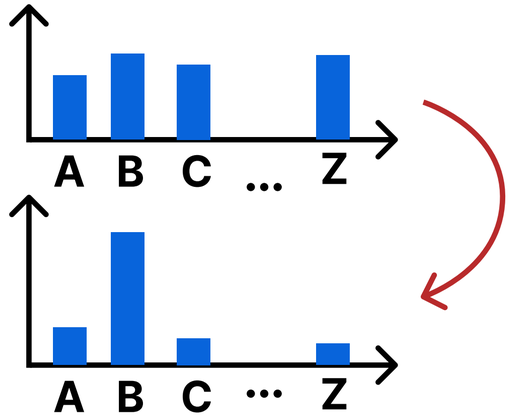
Multi-scale Loss Function

$$L_{adapt} = \alpha L_{ent} + \beta L_{utt} + cL_{token}$$



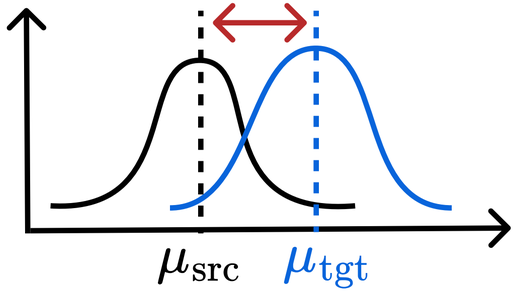
- **Entropy Loss**

$$L_{ent} = -\frac{1}{|\tilde{X}_t|} \sum_{x_t^i \in \tilde{X}_t} \mathcal{H}(\Theta(x_t^i))$$



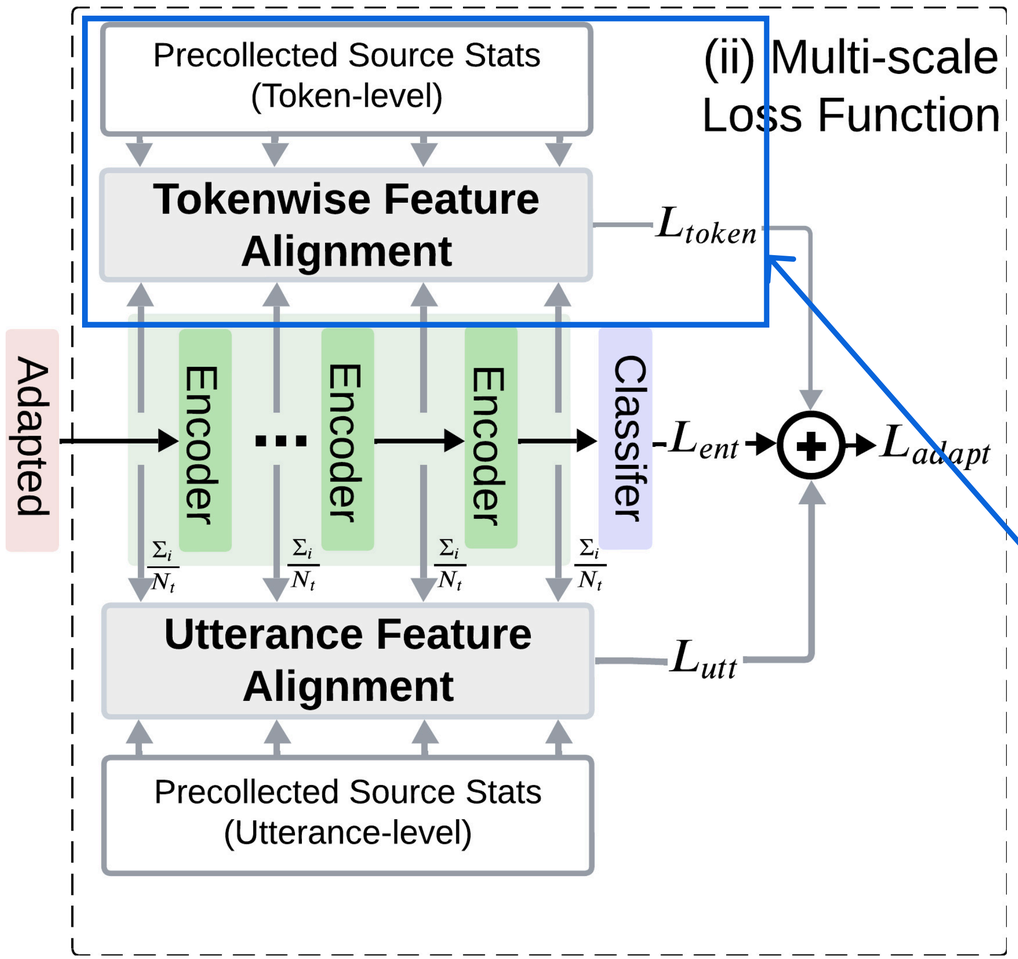
- **Utterance-level Latent Embedding Alignment Loss**

$$L_{\text{utt}} = \frac{1}{L} \sum_{l=0}^{L} \|\mu_{\text{tgt}}^l - \mu_{\text{src}}^l\|_2^2$$



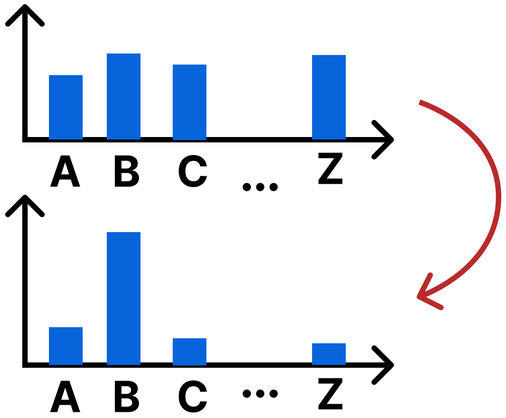
Multi-scale Loss Function

$$L_{adapt} = \alpha L_{ent} + \beta L_{utt} + cL_{token}$$



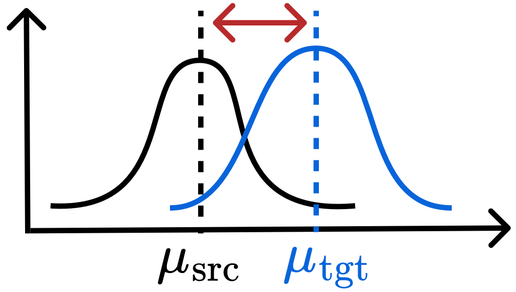
- **Entropy Loss**

$$L_{ent} = -\frac{1}{|\tilde{X}_t|} \sum_{x_t^i \in \tilde{X}_t} \mathcal{H}(\Theta(x_t^i))$$



- **Utterance-level Latent Embedding Alignment Loss**

$$L_{\text{utt}} = \frac{1}{L} \sum_{l=0}^{L} \|\mu_{\text{tgt}}^l - \mu_{\text{src}}^l\|_2^2$$

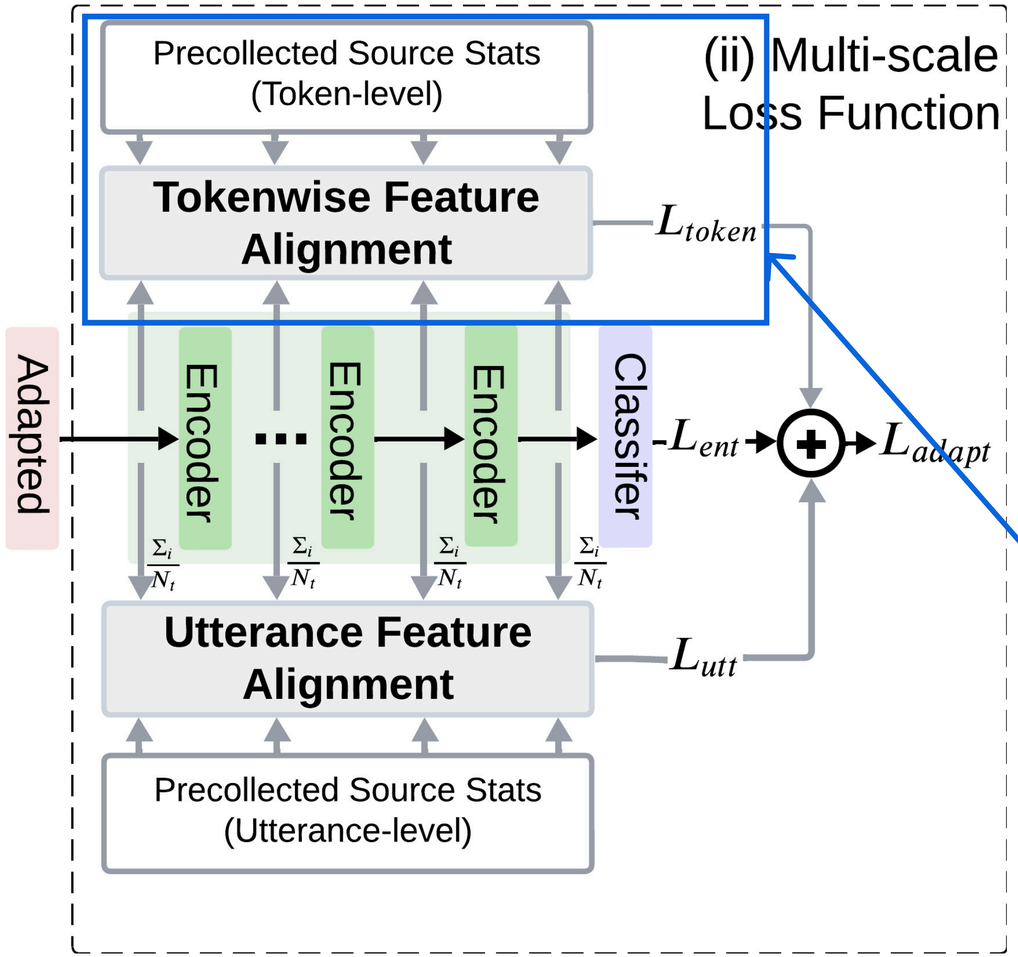


- **Tokenwise Latent Embeddings Alignment**

$$L_{token} = \frac{1}{L} \frac{1}{|\mathcal{V}|} \sum_{l=0}^{L} \sum_{v \in \mathcal{V}} \left(\left\| \boldsymbol{\mu}_{tgt}^{v,l} - \boldsymbol{\mu}_{src}^{v,l} \right\|_2^2 + \left\| \boldsymbol{\sigma}_{tgt}^{v,l} - \boldsymbol{\sigma}_{src}^{v,l} \right\|_2^2 \right)$$

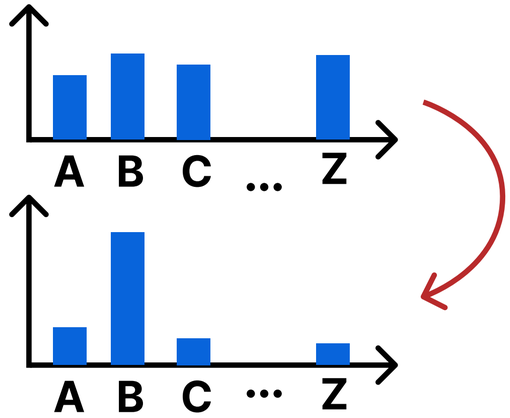
Multi-scale Loss Function

$$L_{adapt} = \alpha L_{ent} + \beta L_{utt} + cL_{token}$$



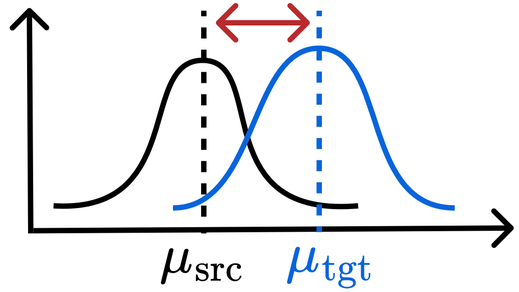
- **Entropy Loss**

$$L_{ent} = -\frac{1}{|\tilde{X}_t|} \sum_{x_t^i \in \tilde{X}_t} \mathcal{H}(\Theta(x_t^i))$$



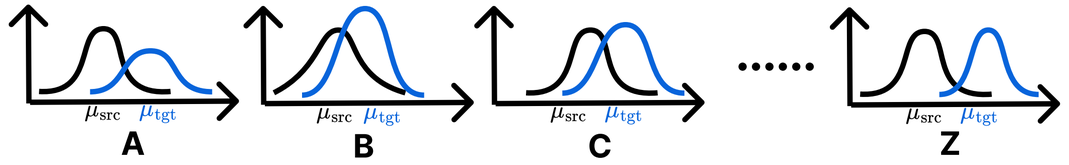
- **Utterance-level Latent Embedding Alignment Loss**

$$L_{\text{utt}} = \frac{1}{L} \sum_{l=0}^{L} \|\mu_{\text{tgt}}^l - \mu_{\text{src}}^l\|_2^2$$



- **Tokenwise Latent Embeddings Alignment**

$$L_{token} = \frac{1}{L} \frac{1}{|\mathcal{V}|} \sum_{l=0}^{L} \sum_{v \in \mathcal{V}} \left(\left\| \boldsymbol{\mu}_{tgt}^{v,l} - \boldsymbol{\mu}_{src}^{v,l} \right\|_2^2 + \left\| \boldsymbol{\sigma}_{tgt}^{v,l} - \boldsymbol{\sigma}_{src}^{v,l} \right\|_2^2 \right)$$



T-EMA across Utterances

- Stablize adaptation across the utterance streams
- Ensure a smoother CMA-ES parameter updates
- Reduce overfitting and mitigating model drift

T-EMA across Utterances

- **Stablize adaptation across the utterance streams**
- **Ensure a smoother CMA-ES parameter updates**
- **Reduce overfitting and mitigating model drift**

Update CMA-ES parameters with T-EMA

- Mean vector

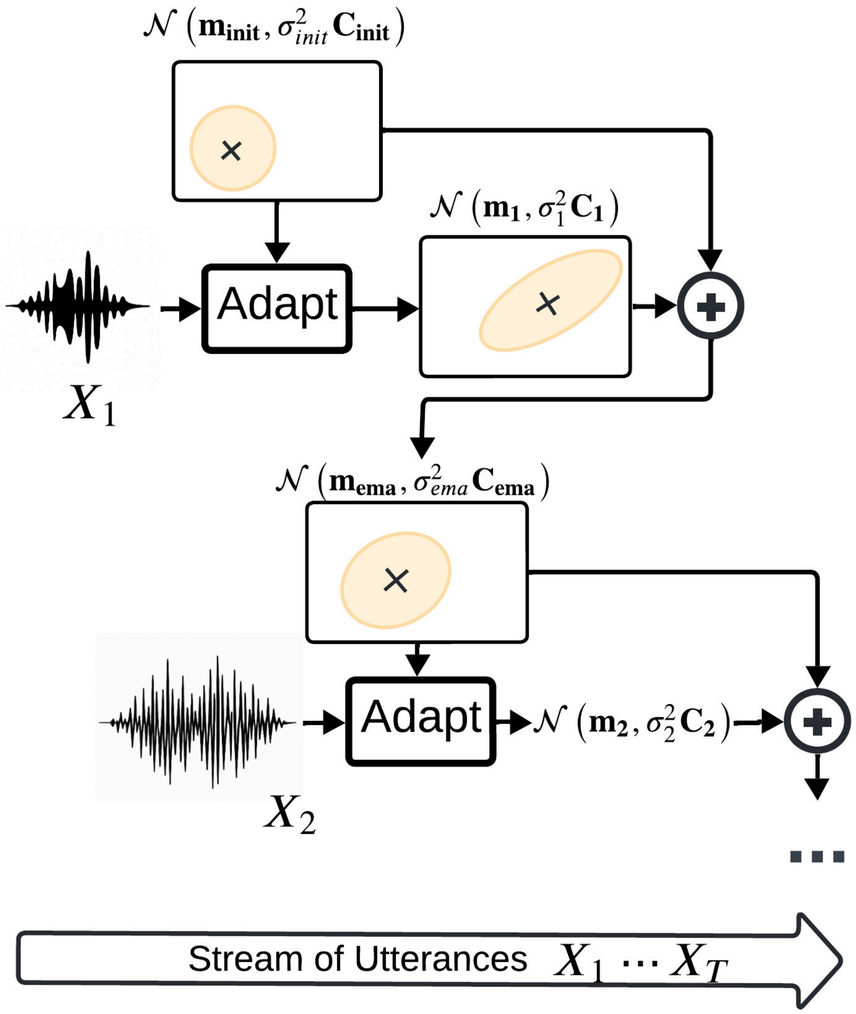
$$\mathbf{m}_{ema} = \gamma\, \mathbf{m}_{ema} + (1-\gamma)\, \mathbf{m}_t$$

- Covariance matrix

$$\mathbf{C}_{ema} = \gamma\, \mathbf{C}_{ema} + (1-\gamma)\, \mathbf{C}_t$$

- Step size

$$\sigma_{ema} = \gamma\, \sigma_{ema} + (1-\gamma)\, \sigma_t$$



Experimental Setup

Model Architecture

- Wav2Vec2ForCTC-Base
- HuBERTForCTC-Large

Datasets

- LibriSpeech
- CHiME-3
- CommonVoice
- TEDLIUM-v2

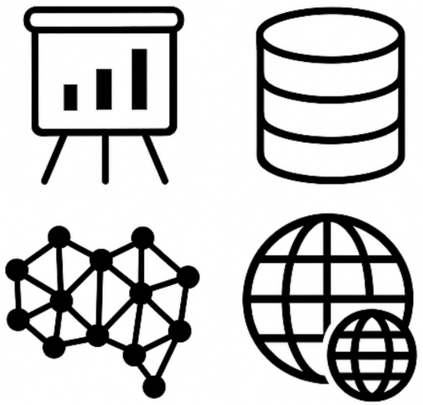
Baselines

- Backpropagation-based TTA:
  (1)TENT (2)EATA (3)SAR
  (4)CoTTA (5)CEA (6)SGEM
  (7)AWMC (8)SUTA (9)CSUTA
  (10)DSUTA

- Backpropagation-Free TTA:
  (11)T3A (12)LAME (13)FOA

Domain Shifts Design

- Synthetic noise (5 severity levels)
- Environmental noise (Bus, Cafe, Pedestrian, Street junction)
  - Single domain
  - Mixed domain
- Mixed variability (accents, devices, environments)
- Mixed variability (accents, styles, syntactic structures)



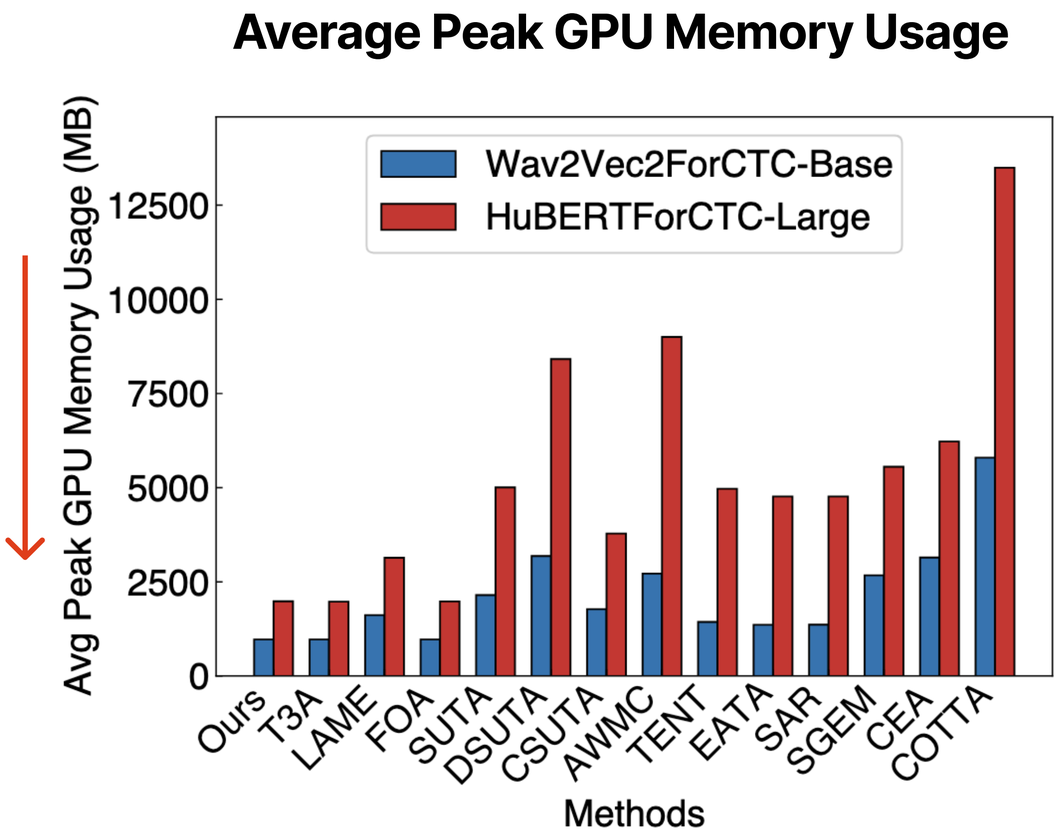
Results - Accuracy

Table 1: Word Error Rate (WER) on various noisy conditions using Wav2Vec2ForCTC-Base. Lower value means better adaptation performance. **Bold** represents the best performance for BP-free TTA, while underlined means the best for both BP-based and BP-free TTA.

Method	BP-free	Gaussian noise 0.0	0.005	0.01	0.015	0.02	Avg	CHiME3 (Single)	CHiME3 (Mixed)	TED	Common Voice
Source	—	8.6	13.9	24.4	39.5	54.5	28.2	34.2	34.2	13.2	36.8
TENT	✗	8.5	14.0	24.1	39.2	54.3	28.0	34.1	34.1	13.1	36.8
EATA	✗	14.1	18.1	27.0	37.9	51.3	29.7	33.1	39.9	14.1	61.3
SAR	✗	8.4	13.6	22.9	36.0	49.9	26.2	33.6	34.7	13.0	38.2
CoTTA	✗	9.2	12.6	18.1	39.3	54.5	26.7	32.9	34.3	12.8	36.9
CEA	✗	7.5	11.1	16.4	23.8	33.6	18.5	26.8	26.8	12.0	31.5
SGEM	✗	7.3	10.9	16.4	23.8	33.9	18.5	27.2	27.1	11.9	31.2
AWMC	✗	9.5	11.7	16.6	23.9	31.8	18.7	34.0	33.9	13.6	37.9
SUTA	✗	7.3	10.9	16.5	24.1	34.1	18.6	26.8	26.8	11.9	31.5
CSUTA	✗	13.1	17.5	24.5	31.4	37.0	24.7	26.5	32.6	15.6	135.0
DSUTA	✗	9.0	11.7	16.1	21.1	24.1	16.4	24.0	24.1	12.7	36.1
T3A	✓	10.0	15.9	26.8	42.7	58.6	30.8	35.9	35.8	14.6	38.8
LAME	✓	9.1	15.0	26.0	42.4	58.2	30.1	36.0	36.0	14.0	38.8
FOA	✓	8.7	13.9	22.7	33.3	45.3	24.8	31.7	31.1	13.3	38.2
Ours	✓	**7.7**	**10.5**	**14.8**	**19.9**	**25.3**	**15.6**	**24.0**	**24.3**	**12.5**	**30.6**

E-BATS achieves up to 20.0% lower word error rate (WER) compared to BP-free TTA baselines (Wav2Vec2ForCTC-Base)

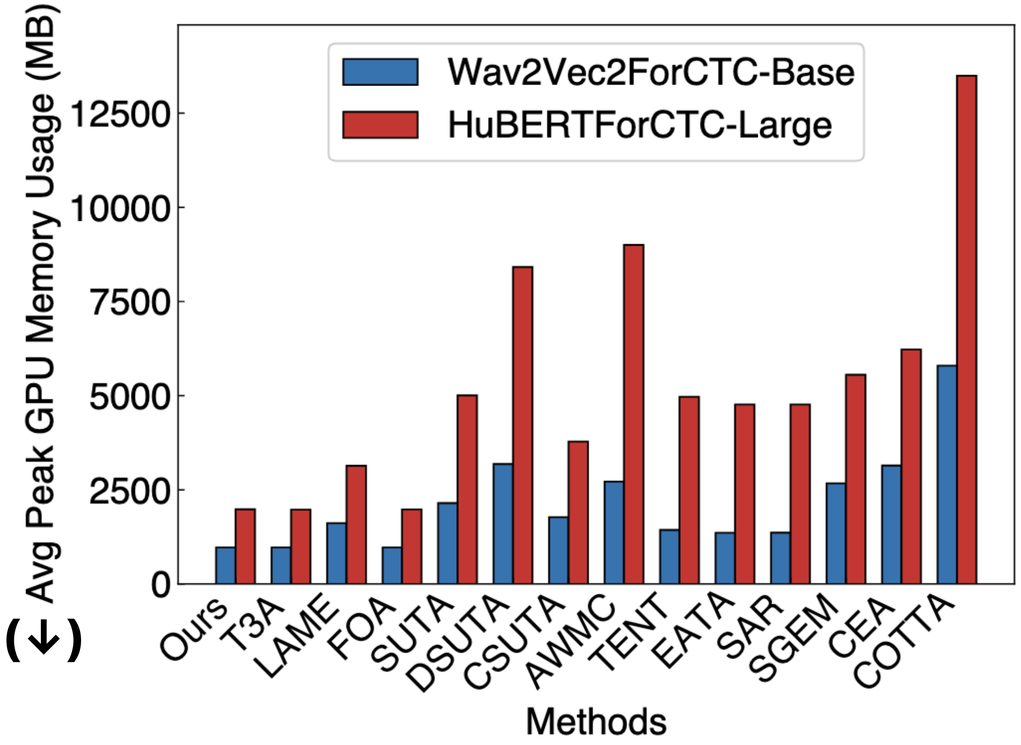
Results - Memory



E-BATS achieves up to 6.8x lower peak GPU memory usage compared to BP-based TTA baselines

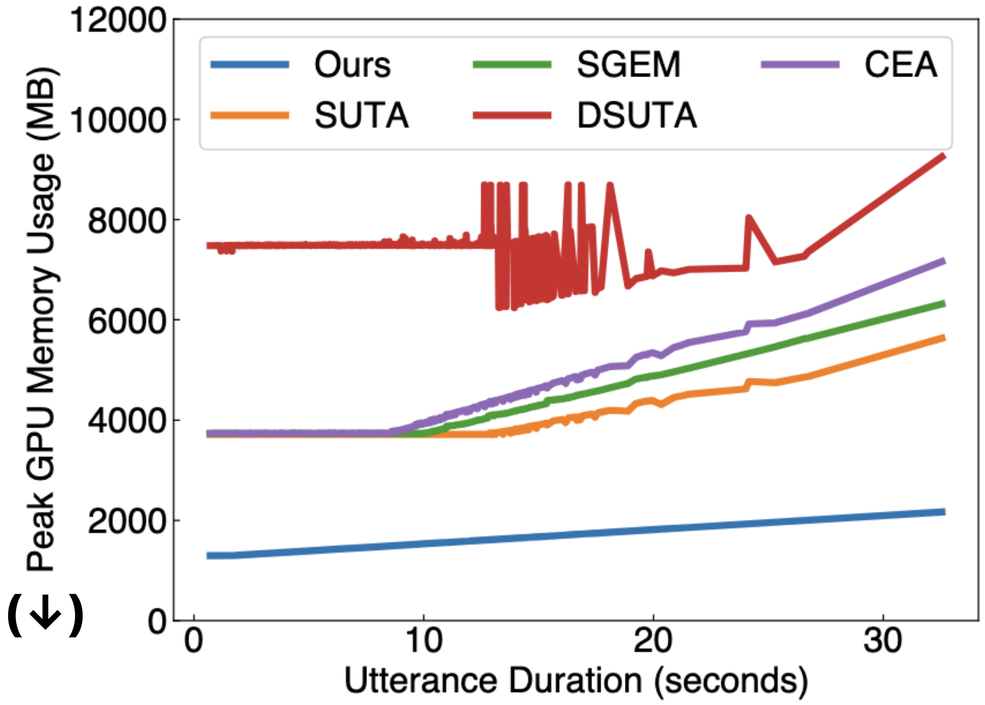
Results - Memory

Average Peak GPU Memory Usage



E-BATS achieves up to 6.8x lower peak GPU memory usage compared to BP-based TTA baselines

Memory usage vs. Utterance duration



E-BATS displays a stable and near-linear profile for increasing utterance durations

Summary

Key Contributions

- First backpropagation-free TTA approach for SFMs
- Three novel modules address unique challenges of speech tasks
- Robust performance across diverse acoustic domain shifts

Thank you !

Paper

Code

Jiaheng.dong@student.unimelb.edu.au