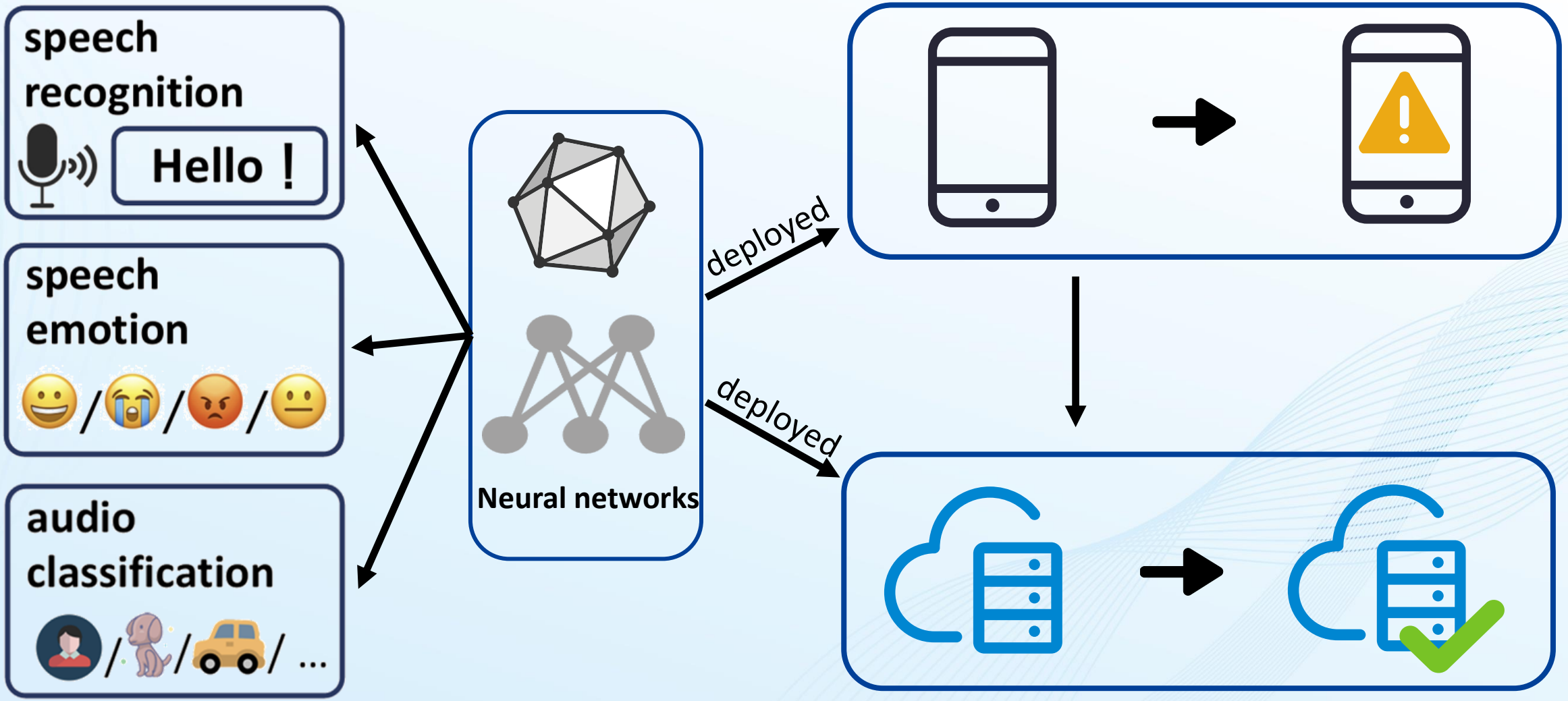
2025 IEEE INTERNATIONAL CONFERENCE ON
ACOUSTICS, SPEECH, AND SIGNAL PROCESSING (ICASSP 2025)
April 06 - 11, 2025 Hyderabad, India
2025.ieeeicassp.org

AMSER: Accelerate Mobile Speech Emotion Recognition with Signal Compression

Yu Lu[&], Ran Wang[&], Dian Ding[*], Han Zhang, Liyun Zhang,
Lanqing Yang, Yi-Chao Chen, Guangtao Xue[*]

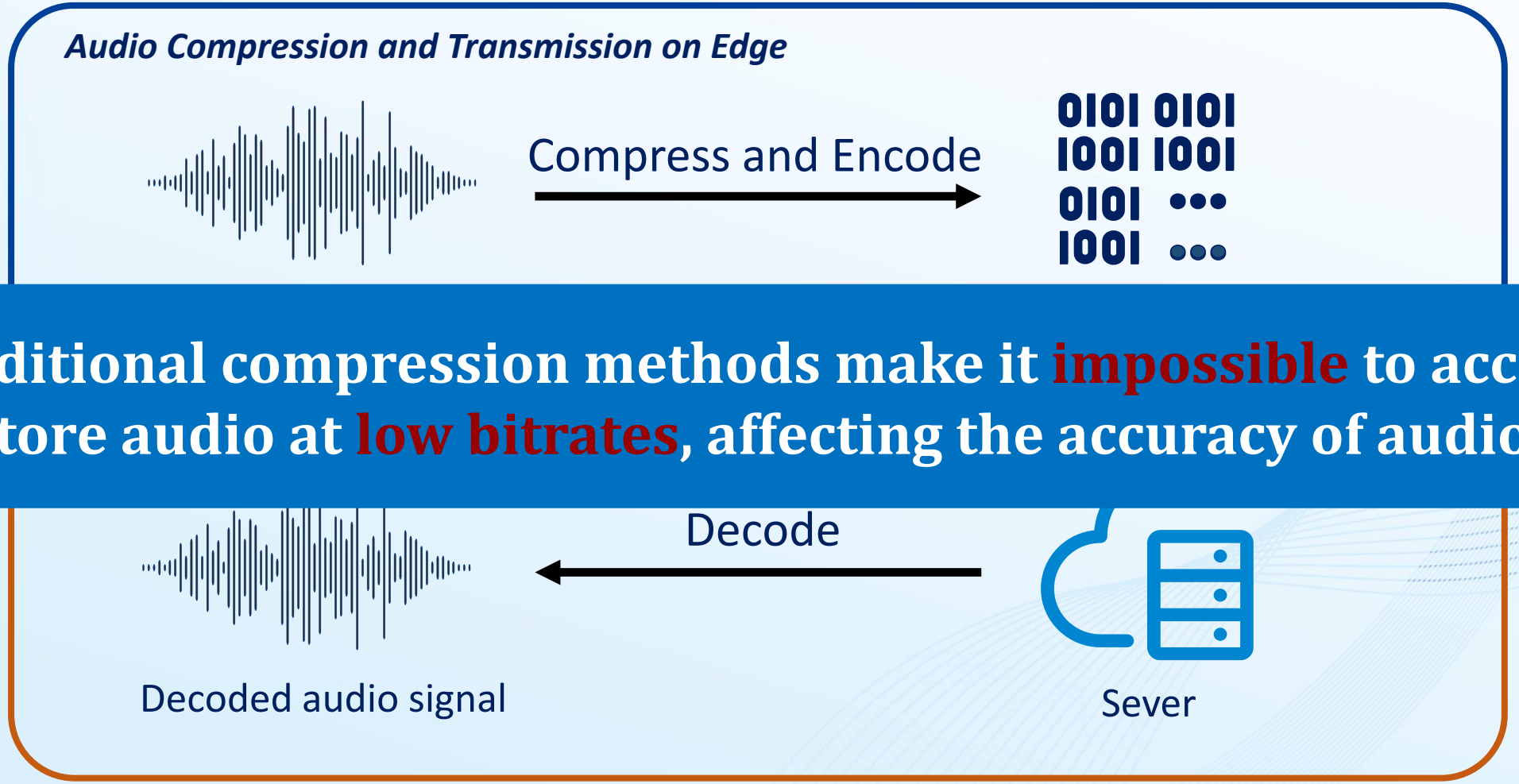
Department of Computer Science and Engineering
Shanghai Jiao Tong University

Background: Numerous neural network–based speech tasks are deployed on servers.

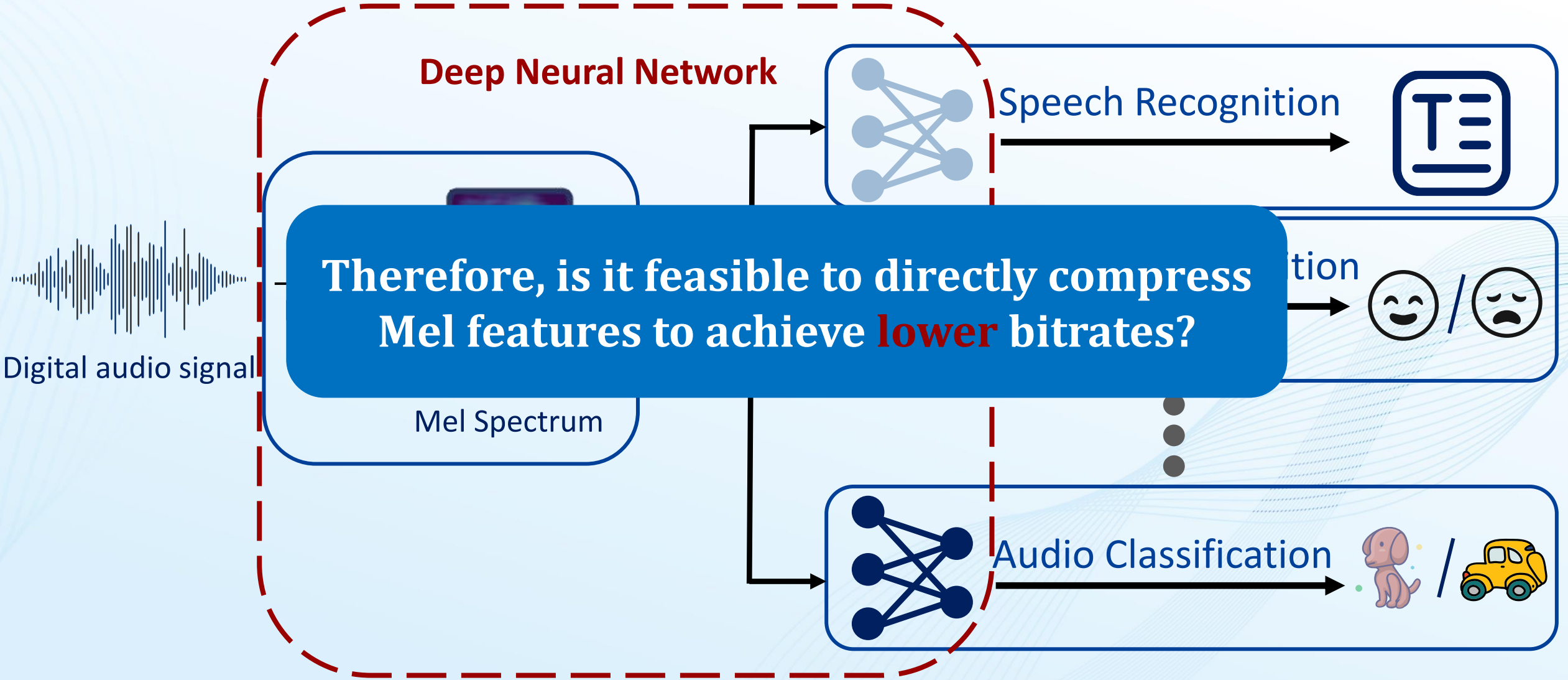


Challenge: Current cloud-edge systems struggle to handle large audio streams in real time.

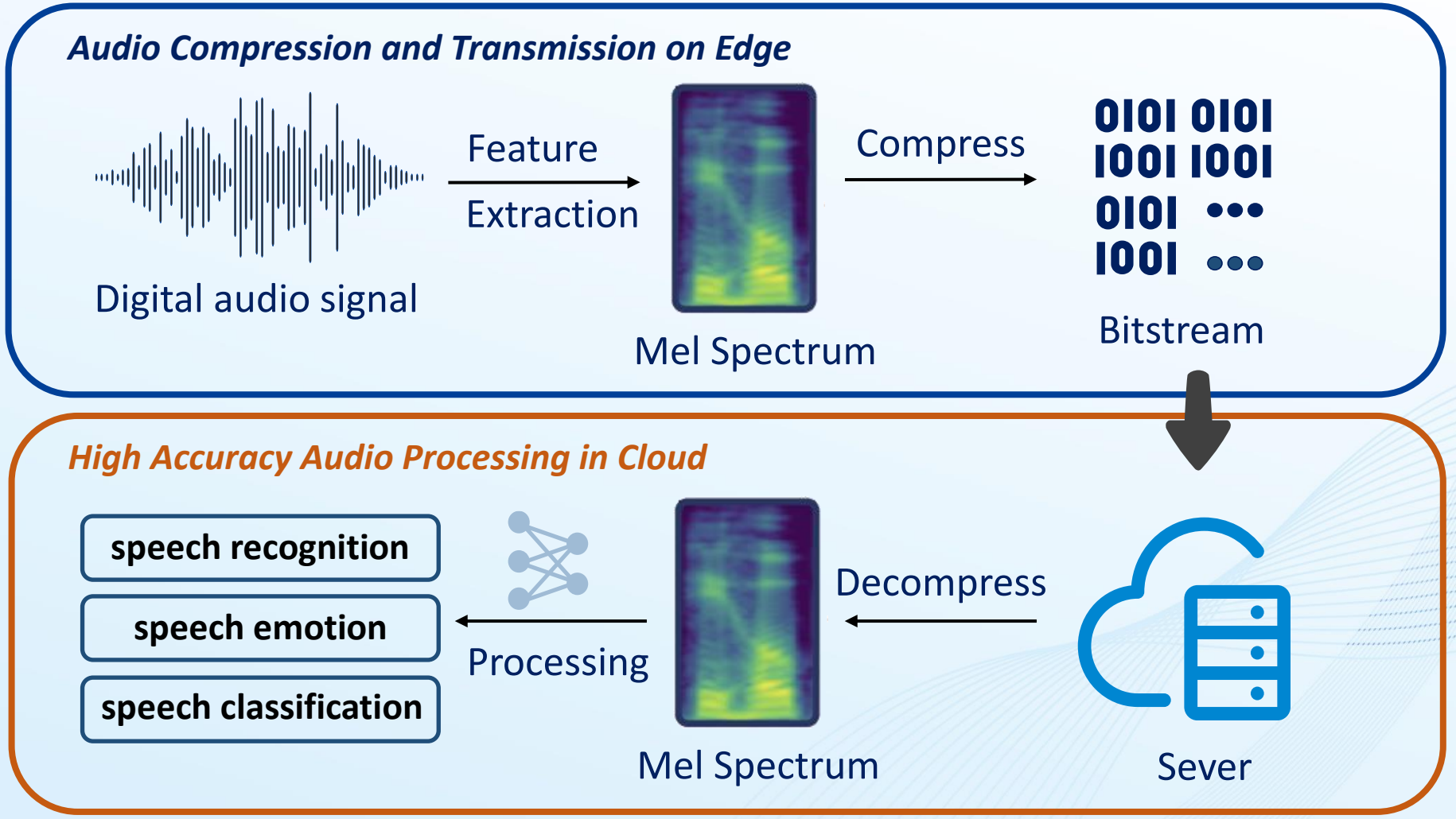
Solution: Audio Compression, Traditional Compress Methods (MP3, FLAC, …).



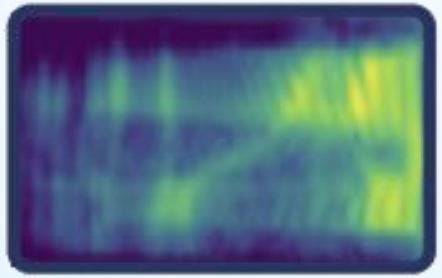
Intuition: Most DNN models for audio tasks extract Mel features at first.



Our Solution: Mel Feature Compression.

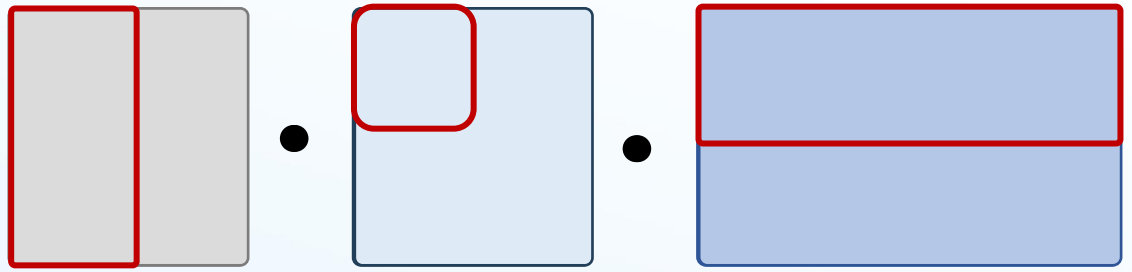


Our Compression Method: SVD Decomposition.



Mel Features $M \in R^{80\times T}$

Reduced SVD

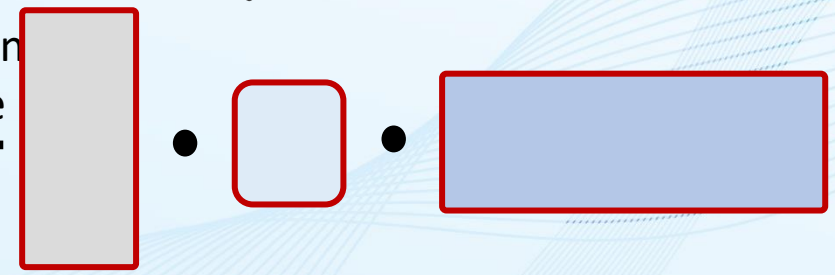


$U \in R^{80\times k}$
$k = \min(80, T)$

$diag(S)$
$S \in R^{k}$

$V \in R^{k\times T}$

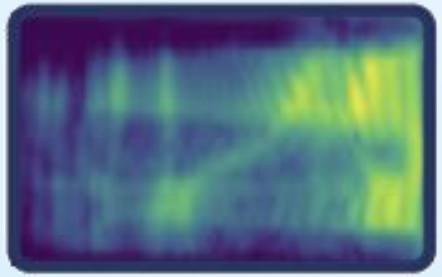
γ *is compression rate* $k_{new} = k \cdot \gamma$



$U_{\text{new}} \in R^{80\times k_{new}}$ $diag(S_{new})$ $V_{\text{new}} \in R^{k_{new}\times T}$
$S_{new} \in R^{k_{new}}$

Transmission and Storage

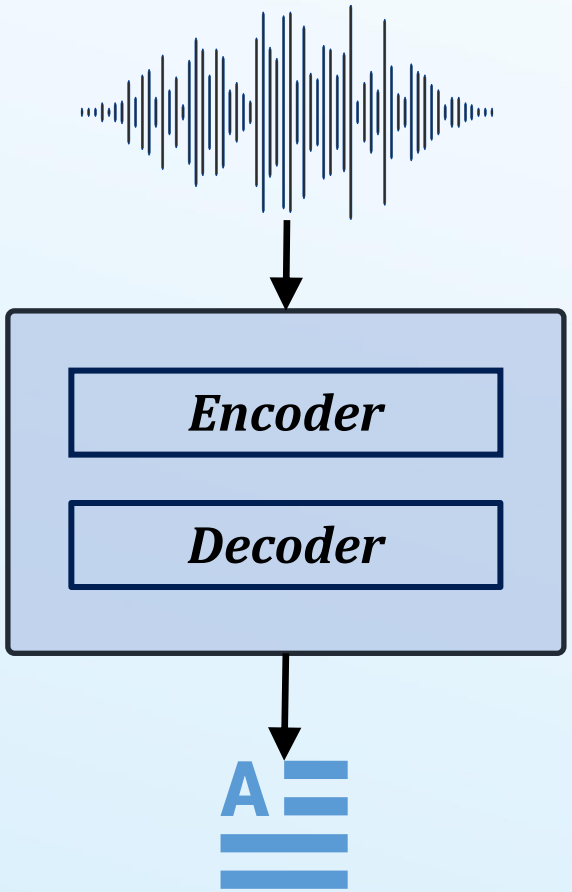
Inversed SVD



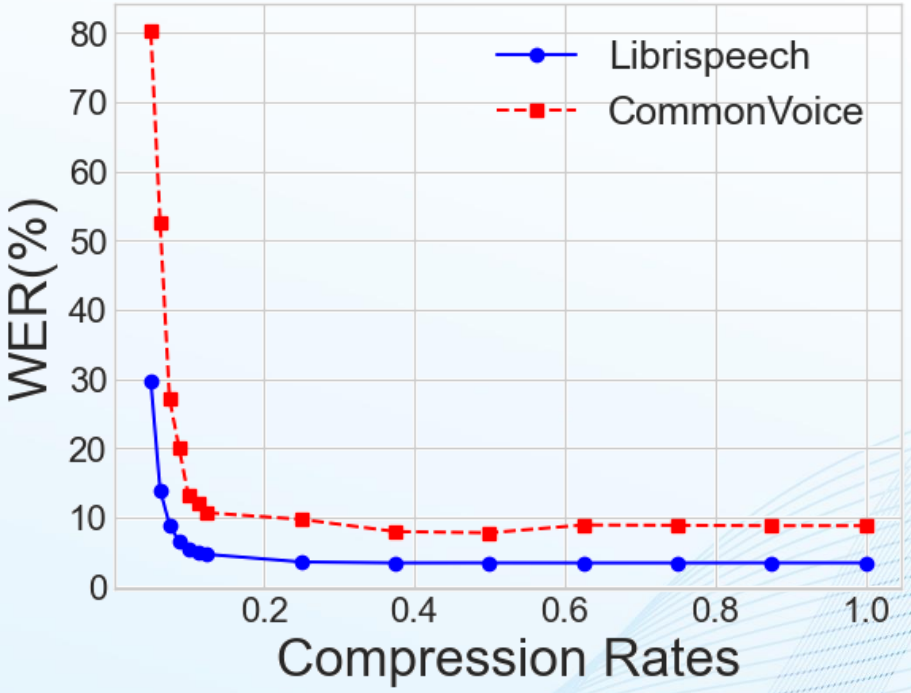
Decompressed Mel Features

$M_{new} = \mathrm{U}_{\text{new}} \cdot diag(S_{new}) \cdot V_{new}, M_{new} \in R^{80\times T}$

Evaluation: Speech Recognition.

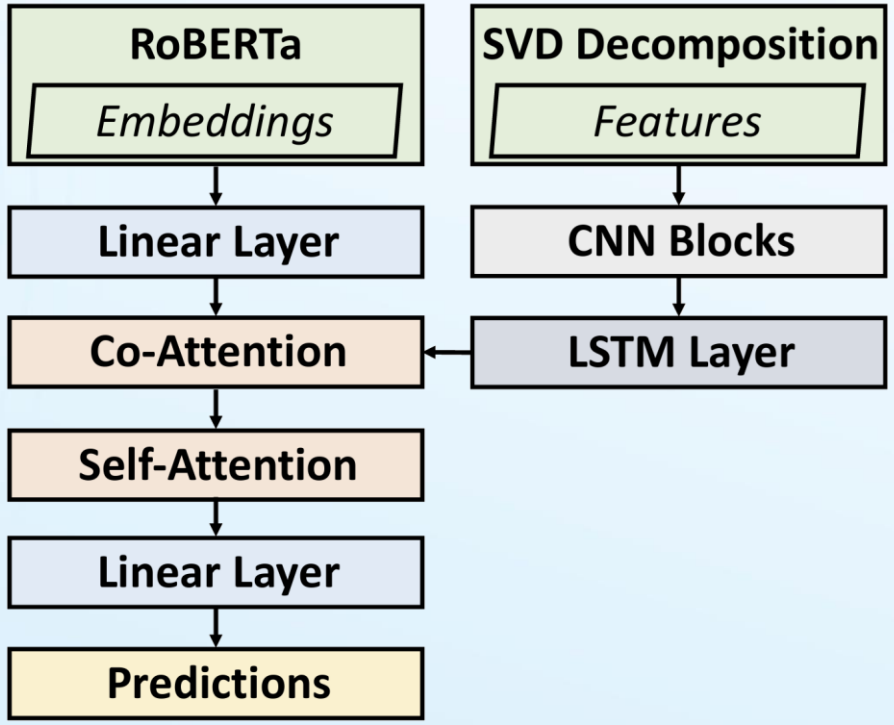


ASR Model： Whisper

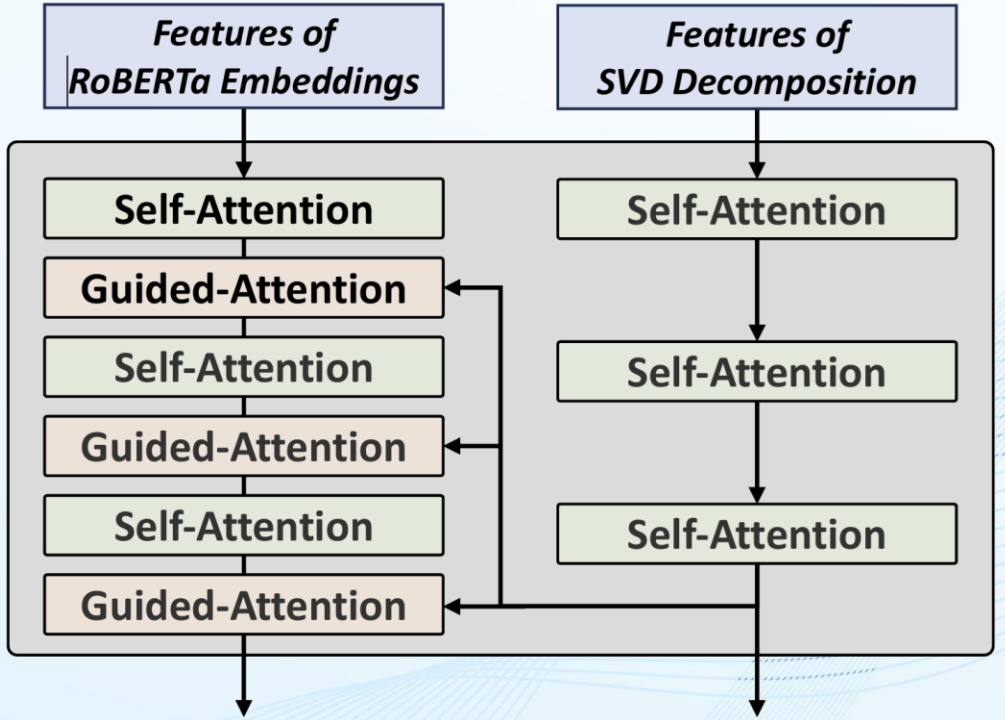


In the Whisper mode, WER inversely correlates with compression rates across the two datasets.

SER Model: Multimodal fusion network.



New speech emotion recognition model



Co-attention Architecture

Evaluation: Speech emotion recognition and system overhead.

Compress Rate	Ours	UMONS	Xu	Yoon
12.50%	**0.69833**	0.67840	0.63343	0.55523
18.75%	**0.69540**	0.67644	0.63636	0.55914
25.00%	**0.69735**	0.67644	0.63742	0.56207
50.00%	**0.69840**	0.67742	0.63832	0.56891
100.00%	**0.70126**	0.67644	0.64321	0.58260

Accuracy on IEMOCAP of our model and other emotion recognition model

Compress Rate	Ours	UMONS	Xu	Yoon
12.50%	**0.69786**	0.67713	0.62987	0.54849
18.75%	**0.69486**	0.67539	0.63329	0.55306
25.00%	**0.69696**	0.67560	0.63298	0.55639
50.00%	**0.69688**	0.67654	0.63487	0.56381
100.00%	**0.70089**	0.67540	0.63968	0.57749

F1 Score on IEMOCAP of our model and other emotion recognition model

Model	Raw	AMSER
transmission time	406.58s	180.75s
transmission energy overhead	0.0056kWh	0.0025 kWh

On-device transmission time and energy consumption for over 20,000 speech files

Thanks for Listening!

Any questions, you can contact:
Dian Ding,
dingdian94@sjtu.edu.cn