Generative Error Correction for Speech-to-Text Transcription and Translation

Zhengdong Yang

June 24, 2026

KYOTO UNIVERSITY

京都大学

Outline

- Introduction
- Challenges & Solutions
  - First-pass Diversity → *CoVoGER*
  - Low-resource Language → *H-Softmax*
  - Emotion → *EmoST*
- Conclusion & Future Prospects

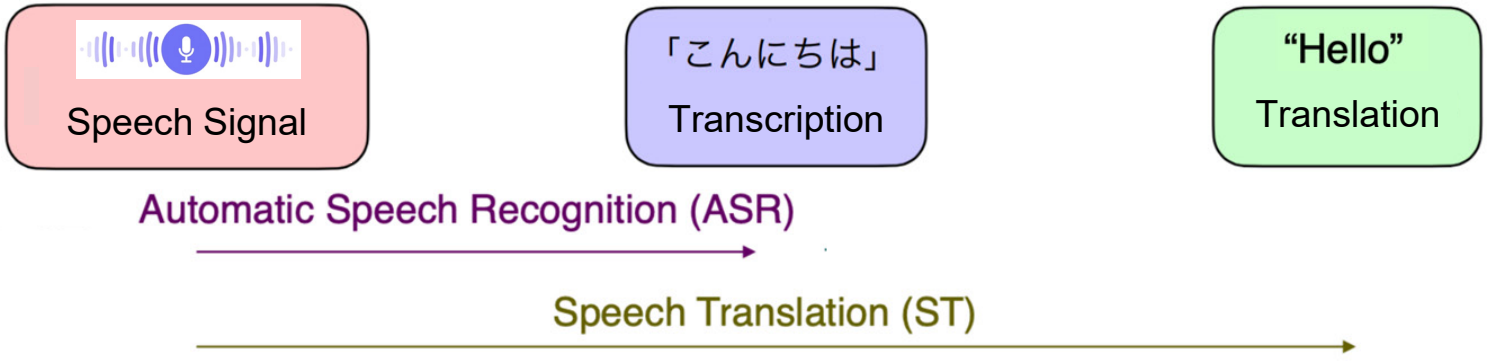
京都大学

Introduction

京都大学

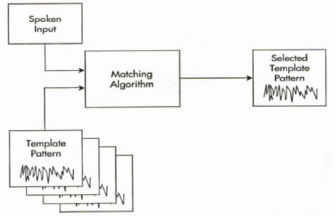
Speech-to-Text: ASR & ST

- Automatic speech recognition (ASR) and speech-to-text translation (ST) offer instant, low-cost transcriptions and translations of human speech

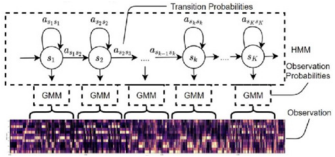


History of ASR and ST

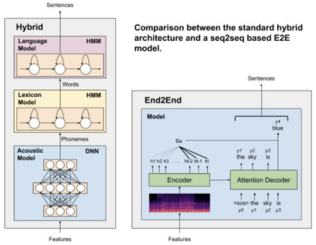
- Methods of ASR and ST have been constantly evolving throughout the history with the development of hardware, data and algorithm:



Template-based ASR



Statistical ASR (HMM–GMM)



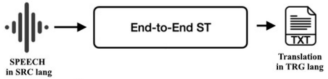
Neural ASR (DNN-HMM & End-to-End)

LLM supply external semantic knowledge to boost ASR/ST performance

LLM-based ASR & ST

Our focus!

Cascaded ST

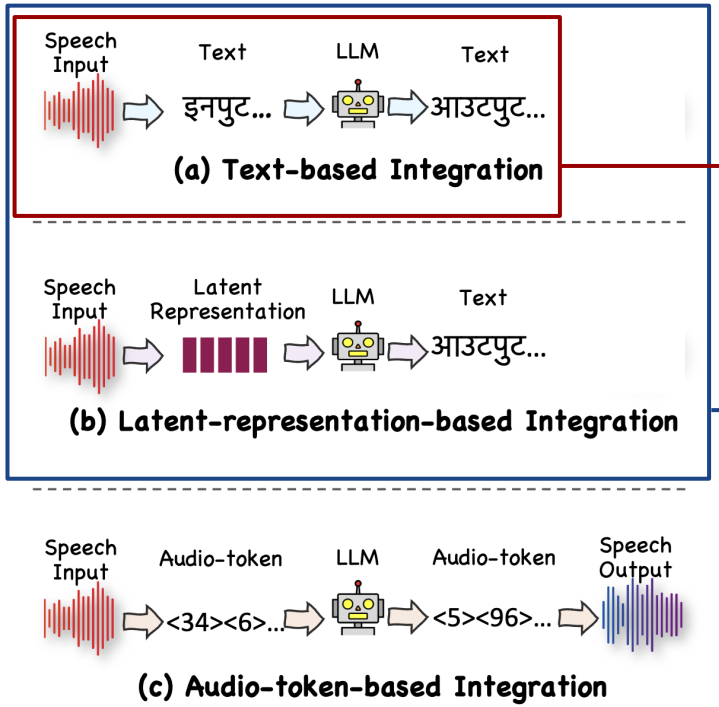


End-to-End ST

京都大学

Speech-LLM Integration

Three approaches for speech-LLM integration



Our focus! (GER)

Better interpretability than (b)

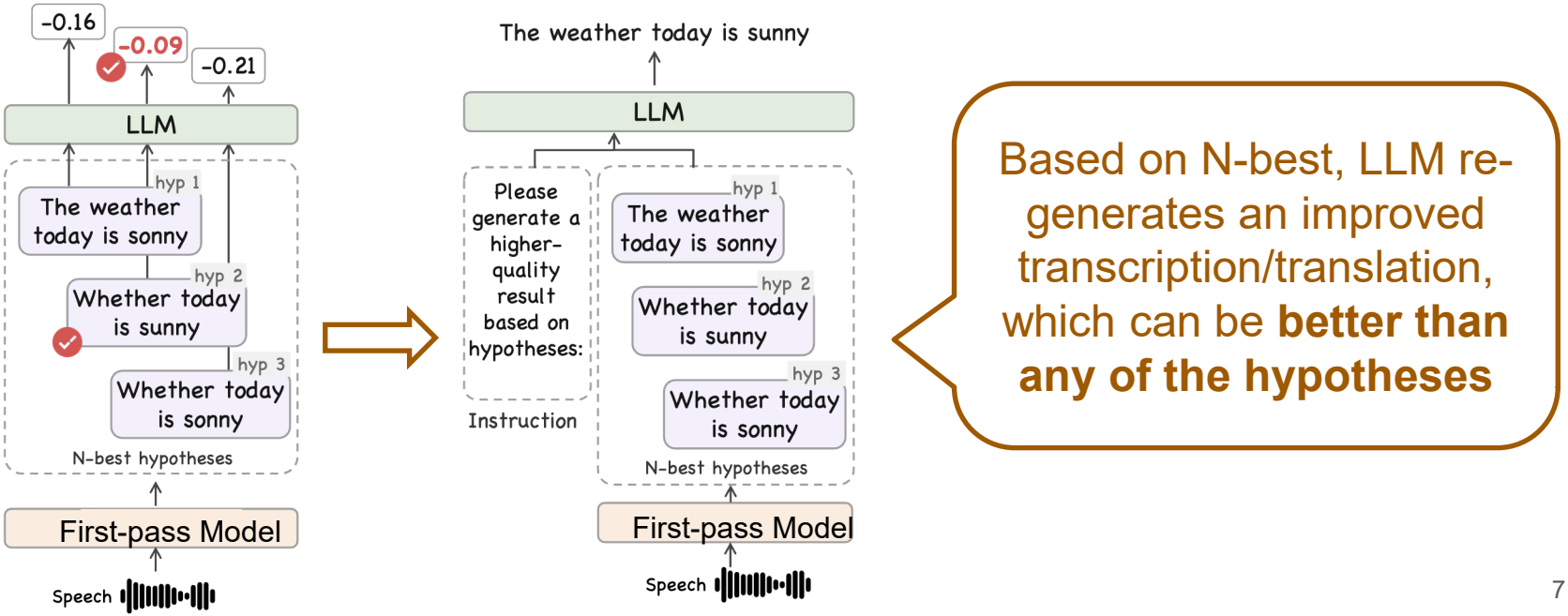
Commonly used for S2T tasks

Commonly used for S2S tasks

京都大学

Generative Error Correction

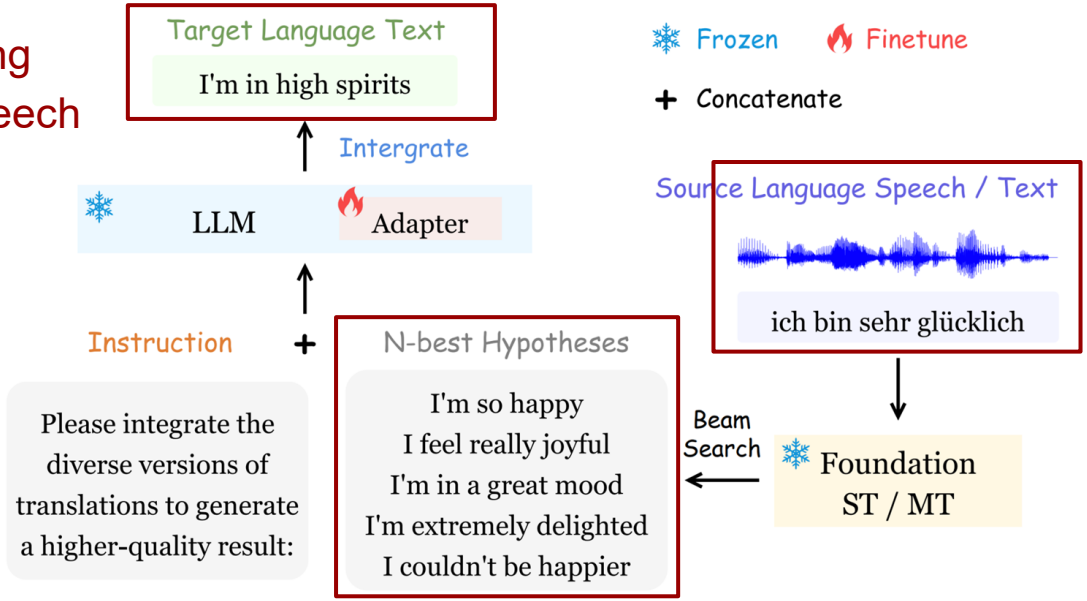
- With the development of LLMs, **Generative Error coRrection** (**GER**) has emerged based on the rescoring mechanism in ASR systems



Existing Challenges in GER

- An example of prior GER study [Hu+, 2024]:

Not considering emotion in speech



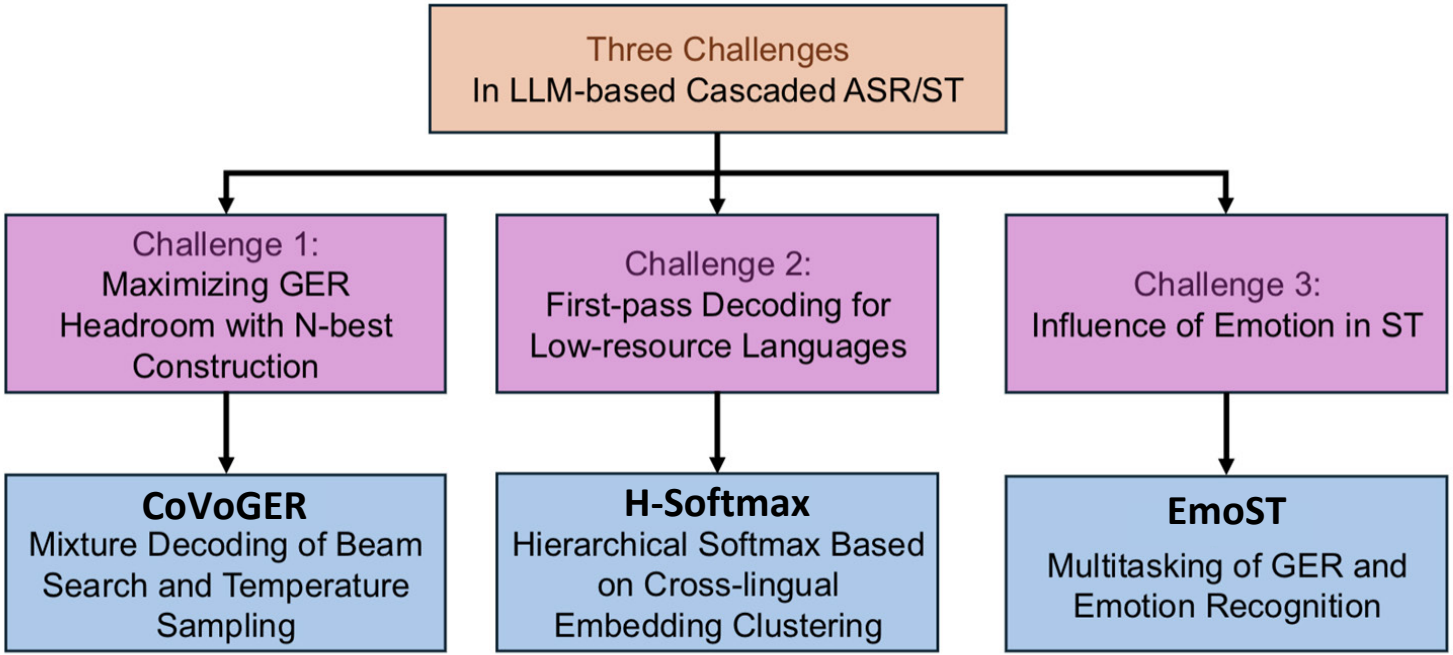
Mainly focus on high-resource languages

Only using beam search for N-best decoding with limited diversity

GenTranslate: Large Language Models are Generative Multilingual Speech and Machine Translators [Hu+, 2024]

京都大学

Our Solutions to Challenges



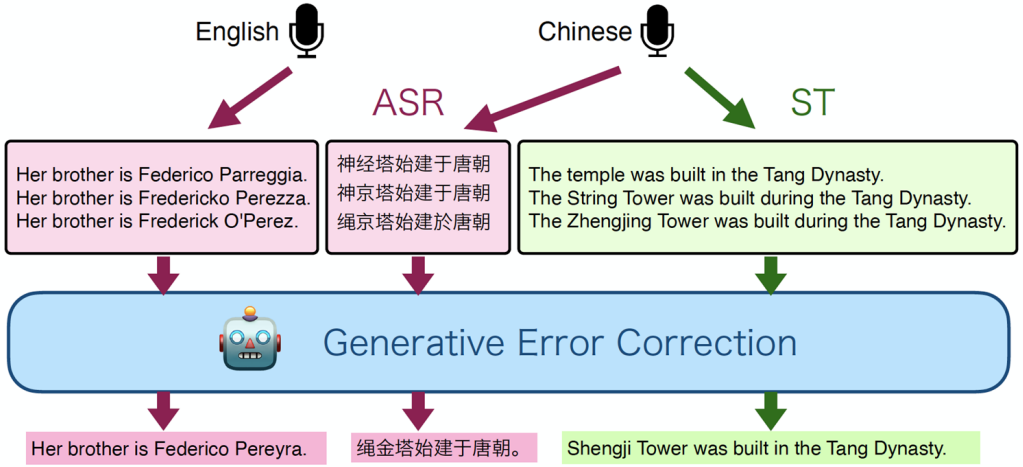
CoVoGER: Multilingual Multitask Benchmark for Speech-to-text Generative Error Correction

EMNLP 2025

京都大学

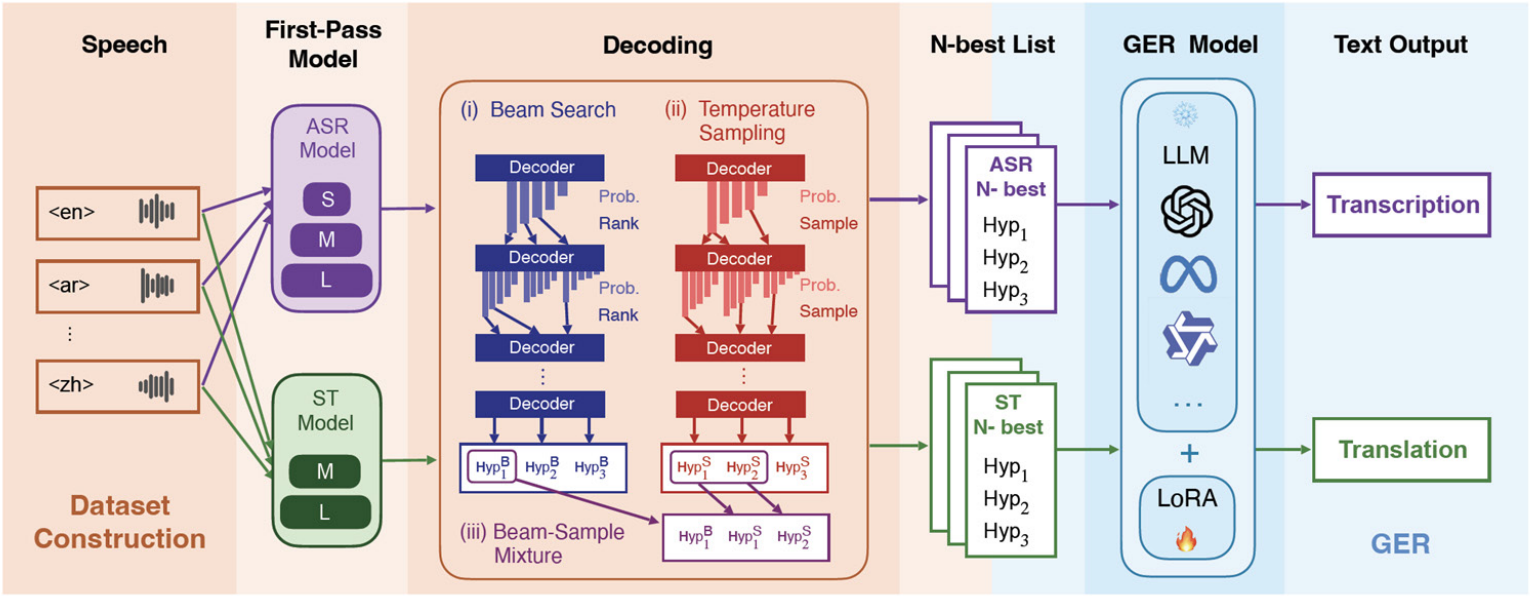
Motivation

- Most GER studies concentrate on English
- Most studies address ASR & ST in isolation, overlooking their synergies
- Previous GER studies only adopt beam search for N-best decoding, leading to limited diversity



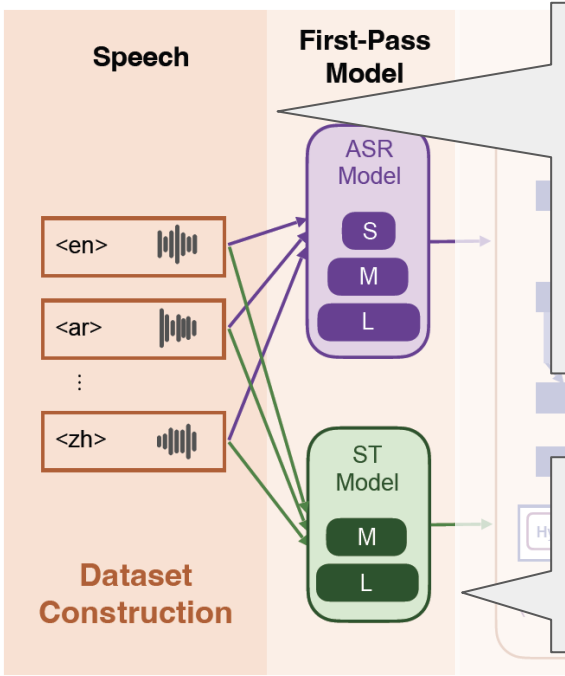
Example of multilingual multitask GER system

Overview of CoVoGER



We constructed a dataset with various first-pass setups for multilingual ASR/ST GER evaluation

Source Data & First–Pass Models

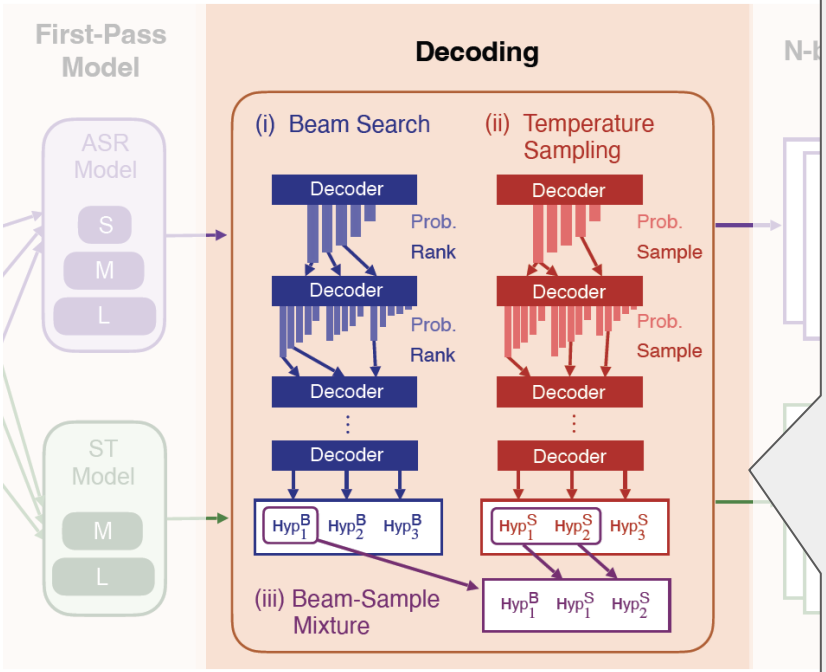


Decode speech from 2 datasets:

- **CoVoST-2 for ST:**
  Select 14 bidirectional pairs (En-X, X-En)
- **Common Voice 20.0 for ASR**:
  Select 15 languages corresponding to the ST data

N-best lists are generated by **Whisper** and **SeamlessM4T** with different sizes

First-Pass Decoding Strategies



We examine two complementary decoding strategies and their combination:

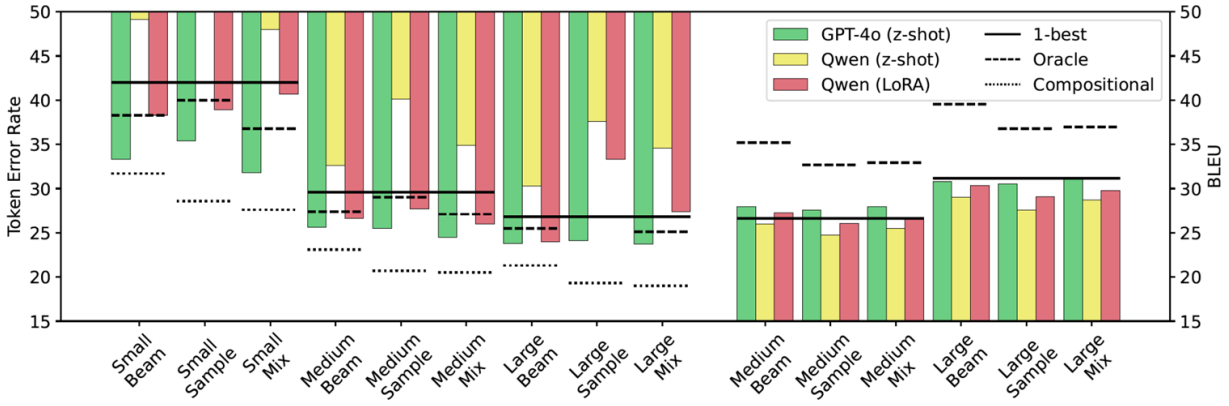
- Beam search (**Beam**):
  High accuracy, low diversity
- Temperature sampling (**Sample**):
  Low accuracy, high diversity
- Beam–sampling mixture (**Mix**):
  1-best of Beam + other candidates of Sampling, balanced accuracy & diversity

(Temperature tuned on oracle TER/BLEU:
0.8 gives the most GER-friendly N-best)

京都大学

Experiments: GPT-4o & Qwen

- Commercial vs open model on a small valid set:



- GPT-4o (zero-shot) > Qwen (LoRA) > Qwen (zero-shot)
- Mix holds the advantage over Beam for GPT-4o

  (significantly better on small/medium ASR, comparable on other settings)

  unlike Qwen (Beam is the best overall)

  -> Stronger GER model can better utilize the diversity

京都大学

Experiments: Multi-task Training with Various LLMs

- **7 open models** with LoRA on full test set: Qwen2.5-7B-Instruct, Qwen2.5-7B, Qwen2.5-3B-Instruct, Meta-Llama-3-8B-Instruct, DeepSeek-R1-Distill-Llama-8B, Platypus2-7B, Falcon3-7B-Instruct
- N-best lists decoded with **large models / Mix decoding** for both ASR / ST

ASR (TER↓)

GER	AVG
Q2.5-7B-i	28.2
Q2.5-7B	26.5
Q2.5-3B-i	29.7
L3-8B-i	**26.3**
DS-8B	27.8
P2-7B	26.4
F3-7B-i	29.3

ST (BLEU↑)

GER	X–En	En–X	AVG
Q2.5-7B-i	36.03	32.12	34.08
Q2.5-7B	**36.51**	32.15	34.33
Q2.5-3B-i	35.79	31.56	33.68
L3-8B-i	36.32	32.64	34.48
DS-8B	36.00	31.89	33.95
P2-7B	36.34	**33.18**	**34.76**
F3-7B-i	35.66	29.80	32.73

- For Qwen, No-instruct > Instruct, 7B > 3B
- Llama & Platypus are better models
- Fail to outperform single-task finetuning (TER 27.8, BLEU 34.09 for Qwen2.5-7B-i)

京都大学

H-Softmax: Cross-lingual Embedding Clustering for Hierarchical Softmax in Low-resource Multilingual Speech Recognition

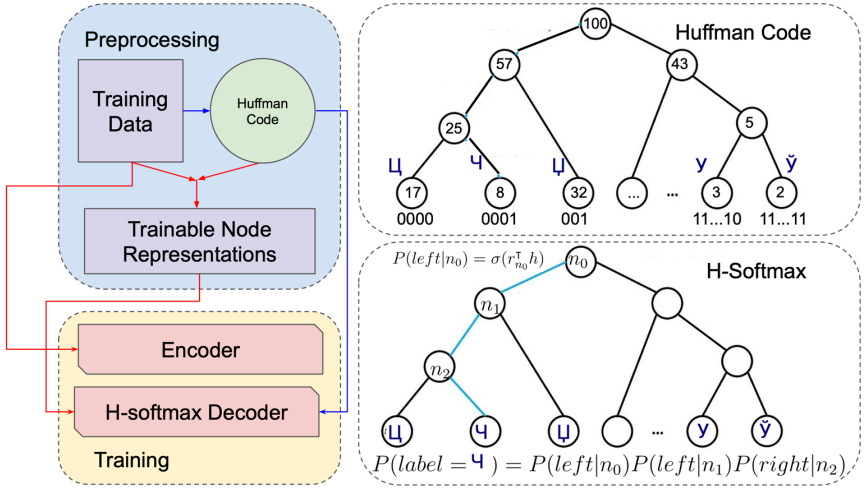
TASLP (2025)

京都大学

Motivation

- ASR for low-resource languages can be improved by **transferring knowledge across different language**

- Previous study [Liu+, 2023] proposes cross-lingual representations for decoding stage, by leveraging **Huffman code** to construct tree structure for **Hierarchical Softmax (H-Softmax)**
  →However, shallow features like **token frequency** could struggle to effectively capture cross-lingual correlations



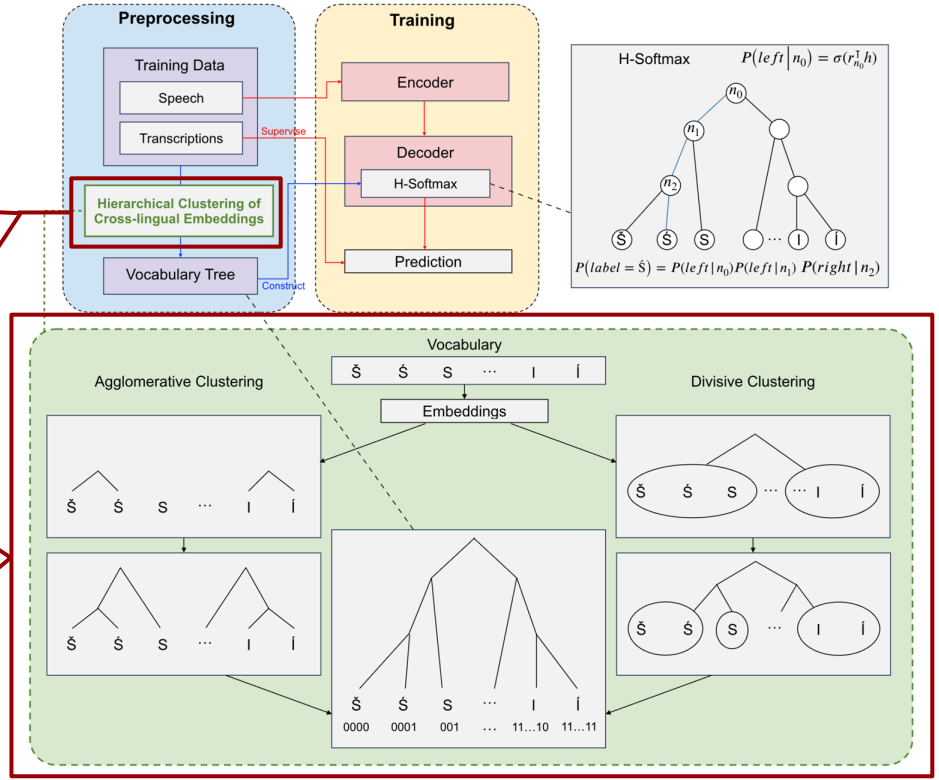
Hierarchical Softmax for End-to-End Low-resource Multilingual Speech Recognition [Liu+, 2023]

京都大学

Proposed Method

We propose to utilize **cross-lingual embedding**, which contains richer semantic information, to construct the tree structure for H-Softmax

Hierarchical clustering is adopted to build the binary tree based on cross-lingual embeddings



Dataset Simulating Low-resource Languages

- We created the dataset based on **Common Voice Corpus 11.0**
- We selected **15 languages** from **3 language families**
- To simulate low-resource scenario, we **downsample** the dataset to 10~80 hours for each language

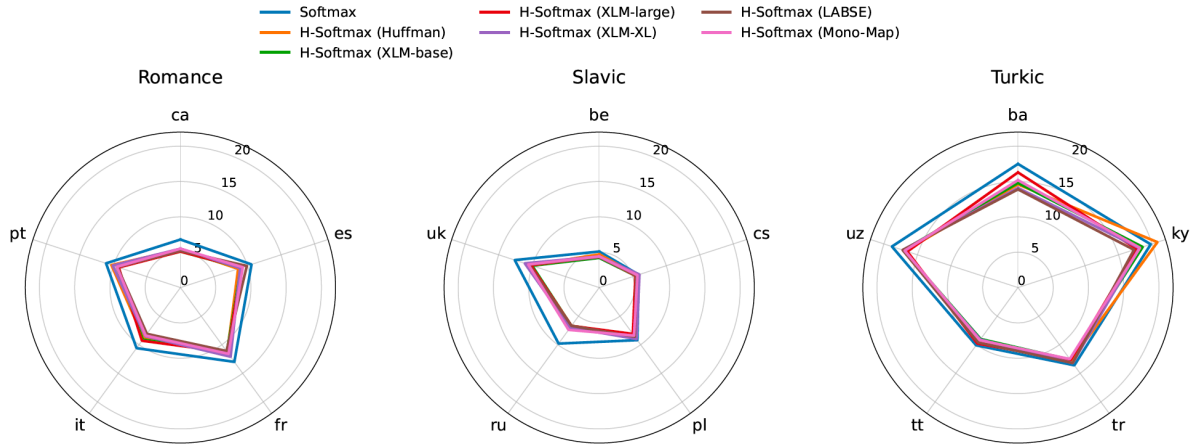
Group	Language	Training	Validation	Test
Romance	Catalan (ca)	76.3	9.5	15.3
	Spanish (es)	37.6	4.7	14.4
	French (fr)	55.7	7.0	13.4
	Italian (it)	33.4	4.2	15.0
	Portugal (pt)	20.7	2.6	11.4
Slavic	Belarusian (be)	63.0	7.9	13.9
	Czech (cs)	42.5	5.3	5.8
	Polish (pl)	37.8	4.7	12.3
	Russian (ru)	27.3	3.4	14.5
	Ukrainian (uk)	23.4	2.9	7.2
Turkic	Bashkir (ba)	29.6	3.7	12.6
	Kyrgyz (ky)	11.3	1.4	3.8
	Turkish (tr)	16.9	2.1	8.4
	Tatar (tt)	10.0	1.2	3.0
	Uzbek (uz)	18.1	2.3	10.4

Data statistics in hours

京都大学

Experiments: First-Pass

On test set, training with 5 languages:



Training	Test	Softmax	Huffman	H-Softmax Embedding: XLM-base 2-med S-Euc	XLM-large 2-med S-Euc	XLM-XL Wtd Corr	LABSE Median Euc	Mono-Map Avg CB
CER↓ Global		11.1	9.5†	**9.3**†‡	9.4†	9.6†	**9.3**†‡	9.5†
CER↓ Average		11.6	10.3	**10.0**	10.1	10.3	**10.0**	10.2
WER↓ Average		29.3	26.4	25.5	25.8	26.8	**25.0**	26.0

- Two of the embedding-based methods (XLM-base, LABSE) **outperforms** the Huffman baseline with significance
- Larger embedding models fails to improve
  ← Possibly due to tokenization mismatch (character for ASR, sub-word for embedding)

京都大学

Experiments: GER

On test set, training with 5 languages:

	First-pass			GER		
	Softmax	Huffman	LABSE	Softmax	Huffman	LABSE
Average	11.6	10.3	**10.0**	14.0	12.1	**10.7**

CER↓

	First-pass			GER		
	Softmax	Huffman	LABSE	Softmax	Huffman	LABSE
Average	29.3	26.4	**25.0**	27.8	25.6	**23.5**

WER↓

Reference	quando o mérito passa pelo jardim entre na lenha atire e feche a porta
First Pass	quando o mérito passa pelo jardim entrei na lenha atire e feche a poita
GER	quando o mérito passa pelo jardim entre na lenha atire e feche a poita
Reference	a propriedade é nove dezenas dos nossos
First Pass	a propriedade é nove desenas dos nossos
GER	a propriedade é nove desenhos dos nossos

Case study (LABSE) (pt)

- **CER** do **not** consistently improve after GER
- However, evaluating with the **word error rate (WER)** demonstrates GER improvement
- This divergence indicates that LLM-based GER predominantly resolves **word-level, semantic mistakes** rather than character-level confusions

京都大学

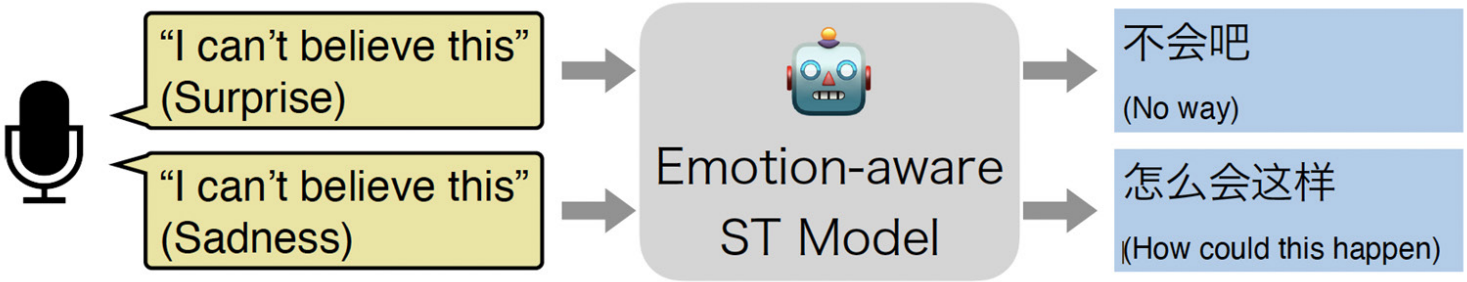
EmoST: Generative Error Correction for Emotion-aware Speech-to-Text Translation

ACL 2025 Findings
JNLP (2026)

京都大学

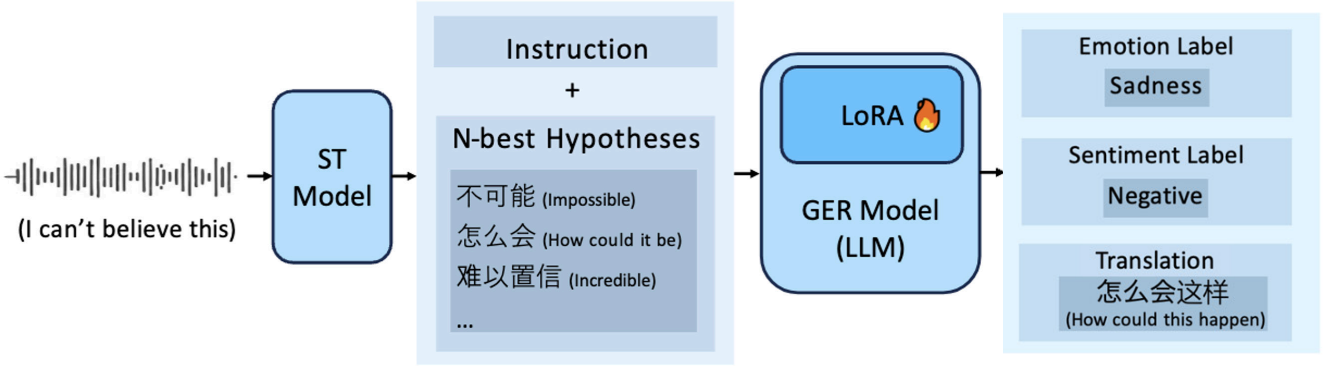
Motivation

- Human speech naturally includes **emotions,** can **significantly influence** ST result
- However, **emotion in ST** has been **overlooked**
- **Challenge**: Speech-text bilingual parallel data is **scarce**, even scarcer with emotion annotations
- **Solution**: Leveraging external models like **LLMs** to help



Integrating Emotion Labels

Label-only Integration

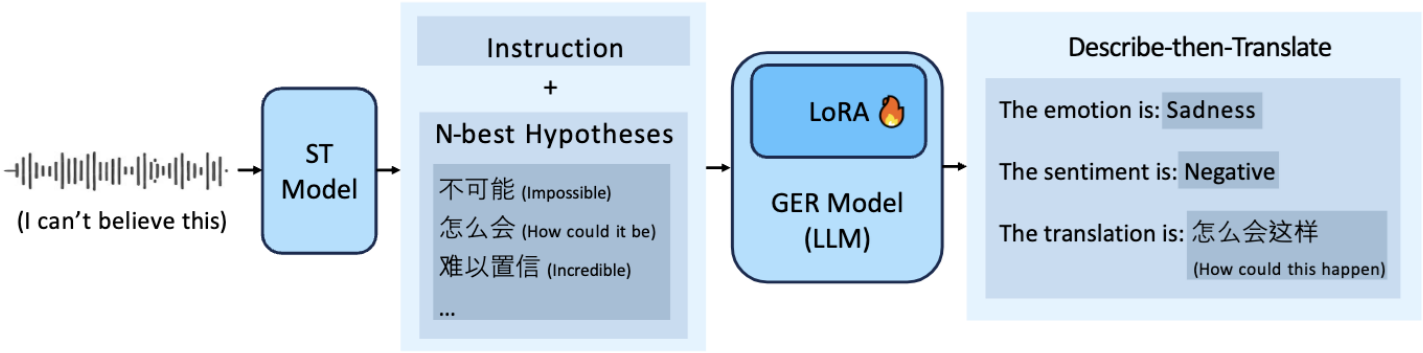


- Propose using the GER model to directly predict **emotion label + translation** auto-regressively
- Can be considered as a type of **multitask learning**

京都大学

Integrating Emotion Labels

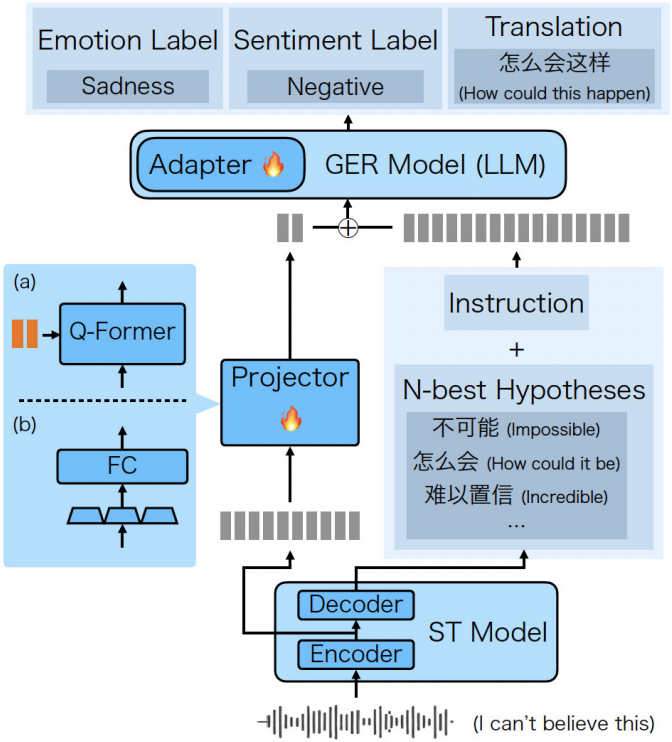
Describe-then-Translate (DtT)



- Simply concatenating discrete emotion labels & translations
  → **an overly coarse design**: Weakly aligned with the **LLM's next-token training objective**, Emotion may fail to influence the translation
- We propose replacing discrete labels with a short **natural-language sentence** predicting the perceived emotion prior to translation

京都大学

Injecting Acoustic Representation



- Only textual N-best hypotheses
  → **lose cues for emotion prediction**
- **Solution:**
  Inject the ST encoder's **acoustic representation** into GER model
- Explore 2 projectors:
  - Q-Former
  - 1-D Convolution Downsampling

京都大学

Experiments: Label-only Integration

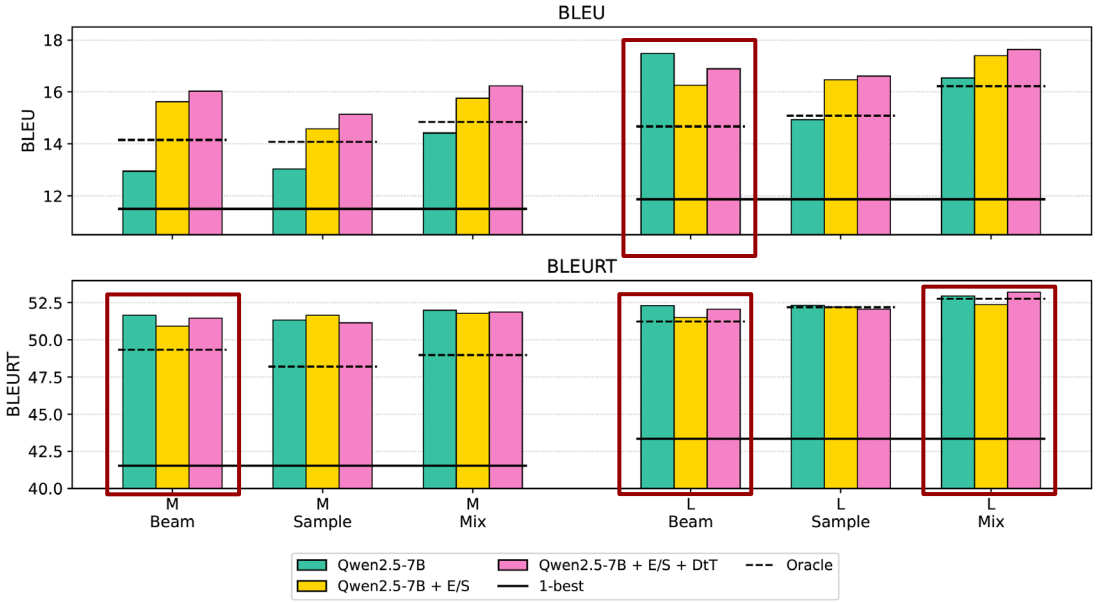
	GER	E/S Labels	Projector	BLEU	BLEURT
SeamlessM4T		-	-	11.87	43.34
Ours	✓	-	-	15.54†	51.57†
	✓	GER Outputs	-	15.61†	51.81†
	✓	GER Outputs	Q-former	15.91†‡	51.86†
	✓	GER Outputs	Conv1D	**15.97**†‡	**52.07**†‡
Ours (Upper-bound)	✓	GER Inputs	-	16.28†‡	52.50†‡

On BMELD dataset (en-zh), with SeamlessM4T-Large + Llama-2-7B

- **GER** >>> No GER
- **GER with emotion** ≈ GER
- Emotion as inputs (upperbound) > **Emotion as outputs with projector** > Emotion as outputs

京都大学

Experiments: Describe-then-Translate



- **DtT** brings **significant improvement** over label-only on 4 comparisons (L+Beam BLEU, M+Beam BLEURT, L+Beam/Mix BLEURT), and comparable results on others
- The effectiveness of Mix decoding is confirmed

On BMELD dataset, with SeamlessM4T-Medium/Large + Qwen2.5-7B

京都大学

Conclusion & Future Prospects

京都大学

Conclusion

- We tackled three main challenges in LLM-based GER for speech-to-task recognition and translation:
  - Constructing informative N-best lists
  - Building efficient multilingual first-pass decoders under data scarcity
  - Achieving emotion-aware GER for ST
- Overall, we delivers a practical recipe for LLM-based GER: Generate diverse but plausible hypotheses, use structured sharing in first-pass decoding, and condition correction on emotion when translating speech

京都大学

Future Prospects

- End-to-End Training of Decoding Policies with GER Supervision
  - Instead of tune first-pass decoding by hand, parameterizing the decoding setup and learn it jointly with GER
- Bridging Text-Based and Latent-Representation-Based Integration
  - A systematic comparison remains missing under matched backbones, data, and budgets

京都大学