Machine Translation Marathon 2024 Talks

Speech Translation: From Basics to Recent Advances

Tsz Kin Lam

School of Informatics
University of Edinburgh

Outline

1. **Introduction: What to expect**
2. **I: How is the translation of speech different from its text?**
3. **II: Existing approaches**
4. **Summary and the future works**

Introduction: What to expect

What to expect:

- data and modelling aspects of ST for MT researchers who are interested in the translation of speech
- a recent development of ST in the field, e.g., integrating speech into LLM

Introduction: What to expect

What to expect:

- data and modelling aspects of ST for MT researchers who are interested in the translation of speech
- a recent development of ST in the field, e.g., integrating speech into LLM

What is not included:

- not much about speech-to-speech translation
- specific applications, e.g., simultaneous ST, subtitling and dubbing

Introduction: What to expect

What to expect:

- data and modelling aspects of ST for MT researchers who are interested in the translation of speech
- a recent development of ST in the field, e.g., integrating speech into LLM

What is not included:

- not much about speech-to-speech translation
- specific applications, e.g., simultaneous ST, subtitling and dubbing
- (linguistic) analysis of the ST errors
- the mathematical details, e.g., Fast-Fourier Transform (FFT) and Connectionist Temporal Classification (CTC)

I: How is the translation of speech different from its text?

Speech translation is cross-lingual and cross-modal

Unique properties of speech signals

- Speech signal is sparse, i.e., low information content per unit time (An audio file of 2.2 seconds in 16kHz has ≈ 35K time steps).

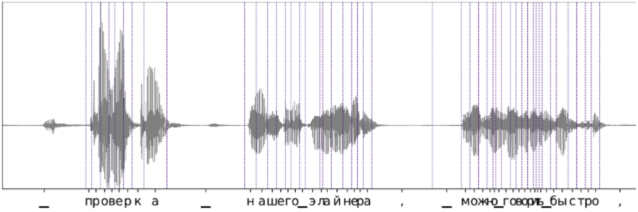


Figure: Speech-to-text alignment [Barrault et al. 2023]

Unique properties of speech signals

- Speech signal is sparse, i.e., low information content per unit time (An audio file of 2.2 seconds in 16kHz has $\approx$ 35K time steps).

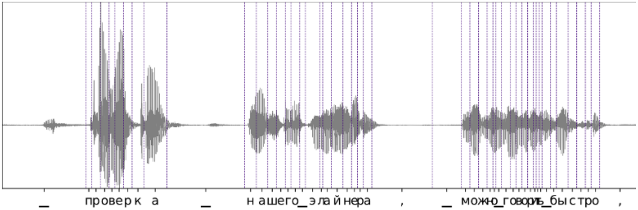


Figure: Speech-to-text alignment [Barrault et al. 2023]

- The file format matters, e.g., the sampling rate and the bit depth.

Unique properties of speech signals

- Speech signal is sparse, i.e., low information content per unit time (An audio file of 2.2 seconds in 16kHz has ≈ 35K time steps).

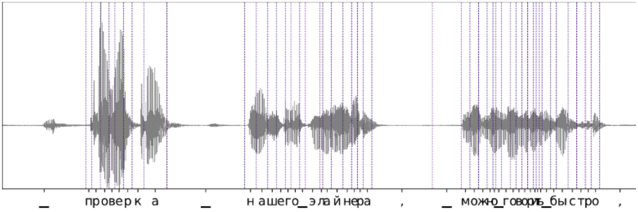
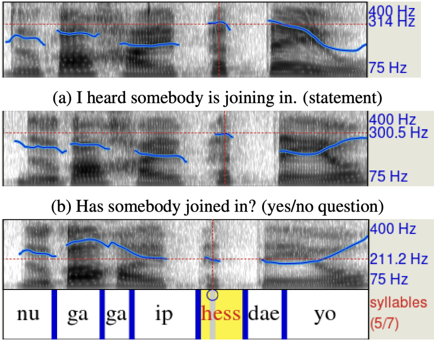


Figure: Speech-to-text alignment [Barrault et al. 2023]

- The file format matters, e.g., the sampling rate and the bit depth.
- Background noises may appear in the speech.

Unique properties of speech signals

- Paralinguistic signals, such as prosody and accents, matter



(a) I heard somebody is joining in. (statement)

(b) Has somebody joined in? (yes/no question)

(c) Who joined in? (*wh*-question)

Figure: Prosody is crucial in the translation of Korean speech [Zhou et al. 2024]

Unique properties of speech signals

- Speech is often disfluent, esp. in spontaneous speech:

Hesitation	eh, eh, eh, um, yo pienso que es así.
	uh, uh, uh, um, i think it's like that.
Repetition	Y, y no cree que, que, que,
	And, and I don't believe that, that, that
Correction	no, no puede, no puedo irme para ...
	no, it cannot, I cannot go there ...
False start	porque qué va, mja ya te acuerda que ...
	because what is, mhm do you recall now that ...

Figure: Types and examples of disfluency [Salesky et al. 2018]

ST is low-resource

Data	X-Y	#utterances	#words (src+tgt)
MuST-C	En-Fr	280K	10.6M
	En-De	234K	8.3M
CoVoST2	Fr-En	207K	4M
	De-En	127K	2M
(MT) Wiki-Matrix	De-En	6.2M	196M

Table: Training data statistics of two common S2TT data and a MT data

ST is low-resource: CoVoST-2

- Based on CommonVoice (v4) $\implies$ read speech
- The sentence structure is simple (De-En: 127K sentences with only 2M words (src+tgt):
  1. "I am going to shower now."
  2. "I am happy when i can make others happy."

ST is low-resource: CoVoST-2

- Based on CommonVoice (v4) $\Longrightarrow$ read speech
- The sentence structure is simple (De-En: 127K sentences with only 2M words (src+tgt):
  1. "I am going to shower now."
  2. "I am happy when i can make others happy."
  3. "Punishments of this kind are a means of targeted terror, if they are carried out in such a way as to have an effect on the public."
  4. "The plans of the head of the municipal planning and building control office Erich Heinicke will be defining for the townscape of the post-war era."

ST is low-resource: MuST-C

- Based on English TED talks $\implies$ more realistic

ST is low-resource: MuST-C

- Based on English TED talks $\implies$ more realistic
- Inconsistent annotations, existence of non-verbal symbols, segmentation error...:

```
[En] we are 12 billion lightyears from the edge
[De] ♫ Wir sind 12 Milliarden Lichtjahre entfernt vom Rand ♫

[En] ♫ everyones out in merry manhattan in January
[De] ♫ Ganz Manhattan ist draußen und wunderbar – im Januar. ♫

[En] and the second one thats a violin
[De] Und nun den zweiten. (♪ Violine) Das ist eine Geige.

[En] kb thank you
[De] SJ: Oh! (Applaus) KB: Danke.
```

Inference: audio segmentation (I)

Can we translate the entire recorded lecture (audio) in one forward-pass?

1. It is an audio sequence of >60 minutes
2. In training, the sequence length rarely exceed 30[1] seconds.

[1] It is about 3K time steps if log Mel spectrogram features are used

Inference: audio segmentation (I)

Can we translate the entire recorded lecture (audio) in one forward-pass?

1. It is an audio sequence of >60 minutes
2. In training, the sequence length rarely exceed 30[1] seconds.
3. We need to segment the audio sequence into smaller chunks!

[1] It is about 3K time steps if log Mel spectrogram features are used

Inference: audio segmentation (II)

Some common segmentation methods are:

- Length-based, e.g., for every 3s.

Inference: audio segmentation (II)

Some common segmentation methods are:

- Length-based, e.g., for every 3s.
- Content-based, e.g, pause that is detected by voice activity detection.

Inference: audio segmentation (II)

Some common segmentation methods are:

- Length-based, e.g., for every 3s.
- Content-based, e.g, pause that is detected by voice activity detection.
- Hybrid approach that is based on both length-based and content-based.

Inference: audio segmentation (II)

Some common segmentation methods are:

- Length-based, e.g., for every 3s.
- Content-based, e.g, pause that is detected by voice activity detection.
- Hybrid approach that is based on both length-based and content-based.
- Neural-network-based: Supervised Hybrid Audio Segmentation (SHAS) [Tsiamas et al. 2022]

Sentence-level evaluation for S2TT

- Each ST model has their own speech segmentation method, so each model could generate different number of outputs.
- For sentence-level evaluation at IWSLT, we need to re-segment the outputs to match the number of references.

[1] [Matusov et al. 2005]

Sentence-level evaluation for S2TT

- Each ST model has their own speech segmentation method, so each model could generate different number of outputs.
- For sentence-level evaluation at IWSLT, we need to re-segment the outputs to match the number of references.
- The re-segmentation is done by a minimum WER[1] $\Longrightarrow$ re-segmentation error.

[1] [Matusov et al. 2005]

Automatic evaluation metrics

In S2T translation, the automatic metrics are the same as MT:

- n-gram matching: `BLEU` and `chrF`
- neural metrics: `COMET`
  - might require ASR/BT to get the transcripts
  - punctuation insertion or not to the transcripts

[1] [Chen et al. 2023]

Automatic evaluation metrics

In S2T translation, the automatic metrics are the same as MT:

- n-gram matching: `BLEU` and `chrF`
- neural metrics: `COMET`
  - might require ASR/BT to get the transcripts
  - punctuation insertion or not to the transcripts

In S2S translation,

- transcribe and MT-evaluate: `ASR-BLEU` and `ASR-chrF`
- neural metrics: `BLASER`[1]

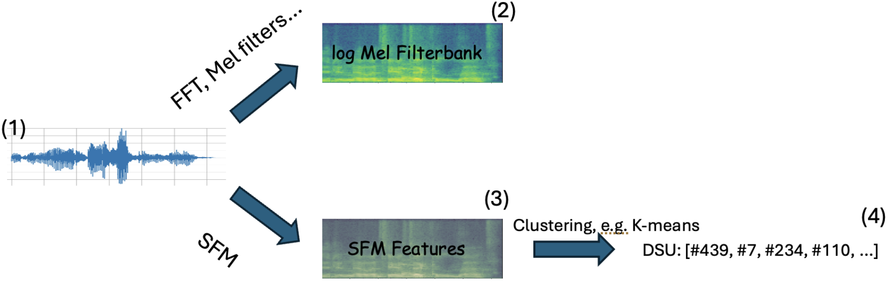
[1] [Chen et al. 2023]

Two major issues

- Data scarcity
- Modality gap between speech and text signal

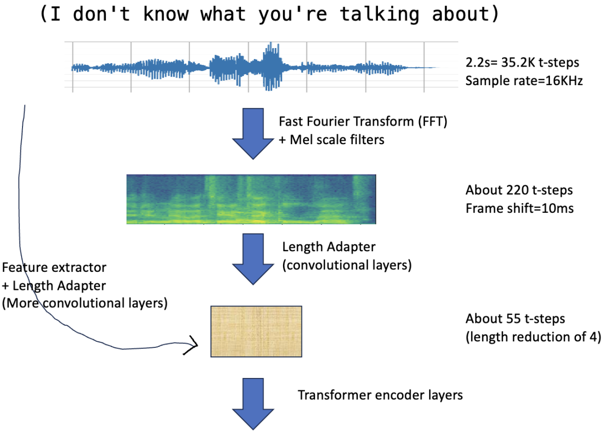
II: Existing approaches

Feeding speech into Transformer: Speech input formats



Feeding speech into Transformer: Length adapter

Recap: Speech is sparse and long.



Cascaded model (I)

Recap: ST is a cross-lingual and probably a cross-modal problem.

- Can we decompose ST into simpler related sub-tasks?

Cascaded model (I)

Recap: ST is a cross-lingual and probably a cross-modal problem.

- Can we decompose ST into simpler related sub-tasks?

Cascaded ST: It converts ST into a task of running ASR and MT tasks sequentially (Text-To-Speech is required in S2S).

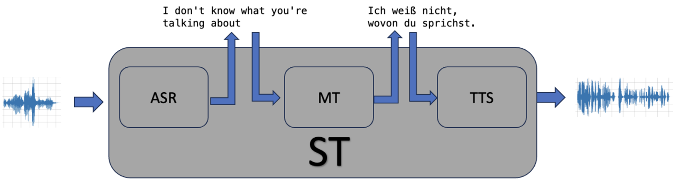


Figure: An illustration of the cascaded ST model.

Cascaded model (II)

Advantage:

- The training is easier since the cross-lingual and -modal parts are learnt independently.
  - There are more training data for the sub-tasks.

Cascaded model (II)

Advantage:

- The training is easier since the cross-lingual and -modal parts are learnt independently.
  - There are more training data for the sub-tasks.
- Output correction and Human-in-the-loop become simpler by inspecting the intermediate transcripts(/translations).

Cascaded model (II)

Advantage:

- The training is easier since the cross-lingual and -modal parts are learnt independently.
  - There are more training data for the sub-tasks.
- Output correction and Human-in-the-loop become simpler by inspecting the intermediate transcripts(/translations).
- It can leverage foundation models easily.

Cascaded model (III)

Disadvantage:

- The translation pipeline is lengthy. This might cause
  - higher inference cost
  - error propagation from the ASR(/MT) model(s).
  - loss of speech information, e.g., prosody in the ASR step.
- Cascaded model is not very parameter efficient.

Direct end-to-end (E2E) model (I)

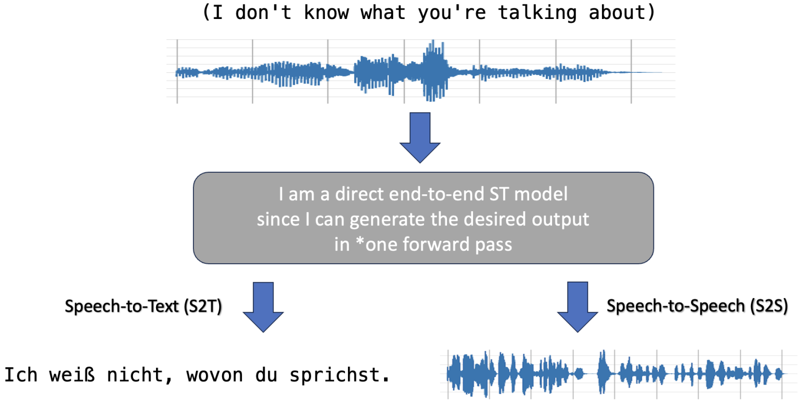


Figure: An illustration of the direct end-to-end (E2E) ST model.

Direct E2E model (II)

Advantage:

- Translation is done in one forward-pass. This helps to
  - give lower latency in translation, e.g., (very important) in real-time speech translation.
  - avoid error propagation.
  - preserve speech information for translation.
- End-to-end ST is more parameter efficient.

Direct E2E model (III)

Disadvantage:

- The amount of paired ST data is limited.
- End-to-end ST model is harder to optimise.

Direct E2E model (III)

Disadvantage:

- The amount of paired ST data is limited.
- End-to-end ST model is harder to optimise.

Regardless, E2E model is the main ~~publication~~ research direction now!

Improving E2E ST: data augmentation (DA)

We can generate more data via related task's model(s) and paired data:

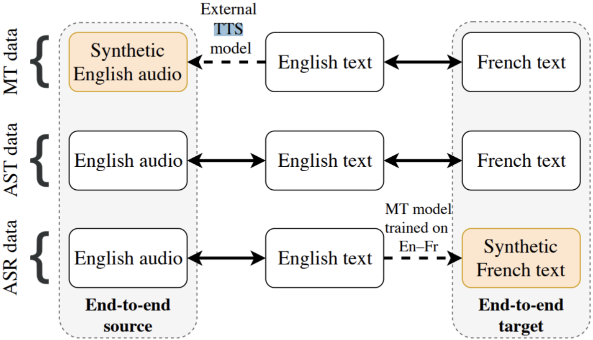


Figure: Pseudo ST data generation [Pino et al. 2019]

Alignment helps, even in data augmentation

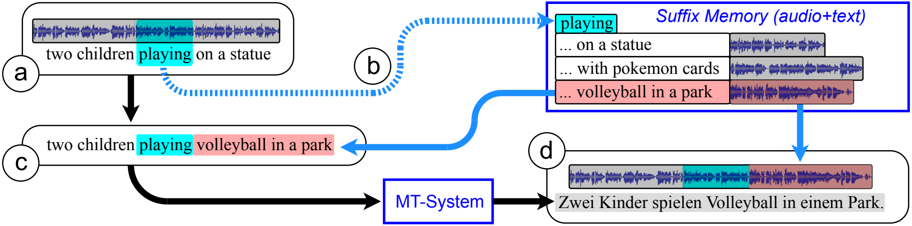
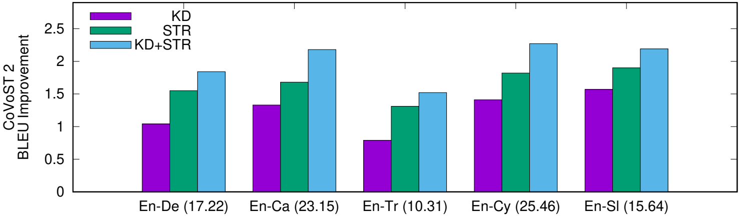


Figure: Acoustic alignment for data augmentation [Lam et al. 2022]

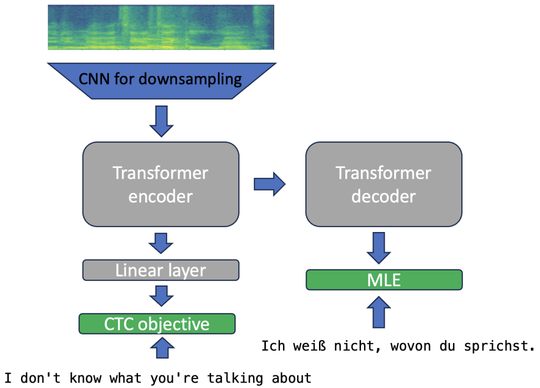
Seq-KD also helps: Results on the CoVoST 2 dataset



- KD: baseline ∪ KD-training data
- STR: baseline ∪ STR-training data
- KD+STR: baseline ∪ KD ∪ STR

Improving E2E ST: multi-task learning

Training ST with other sub-tasks in parallel instead of using them sequentially, e.g., CTC[1] loss on ASR task



[1] Connectionist Temporal Classification [Graves et al. 2006]

Improving E2E ST: using pre-trained models

We can use pre-trained models to initialise the ST model, e.g.,

- wav2vec 2.0[1] for initialising the acoustic encoder
- mBART for initialising the translation decoder

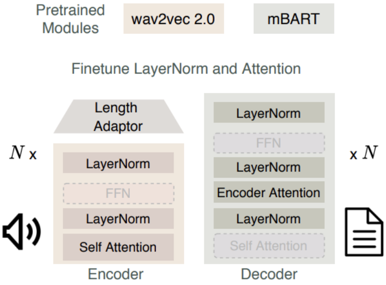


Figure: ST model initialisation via pre-trained SSL models [Li et al. 2020]

[1] Baevski et al. 2020

Mixed modality training

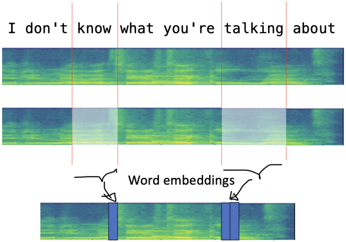


Figure: a sequence of alternating speech and text embeddings.

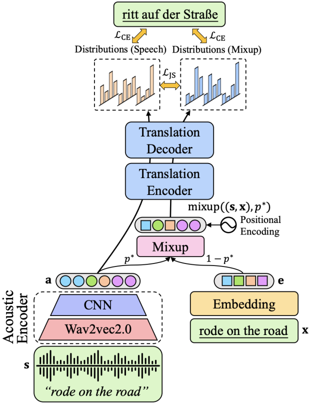
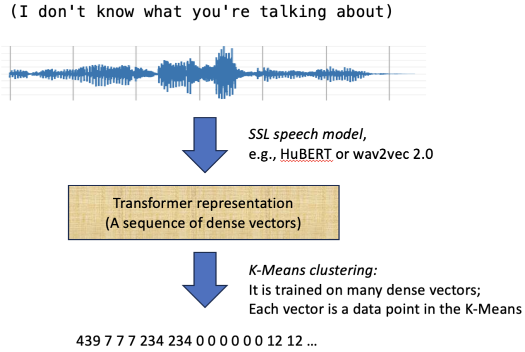


Figure: [Fang et al. 2022]

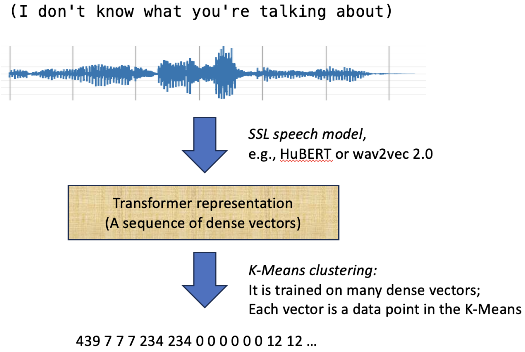
Improving E2E ST: bridging the modality gap (II)

Speech quantisation [Lakhotia et al. 2021]



Improving E2E ST: bridging the modality gap (II)

Speech quantisation [Lakhotia et al. 2021]



DSU are the centroid indexes of the SFM model's dense representations.

Improving E2E ST: bridging the modality gap (II)

Heuristics for length reduction in Discrete Speech Units (DSU):

1. Merging sequential repetitions, e.g.,
   "439 7 7 7 234 234 0 0 0 12 12" $\Longrightarrow$ "439 7 234 0 12"
2. Byte pair encoding, e.g.,
   "439 7 234 0 12" $\Longrightarrow$ "4397234 012"

Improving E2E ST: speech quantisation

Advantages

- Data storage and transmission becomes easier, e.g., can feed more instances to the GPUs

Improving E2E ST: speech quantisation

Advantages

- Data storage and transmission becomes easier, e.g., can feed more instances to the GPUs
- Speech generation becomes more feasible
  - e.g., speech-to-unit, unit-based LM and a unit-based vocoder
  - no text data is required for speech-to-speech translation

Improving E2E ST: speech quantisation

Disadvantages

- The translation pipeline gets lengthy (quantisation and clustering)
- The information lost in quantisation is quite unclear

Improving E2E ST: speech quantisation

Disadvantages

- The translation pipeline gets lengthy (quantisation and clustering)
- The information lost in quantisation is quite unclear
- There are more hyper-parameters to tune, e.g.,
  1. The hyper-parameters in the K-Means model:
     - ★ Its training data size and clustering size.
     - ★ It also require storing the high-dimensional features.
  2. The representation layer of the SSL model/SFM to be used for quantisation

Improving E2E ST: putting all together

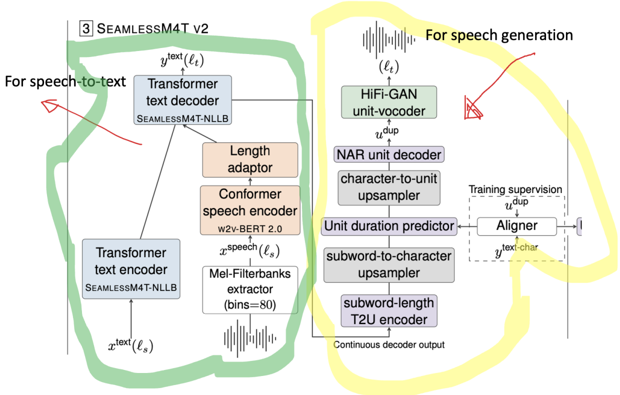


Figure: Seamless-M4T v2 model [Barrault et al. 2023]

II: Existing approaches

Integrating Speech into LLM: Discrete Units or Dense Features?

Speech → LLM: Discrete (Speech) Units

Quantise the speech inputs (choose your DSU symbols wisely),

1. Update the tokenizer, e.g., BPE on the DSU
2. Expand the vocabulary size of your LLM
3. Train the model on the DSU (might need training in multi-stages)

Speech → LLM: Discrete (Speech) Units

Quantise the speech inputs (choose your DSU symbols wisely),

1. Update the tokenizer, e.g., BPE on the DSU
2. Expand the vocabulary size of your LLM
3. Train the model on the DSU (might need training in multi-stages)
    - 1st-stage: next-token prediction on the DSU only
    - 2nd-stage: instruction-like tuning on the DSU-text data which the DSU are the part of the prompts.

Speech → LLM: AudioPaLM

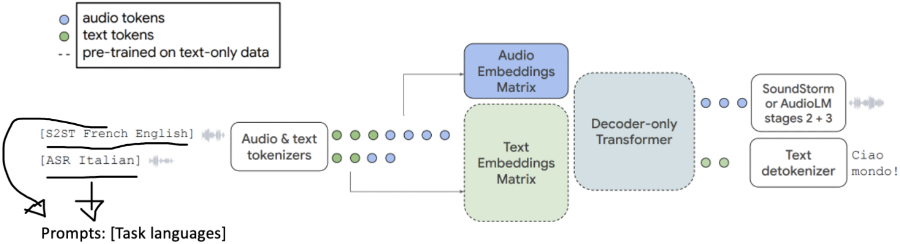
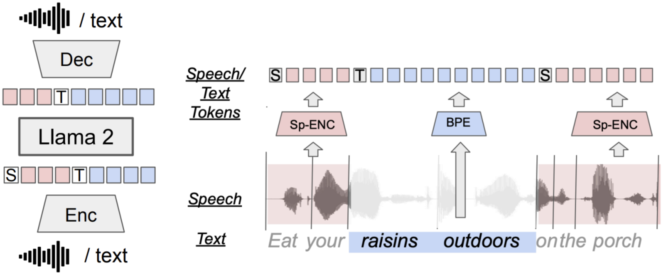


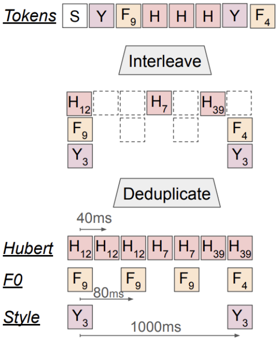
Figure: Illustration of the AudioPaLM model [Rubenstein et al. 2023]

Speech → LLM: Spirit-LM



- HiFi GAN unit-vocoder (decoder) is used to support TTS
- Interleaved speech-text sequences are helpful in DS

Speech → LLM: Spirit-LM II



- HuBERT token for linguistic signals
- Pitch token extracted from VQ-VAE on F0 of speeches
- Style token from SONAR expressive

Speech $\rightarrow$ LLM: Dense Features (I)

Apart from the DSU integration, we can directly prepend the dense (acoustic) features:

1. Find an acoustic encoder, e.g, mHuBERT and Whisper-encoder.

Speech → LLM: Dense Features (I)

Apart from the DSU integration, we can directly prepend the dense (acoustic) features:

1. Find an acoustic encoder, e.g, mHuBERT and Whisper-encoder.
2. Convert the dimension of the speech embedding to suit the embedding dimension of the LLM, e.g., via a linear layer

Speech → LLM: Dense Features (I)

Apart from the DSU integration, we can directly prepend the dense (acoustic) features:

1. Find an acoustic encoder, e.g, mHuBERT and Whisper-encoder.
2. Convert the dimension of the speech embedding to suit the embedding dimension of the LLM, e.g., via a linear layer
3. Prepend the dense (speech) embeddings to the word embeddings.

Speech → LLM: Dense Features (I)

Apart from the DSU integration, we can directly prepend the dense (acoustic) features:

1. Find an acoustic encoder, e.g, mHuBERT and Whisper-encoder.
2. Convert the dimension of the speech embedding to suit the embedding dimension of the LLM, e.g., via a linear layer
3. Prepend the dense (speech) embeddings to the word embeddings.
4. Train the model, typically via LoRA or its variations.
   - Unlike DSU, we don't compute the losses on the audio representations.

Speech → LLM: Dense Features (II)

Prompting in dense feature integration

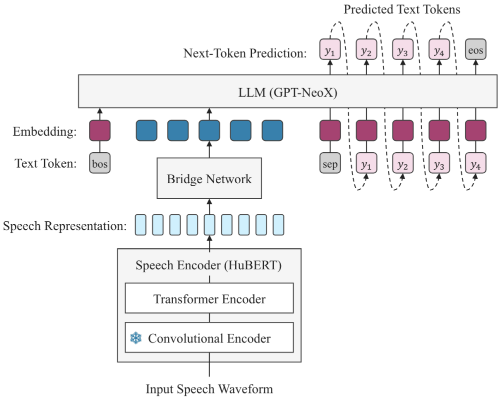


Figure: Hono et. al 2024

Speech → LLM: Dense Features (III)

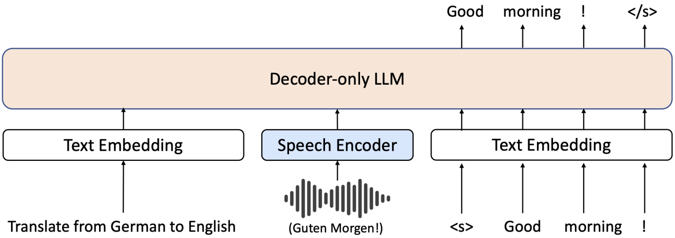


Figure: Huang et. al 2024

Speech → LLM: WavLLM

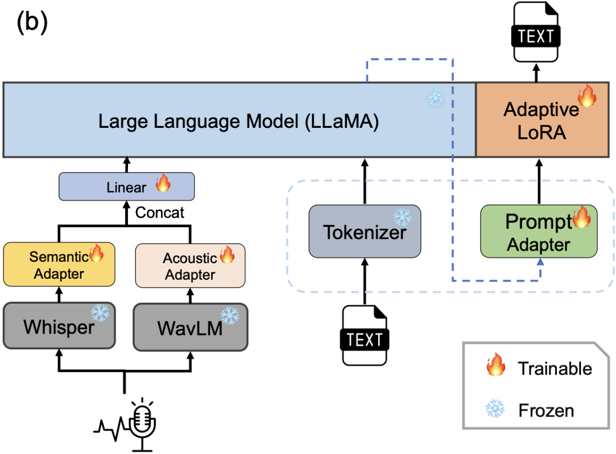
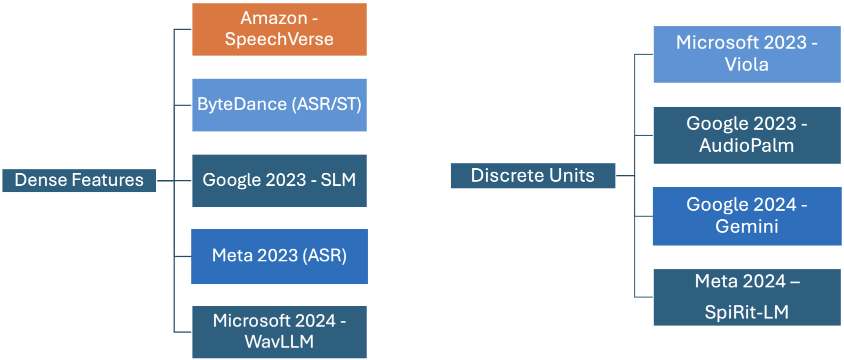


Figure: Hu et. al 2024

Speech → LLM: Hints From The Tech Giants?



and many more ...

Speech → LLM: Discrete or Dense features?

Including speech into LLM is a hot topic, but most works lack comparability [Gaido et al. 2024], e.g.,

- The SFM and the LLM, e.g.,
  - AudioPaLM used Universal Speech Model (USM)[1] as its SFM, but USM is not openly accessible.

[1] [Zhang et al. 2023]

Speech → LLM: Discrete or Dense features?

Including speech into LLM is a hot topic, but most works lack comparability [Gaido et al. 2024], e.g.,

- The SFM and the LLM, e.g.,
  - AudioPaLM used Universal Speech Model (USM)[1] as its SFM, but USM is not openly accessible.
  - the speech quantisation hyper-parameters are different.

[1] [Zhang et al. 2023]

Speech → LLM: Discrete or Dense features?

Including speech into LLM is a hot topic, but most works lack comparability [Gaido et al. 2024], e.g.,

- The SFM and the LLM, e.g.,
  - AudioPaLM used Universal Speech Model (USM)[1] as its SFM, but USM is not openly accessible.
  - the speech quantisation hyper-parameters are different.
  - There are no direct empirical comparison between these discrete and dense methods

[1] [Zhang et al. 2023]

Speech → LLM: Discrete or Dense features?

Including speech into LLM is a hot topic, but most works lack comparability [Gaido et al. 2024], e.g.,

- The SFM and the LLM, e.g.,
  - AudioPaLM used Universal Speech Model (USM)[1] as its SFM, but USM is not openly accessible.
  - the speech quantisation hyper-parameters are different.
  - There are no direct empirical comparison between these discrete and dense methods
- The training and evaluation data, e.g.,
  - the amount, the language directions and the number of tasks.
  - the instruction data used in training and inference.

[1] [Zhang et al. 2023]

Cascade or E2E: Which one is better?

The winning systems at IWSLT[1] (Offline track) S2T (En-De) in the last 5 years:

Year	2020	2021	2022	2023	2024
Winner	End-to-end	End-to-end	Cascaded	Cascaded	*Only Cascaded

[1] The International Conference on Speech Language Translation

MBR Decoding: automatic evaluation

System	D	Joint		TED 2024		ITV		Peloton		Accent	
		COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
CMU	U	0.743	18.3	0.862	25.7	0.735	17.3	0.670	11.5	0.705	18.5
HW-TSC	C^+	0.731	19.3	0.851	27.4	0.728	17.2	0.652	11.9	0.691	20.7
HW-TSC	U	0.727	19.1	0.849	27.1	0.723	17.3	0.646	11.0	0.690	20.8
HW-TSC	C	0.717	18.5	0.841	26.6	0.712	16.7	0.637	10.4	0.678	20.2
NYA	U	0.695	19.5	0.837	28.1	0.648	15.8	0.616	12.2	0.677	21.7
KIT	C^+	0.677	17.5	0.832	27.5	0.618	13.2	0.600	10.2	0.656	19.1

Figure: Official results of the automatic evaluation for the Offline ST Task, English to German.

MBR Decoding: human (src) direct assessment

A flip in ranking for the CMU team in the DA results

	All		TED		ITV		Accent		Peloton	
System	Rank	DA	Rank	DA	Rank	DA	Rank	DA	Rank	DA
HWTSC-LLM	1	84.8	1-2	94.9	1-2	84.7	1-4	76.1	1-4	82.6
HWTSC	2-3	84.2	3-5	92.8	1-3	84.0	1-4	76.8	1-4	81.6
CMU	2-4	83.3	3-5	92.5	2-3	83.1	1-4	75.4	1-4	81.2
NYA	3-4	81.0	1-2	94.7	4	73.9	1-4	77.9	1-4	80.2
KIT	5	76.7	3-5	91.8	5	69.3	5	72.8	5	74.6

Figure: Official DA results for the Offline ST Task, English to German.

MBR Decoding: human (src) direct assessment

A flip in ranking for the CMU team in the DA results

	All		TED		ITV		Accent		Peloton	
System	Rank	DA	Rank	DA	Rank	DA	Rank	DA	Rank	DA
HWTSC-LLM	1	84.8	1-2	94.9	1-2	84.7	1-4	76.1	1-4	82.6
HWTSC	2-3	84.2	3-5	92.8	1-3	84.0	1-4	76.8	1-4	81.6
CMU	2-4	83.3	3-5	92.5	2-3	83.1	1-4	75.4	1-4	81.2
NYA	3-4	81.0	1-2	94.7	4	73.9	1-4	77.9	1-4	80.2
KIT	5	76.7	3-5	91.8	5	69.3	5	72.8	5	74.6

Figure: Official DA results for the Offline ST Task, English to German.

The submitted ST models

- performs well on the TED dataset
- struggle on speeches which are spontaneous, accent-heavy and mixed with background noises

Summary and the future works

Summary I

I: How is the translation of speech different from its text?

- Speech is sparse with acoustic variations $\Longrightarrow$ modality gap
- Existing publicly available datasets are small, noisy or rather unrealistic $\Longrightarrow$ data scarcity

Summary II

II: Existing solutions

- Using data augmentation, multi-task learning, large pretrained models, mixed modality training help to improve E2E ST

Summary II

II: Existing solutions

- Using data augmentation, multi-task learning, large pretrained models, mixed modality training help to improve E2E ST
- In the case of LLM, speech can be integrated via quantisation or dense feature prepending

Summary II

II: Existing solutions

- Using data augmentation, multi-task learning, large pretrained models, mixed modality training help to improve E2E ST
- In the case of LLM, speech can be integrated via quantisation or dense feature prepending
- There are more interesting research directions in E2E model, but cascaded model still remains competitive

Future works

Many...

- LLM for simultaneous ST
  - ByteDance AI: Towards Achieving Human Parity on End-to-end Simultaneous Speech Translation via LLM Agent

Future works

Many...

- LLM for simultaneous ST
  - ByteDance AI: Towards Achieving Human Parity on End-to-end Simultaneous Speech Translation via LLM Agent
- Linguistics or phonetic analysis on ST errors, e.g., Homophones
  - ACL 2024: Speech Sense Disambiguation: Tackling Homophone Ambiguity in End-to-End Speech Translation

Future works

Many...

- LLM for simultaneous ST
  - ByteDance AI: Towards Achieving Human Parity on End-to-end Simultaneous Speech Translation via LLM Agent
- Linguistics or phonetic analysis on ST errors, e.g., Homophones
  - ACL 2024: Speech Sense Disambiguation: Tackling Homophone Ambiguity in End-to-End Speech Translation
- From the modelling POV, it might not be ST specific, e.g., effect of prosody to Q&A task
  - ACL 2024: Advancing Large Language Models to Capture Varied Speaking Styles and Respond Properly in Spoken Conversations

Bibliography I

Baevski, Alexei et al. (2020). “wav2vec 2.0: A framework for self-supervised learning of speech representations”. In: Advances in neural information processing systems 33, pp. 12449–12460.

Barrault, Loïc et al. (2023). “Seamless: Multilingual Expressive and Streaming Speech Translation”. In: arXiv preprint arXiv:2312.05187.

Chen, Mingda et al. (July 2023). “BLASER: A Text-Free Speech-to-Speech Translation Evaluation Metric”. In: Proceedings of the 61st Annual Meeting of the Association for Compu Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 9064–9079. DOI: 10.18653/v1/2023.acl-long.504. URL: https://aclanthology.org/2023.acl-long.504.

Bibliography II

Fang, Qingkai et al. (2022). "Stemm: Self-learning with speech-text manifold mixup for speech translation". In: arXiv preprint arXiv:2203.10426.

Gaido, Marco et al. (2024). "Speech Translation with Speech Foundation Models and Large Language Models: What is There and What is Missing?" In: arXiv preprint arXiv:2402.12025.

Graves, Alex et al. (2006). "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks". In: Proceedings of the 23rd international conference on Machine learning, pp. 369–376.

Lakhotia et al. (2021). "On generative spoken language modeling from raw audio". In: Transactions of the Association for Computational Linguistics 9, pp. 1336–1354.

Bibliography III

Lam, Tsz Kin, Shigehiko Schamoni, and Stefan Riezler (May 2022). "Sample, Translate, Recombine: Leveraging Audio Alignments for Data Augmentation in End-to-end Speech Translation". In: Proceedings of the 60th Annual Meeting of the Association for Compu Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 245–254. DOI: 10.18653/v1/2022.acl-short.27. URL: https://aclanthology.org/2022.acl-short.27.

Li, Xian et al. (2020). "Multilingual speech translation with efficient finetuning of pretrained models". In: arXiv preprint arXiv:2010.12829.

Bibliography IV

- Matusov, Evgeny et al. (2005). "Evaluating Machine Translation Output with Automatic Sentence Segmentation". In: Proceedings of the Second International Workshop on Spoken Languag Pittsburgh, Pennsylvania, USA. URL: https://aclanthology.org/2005.iwslt-1.19.
- Pino, Juan et al. (2019). "Harnessing indirect training data for end-to-end automatic speech translation: Tricks of the trade". In: arXiv preprint arXiv:1909.06515.
- Rubenstein, Paul K et al. (2023). "Audiopalm: A large language model that can speak and listen". In: arXiv preprint arXiv:2306.12925.
- Salesky, Elizabeth et al. (2018). "Towards fluent translations from disfluent speech". In: 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, pp. 921–926.

Bibliography V

Tsiamas, Ioannis et al. (2022). "Shas: Approaching optimal segmentation for end-to-end speech translation". In: arXiv preprint arXiv:2202.04774.

Zhang, Yu et al. (2023). "Google usm: Scaling automatic speech recognition beyond 100 languages". In: arXiv preprint arXiv:2303.01037.

Zhou, Giulio et al. (2024). "Prosody in Cascade and Direct Speech-to-Text Translation: a case study on Korean Wh-Phrases". In: arXiv preprint arXiv:2402.00632.