Natural Language Understanding, Generation, and Machine Translation

Lecture 28: Spoken Language Translation

Tsz Kin Lam

School of Informatics
University of Edinburgh

Introduction: When the model is a ”black box”

What is Spoken Language Translation?

Spoken Language/Speech Translation (SLT/ST) is cross-lingual and (most likely) cross-modal.

How is speech different from text? (I)

Major differences are:

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

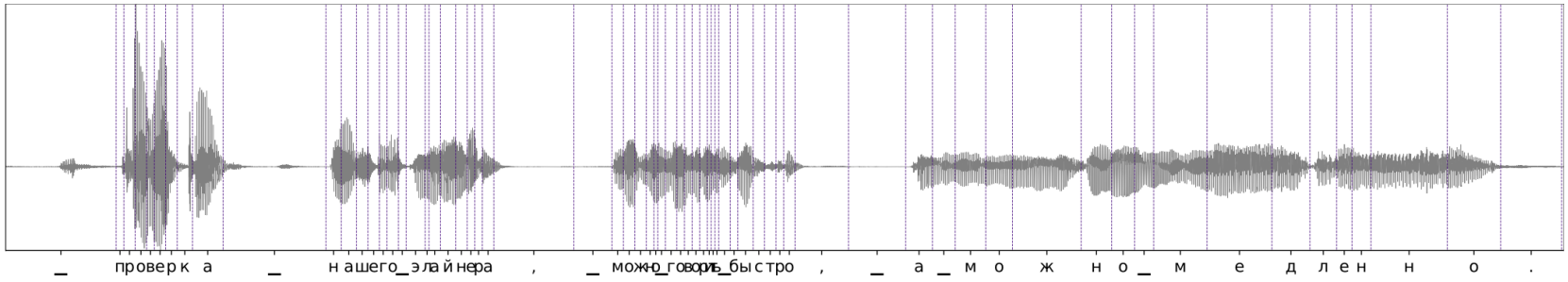
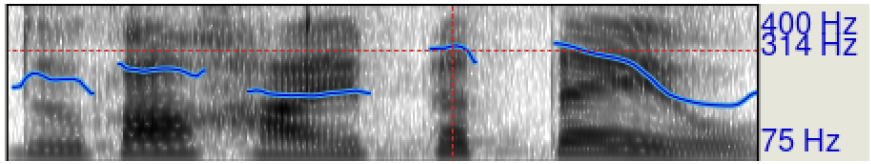


Figure: Speech-to-text alignment [Barrault et al. 2023]

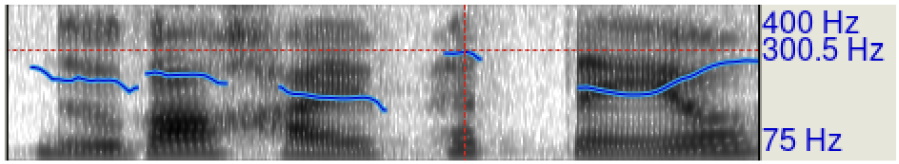
- The file format matters (sampling rate, bit depth and bit rate).
- Background noises may appear in the speech.

How is speech different from text? (II)

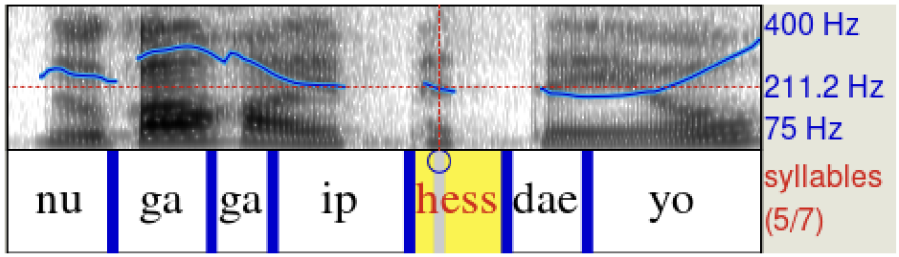
- Paralinguistic signals, such as prosody and accents, matter



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



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



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

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

How is speech different from text? (III)

- 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]

Should we keep the disfluencies in the translation?

How is ST different from text translation?

Compared to text translation, *ST is low-resource* even in high-resource language pairs.

Data	X-Y	#utterances	#words (src+tgt)
MuST-C[1]	En-Fr	280K	10.6M
"	En-De	234K	8.3M
CoVoST-2[2]	Fr-En	207K	4M
"	De-En	127K	2M
(MT) Wiki-Matrix[3]	De-En	6.2M	196M

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

[1] Di Gangi et al. 2019
[2] Wang et al. 2020
[3] Schwenk et al. 2019

Inference: audio segmentation (I)

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

1. It is an audio sequence of 50 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 spectrogram is used

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]

Evaluation: (automatic) metrics

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

- n-gram matching: `BLEU` and `chrF`
- neural metrics: `COMET`[1] (may require back-translation to get the transcripts)

In S2S translation,

- transcribe and MT-evaluate: `ASR-BLEU` and `ASR-chrF`
- neural metrics: `BLASER`[2]

[1] Rei et al. 2020

[2] Chen et al. 2022

Modeling: Unfolding the "black box"

Feeding audio input into Transformer

Recap: Speech signal is very sparse (very long).

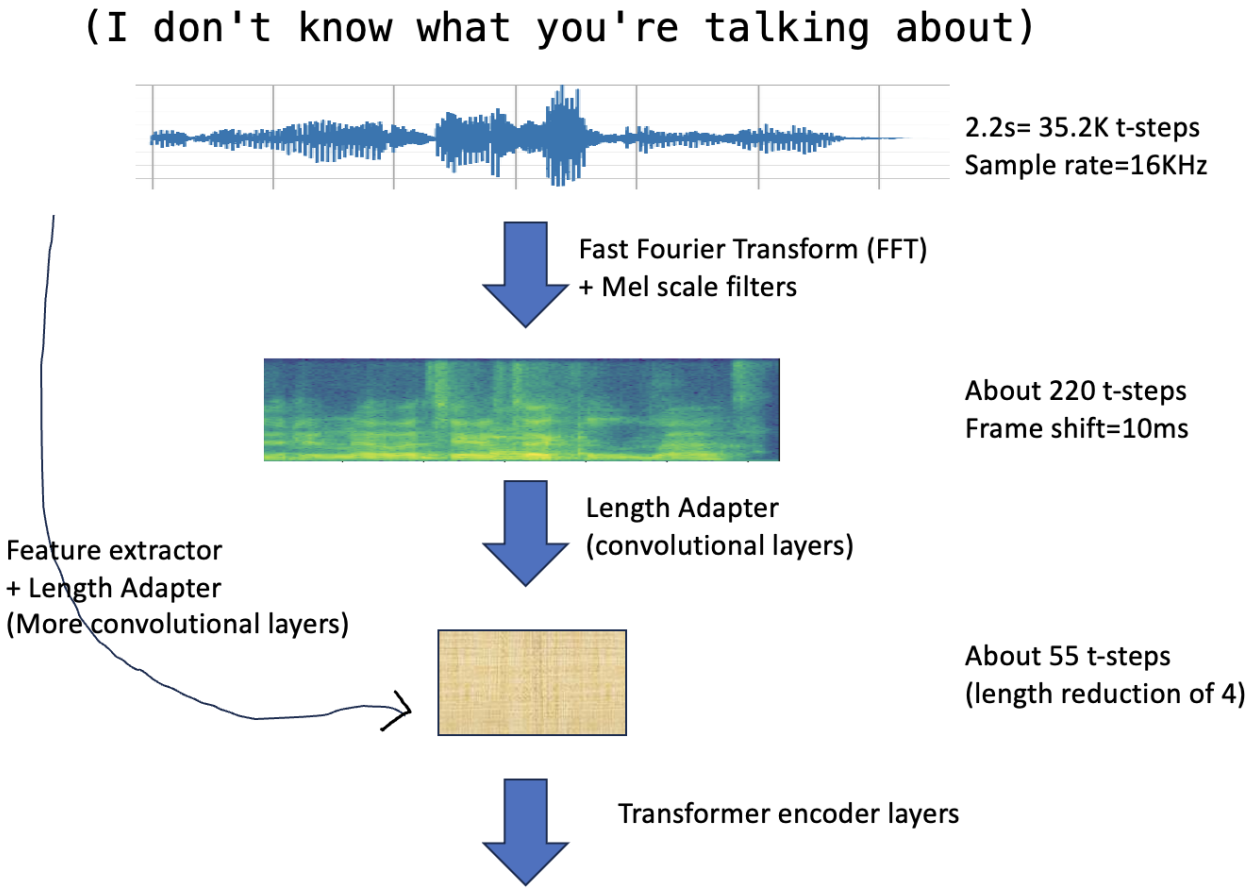


Figure: An illustration of sequence-length reduction of the speech signal.

Cascaded model (I)

Recap: ST is a (most likely) cross-model and cross-lingual 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).

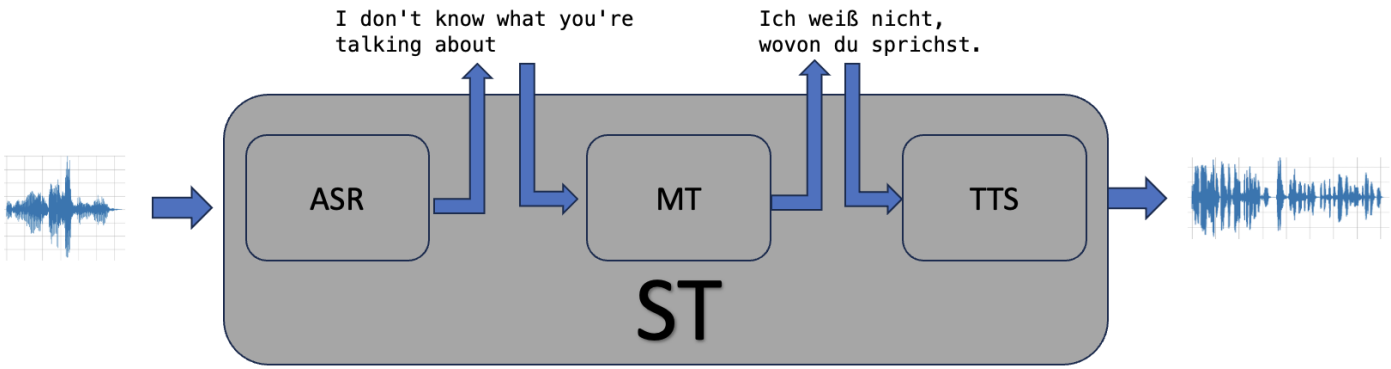


Figure: An illustration of the cascaded ST model.

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.
- *Output correction is simpler* by inspecting the intermediate transcripts(/translations).
- It can leverage foundation models easily.

Cascaded model (III)

Disadvantage:

- The inference 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 model (I)

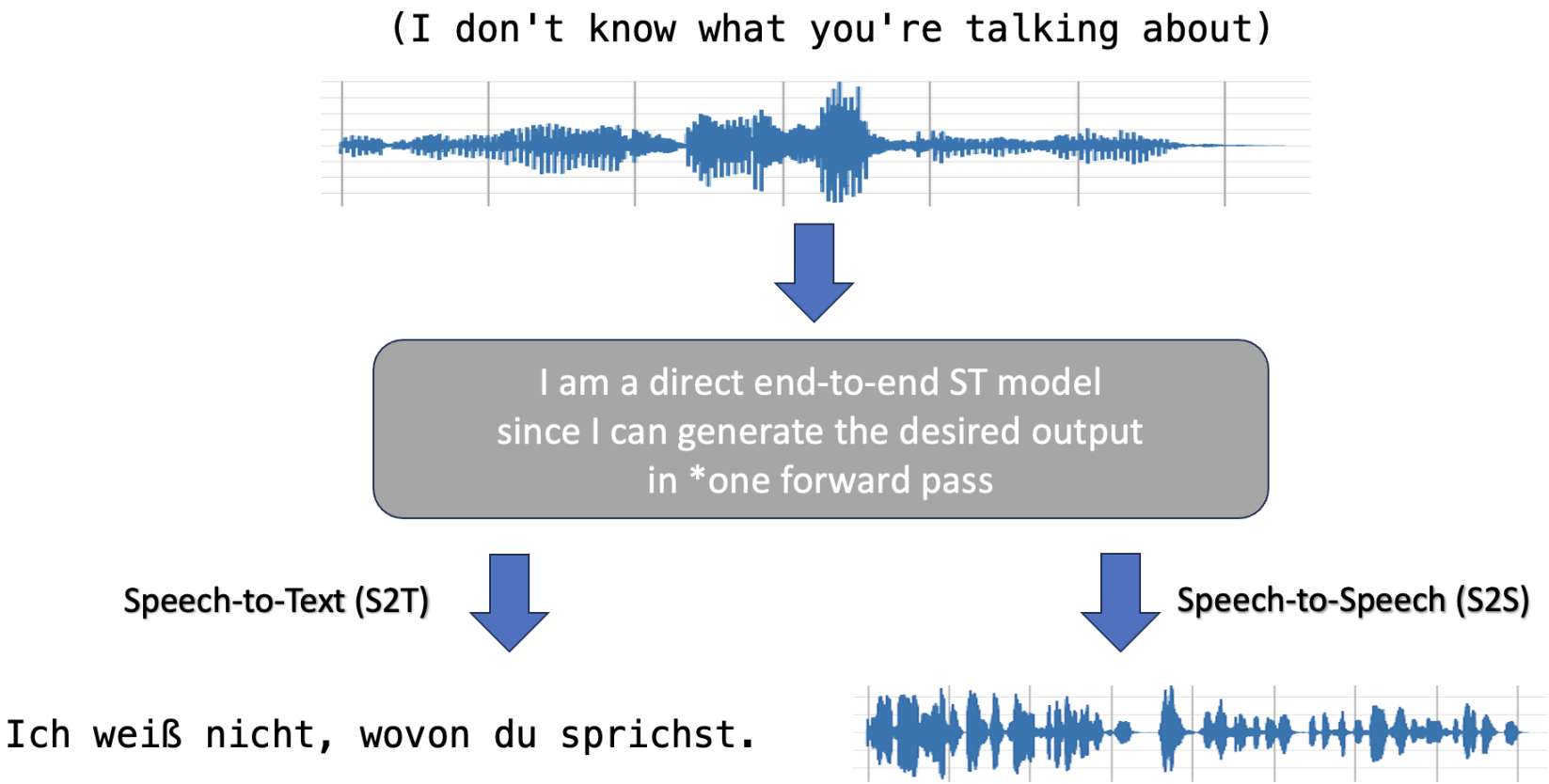


Figure: An illustration of the direct end-to-end ST model.

Direct end-to-end model (II)

Advantage:

- The inference is simply *one forward-pass*. This helps to
  - give *lower latency in inference*, 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 end-to-end model (III)

Disadvantage:

- The amount of *paired ST data is limited*. (**Recap**: ST is low-resource)
- End-to-end ST model *is harder to optimise*.

Regardless, end-to-end model is the main research direction now!

Improving end-to-end ST: data augmentation

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

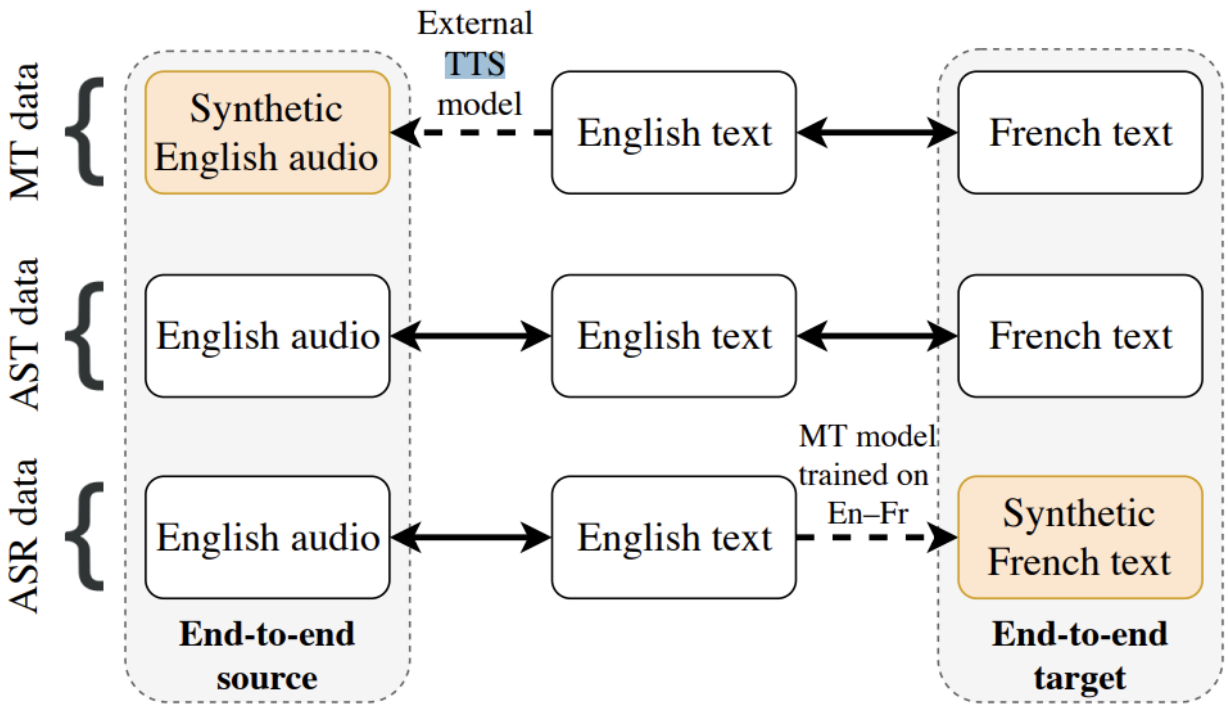


Figure: Pseudo ST data generation [Pino et al. 2019]

Improving end-to-end ST: multi-task learning

Training ST with other sub-tasks in parallel instead of using them sequentially, e.g., in [Gaido et al. 2020]

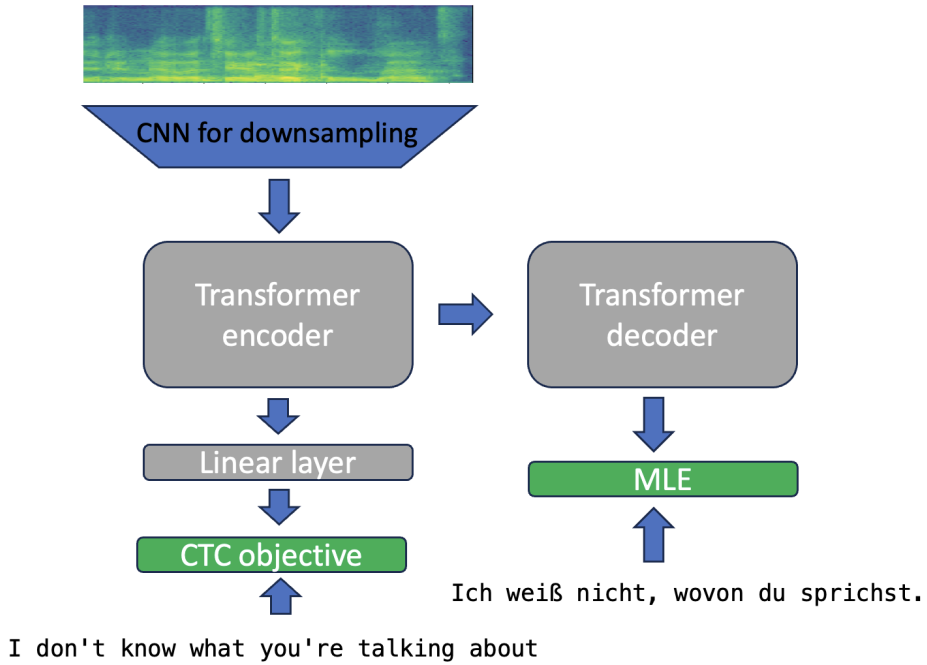


Figure: ST training with CTC[3] on ASR task

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

Improving end-to-end ST: using pre-trained models

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

- wav2vec 2.0[4] for initialising the acoustic encoder
- mBART for initialising the translation decoder

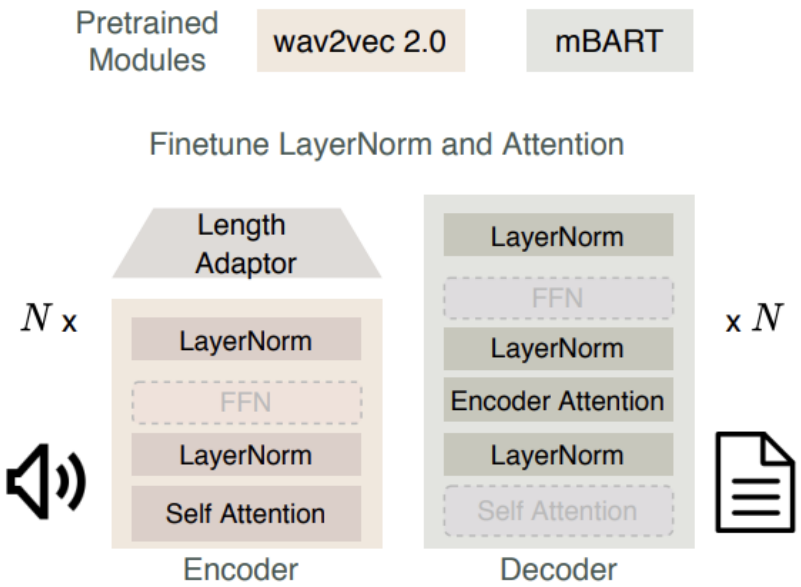


Figure: ST model initialisation via pre-trained SSL models [Li et al. 2020]

[4] Baevski et al. 2020

Improving end-to-end ST: bridging the modality gap (I)

- Bi-encoder architecture:
  1. for speech input
  2. for text input
- Mixed-modality training:

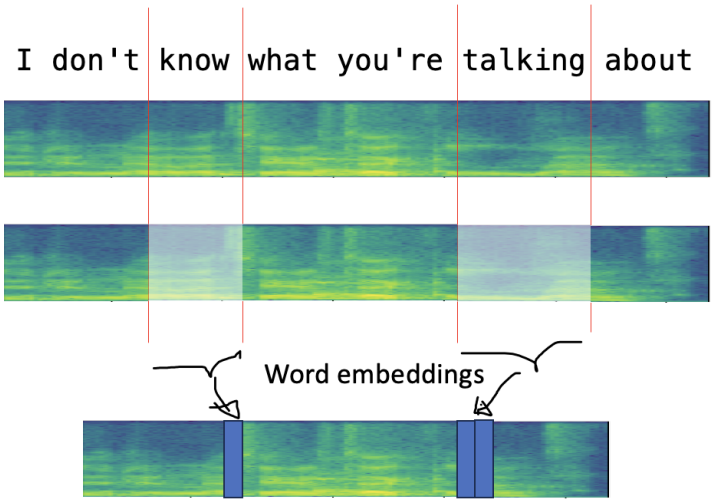


Figure: Mixed modality: a sequence of alternating speech and text embeddings.

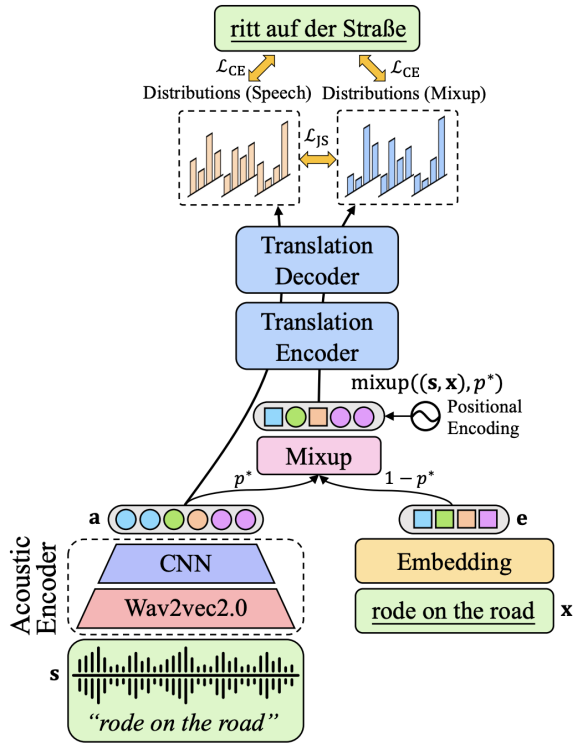


Figure: STEMM [Fang et al. 2022]

Improving end-to-end ST: bridging the modality gap (II)

Speech quantisation: to transform speech signal into a sequence of tokens [Lakhotia et al. 2021]

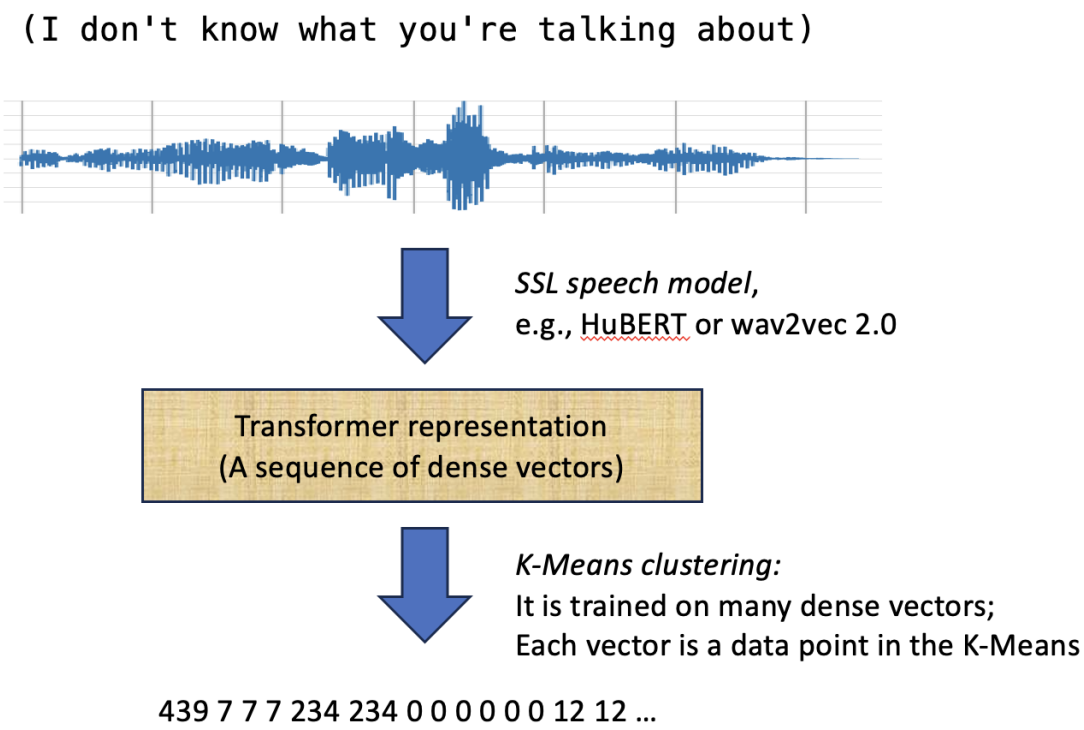


Figure: An illustration of speech quantisation using K-Means clustering.

Improving end-to-end ST: speech quantisation

Advantages

- Data storage and transmission becomes easier
- Speech generation becomes more feasible
  - e.g., speech-to-unit, unit-based LM and a unit-based vocoder[1])
  - Textless NLP (*Link*)

[1] Lee et al. 2021

Improving end-to-end ST: speech quantisation

Disadvantages

- The inference pipeline gets lengthy (quantisation and clustering)
- There are more hyper-parameters to tune, e.g.,
  1. The hyper-parameters in the K-Means model:
     - ★ Its training data size
     - ★ Its clustering size
  2. The representation layer of the SSL model to be used for quantisation

Improving end-to-end ST: putting all together

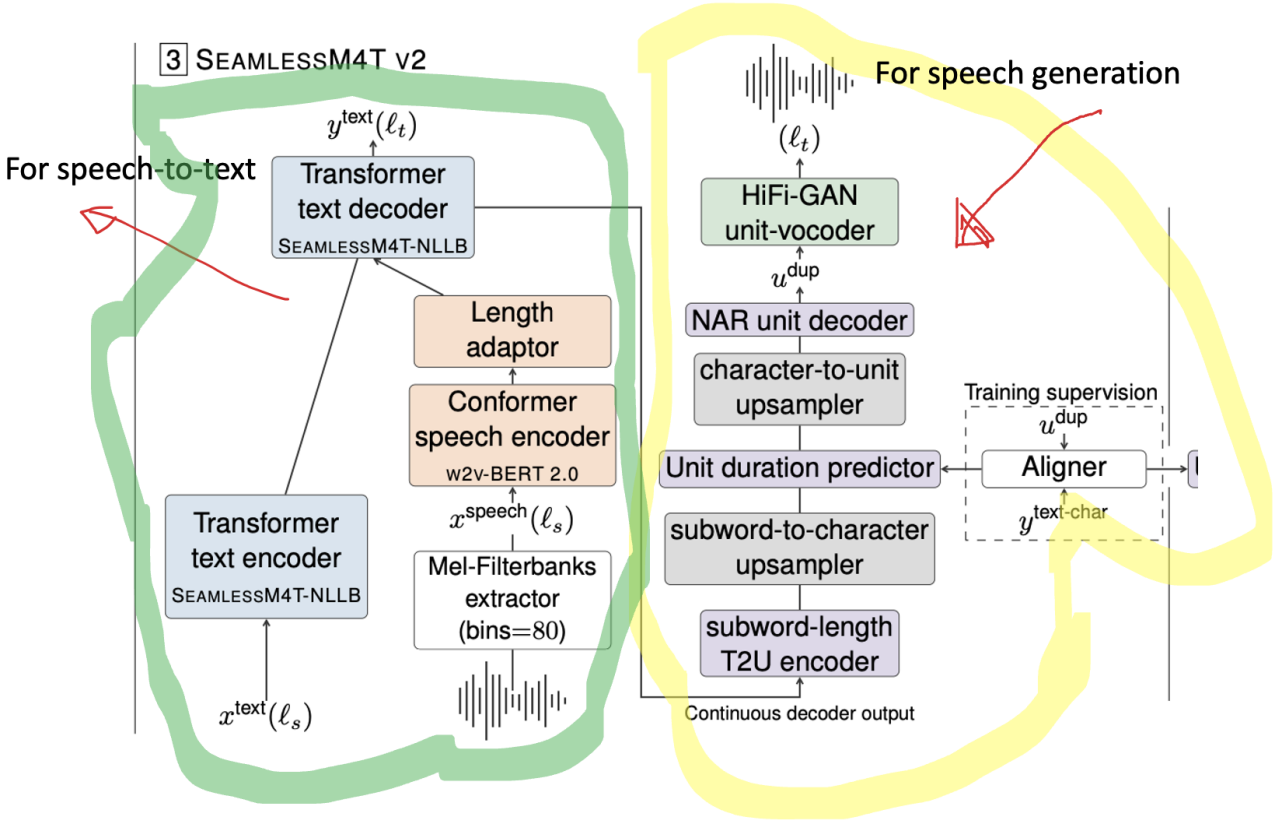


Figure: Seamless-M4T v2 model [Barrault et al. 2023]

Improving end-to-end ST with LLM: AudioPaLM

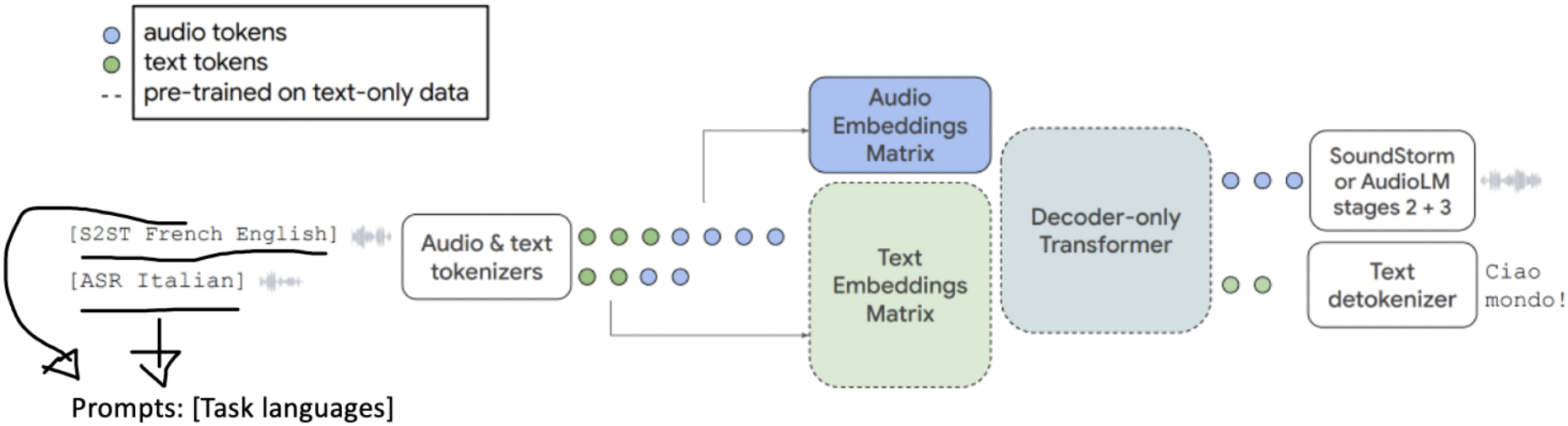


Figure: Illustration of the AudioPaLM model [Rubenstein et al. 2023]

Improving end-to-end ST with LLM (II)

Apart from discrete speech units, we can also work on dense feature integration:

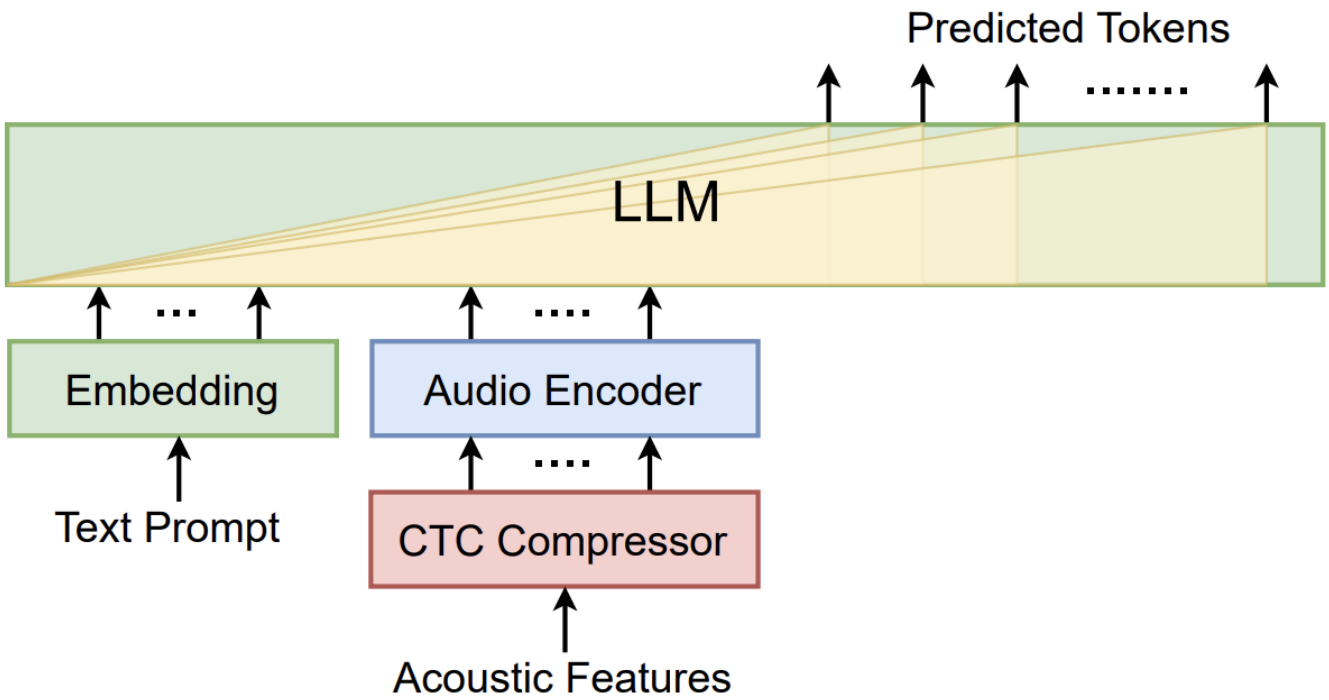


Figure: Dense feature integration into LLaMA-2 [Wu et al. 2023]

Improving end-to-end ST with LLM (III)

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

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

Which one is better? (I)

In the recent IWSLT[2] (Offline track) for S2T translation,

Year	2020	2021	2022	2023
Winner	End-to-end	End-to-end	Cascaded	Cascaded

Figure: Winning architectures in the last 4 years in IWSLT (Offline track)

We need more *diverse test sets*, not only limited to TED talks.

[2]The International Conference on Spoken Language Translation

Which one is better? (II)

Model	size	S2TT FLEURS (↑BLEU) X–eng (n=81)	S2TT FLEURS (↑BLEU) eng–X (n=88)	S2ST FLEURS (↑ASR-BLEU) X–eng (n=81)	S2ST FLEURS (↑ASR-BLEU) eng–X (n=26)	S2ST CVSS (↑ASR-BLEU) X–eng (n=21)
WL-v2 (S2TT)	1.5B	17.9	–	17.8	–	29.6
WL-v3 (S2TT)	1.5B	16.9[8]	–			
A8B (S2TT)	8B	19.7	–			
WM (ASR) + NLLB-1.3B	2B	19.7	20.7	20.7	21.5	
WM (ASR) + NLLB-3.3B	4B	20.4	22.0	21.4	22.4	
WL-v2 (ASR) + NLLB-1.3B	2.8B	22.0	21.2	22.9	21.8	
WL-v2 (ASR) + NLLB-3.3B	4.8B	22.7	**22.4**	23.7	22.7	
SeamlessM4T-Medium	1.2B	20.9	19.4	20.2	15.8	30.6
SeamlessM4T-Large	2.3B	24.1	21.8	25.8	20.9	35.7
SeamlessM4T v2	2.3B	**26.6**	22.2	**29.7**	**26.1**	**39.2**

Figure: Comparison between SEAMLESS-M4T, cascaded system and AudioPaLM (8B) [Barrault et al. 2023].

Summary

- There are extra complexities in spoken language, e.g., its sparsity and acoustic variations.
- Two common approaches for ST are 1) cascaded modeling and 2) end-to-end modeling.
- There are more interesting things going on in end-to-end model, e.g., integrating speech into LLM.
- Cascaded model, however, still remains very competitive.

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. (2022). "BLASER: A text-free speech-to-speech translation evaluation metric". In: *arXiv preprint arXiv:2212.08486*.
- Di Gangi, Mattia A. et al. (June 2019). "MuST-C: a Multilingual Speech Translation Corpus". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 2012–2017. DOI: 10.18653/v1/N19-1202. URL: https://aclanthology.org/N19-1202.

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. (2020). "End-to-end speech-translation with knowledge distillation: FBK@ IWSLT2020". In: *arXiv preprint arXiv:2006.02965*.
- 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.

Bibliography III

- Lakhotia et al. (2021). "On generative spoken language modeling from raw audio". In: *Transactions of the Association for Computational Linguistics* 9, pp. 1336–1354.
- Lee, Ann et al. (2021). "Direct speech-to-speech translation with discrete units". In: *arXiv preprint arXiv:2107.05604*.
- Li, Xian et al. (2020). "Multilingual speech translation with efficient finetuning of pretrained models". In: *arXiv preprint arXiv:2010.12829*.
- 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*.
- Rei, Ricardo et al. (2020). "COMET: A neural framework for MT evaluation". In: *arXiv preprint arXiv:2009.09025*.
- Rubenstein, Paul K et al. (2023). "Audiopalm: A large language model that can speak and listen". In: *arXiv preprint arXiv:2306.12925*.

Bibliography IV

- Salesky, Elizabeth et al. (2018). "Towards fluent translations from disfluent speech". In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 921–926.
- Schwenk, Holger et al. (2019). "Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia". In: *arXiv preprint arXiv:1907.05791*.
- Tsiamas, Ioannis et al. (2022). "Shas: Approaching optimal segmentation for end-to-end speech translation". In: *arXiv preprint arXiv:2202.04774*.
- Wang, Changhan, Anne Wu, and Juan Pino (2020). "Covost 2 and massively multilingual speech-to-text translation". In: *arXiv preprint arXiv:2007.10310*.

Bibliography V

- Wu, Jian et al. (2023). "On decoder-only architecture for speech-to-text and large language model integration". In: *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, pp. 1–8.
- 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*.