11492/11692/18495
Speech Processing

Lecture 14: Speech-to-Speech Translation

Jiatong Shi

Carnegie Mellon University
Language Technologies Institute

TA Introduction

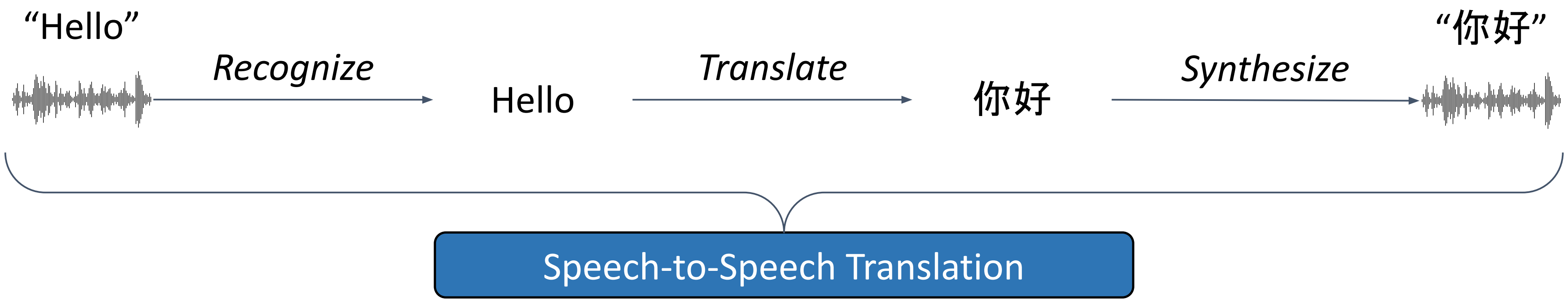
- 3rd Year Ph.D. Student
- Main research focus:
  - speech representation learning and its application
- Broad interests in many downstream tasks:
  - Typical speech tasks: ASR & TTS & ST & SLU
    - architectures
    - decoding
    - aspects in low-resource and multilingual
  - Related music tasks
    - singing voice synthesis
    - singing voice conversion
    - music generation

Table of Contents

Table of Contents

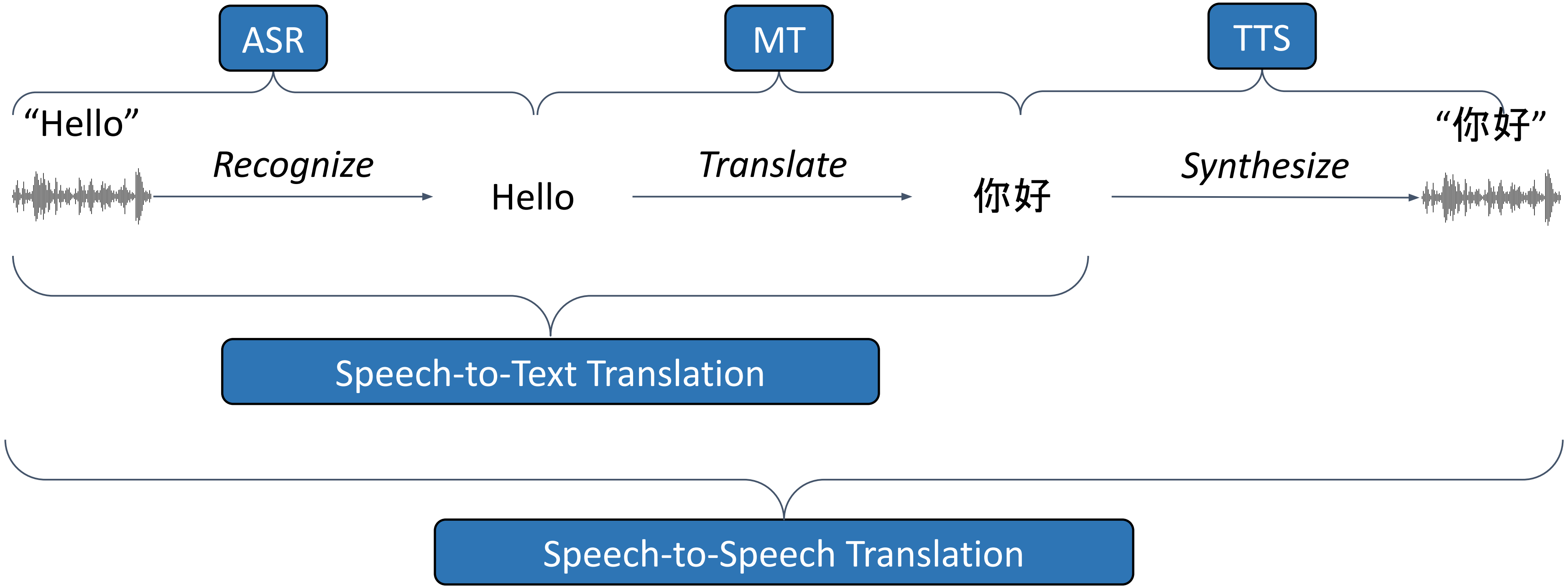
Speech-to-speech Translation

- Converts source language speech into target language text / speech
  - **Sequence Transduction Task**: sequence in, sequence out
  - **Compositional Task:** naturally decomposes into subtasks



Speech Translation

- Converts source language speech into target language text / speech
  - **Sequence Transduction Task**: sequence in, sequence out
  - **Compositional Task:** naturally decomposes into subtasks



More on system construction

- Shinji has introduced the general concepts of speech translation, including a section for speech-to-speech translation.

- Today, we will focus more on how to build the system of speech-to-speech translation (S2ST)

Table of Contents

Speech-to-Speech Translation Metrics

- For speech-to-speech translation, we want to know **the translation quality and the synthesis quality**

- Metrics
  - ASR-BLEU (objective)
  - Naturalness (subjective)
  - Speaker similarity (subjective)
  - EER on speaker (objective)
  - …

Speech-to-Speech Translation Metrics

- For speech-to-speech translation, we want to know **the translation quality and the synthesis quality**

- Metrics
  - ASR-BLEU (objective)
  - Naturalness (subjective)
  - Speaker similarity (subjective)
  - EER on speaker (objective)
  - …

Note that we do not have monotonic assumption in speech-to-speech translation, so we can not use:

- **WER**
- **MCD**
- **F0 RMSE**

ASR-BLEU

- Automatically evaluate speech-to-speech translation models by
  - 1) feeding speech outputs to an **ASR model**,
  - 2) then scoring **BLEU** on the ASR model's transcriptions

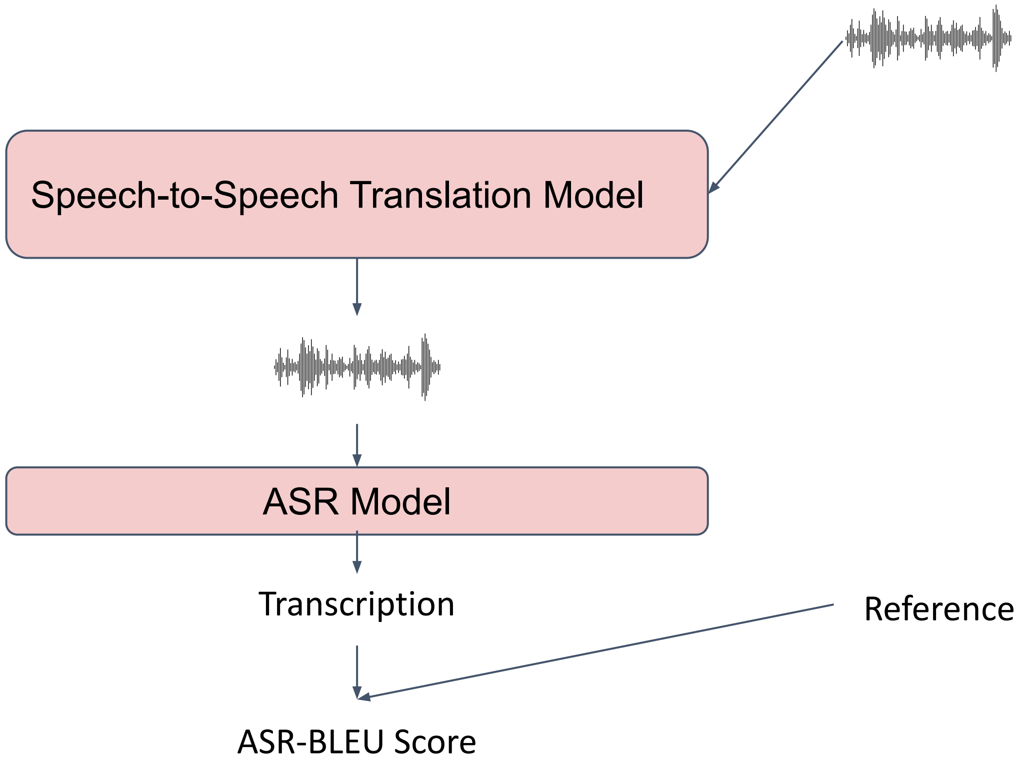


Table of Contents

Frequently Used Corpora

- CVSS (synthesized) - https://github.com/google-research-datasets/cvss
  - X-to-En (21 languages); speech from CommonVoice
  - 2 versions: CVSS-C (easier) and CVSS-T (harder); former has a single target speaker voice, latter has multiple and each target matches the source speaker voice
- LibriS2S (synthesized) - https://github.com/PedroDKE/LibriS2S
  - En-to-De and De-to-En: speech from librivox audio books
  - Note: no speaker matching
- Voxpopuli (real-world) - https://github.com/facebookresearch/voxpopuli
  - 15x15 directions of language pairs: speech from parliament speech
  - Note: no speaker matching
- SpeechMatrix (real-world) - https://github.com/facebookresearch/fairseq/tree/ust/examples/speech_matrix
  - 17x17 directions of langauge pairs: speech from parliament speech
  - Note: no speaker matching
- SeamlessM4T (real-world) - https://github.com/facebookresearch/seamless_communication/blob/main/docs/m4t/seamless_align_README.md
  - Note: no speaker mathcing
  - Initially ~100 directions, now expanding to ~160 directions

Shared Tasks

- The International Conference on Spoken Language Translation (**IWSLT**) is an annual scientific conference, associated with an **open evaluation campaign on spoken language translation**, where both scientific papers and system descriptions are presented.

- https://iwslt.org/2024/
- The speech-to-speech translation is a very hot topic in research community and it is included in **three** tracks:
  - Simultaneous track
  - Speech-to-speech track
  - Dubbing track

Shared Tasks

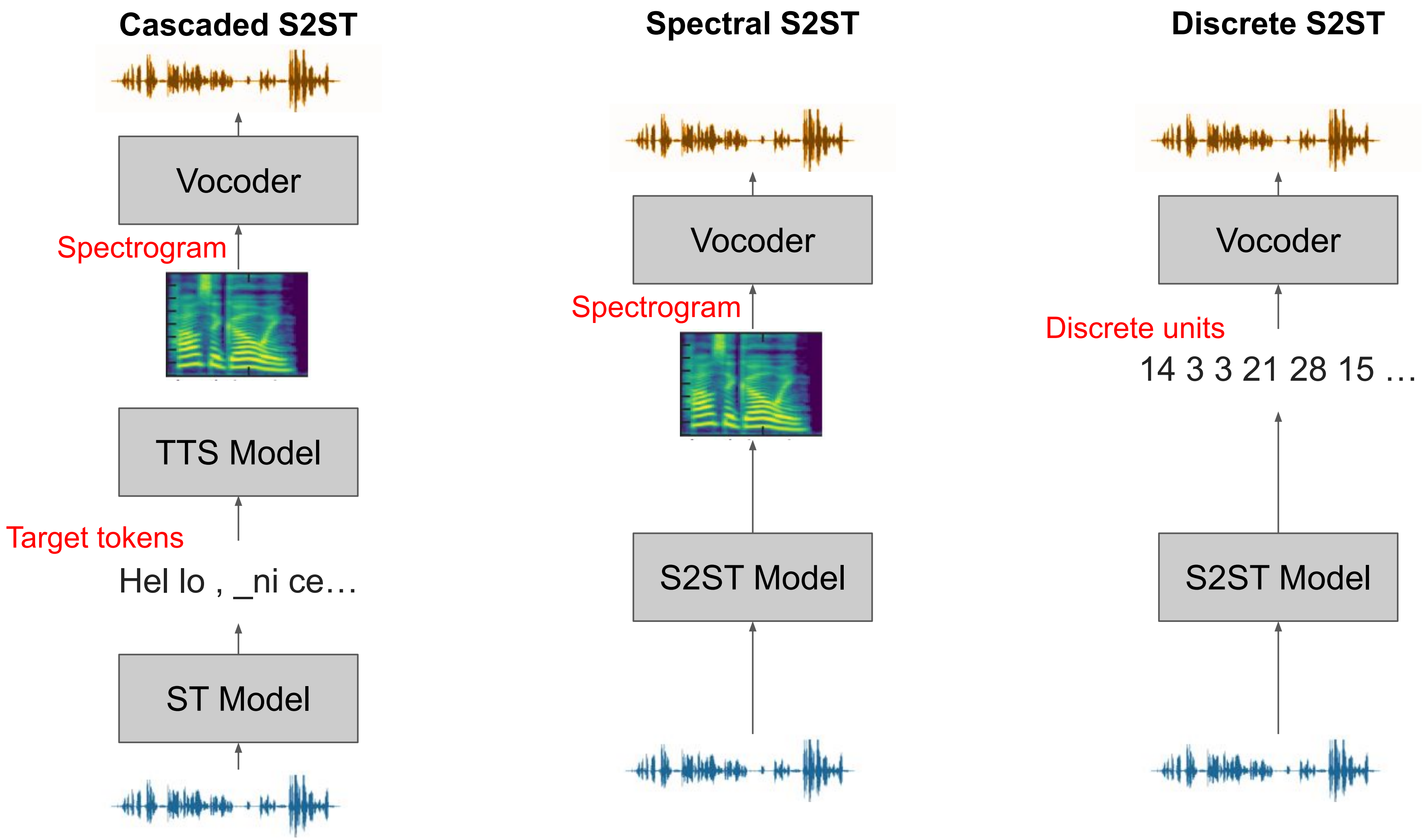
The **IWSLT 2024 Evaluation Campaign** will host shared tasks featuring the following focus areas:

- **Speech translation campaign tracks:**
  - **Speech-to-speech track** (Qianqian Dong, Bytedance, China)
  - **Simultaneous track** (Katsuhito Sudoh, NAIST, Japan)
  - **Subtitling track** (Mauro Cettolo, FBK, Italy; Evgeny Matusov, AppTek, Germany)
  - **Offline track** (Marco Turchi, Zoom, Germany; Matteo Negri, FBK, Italy)
  - **Dubbing track** (Brian Thompson, Amazon, USA; Prashant Mathur, AWS AI Labs, USA)
  - **Low-resource track** (Antonios Anastasopoulos, George Mason University)
  - **Indic track** (Chandresh Kumar Maurya, IIT Indore, India)

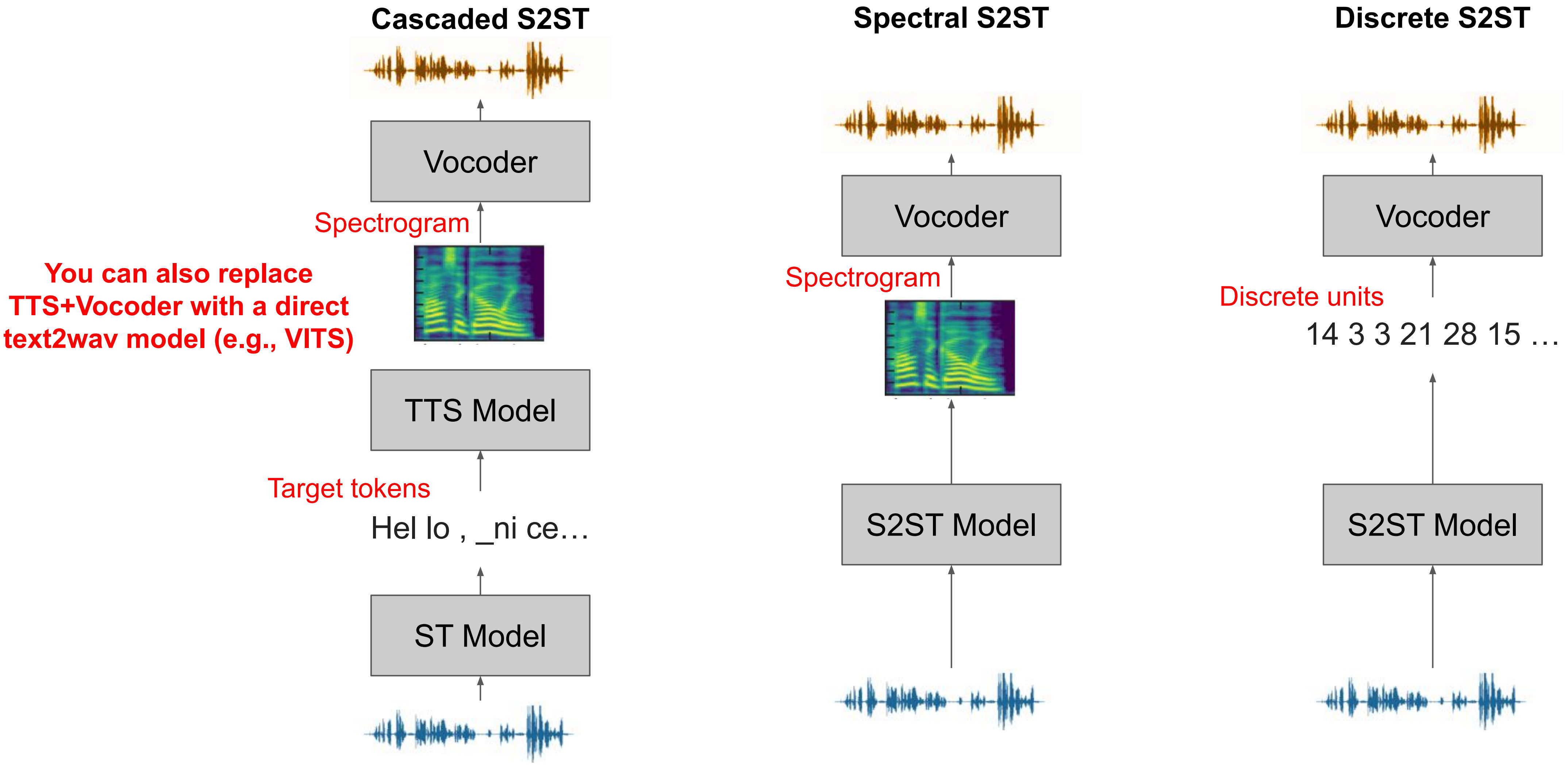
Table of Contents

- Speech-to-speech translation
  - Introduction
  - Evaluation metrics
  - Famous datasets and benchmarks
  - **Technical overviews**
  - References

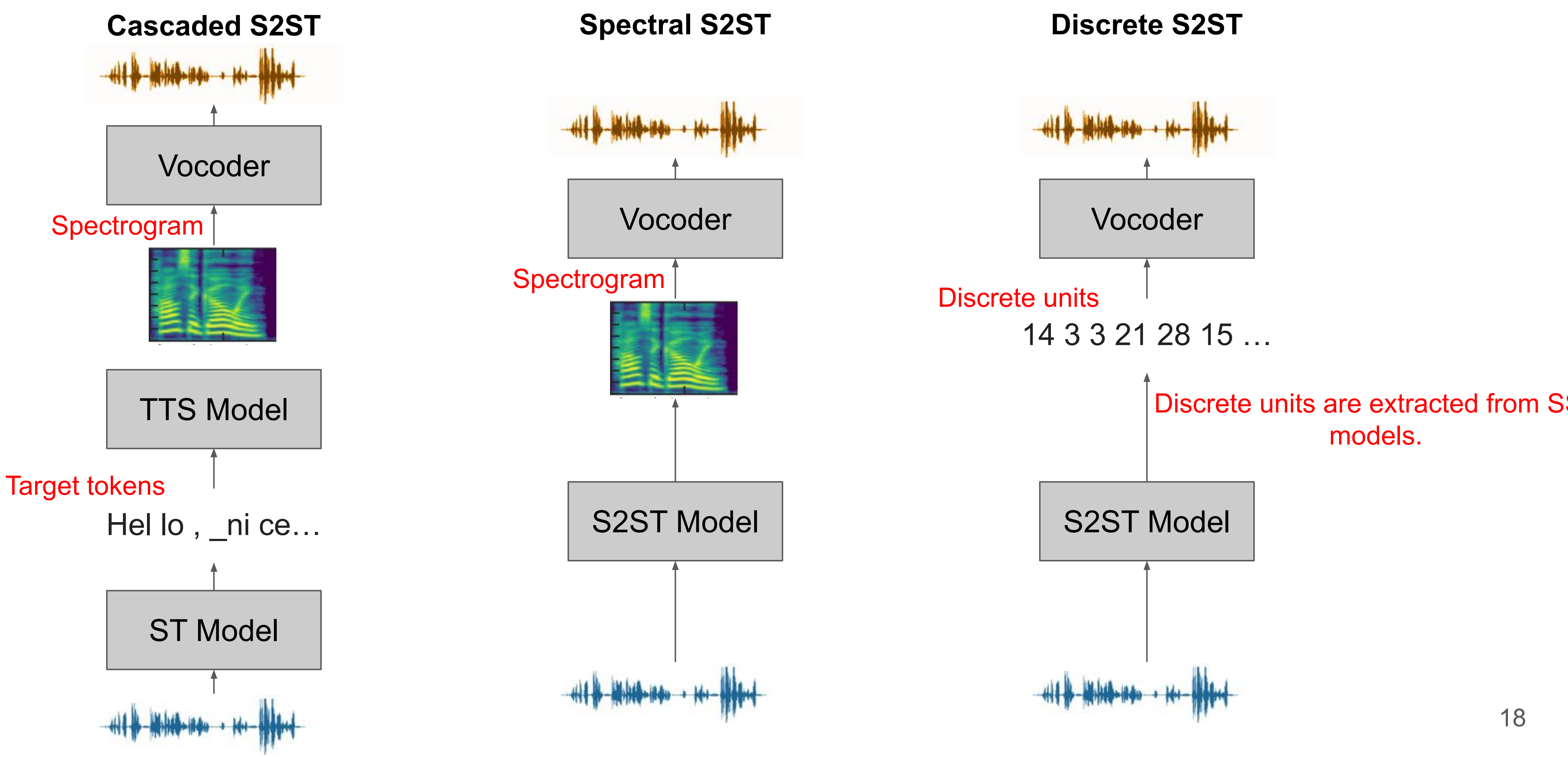
Cascaded S2ST vs. Spectral S2ST vs. Discrete S2ST



Cascaded S2ST vs. Spectral S2ST vs. Discrete S2ST

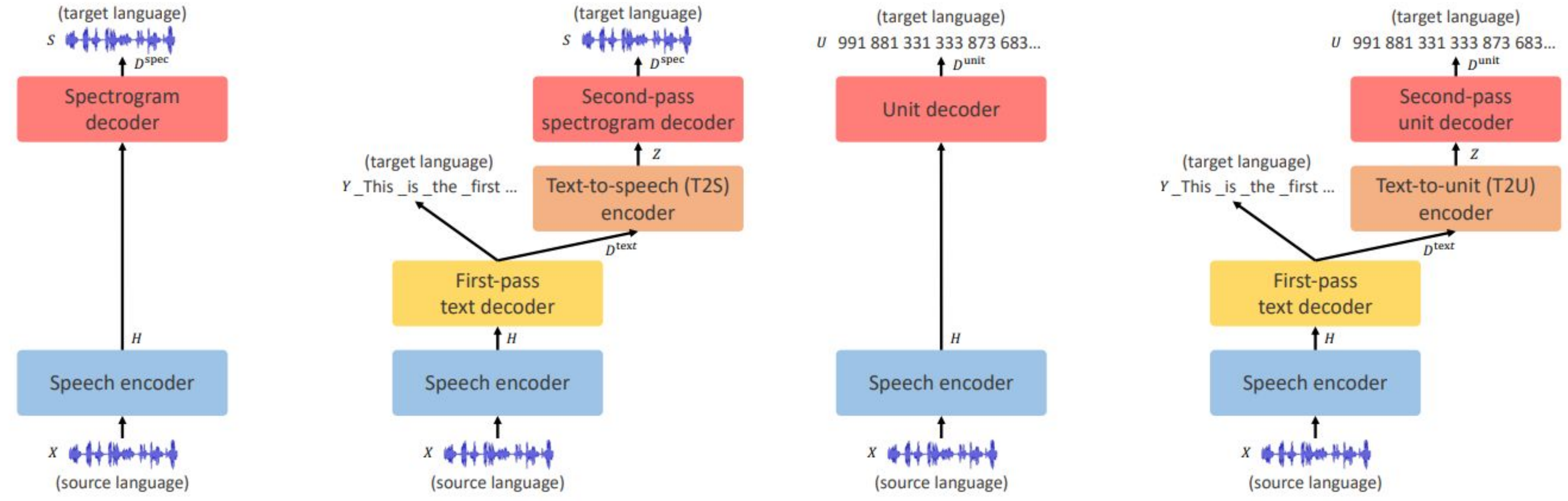


Cascaded S2ST vs. Spectral S2ST vs. Discrete S2ST



Multi-Decoder Speech-to-Speech Translation

- Use Multi-Decoders to build S2ST models with **speech-to-text intermediates**



(a) Translatotron | (b) Translatotron2+ | (c) S2UT | (d) UnitY

Spectral Direct | Spectral Multi-Decoder | Discrete Direct | Discrete Multi-Decoder

Spectral vs. Discrete unit

- Spectral
  - Pros
    - Standardized feature (easy to build common interface)
    - Expressiveness
    - Flexibility
    - Techniques in TTS pre-training
  - Cons
    - Training takes longer time
    - Difficult to converge
    - No good training monitor

- Discrete unit
  - Pros
    - Training is faster
    - Convergence is fast and stable
    - Accuracy is a good indicator of performance
    - Techniques in text pretraining
  - Cons
    - Dependency in SSL
    - Bad generalization to language, data sources, speakers, prosody
    - Feature is not standardized (difficult to share/reuse)

Typical steps of building cascaded S2ST

- Step1: Data Preparation
  - Data
    - Source language speech
    - Source language text
    - Target language text (not necessary but strongly recommend)
    - Target language speech

Typical steps of building cascaded S2ST

- Step2: Model training
    - End-to-end ST or ASR+MT
        - Input: Source language speech
        - Intermediate output: source language text
        - Output: target language text
    - TTS model
        - Input: target language text
        - Output: target language speech (spectral features)
    - Vocoder
        - Input: target language speech (spectral features)
        - Output: target language speech (waveform)

Typical steps of building cascaded S2ST

- Step3: Inference
    - First use the ST model to translate source language speech
    - Then use TTS model to convert source language speech into spectral features
    - Then use vocoder to convert spectral features into waveform.

Typical steps of building discrete direct S2ST

- Step1: Data Preparation
  - Data
    - Source language speech
    - Source language text (not necessary but strongly recommend)
    - Target language text (not necessary but strongly recommend)
    - Target language speech
  - Pre-trained model
    - Speech SSL model pre-trained on target language speech
      - (can be obtained by either pre-trained from scratch or fine-tune from existing SSLs in other languages)

Typical steps of building discrete direct S2ST

- Step 2: Model Preparation
    - Unit extraction model
        - Extract SSL feature (usually just a certain layer) from target speech
        - Conduct Kmeans clustering over the extracted features (unit size can be 100 -> 1000)
        - The Kmeans model is the unit extract model

Typical steps of building discrete direct S2ST

- Step 2: Model Preparation
  - Unit extraction model
    - Extract SSL feature (usually just a certain layer) from target speech
    - Conduct Kmeans clustering over the extracted features (unit size can be 100 -> 1000)
    - The Kmeans model is the unit extract model
  - Unit vocoder
    - Using extracted unit sequences to train a unit-based vocoder
    - Responsible for converting the unit sequence to final waveform

Typical steps of building discrete direct S2ST

- Step 2: Model Preparation
  - Unit extraction model
    - Extract SSL feature (usually just a certain layer) from target speech
    - Conduct Kmeans clustering over the extracted features (unit size can be 100 -> 1000)
    - The Kmeans model is the unit extract model
  - Unit vocoder
    - Using extracted unit sequences to train a unit-based vocoder
    - Responsible for converting the unit sequence to final waveform

What is tradeoff of different unit size?

Typical steps of building discrete direct S2ST

- Step 3: Model Training
    - Train an end-to-end model
        - Input: source speech
        - Output: target speech in units
- Step 4: Inference
    - First use end-to-end model to predict unit sequence
    - Then use vocoder to convert it into the final waveform
- Step 5: Evaluation
    - Use an existing ASR model to generate ASR transcripts
    - Use ASR transcripts and reference target text to compute ASR-BLEU

More extensions

- Speech-to-speech translation has become one of the major tasks in recent speech foundation models, due to its compositional feature

- AudioPalm

https://arxiv.org/pdf/2306.12925.pdf

Types of tasks We apply our method to the problems of speech recognition, speech synthesis and speech-to-speech translation. All datasets used in this report are speech-text datasets which contain a subset of the following fields.

- Audio: speech in the source language.
- Transcript: a transcript of the speech in Audio.
- Translated audio: the spoken translation of the speech in Audio.
- Translated transcript: the written translation of the speech in Audio.

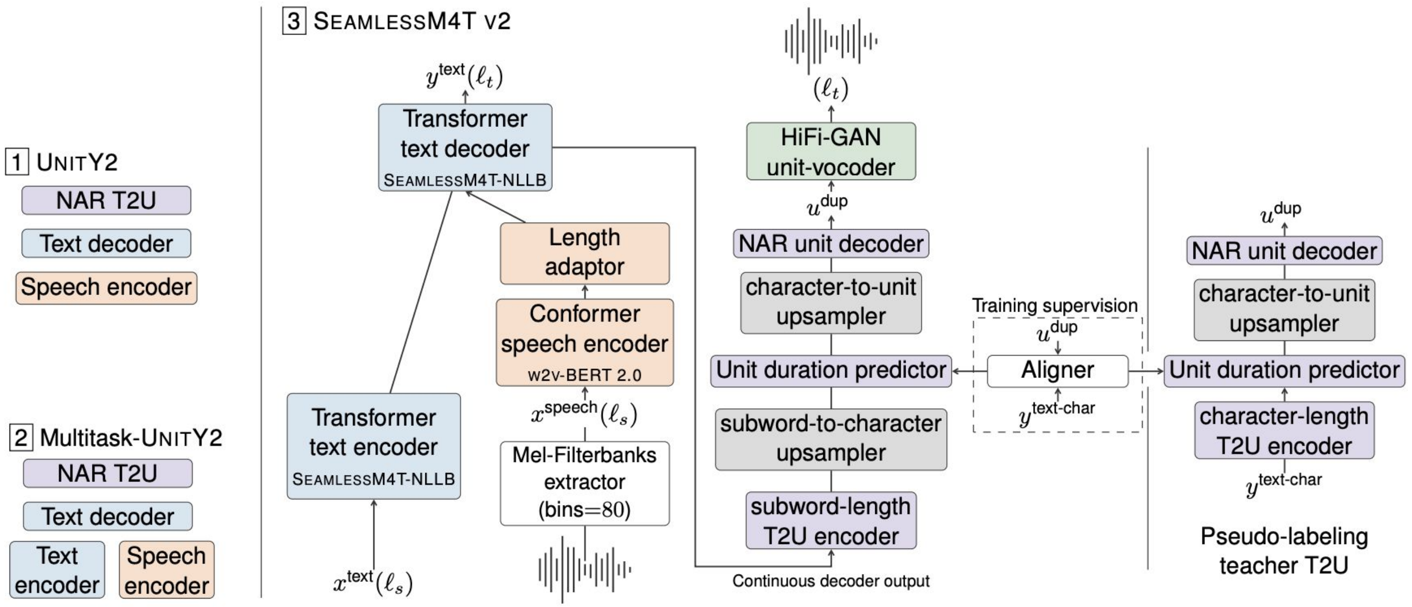
The component tasks that we consider in this report are:

- ASR (automatic speech recognition): transcribing the audio to obtain the transcript.
- AST (automatic speech translation): translating the audio to obtain the translated transcript.
- S2ST (speech-to-speech translation): translating the audio to obtain the translated audio.
- TTS (text-to-speech): reading out the transcription to obtain the audio.
- MT (text-to-text machine translation): translating the transcript to obtain the translated transcript.

More extensions

- Speech-to-speech translation has become one of the major tasks in recent speech foundation models, due to its compositional feature
- Seamless

https://ai.facebook.com/research/publications/seamless-multilingual-expressive-and-streaming-speech-translation/



Summary

- Speech-to-speech translation
  - Recap of basic speech-to-speech translation information
  - Technicals
    - Discrete and spectral approaches
    - Detailed steps of building a speech-to-speech translation system

References

- Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.
- Sellam, Thibault, Dipanjan Das, and Ankur Parikh. "BLEURT: Learning Robust Metrics for Text Generation." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.
- Ma, Xutai, et al. "SIMULEVAL: An Evaluation Toolkit for Simultaneous Translation." EMNLP. 2020.
- Ma, Mingbo, et al. "STACL: Simultaneous Translation with Implicit Anticipation and Controllable Latency using Prefix-to-Prefix Framework." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.
- Dalmia, Siddharth, et al. "Searchable Hidden Intermediates for End-to-End Models of Decomposable Sequence Tasks." Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021.
- Yan, Brian, et al. "CMU’s IWSLT 2022 dialect speech translation system." Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022). 2022.
- Yan, Brian, et al. "CTC alignments improve autoregressive translation." arXiv preprint arXiv:2210.05200 (2022).
- Yan, Brian, et al. "ESPnet-ST-v2: Multipurpose Spoken Language Translation Toolkit." preprint (2023).
- Inaguma, Hirofumi, et al. "UnitY: Two-pass Direct Speech-to-speech Translation with Discrete Units." arXiv preprint arXiv:2212.08055 (2022).