Zero-shot Speech Translation

Student: Tu Anh Dinh – i6164898

Supervisor: Dr. Jan Niehues

Examiners: Dr. Jan Niehues, Dr. Christof Seiler

Overview

- Problem definition
- Research questions
- Methods
- Experiments + Results
- Analysis
- Conclusions

Problem definition

Speech Translation

Speech Translation (ST):

Translating speech in one language into text in another language

"Some storms are worth the wreckage."

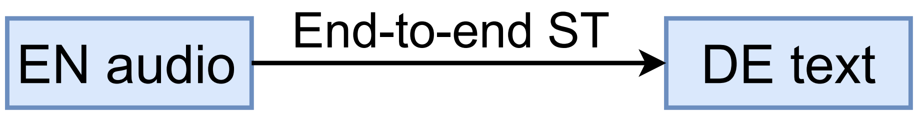
"Einige Stürme sind das Wrack wert."

Problem definition

Literature research

- Cascaded Speech Translation
  - Use 2 systems:
    - Automatic Speech Recognition (ASR)
    - Machine Translation (MT)
  - Problem: **error propagation**

- End-to-end Speech Translation
  - Use 1 system (*avoid error propagation*)
  - Problem: **lack of end-to-end ST data**



Problem definition

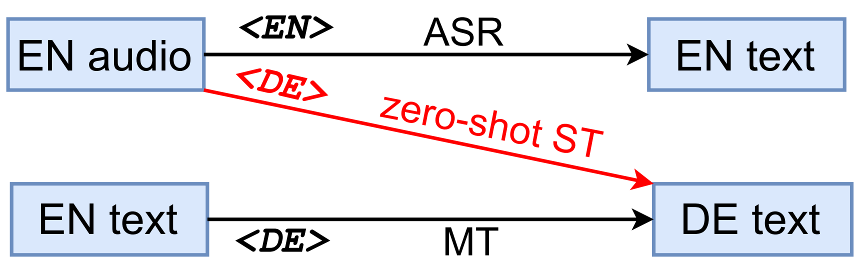
Proposed approach: Zero-shot Speech Translation

- Zero-shot:

  Enables translating a pair of languages *unseen during training*
  → **No end-to-end ST data needed**

- Requirement:

  Similar representation of EN audio and EN text



Research questions

1) How data-efficient are end-to-end and cascaded models?

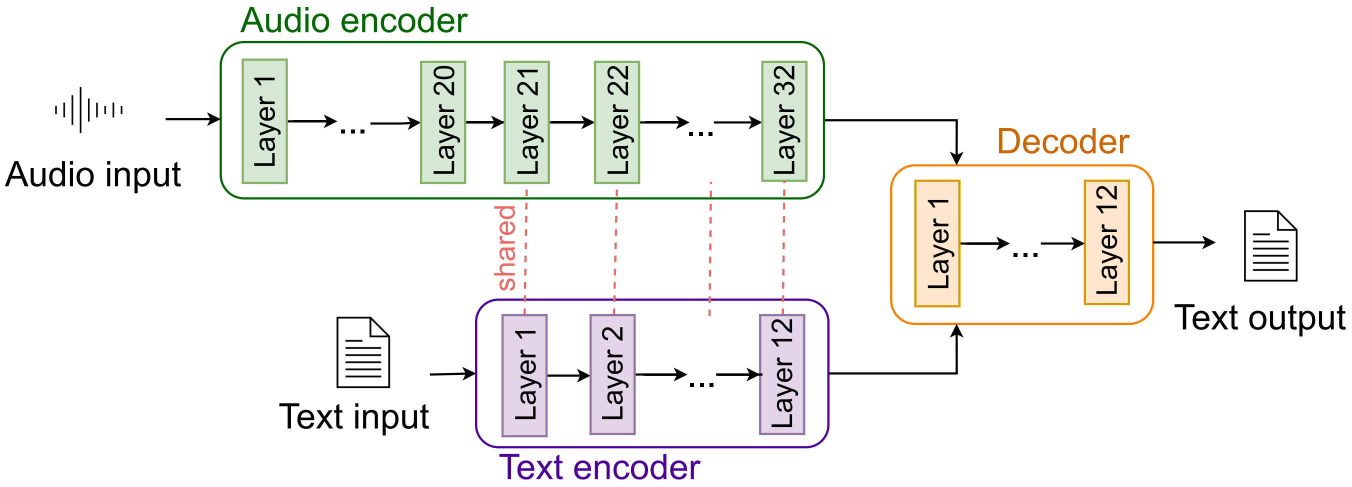
2) Can techniques from zero-shot multilingual machine translation be applied to end-to-end speech translation?

3) How can we model the different modalities in zero-shot speech translation?

Methods

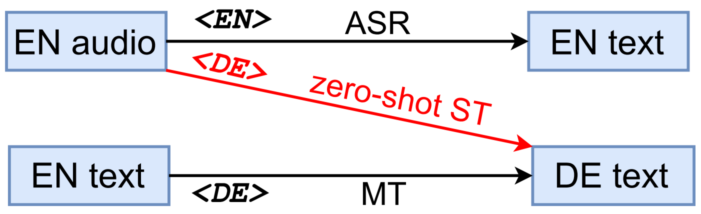
Zero-shot Speech Translation

- Model architecture: Transformer
- 2 parallel encoders:
  - Text encoder + Audio encoder
  - Share parameters



- Training data: ASR + MT
- Which language to output:

  Add **target-language tokens** to:
  - the beginning of input sequences
  - every decoder inputs



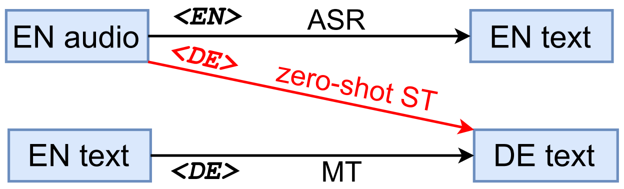
Challenge: zero-shot ST output wrong language (EN instead of DE)

Methods

Encourage Zero-shot Speech Translation

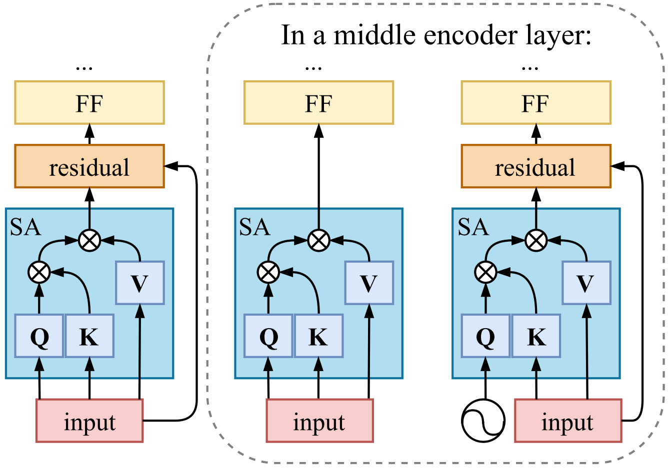
- Disentangling Positional Information
- Auxiliary loss function
- Data augmentation
- Additional opposite data

Methods



Encourage Zero-shot Speech Translation

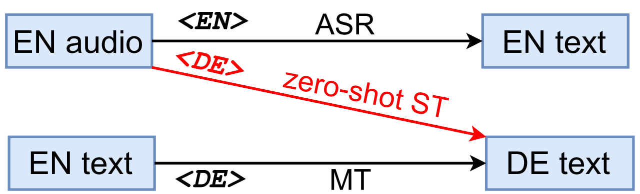
- Disentangling Positional Information
    - Originally used for zero-shot multilingual MT
    - Encourage **language-independent** representation
    - Idea: Remove residual connections in a middle encoder layer
      → Relax the strong positional correspondence of the output to input tokens
      → More freedom on word reordering
      → More language-independent



(a) Original (b) Drop Residual (c) Position-Based Query

(Image borrowed from the original paper)

Methods

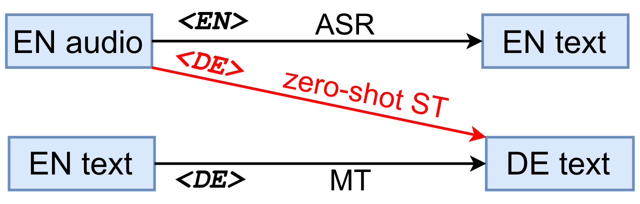


Encourage Zero-shot Speech Translation

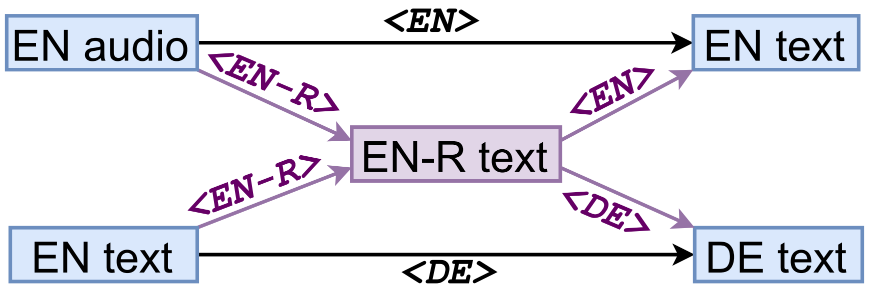
- Disentangling Positional Information
- Auxiliary loss function
  - Minimize text-audio encoder output difference
    → **Modality-independent** representation
  - Metrics for difference: squared error of mean-pool over time

Methods

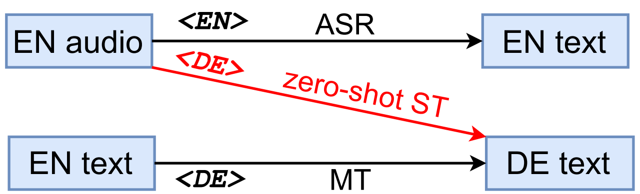
Encourage Zero-shot Speech Translation



- Disentangling Positional Information
- Auxiliary loss function
- Data augmentation
  - Aim: learn the **target-language tokens**
    More than 1 target language output for each modality
  - Artificial language: character-wise-reversed English (EN-R)
    E.g. “Hello world!” → “Dlrow olleh!”
  - Require no additional real dataset



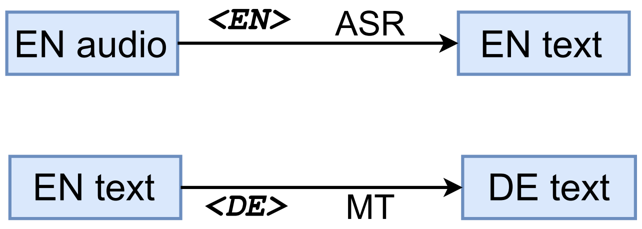
Methods



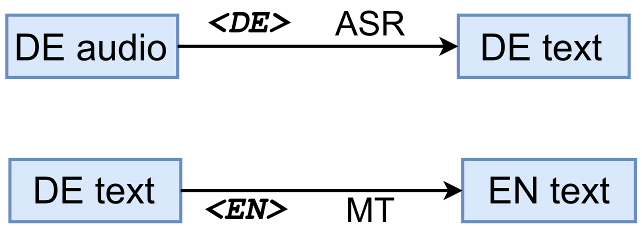
Encourage Zero-shot Speech Translation

- Disentangling Positional Information
- Auxiliary loss function
- Data augmentation
- Additional opposite data
  - Aim: learn the **target-language tokens**
    More than 1 target language output for each modality
  - Additionally require DE audio data and transcription

EN → DE



DE → EN



Methods

Few-shot models

- Zero-shot models: use **no** ST data
  Few-shot models: use **limited** amount of ST data
- Motivation:
  - Investigate the low-resource setting for ST data
- Building few-shot models:
  - Fine-tune zero-shot models with a small amount of ST data

Experiment setups

- Data: CoVoST 2
  - A large-scale multilingual ST corpus
  - Focus of the thesis: EN audio → DE text
    - 289K samples for training
    - 15K samples for validation
    - 15K samples for testing
- Metrics:
  - For translation tasks: BLEU score (the higher the better)
  - For ASR tasks: WER (the lower the better)

Experiments + Results

Data-efficiency of individual tasks

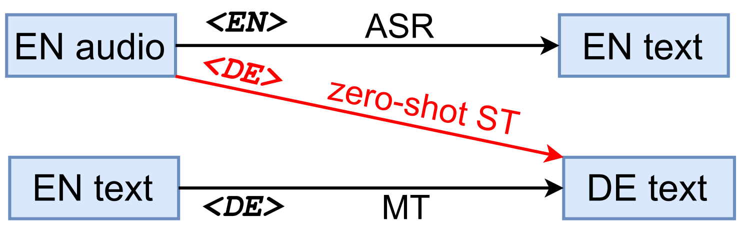
- The models are trained on single tasks independently
- Observations
  - ASR and MT tasks need less data than ST task
    → Motivation for zero-shot ST
  - Cascaded ST is more data-efficient than end-to-end ST
    → ✅ : Research question 1) answered

PERFORMANCE OF MODELS TRAINED ON SINGLE TASKS.

Data portion	ASR	MT	Cascaded ST	Direct end-to-end ST
25%	43.6	23.1	11.5	0.8
33%	37.5	26.3	14.0	1.6
100%	25.8	33.0	20.6	14.9

Experiments + Results

Plain Zero-shot Speech Translation



- Zero-shot model:
  - Can learn supervised tasks: ASR and MT
  - Output wrong language for zero-shot ST (EN instead of DE)
    Reason: text-audio difference
- Few-shot model: **9.8** BLEU points
  - Vs. direct end-to-end ST: **+9.3** BLEU points
  - Vs. fine-tuned ASR: **+1.4** BLEU points

Experiments + Results

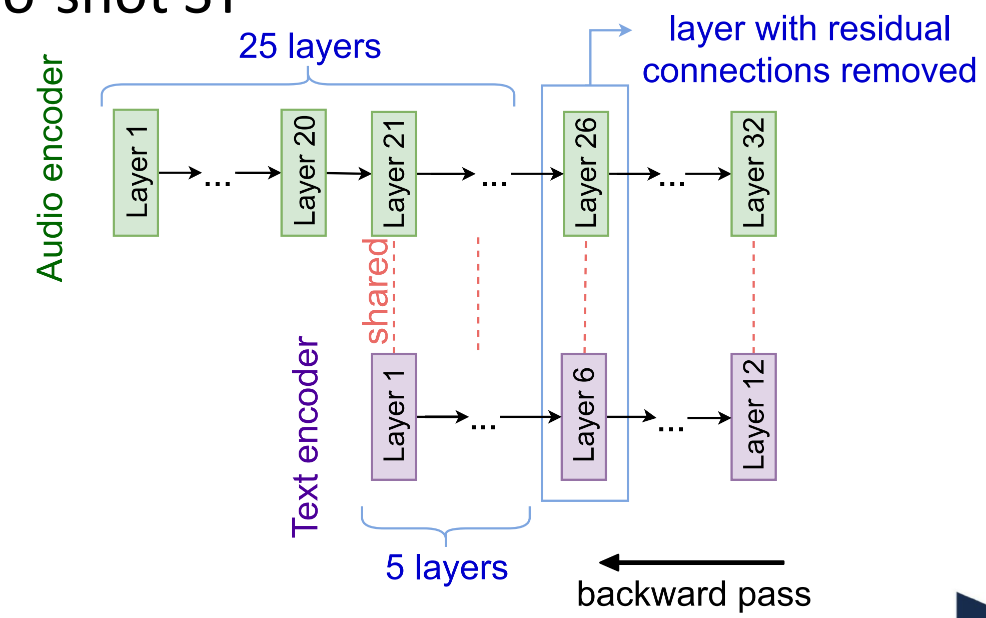
Encourage Zero-shot Speech Translation

- Disentangling Positional Information
- Auxiliary loss function
- Data augmentation
- Additional opposite data

Experiments + Results

Encourage Zero-shot Speech Translation

- Disentangling Positional Information
  - Fail to learn ASR task
    Worse performance on MT task
    Still output wrong language for zero-shot ST
  - Reason:
    - Audio encoder depth
    - Text-audio difference
  - ✅ : Answer research question 2)
    No, not all zero-shot multilingual MT techniques can be applied to zero-shot ST



Experiments + Results

Encourage Zero-shot Speech Translation

- Disentangling Positional Information
- Auxiliary loss function
  - Zero-shot model
    - Can learn supervised tasks (better ASR)
    - Zero-shot ST: **wrong** language
  - Few-shot model
    Vs. plain model: **+0.8** BLEU points
- Data augmentation
  - Zero-shot model
    - Can learn supervised tasks
    - Zero-shot ST: **some** correct language
  - Few-shot model
    Vs. plain model: **+1.7** BLEU points
- Additional opposite data
  - Zero-shot model
    - Can learn supervised tasks
    - Zero-shot ST: **mostly** correct language
  - Few-shot model
    Vs. plain model: **+0.5** BLEU points

Experiments + Results

Encourage Zero-shot Speech Translation

- Data augmentation + Auxiliary loss

  Same performance as without auxiliary loss

- Additional opposite data + Auxiliary loss: **best performance**
  - Best zero-shot ST: **1.5** BLEU points
  - Few-shot model: **12.3** BLEU points (**+2.5** vs. plain)

Analysis

Sentence level: SVCCA analysis

- Singular Vector Canonical Correlation Analysis (SVCCA)
- *EN audio – EN text* meanpooled encoder output
- Higher SVCCA score ↔ More text-audio similarity in **sentence level** ↔ More modality independent

Analysis

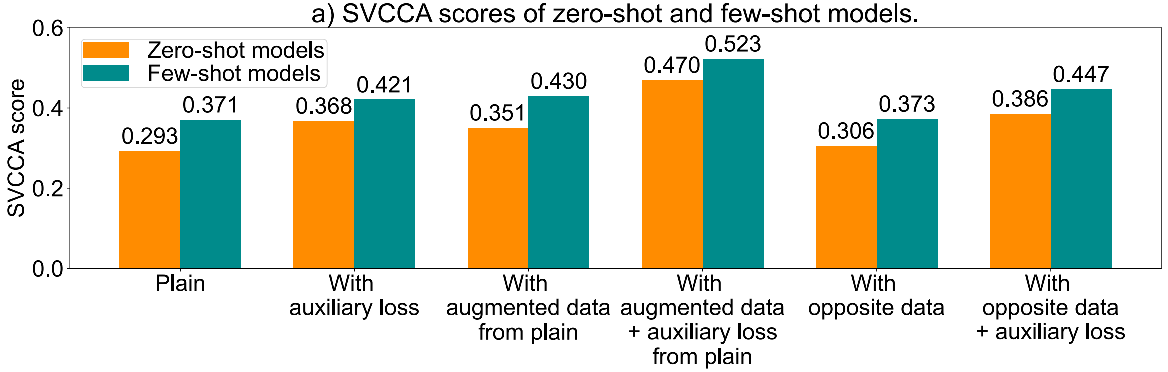
Sentence level: SVCCA analysis

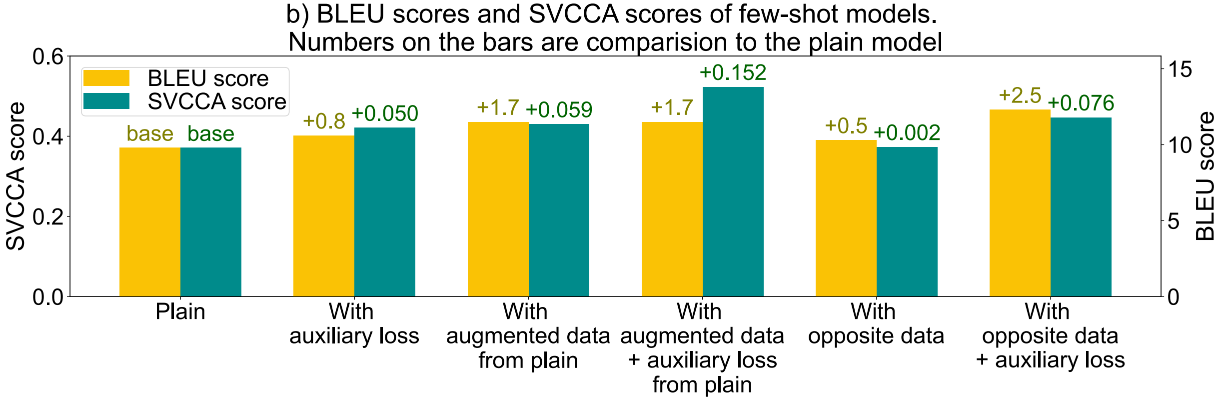
- Most approaches helps improve **modality independent**
  - ✅ : Research question 3) answered

→ Improve the ST performance

→ Promising approaches for zero-shot ST

- Exceptional case: adding opposite data:
  - Higher SVCCA score of EN text – DE text
    - → Improve **language independent** instead of modality independent
    - → Also help improve ST performance
- Best setting: auxiliary loss + opposite data:
  - Has both **language-independent** and **modality-independent** representation





Analysis

Token level: modality classifier

- Classify encoder output tokens (text/audio)
- Better classification performance → lower text-audio similarity
- Outcome:
    - Models with auxiliary loss:
      Most tokens classified as "audio"
    - Models without auxiliary loss:
      Over 99.9% classification accuracy

→ Auxiliary loss indeed improves **text-audio similarity** in **token level**

Conclusions

- Promissing approaches for zero-shot ST

- Particulary useful in the few-shot setting

- Future work: further enhance text-audio similarity
  Reduce the difference in the **number of time steps** between text and audio

Thank you for your attention!