# Utterance segmentation 

## Sec 5:

Advanced topics
Multilingual ST

Under-resourced languages

Sec 5.1

## Utterance

Segmentation

## Utterance segmentation - Problem

- Mismatch between training and evaluation data
- Training corpora: "sentence-level" split of continuous speech


This is an audio signal.


In the training data it was split using strong punctuation.


Three sentences in total!

## Utterance segmentation - Problem

- Mismatch between training and evaluation data
- Training corpora: "sentence-level" split of continuous speech


This is an audio signal.


In the training data it was split using strong punctuation.


Three sentences in total!

- At run-time: unsegmented continuous speech



## How to split continuous speech in cascade ST?

thisisanaudiosignalinthetrainingdataitwassplit

usingstrongpunctuationthreesentencesintotal


## How to split continuous speech in e2e ST?



[^0]
## Solution 1: Split on silences (via VAD)


thisisanaudiosignalinthetrainingdataitwassplitusingstrongpunctuationthreesentencesintotal

this is an audio signal


## Solution 1: Split on silences (via VAD)


thisisanaudiosignalinthetrainingdataitwassplitusingstrongpunctuationthreesentencesintotal

this is an audio signal


Advantage: silences as a proxy of sentence boundaries
Drawback: variable segments' length (including very short and very long ones)

## Solution 2: Split based on fixed audio duration


thisisanaudiosignalinthetrainingdataitwassplitusingstrongpunctuationthreesentencesintotal


## Solution 2: Split based on fixed audio duration


thisisanaudiosignalinthetrainingdataitwassplitusingstrongpunctuationthreesentencesintotal

ining data it was split using strong pu


Advantage: uniform segment length
Drawback \#1: split points are likely to break the input in critical positions Drawback \#2: non-speech frames are kept in the input

## Solution 3: Split on silences \& segments' length

Potapczyk and Przybysz: "SRPOL’s system for the IWSLT 2020 end-to-end speech translation task", IWSLT 2020

this is an audio signal


using strong punctuation three sentences in total

using strong punctuation

three sentences in total

## Solution 3: Split on silences \& segments' length

Potapczyk and Przybysz: "SRPOL’s system for the IWSLT 2020 end-to-end speech translation task", IWSLT 2020
this is an audio signal in the training data it was split

this is an audio signal

in the training data it was split

using strong punctuation three sentences in total

using strong punctuation

three sentences in total

Advantages: closer to sentence-like splits, uniform segment length
Drawback \#1: manually-detected silences (non scalable/reproducible)
Drawback \#2: full audio required for splitting (not applicable to audio streams)

## Utterance segmentation - An open problem

- Manual
- VADFixedHybrid (VAD-based)


Large room for improvement compared to manual segmentation

## Utterance segmentation - An open problem

Manual

- VADFixedHybrid (VAD-based)


FIXED length surprisingly good
$\rightarrow$ segments' length is more important than precise split times

## Utterance segmentation - An open problem

- ManualHybrid (VAD-based)


Fully automatic hybrid segmentation?
$\rightarrow$ better than VAD, better than FIXED on one language pair


[^0]:    thisisanaudiosignalinthetrainingdataitwassplitusingstrongpunctuationthreesentencesintotal

