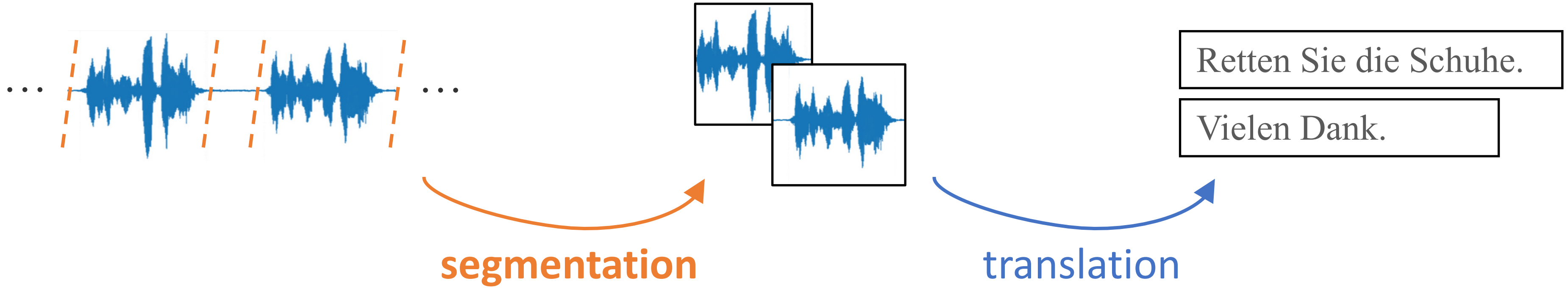
Speech Segmentation Optimization using Segmented Bilingual Speech Corpus for End-to-End Speech Translation

Ryo Fukuda, Katsuhito Sudoh, Satoshi Nakamura
Nara Institute of Science and Technology, Japan

Segmentation for Speech Translation

Speech segmentation is essential for automatic speech translation (ST)

- splits continuous speech into short segments before translation

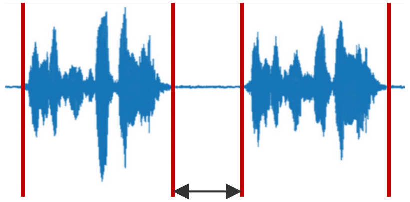


- existing ST systems cannot directly translate long continuous speech
- explicit segment boundaries are unavailable in speech

Related work: Pause-based segmentation

Voice Activity Detection (VAD) - traditional approach [Sohn+1999][Bangalore+2012]

- splits speech based on detected **silences**



- large gap with the manual segmentation

Segm. method	MuST-C en-de		Europarl en-de		MuST-C en-it		Europarl en-it	
	BLEU (↑)	TER (↓)	BLEU (↑)	TER (↓)	BLEU (↑)	TER (↓)	BLEU (↑)	TER (↓)
Manual segm.	27.55	58.84	26.61	60.99	27.70	58.72	28.79	59.16
Best VAD	21.87	66.72	18.51	78.12	22.34	66.12	20.90	69.54
Best Fixed (20s)	23.86	**61.29**	23.27	64.01	23.20	64.24	22.28	64.57
SRPOL-like	22.26	71.10	20.49	77.61	23.12	66.27	23.26	66.19
Pause in 17-20s	**24.39**	61.35	**23.78**	**63.15**	**23.50**	**63.76**	22.86	63.44
+ force split	23.17	66.20	22.52	68.56	23.45	63.79	**24.15**	**63.31**

Table 3: Comparison between manual and automatic segmentations: VAD, fixed-length and hybrid approaches.

(quote from [Papi+2021])

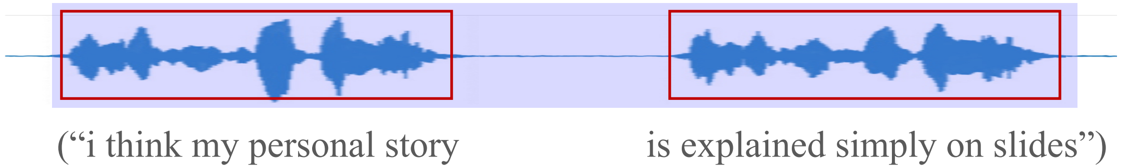
Related work: Pause-based segmentation

Voice Activity Detection (VAD) - traditional approach [Sohn+1999][Bangalore+2012]

- pauses do not always match sentence boundaries

  → Over-/Under-segmentation problem

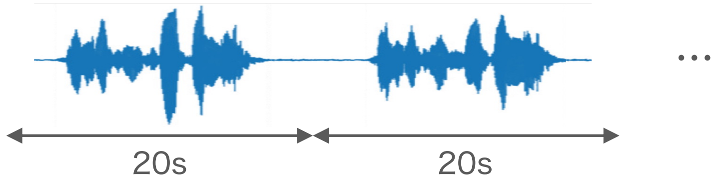
■ Long silences in a (oracle) segment → Over-segmentation



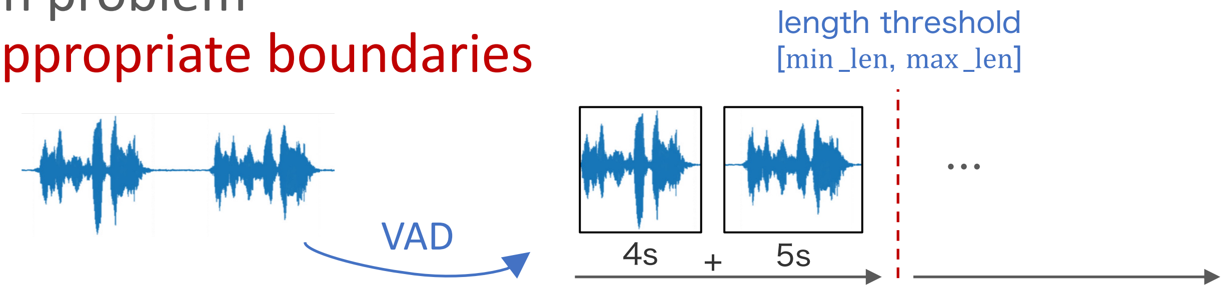
■ Short silences between (oracle) segments → Under-segmentation

Related work: Length-based segmentation

- **Fixed-length segmentation** [Sinclair+2014]
  - simple but works better than pause-based segmentation [Gaido+2021]
  - does not take acoustic and linguistic clues into account

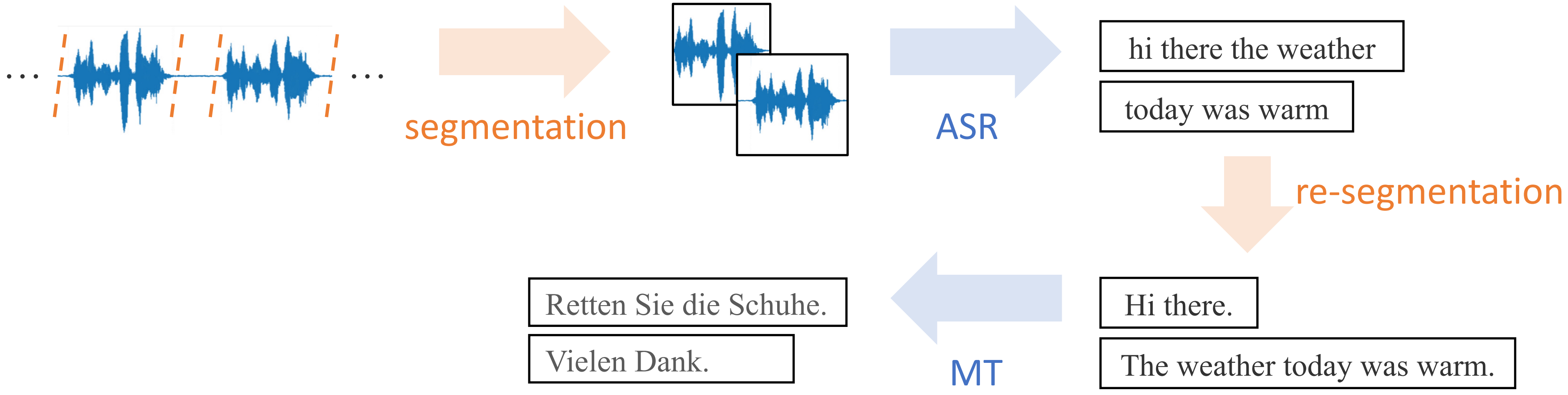


- **Hybrid of pause- and length-based segmentation** [Potapczyk+2020][Gaido+2021][Inaguma+2021]
  - heuristic concatenation of VAD segments up to a fixed length to address the over-segmentation problem
  - still splits audio at inappropriate boundaries



Related work: Re-segmentation of transcripts

- **Re-segmenting ASR results**
  - punctuation restoration [Lu+2010][Rangarajan Sridhar+2013][Niehues+2015]
  - language model [Stolcke and Shriberg, 1996][Wang+2016]
  - corpus-based segmentation model [Wan+2021][Wang+2019][Iranzo-Sánchez+2020]



- difficult to use in end-to-end ST, and cannot recover ASR errors due to improper segmentation

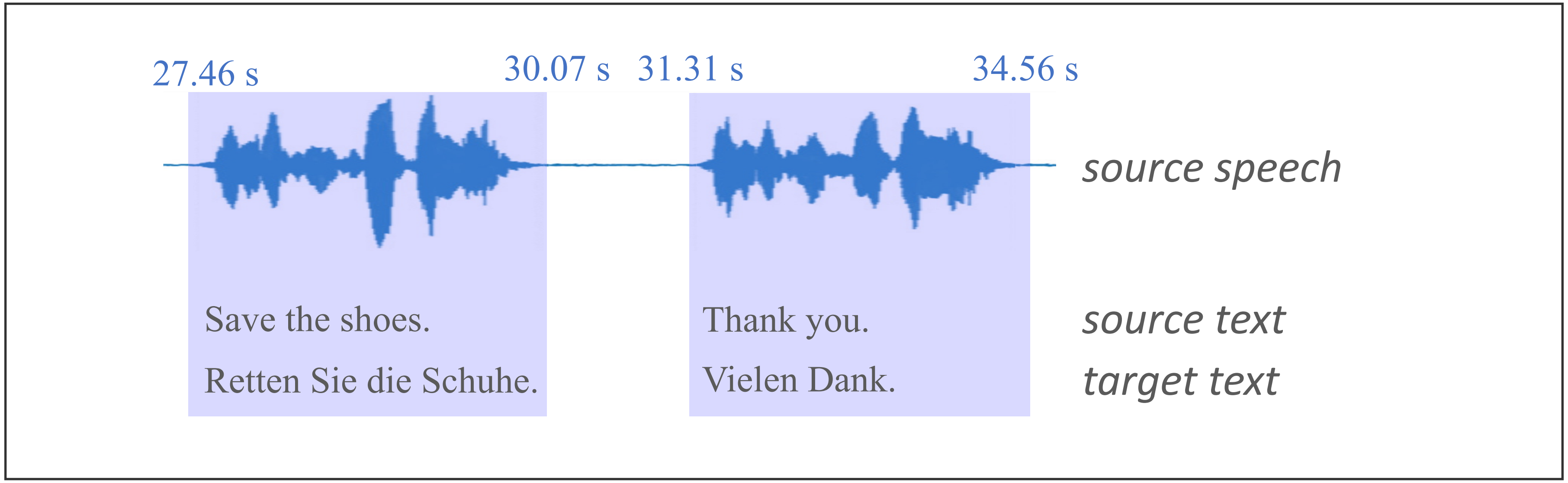
Proposed Method

Corpus-based segmentation

Speech segmentation as a frame-level sequence labeling task

- use a bilingual speech corpus as training data for speech segmentation
- bilingual speech corpus includes **speech segments aligned to sentence-like unit**

Bilingual speech corpus

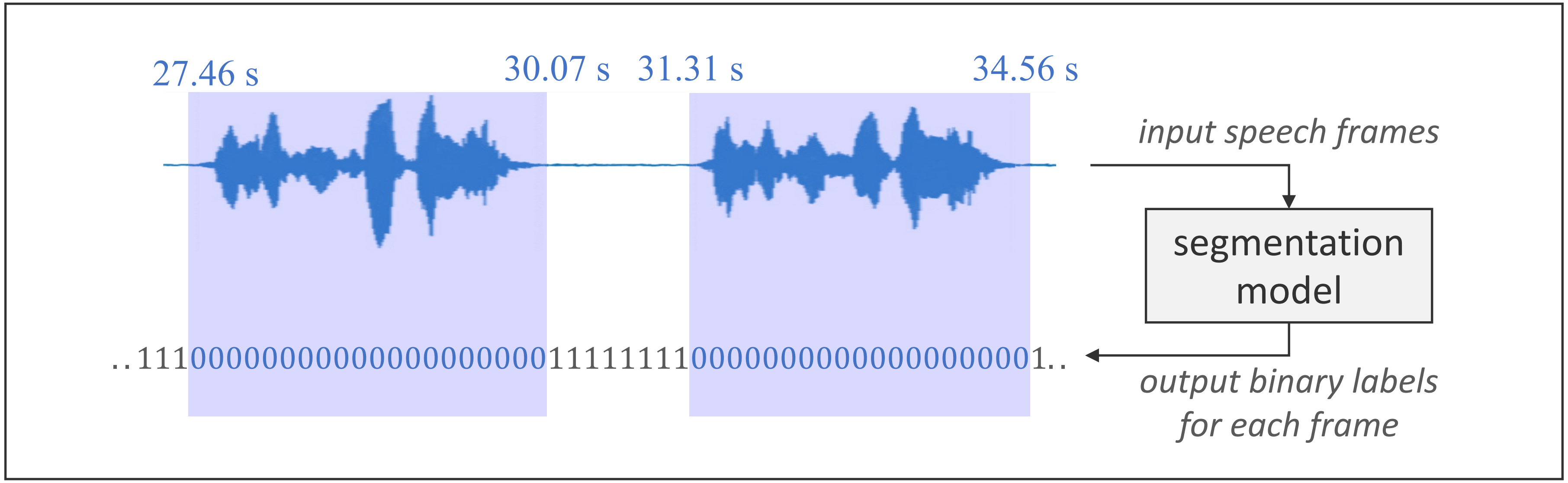


Corpus-based segmentation

Speech segmentation as a frame-level sequence labeling task

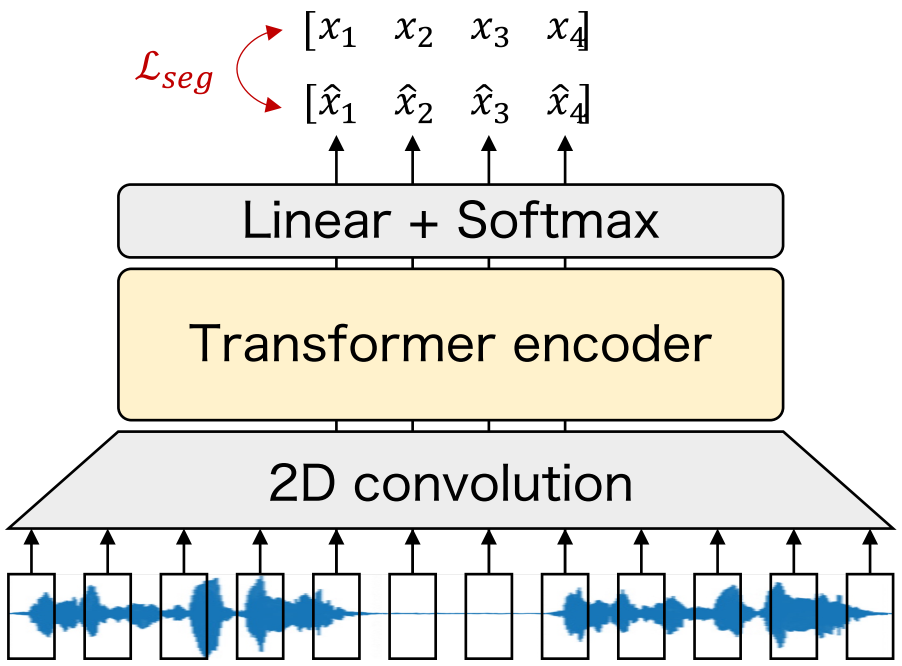
- use a bilingual speech corpus as training data for speech segmentation
- bilingual speech corpus includes **speech segments aligned to sentence-like unit**

Bilingual speech corpus



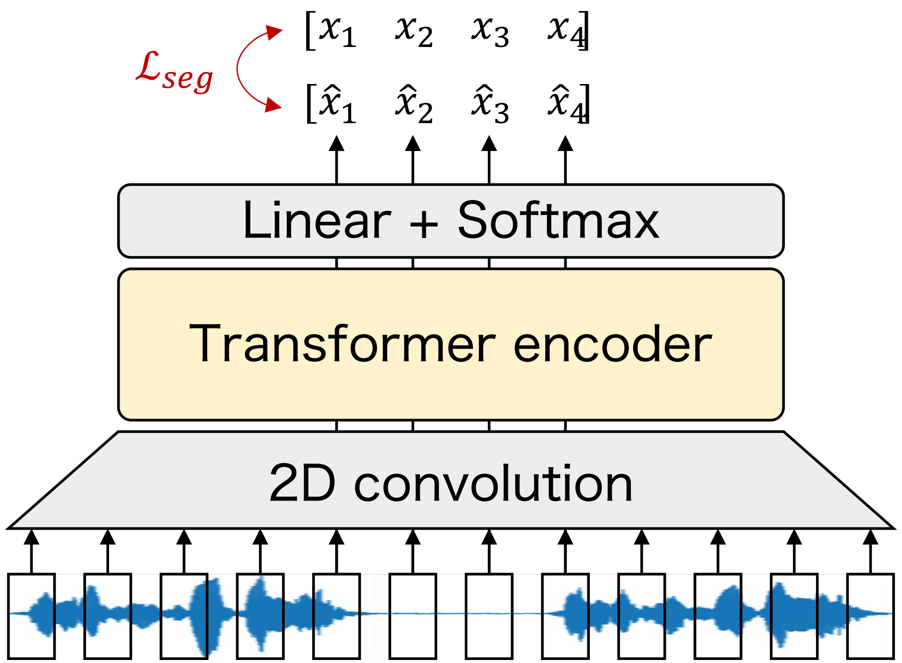
Speech segmentation model

- Model: 2D convolution ＋ Transformer Encoder



Speech segmentation model

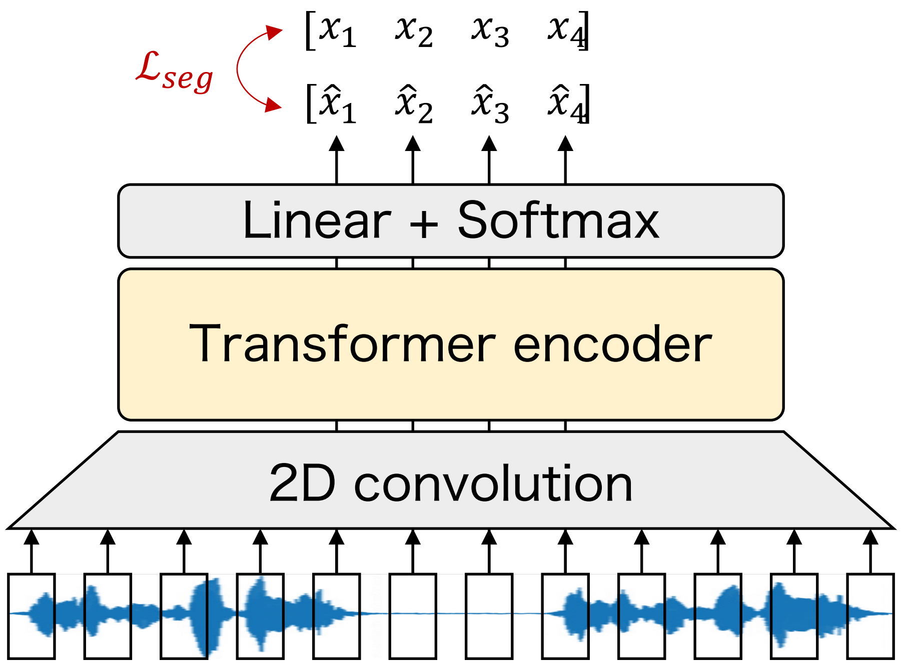
- Model: 2D convolution + Transformer Encoder
- Data: two consecutive segments are concatenated and assigned a sequence of label $x \in \{0,1\}$



Speech segmentation model

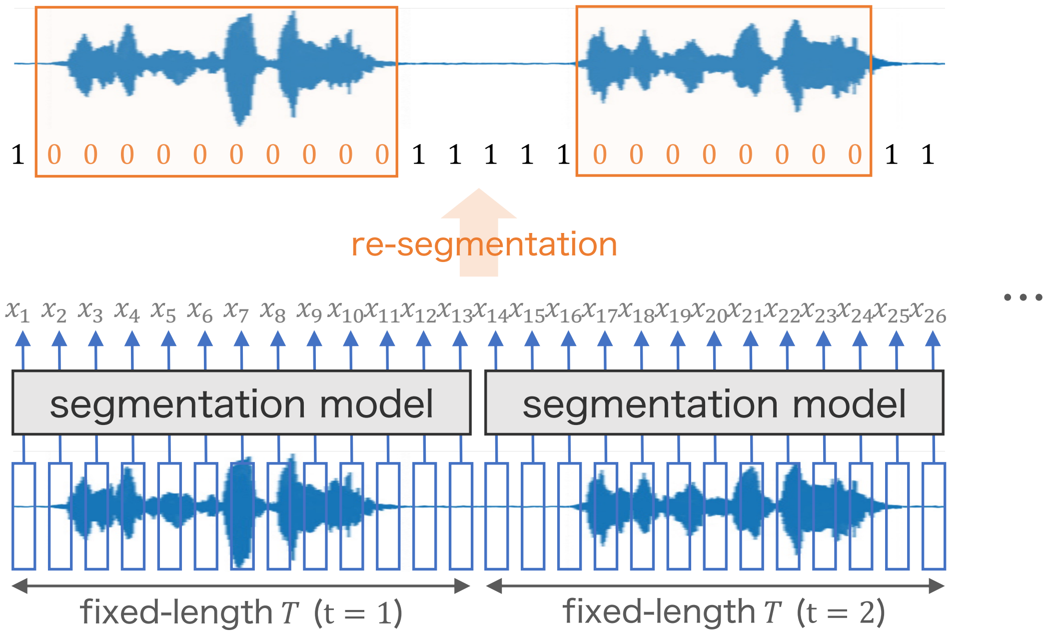
- Model: 2D convolution + Transformer Encoder
- Data: two consecutive segments are concatenated and assigned a sequence of label $x \in \{0,1\}$
- Training objective: cross-entropy $\mathcal{L}_{seg}(\hat{x}, x)$ with rescaling weight w_s

$$\mathcal{L}_{seg}(\hat{x}, x) = -\sum_{n=1}^{N} \left\{ w_s \log \frac{\exp(\hat{x}_{n,1})}{\exp(\hat{x}_{n,0} + \hat{x}_{n,1})} x_{n,1} + (1 - w_s) \log \frac{\exp(\hat{x}_{n,0})}{\exp(\hat{x}_{n,0} + \hat{x}_{n,1})} x_{n,0} \right\}$$



Inference Process

1. speech is segmented at a fixed-length T and input into the segmentation model
2. fixed-length segments are re-segmented according to the labels predicted by the segmentation model



Prediction

- segmentation model selects label $l_n \in \{0,1\}$ with the highest probability at each time n:

$$l_n \coloneqq \mathrm{argmax}(\hat{x}_n) \quad (\hat{x}_n \in R^2)$$

Hybrid method

- combine the model predictions with the VAD results $\mathrm{vad}_n \in \{0,1\}$:

$$l_n \coloneqq \begin{cases} \mathrm{argmax}(\hat{x}_n) \ \wedge \mathrm{vad}_n \ (segm\ len < maxlen) \\ \mathrm{argmax}(\hat{x}_n) \ \vee \mathrm{vad}_n \ (segm\ len \geq maxlen) \end{cases}$$

Experiments

Experimental Settings

Data: Multilingual Speech Translation Corpus (MuST-C) [Gangi+2019]

- English-German: 230k segments from MuST-C v1
- English-Japanese: 330k segments from MuST-C v2

Segmentation methods

- Baseline: WebRTC VAD (**VAD**), pre-defined fixed length (**Fixed-length**)
- Proposal: Segmentation model (**Our model**), hybrid method (**VAD hybrid**)

ST systems

- **Cascade ST**: cascade of ASR and MT models
- **End-to-end ST**: an ST model that directly translates English speech

Evaluation

- WER and BLEU for hypotheses re-segmented by the edit distance-based algorithm [Matusov+2005]

Overall Results

MuST-C v1 English-German

	Cascade ST		End-to-end ST
	WER	BLEU	BLEU
Oracle	12.60	23.59	22.50
Best VAD	30.59	17.02	16.40
Best Fixed-length	20.60	19.29	17.96
Our model	20.99	20.18	19.10
+ VAD hybrid	**19.06**	**20.99**	**19.87**

MuST-C v2 English-Japanese

	Cascade ST		End-to-end ST
	WER	BLEU	BLEU
Oracle	9.30	12.50	10.60
Best VAD	25.81	9.26	8.14
Best Fixed-length	18.89	9.64	8.52
Our model	16.21	9.71	8.77
+ VAD hybrid	**13.67**	**10.60**	**9.24**

- Our model outperformed VAD and the fixed-length baselines for both cascade and end-to-end STs
- The hybrid method with VAD significantly improved translation performance
- room for improvement remains compared to the oracle segments contained by the MuST-C corpus (**Oracle**)

Case study (1)

Example of ASR and MT outputs with segmentation positions (■)

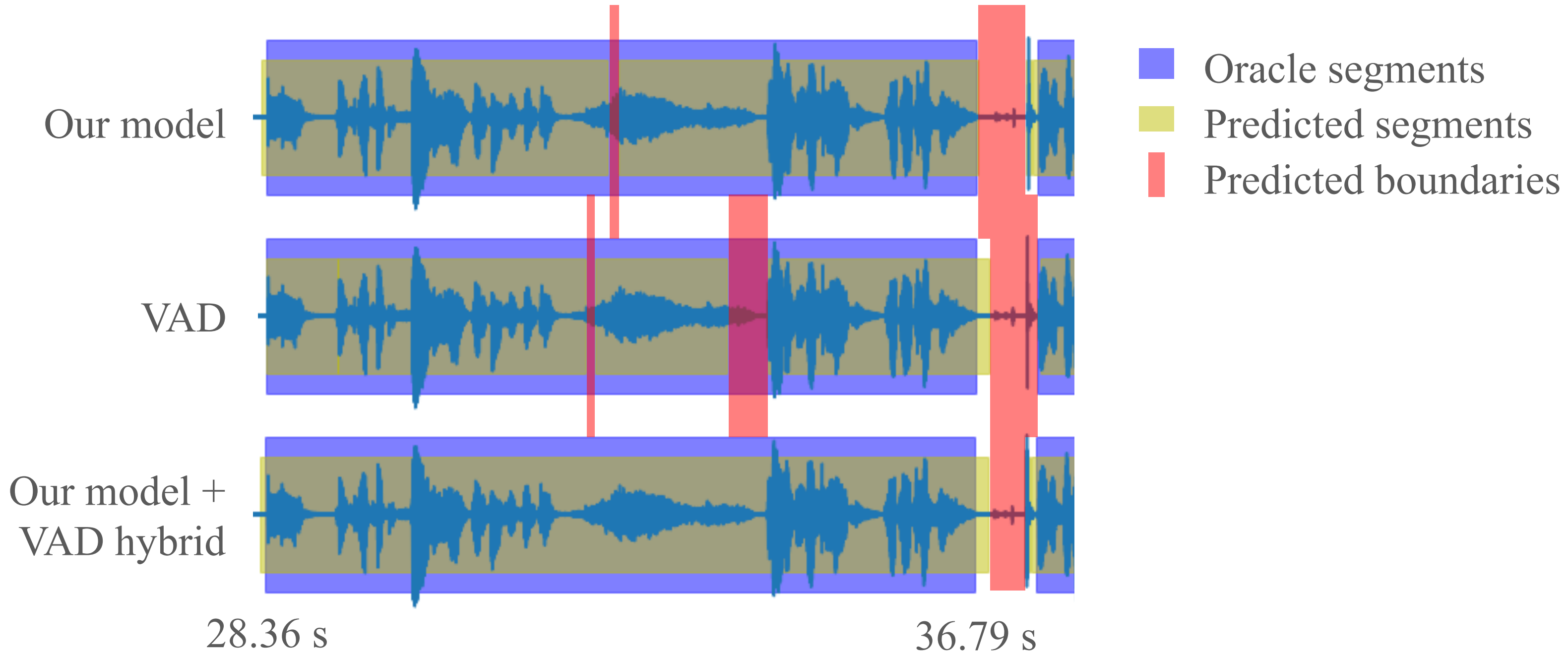
Oracle (ASR)	*bonobos are together with chimpanzees you aposre living closest relative* ■
Oracle (MT)	*Bonobos sind zusammen mit Schimpansen, Sie leben am nächsten Verwandten.* ■
Best VAD (ASR)	*bonobos are* ■ *together with chimpanzees you aposre living closest relative that ...*
Best VAD (MT)	*Bonobos sind es.* ■ *Zusammen mit Schimpansen leben Sie im Verhältnis zum ...*
Our model (ASR)	*bonobos are together with chimpanzees you aposre living closest relative* ■
Our model (MT)	*Bonobos sind zusammen mit Schimpansen, Sie leben am nächsten Verwandten.* ■

- VAD resulted in over-segmentation ("*bonobos are* ■ *together*") and under-segmentation ("*relative that ...*").
- our model split the speech at a boundary close to an oracle segment and obtained the same ASR and MT results

Case study (2)

Visualization of waveforms and segmentation positions

- hybrid decoding alleviated the over-segmentation problem by requiring an agreement between our model and VAD



Conclusions

Speech segmentation method based on bilingual speech corpus

- directly split speech into segments that correspond to sentence-like units

Experimental results

- our method outperformed the existing methods on both cascade and end-to-end STs
- hybrid approach with VAD further improved the translation performance

Future work

- investigation on different domains and noisy environments
- integration of segmentation function into an end-to-end ST

Appendix

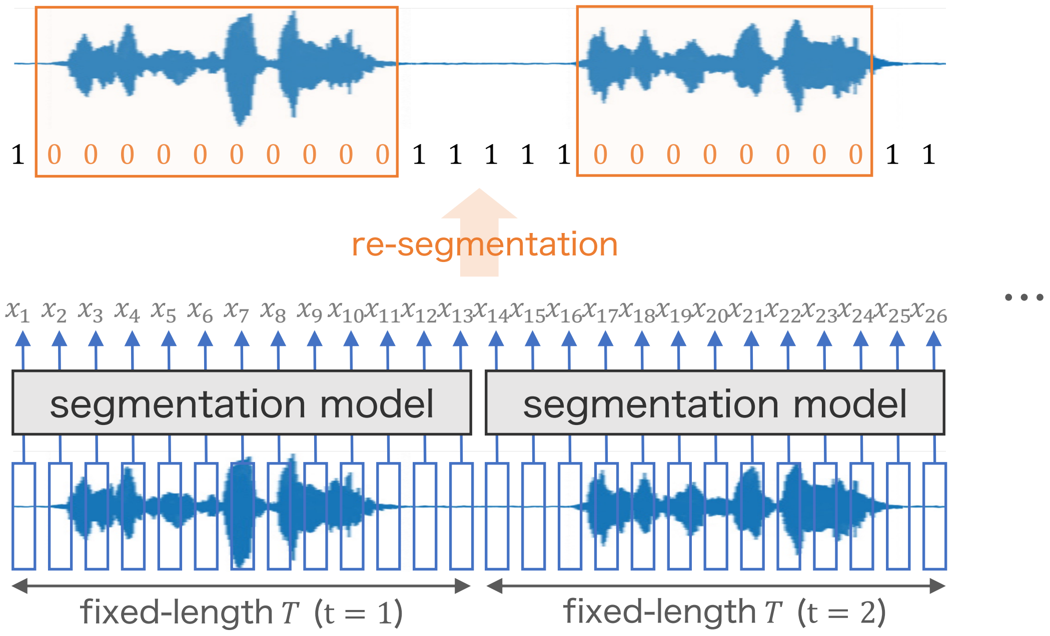
ST model Settings

表 3　Transformer の設定. † バージョン 0.10.3.

設定（ESPnet† の変数名）	ASR	ST	MT
エポック数（epochs）	45	100	
Encoder 層の数（elayers）	12		6
Decoder 層の数（elayers）	6		
FNN の次元数（eunits, dunits）	2048		
Attention の次元数（adim）	256		
Attention のヘッド数（aheads）	4		
ミニバッチ数（batch-size）	64		96
勾配蓄積（accum-grad）	2		1
勾配クリッピング（grad-clip）	5		
学習率（transformer-lr）	5	2.5	1
ウォームアップ（transformer-warmup-steps）	25000		
ラベル平滑化（lsm-weight）	0.1		
ドロップアウト率（dropout-rate）	0.1		

Inference Process

1. speech is segmented at a fixed-length T and input into the segmentation model
2. fixed-length segments are re-segmented according to the labels predicted by the segmentation model



Prediction

- segmentation model selects label $l_n \in \{0,1\}$ with the highest probability at each time n:

$$l_n \coloneqq \mathrm{argmax}(x_n)$$

Hybrid method

- combine the model predictions with the VAD results vad_n:

$$l_n \coloneqq \begin{cases} \mathrm{argmax}(x_n) \wedge \mathrm{vad}_n \; (segm\ len < maxlen) \\ \mathrm{argmax}(x_n) \vee \mathrm{vad}_n \; (segm\ len \geq maxlen) \end{cases}$$