音声機械翻訳のための音声翻訳コーパスに基づく発話分割

◯福田 りょう, 須藤 克仁, 中村 哲
奈良先端科学技術大学院大学
知能コミュニケーション研究室

研究概要

➢ 音声機械翻訳では、適切なセグメントへの自動発話分割が重要

- **問題**：発話区間に基づく自動分割は翻訳に不適切なセグメントを生成
- **本発表**：連続音声を、「文」に対応するセグメントへ直接分割

  音声翻訳コーパスのセグメント境界を学習し、
  音声から「文」の境界を直接予測する発話分割モデルを提案

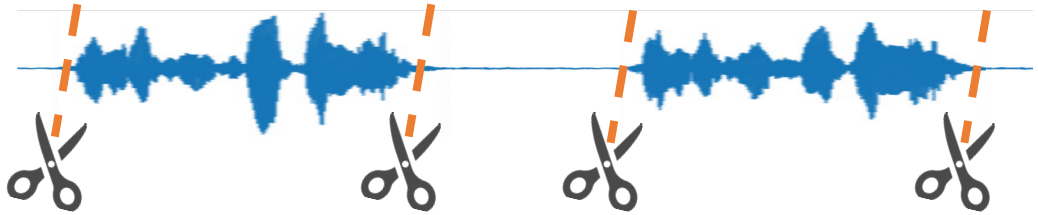
TED talksの英語-ドイツ語 音声翻訳実験において
従来の発話分割手法よりも高い翻訳性能を達成

研究背景

音声機械翻訳（ST）における発話分割

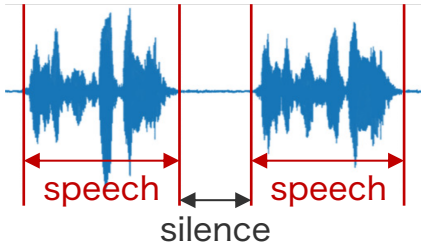
- 音声機械翻訳（ST; Speech Translation）… 音声→テキストの機械翻訳

- **STの前段の処理である、セグメントへの発話分割は性能に大きく影響**
  - 学習時：音声翻訳コーパスに含まれる、分割済みの音声セグメントを使用
  - 実用時：セグメントの境界は未知 → **自動発話分割**



発話区間に基づく分割 - VAD

- 音声区間検出（VAD; Voice Activity Detection）[Sohn+99][Bangalore+12]
  - 最も標準的な手法。無音（silence）を基準とした分割



  - 人手による分割とのギャップが大きい [Papi+21]

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.

（[Papi+21]から引用）

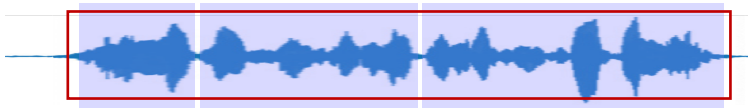
発話区間に基づく分割 - VAD

- 音声区間検出（VAD; Voice Activity Detection）[Sohn+99][Bangalore+12]
  - 無音は必ずしも文境界とは一致しない → 過剰分割・過少分割

■ 長い無音による文の断片化 → **過剰分割（Over-segmentation）**

(“i think my personal story　　is explained simply on slides”)

■ 短い無音を無視した文の一体化 → **過少分割（Under-segmentation）**



(“uh no he is in trouble what is the problem”)

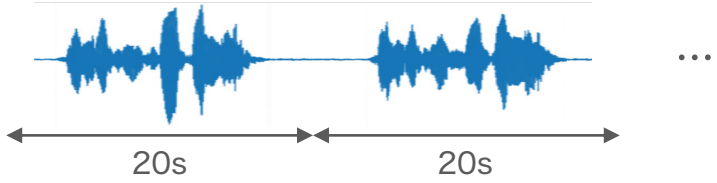
人手による分割

VADによる分割

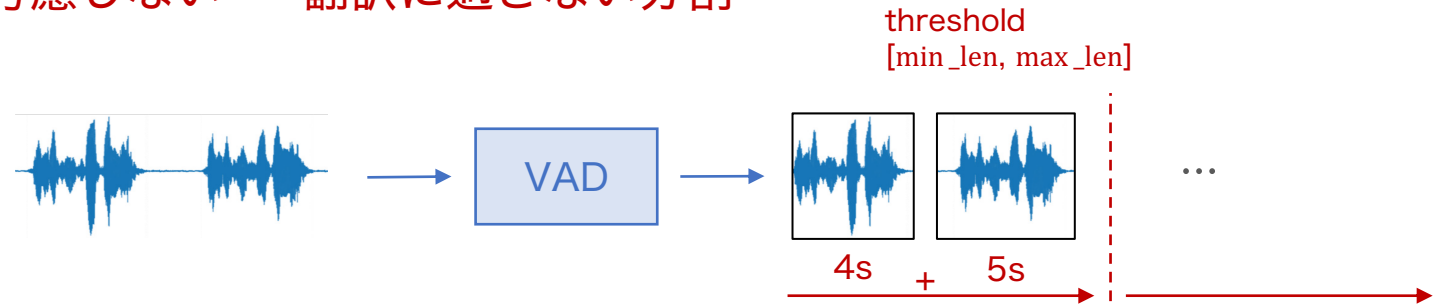
関連研究

長さに基づく分割

- 固定長セグメントへの分割 [Sinclair+14]
  - セグメント長が一定に保たれる利点がある。VADを上回るという報告もある [Gaido+21]
  - 音響や言語情報を考慮しないため不自然な分割が頻繁に起こる

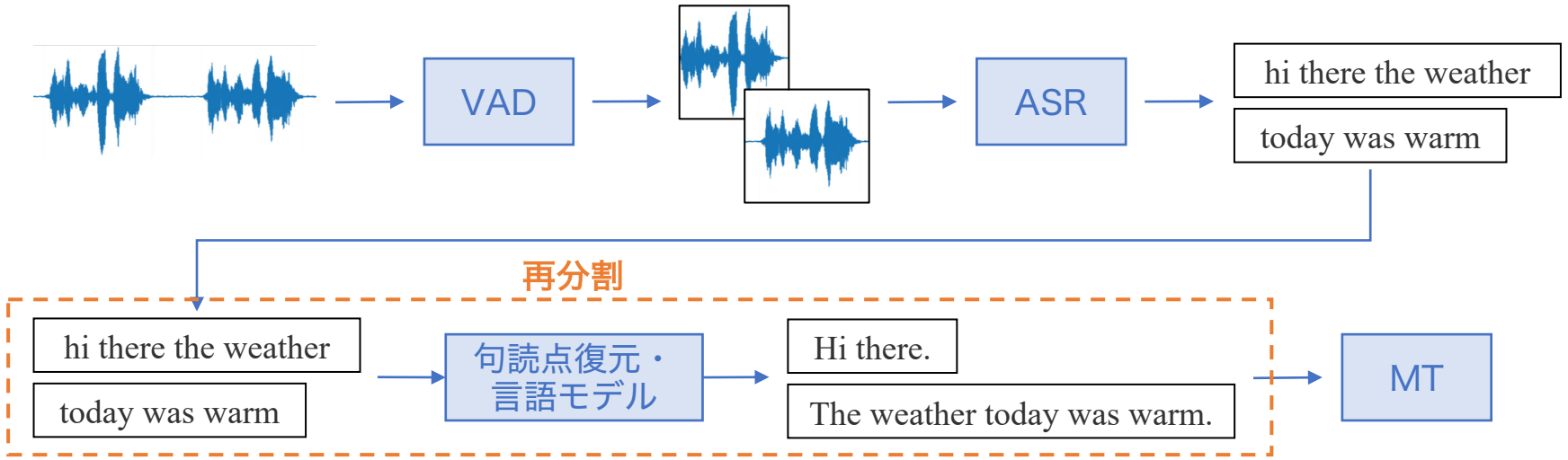


- VADと固定長分割の組み合わせ推論 [Potapczyk+20][Gaido+21][Inaguma+21]
  - 音の強さとセグメントの長さを考慮 → 過剰分割の緩和
  - 言語情報を考慮しない → 翻訳に適さない分割



言語情報に基づく再分割 [Ore+20][Pham+20][Nguyen+21]

- 句読点復元モデルや言語モデルを用いて、音声認識（ASR）出力を再分割



- 文単位に再分割することで機械翻訳（MT）の翻訳精度が向上 → Cascade STで広く利用
- ASRモデルを介さない、End-to-end STモデルへの適用が難しい
- 不適切な分割による、ASRの認識誤りを防げない

本研究の目的

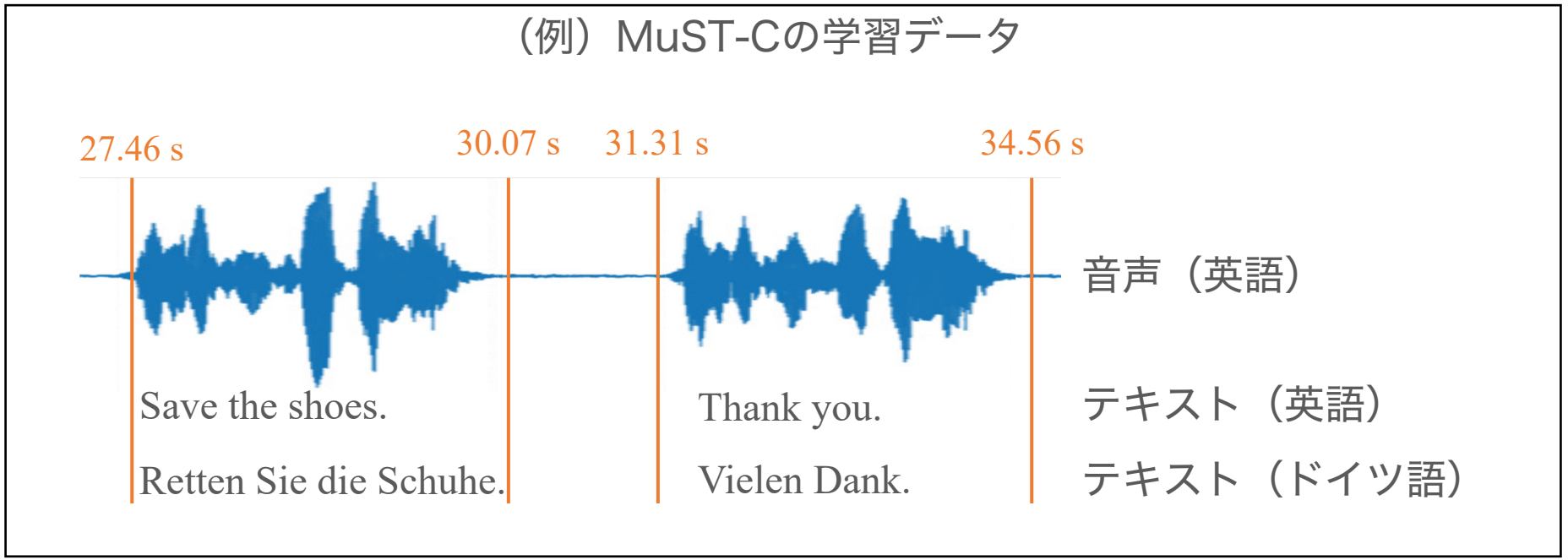
連続音声を「文」に対応するセグメントへ直接分割

- より翻訳に適した音声セグメントへ
- End-to-end STへ適用するため音声レベルで分割

提案手法

音声翻訳コーパスに基づく発話分割学習

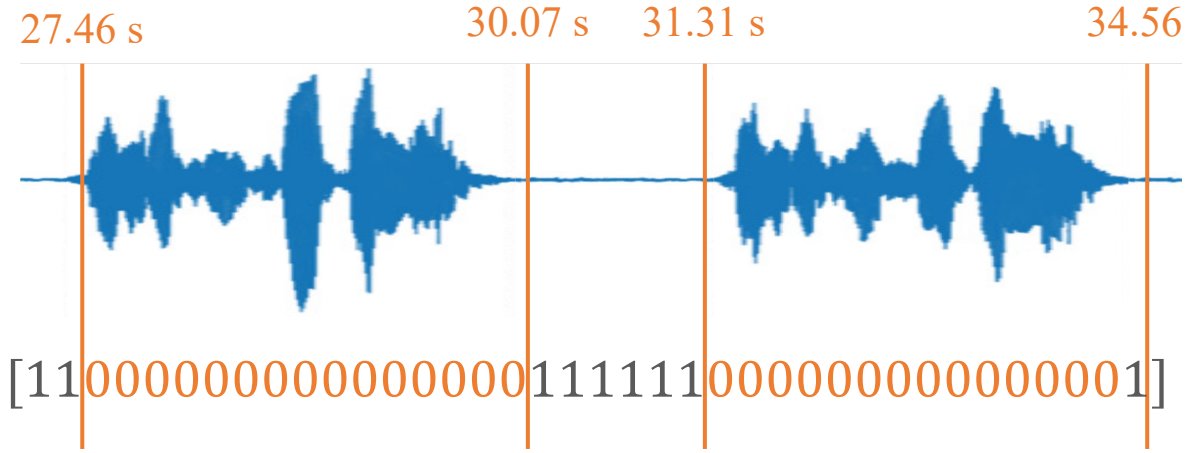
- MuST-C … TED Talksの音声翻訳コーパス
  文単位のテキストに対応付いた音声を入手可能 → **文単位の音声分割の学習**



音声翻訳コーパスに基づく発話分割学習

- MuST-C … TED Talksの音声翻訳コーパス
  文単位のテキストに対応付いた音声を入手可能 → **文単位の音声分割の学習**

提案手法の学習データ。時刻情報を基に、音響特徴量のフレームにラベル$x \in \{0, 1\}$を付与



- 入力：音響特徴量
- 出力：ラベル

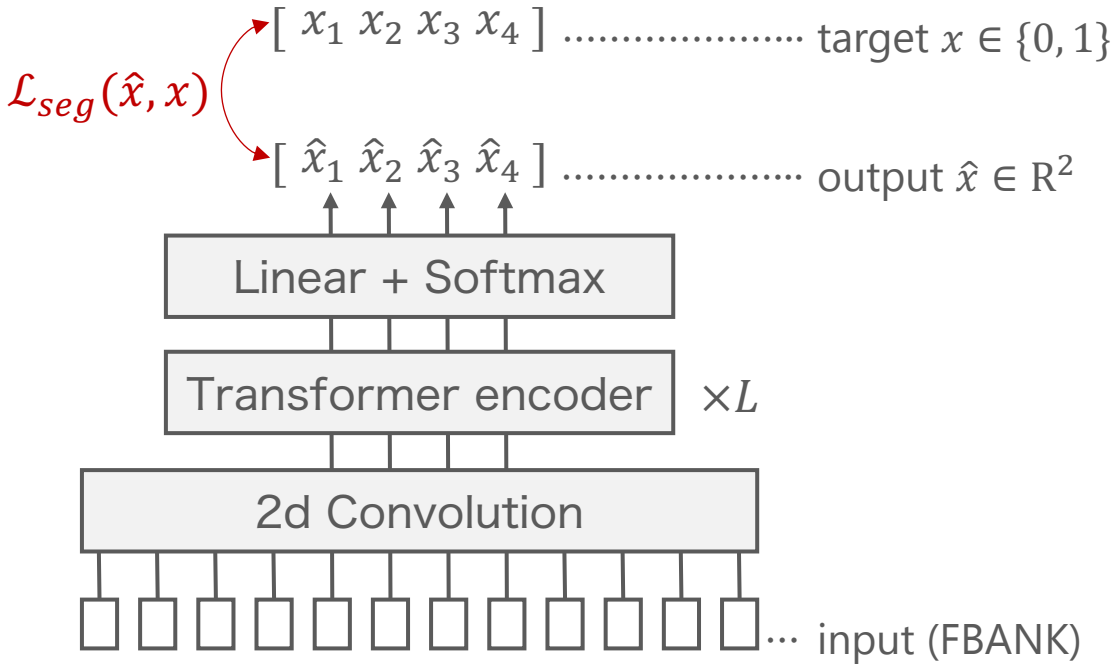
の系列ラベリング

Transformer Encoderを用いた発話分割モデル

- モデル：二次元畳み込み＋Transformer Encoder
- データ：2つの連続するセグメントを連結し、
  FBANKフレームに対してラベル付け
- 学習：交差エントロピー損失$\mathcal{L}_{seg}(\hat{x}, x)$の最小化

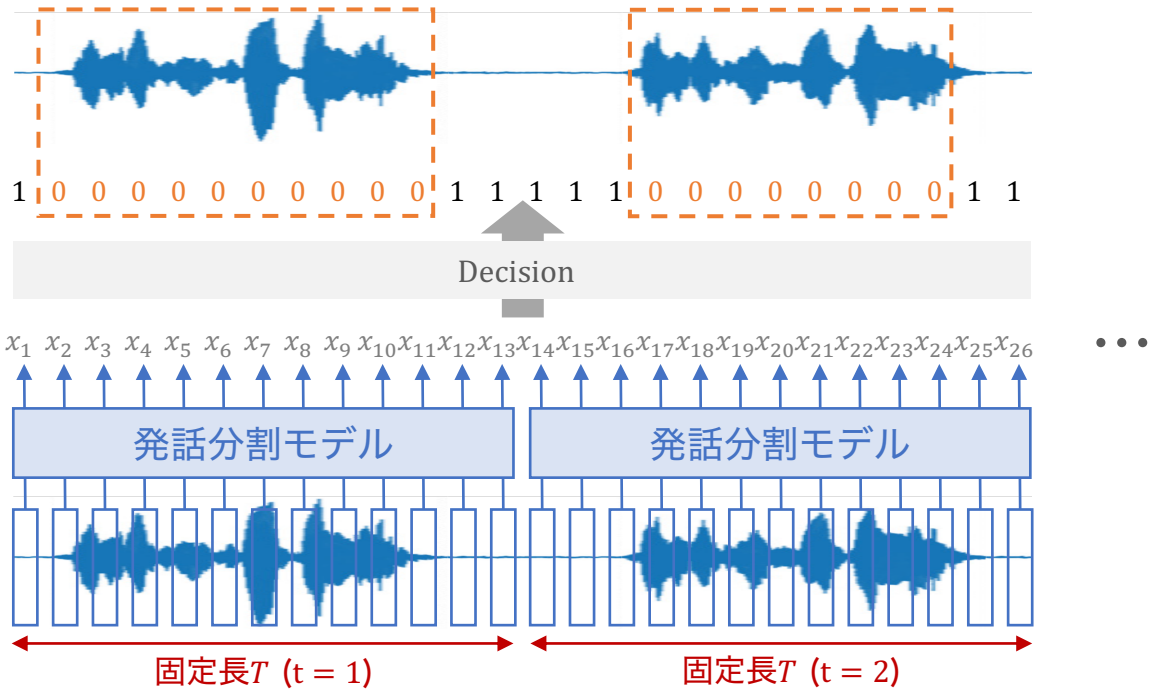
$$\mathcal{L}_{seg}(\hat{x}, x) = -\sum_{n=1}^{N}\left\{ w_s \log \frac{\exp(\hat{x}_{n,1})}{\exp(\hat{x}_{n,0} + \hat{x}_{n,1})} x_{n,1} + (1 - w_s) \log \frac{\exp(\hat{x}_{n,0})}{\exp(\hat{x}_{n,0} + \hat{x}_{n,1})} x_{n,0} \right\}$$

※ rescaling weight $w_s := 0.9$



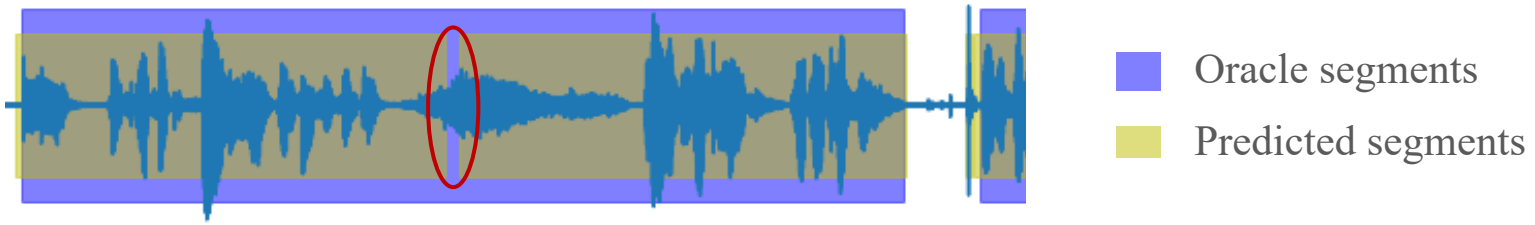
発話分割モデルの推論

(1) 固定長Tのセグメントを発話分割モデルに入力し、各フレームのラベル確率x_iを出力
(2) Decision関数によりラベルを決定（後述）
(3) ラベルを基に音声を再分割



発話分割モデルとVADの組み合わせ推論

- 発話分割モデルは、音のパワーだけを分割の判断根拠にしない
  → パワーが大きい部分で分割することも



- "非音声領域はセグメント境界の必要条件"という仮定のもと、
  発話分割モデルとVADの組み合わせ推論を提案
  → VADで非音声領域を検出して予測を補強

Decision関数

■ 発話分割モデルの推論

$$\text{Decision} = 1 \text{ if } \mathrm{P}(x_i = 1 | f_i) \geq P_{threshold} \text{ else } 0$$

{ 1なら発話外=セグメント境界に含まれる
0なら発話内=セグメントに含まれる

■ 発話分割モデルとVADの組み合わせ推論 追加手法

$$\text{if len(segm)} < maxlen \text{ then}$$
$$\text{Decision} = 1 \text{ if } \mathrm{P}(x_i = 1 | f_i) \geq P_{threshold} \text{ AND } VAD(f_i) == 1$$
$$\text{else}$$
$$\text{Decision} = 1 \text{ if } \mathrm{P}(x_i = 1 | f_i) \geq P_{threshold} \text{ OR } VAD(f_i) == 1$$

※ $P_{threshold}$：発話分割モデルの予測に対する閾値
※ len(segm) ：セグメント長（ラベル“1”の連続数）、$maxlen$：長さ上限を決める閾値

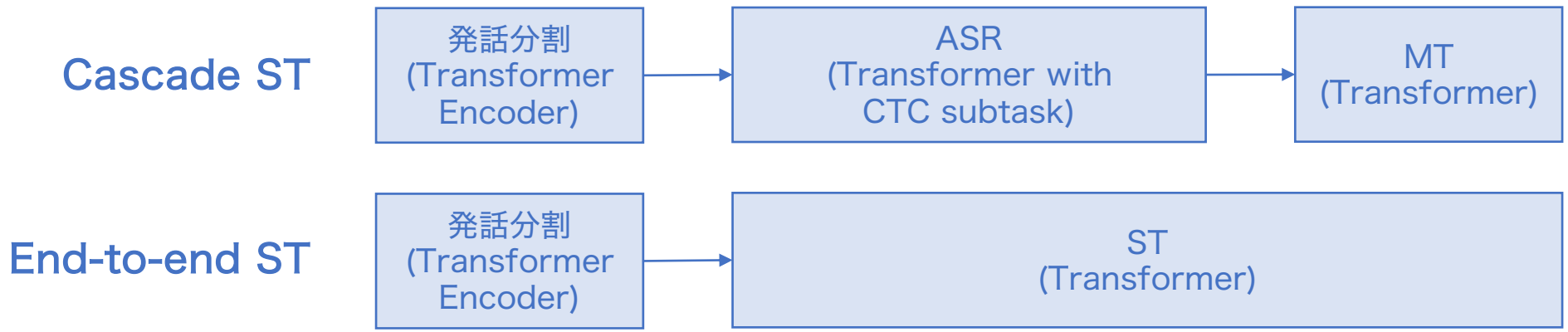
実験

実験設定

- データ：MuST-C English-German [Gangi+19]
  - TED talksの音声翻訳コーパスMuST-Cに含まれる、約408時間の英語の音声と、それに対応づいた書き起こしテキスト、及びドイツ語の翻訳テキスト
- 発話分割手法
  - トップライン：MuST-Cに含まれる分割済みセグメント（**Oracle**）
  - ベースライン：WebRTC VAD（**VAD**）、固定長（**Fixed-length**）
  - 提案手法：発話分割モデル（**Our model**）、組み合わせ推論（**VAD hybrid**）

実験設定

- STシステム
  - Cascade ST、End-to-end ST



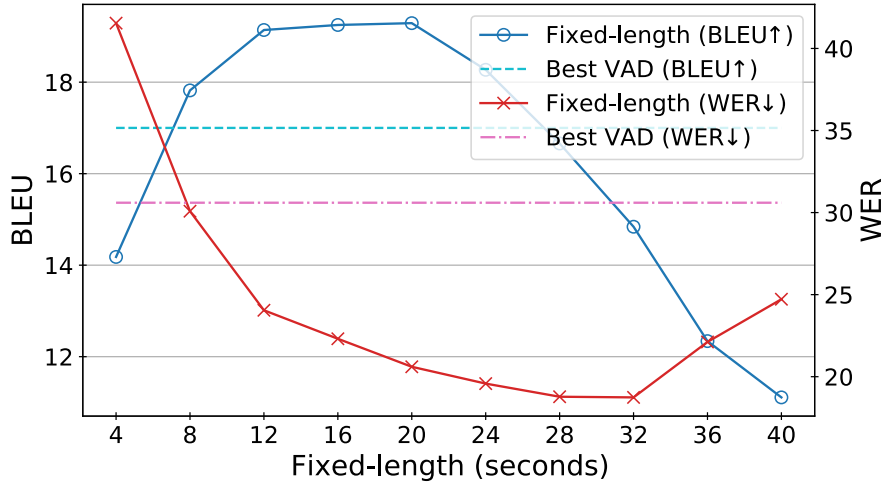
- 評価
  - 編集距離に基づき出力テキストを再分割後、WERやBLEUを測定 [Matusov+2005]

ベースライン手法の結果

複数の試行の中で最も高いBLEUスコアを得た条件の結果

- Best VAD：(aggressiveness, frame size) = (2, 20ms)
- Best Fixed-length：20秒（Cascade ST）、12秒（End-to-end ST）

	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**



Cascade STにおけるベースラインの比較

- VADはFixed-lengthより低いBLEUスコア（17.96 → 16.40）
- OracleとFixed-lengthでは4~5pt程度のギャップ（22.50 → 17.96）
- ある程度まではセグメントの長さを確保することで低下を抑えられる

提案手法の結果

- Best VADに対しては3~4pt、Best Fixed-lengthに対しては~2pt程度の向上
- VADとの組み合わせ推論により~0.8ptの向上
- Oracleと比較して2~3ptのギャップ

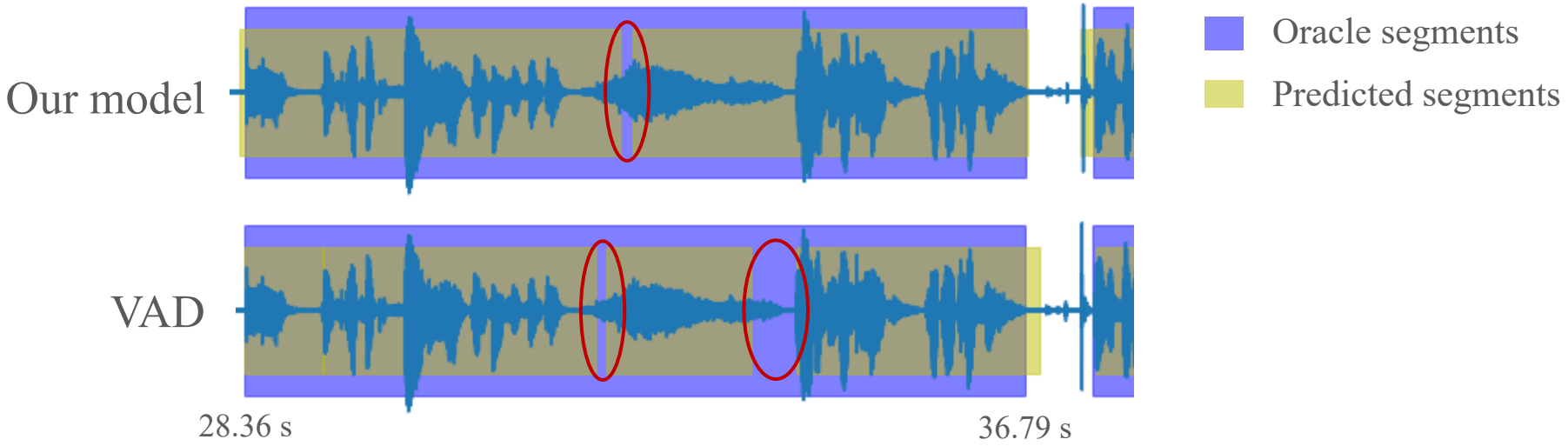
	Cascade ST		End-to-end ST
	WER	BLEU	BLEU
Oracle	12.60	23.59	22.50
Best VAD	30.59	17.02	16.40
Best Fixed-length	20.60	19.29	17.96
Our model	20.99	20.18	19.10
+ VAD hybrid	**19.06**	**20.99**	**19.87**

事例分析：出力テキストと分割位置

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.* ■

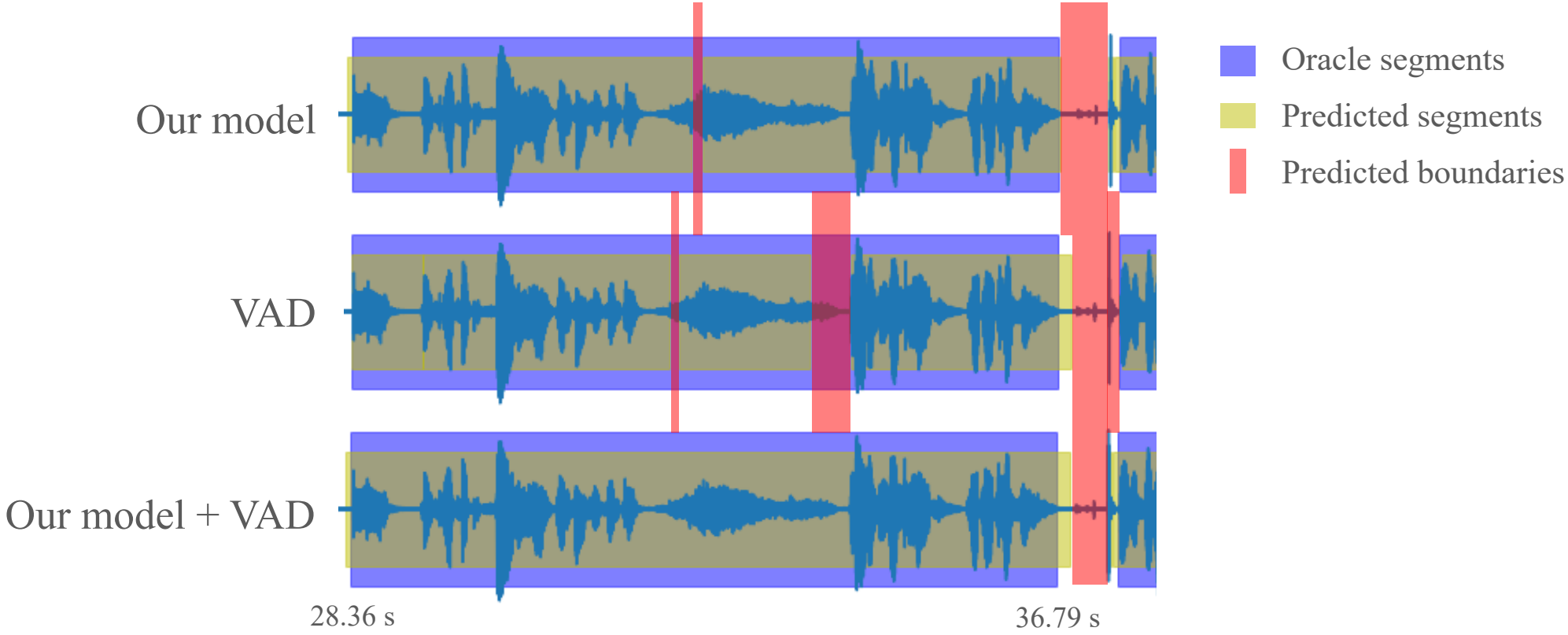
事例分析：音声波形と分割位置

- 発話分割モデルのみや、VADのみによる推論では過剰な分割が目立った



事例分析：音声波形と分割位置

- 発話分割モデルのみや、VADのみによる推論では過剰な分割が目立った
- 組み合わせ推論により2つの手法が相補的に作用 → **過剰分割の軽減**



まとめ

- 本研究では、音声翻訳コーパスに基づく発話分割予測モデルを提案
  - 連続音声を「文」に対応するセグメントへ直接分割
- TED talksの英語-ドイツ語 音声翻訳実験で効果を検証
  - 発話区間や長さに基づく従来手法よりも高い翻訳性能を達成
  - VADとの組み合わせ推論で更に性能が向上

■ 今後の展望

- 翻訳コーパスのセグメント境界は常に最良か？ -> 翻訳性能で最適化
- オンラインSTのための、発話分割機能を備えたEnd-to-end ST

付録

実験設定：STシステム

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		