日本音響学会 2021年春季研究発表会3-2-26

言語モデルによる文横断情報を用いたオーディオブック音声合成の検討

☆**中田 亘**（東大工学部）郡山 知樹，高道 慎之介（東大院・情報理工）
井島 勇祐，増村 亮（NTT）猿渡 洋 (東大院・情報理工)

研究背景

オーディオブック市場の拡大

- オーディオブック作成には多大な労力を伴う
- TTSシステムを用いたオーディオブック音声合成が望まれる

オーディオブック音声合成における課題

- **長い文脈に対応した韻律 （本研究で改善する課題）**
  - e.g. 外を眺めた．ハシがある．（複数文に渡る情報を用いてハシの意味が確定）
- 登場人物による発話スタイルの変化 e.g. 男性，女性，年齢など

関連する研究：TP-GST[Stanton+18]

- Tacotron[Wang+17]の発話スタイルを変化させることができるGlobal Style Token[Wang+18]を当該発話文から推定する構造を提案
- 複数話者データを用いる事により，様々な話者の音声合成が可能

*TTS: Text To Speech

発表概要

目的：長文を考慮したオーディオブック音声合成品質向上

言語モデルの音声合成への利用
- 言語モデルを用いることにより品質が改善[Kenter+20][Hayashi+19]
- 非オーディオブック音声で訓練
  - 韻律の変化は比較的少ない
- 言語モデルへの入力は当該発話文のみ
  - 複数文に渡る文脈である文横断情報は得られない

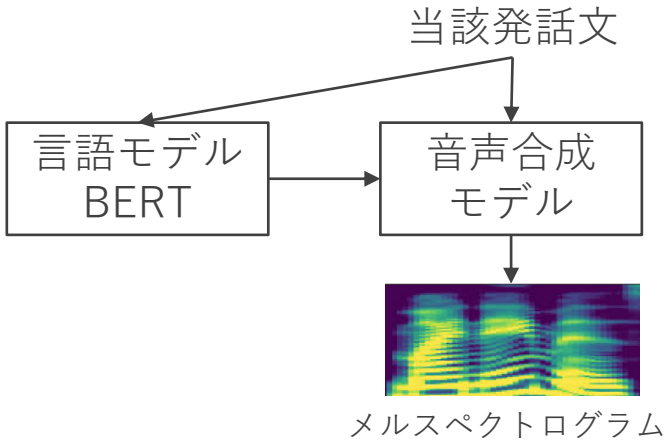
提案法
- オーディオブック音声合成での言語モデルの利用
- 複数文を言語モデル入力することにより得られる文横断情報を用いた音声合成

実験結果
- 文横断情報を用いることにより，合成音声が原音声に近づいた

関連する研究

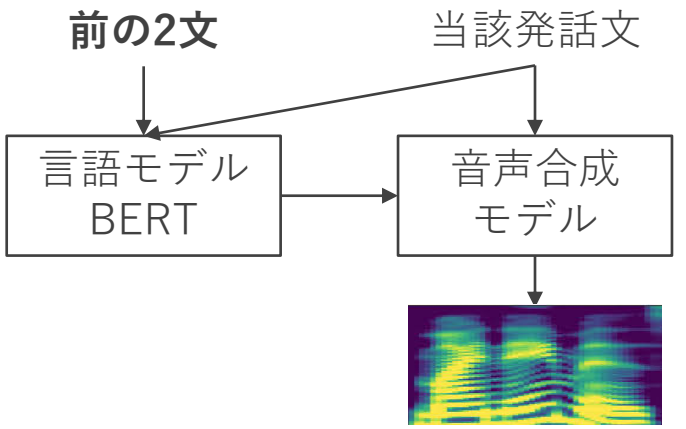
CHiVE-BERT[Kenter+20]

- 言語モデルであるBERTから得られる言語情報を用いてEnd-to-End音声合成モデルの入力を拡張
- 言語モデルを用いる事により，品質向上



CHiVE-BERT

- **長い文脈に対応した韻律を実現する上での問題点**
  - 言語モデルへの入力は当該発話文のみ
  - 複数文に渡る文脈である文横断情報は考慮できない

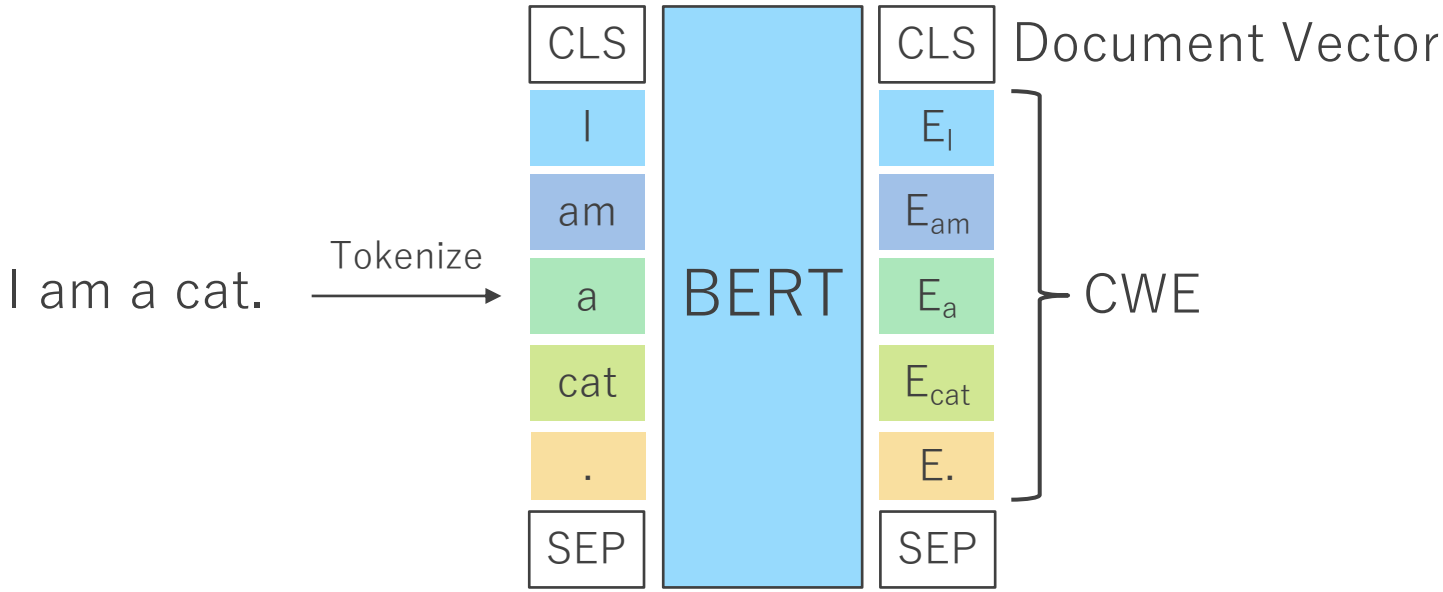


提案法

BERTについて

BERT [Delvin+18]

- Transformerエンコーダ [Vaswani+17] から成る大規模言語モデル
- 複数文から文脈を考慮した単語分散表現（CWE）を得られる
- 複数文から文全体の特徴量である文書ベクトルを得られる
- 自己教師あり学習を用いて大規模なコーパスで訓練
- 複数文を入力することにより文横断情報を得られる

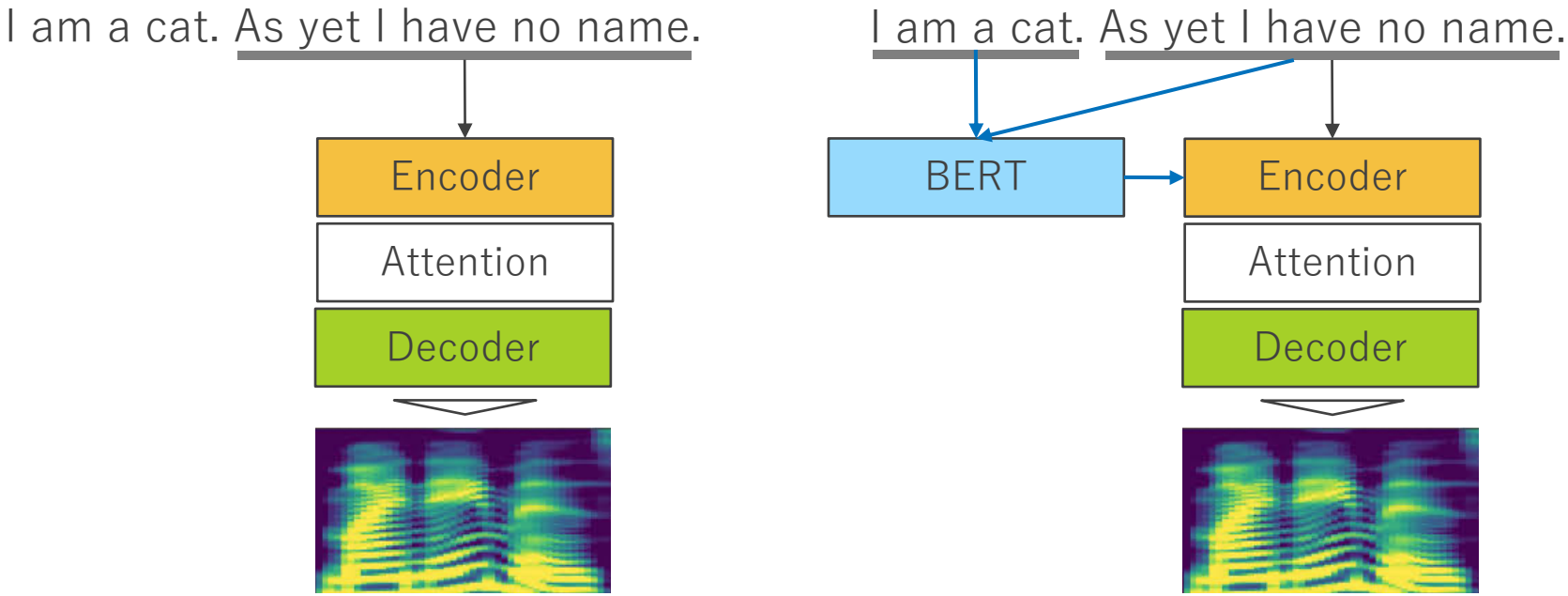


*CWE: Contextualized Word embedding

提案する音声合成モデルの概要

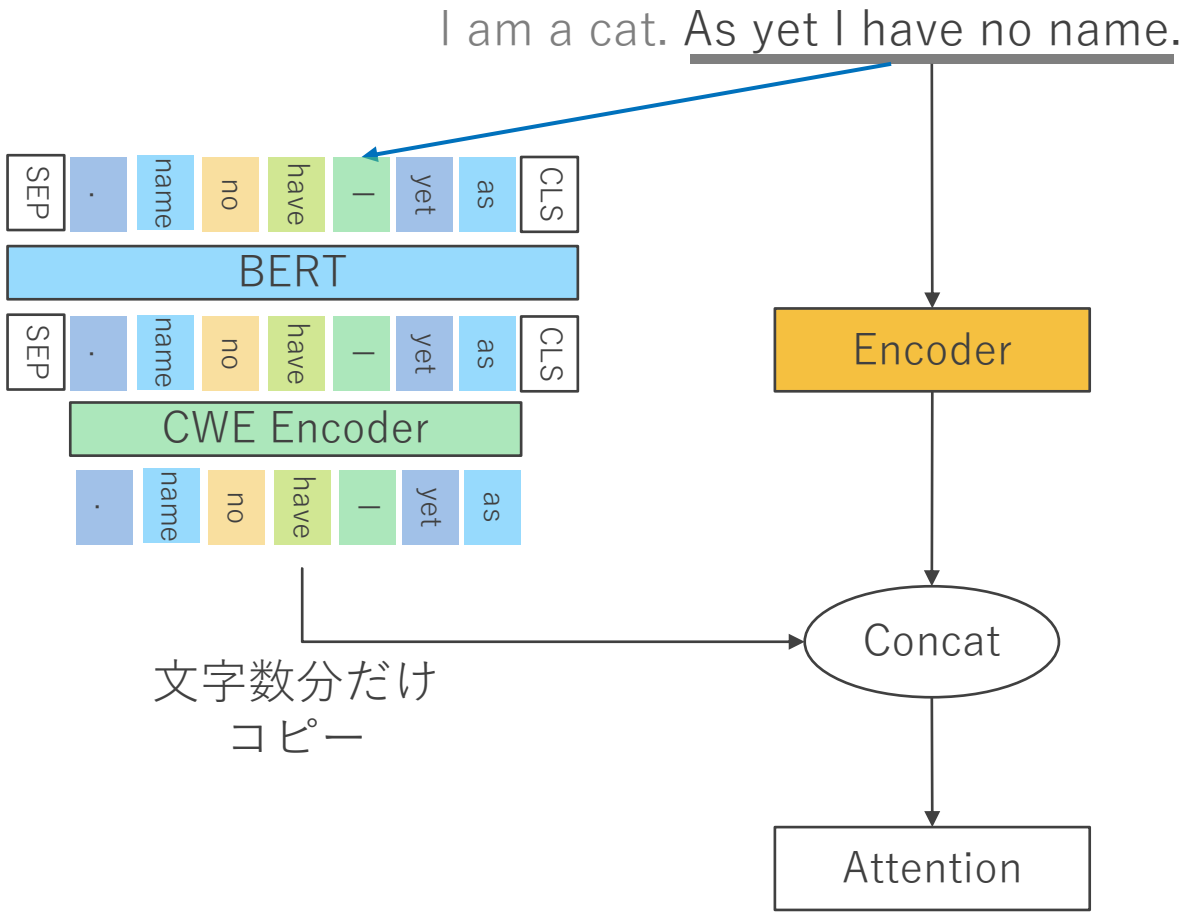
提案する音声合成モデルの概要

- Tacotron2[Shen+18] ＋ BERTから得られる言語情報
- 異なる構成で3つのモデルを検討
  - OneSentence, ThreeSentence, ThreeSentence+文書ベクトル



OneSentenceのモデル構造 (CHiVE-BERT [Kenter+20] と同様)

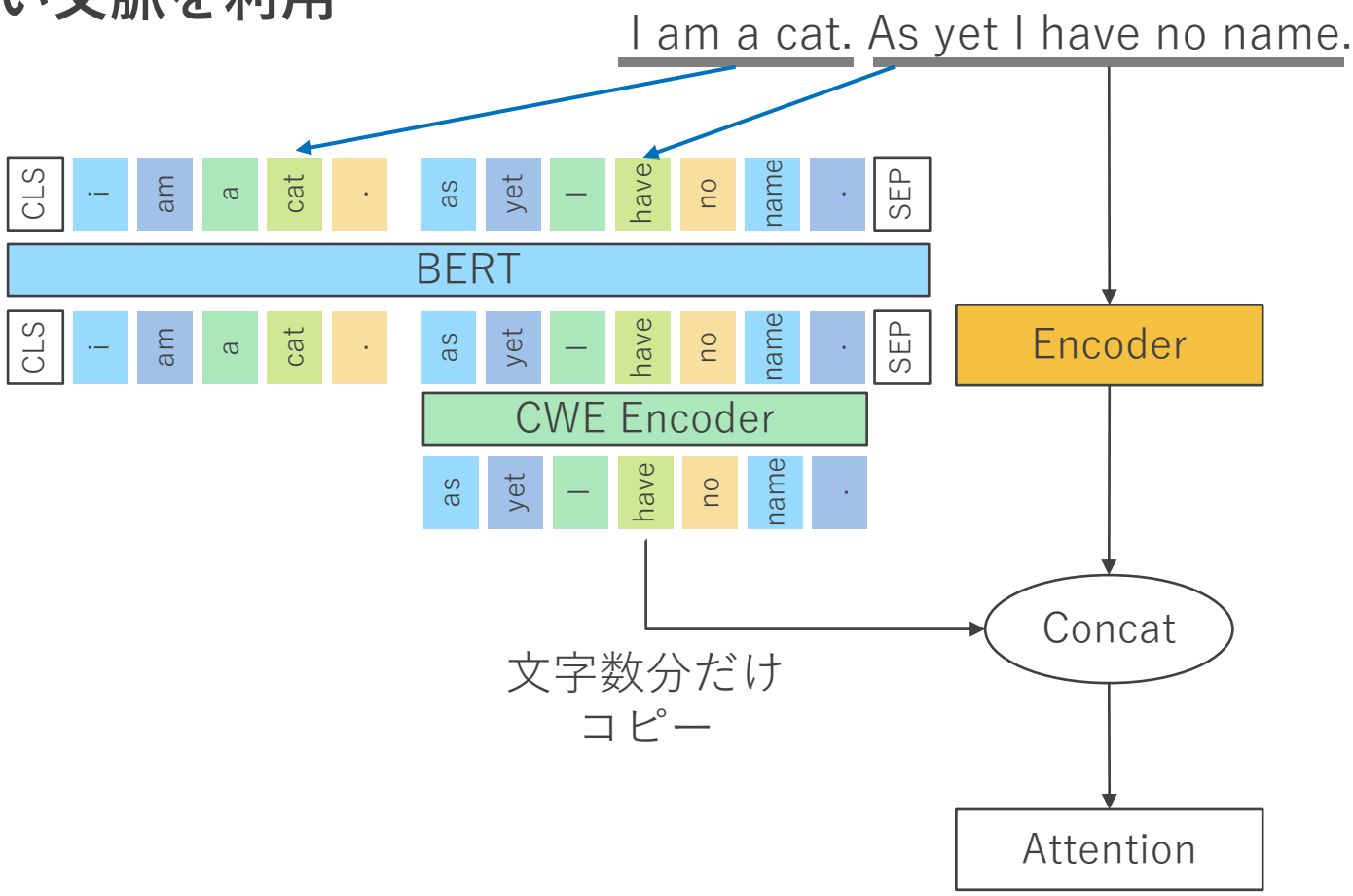
前の文もBERTに入力，当該文のCWEを出力



ThreeSentencesのモデル構造（前の文も考慮）

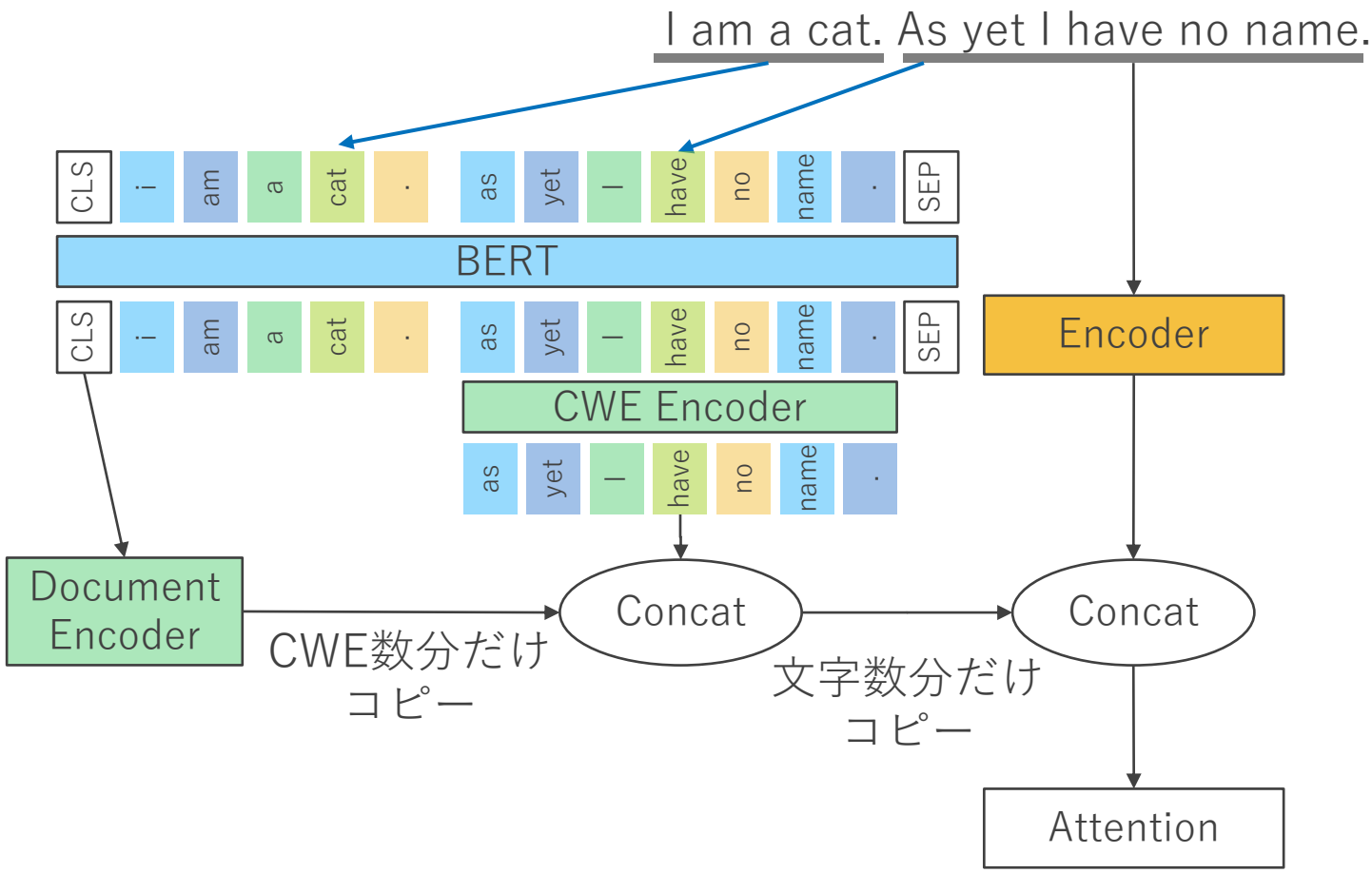
前の文もBERTに入力，当該文のCWEを出力

長い文脈を利用



ThreeSentences+DEのモデル構造（前の文と文書ベクトルを考慮）

前の文もBERTに入力，当該文のCWE及び文書ベクトルを出力



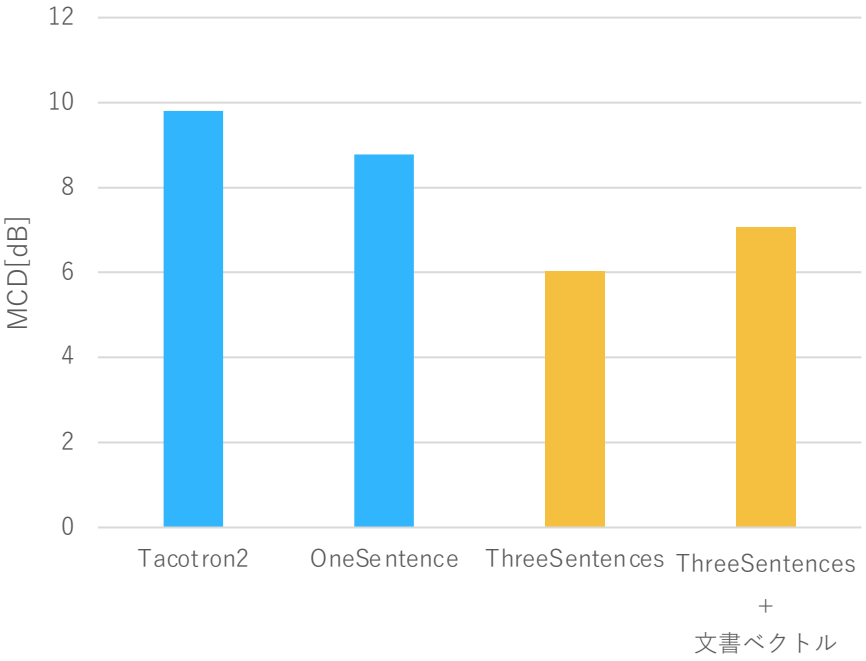
実験条件

データセット	Blizzard Challenge 2013 [King+13] 英語単独話者オーディオブック音声
訓練/確認/テスト データ数	147,249 / 100 / 100 発話
最適化手法	Adam[Kingma+15]
バッチサイズ	32
BERT事前学習済みモデル	distilbert-base-uncased (Hugging Face社)* BERTの重みは固定
学習回数	各モデル収束するまで
Teacher Forcing	デコーダステップごとに50%の確率で適用
評価指標	平均メルケプストラム距離（MCD） グロスピッチエラー(GPE)

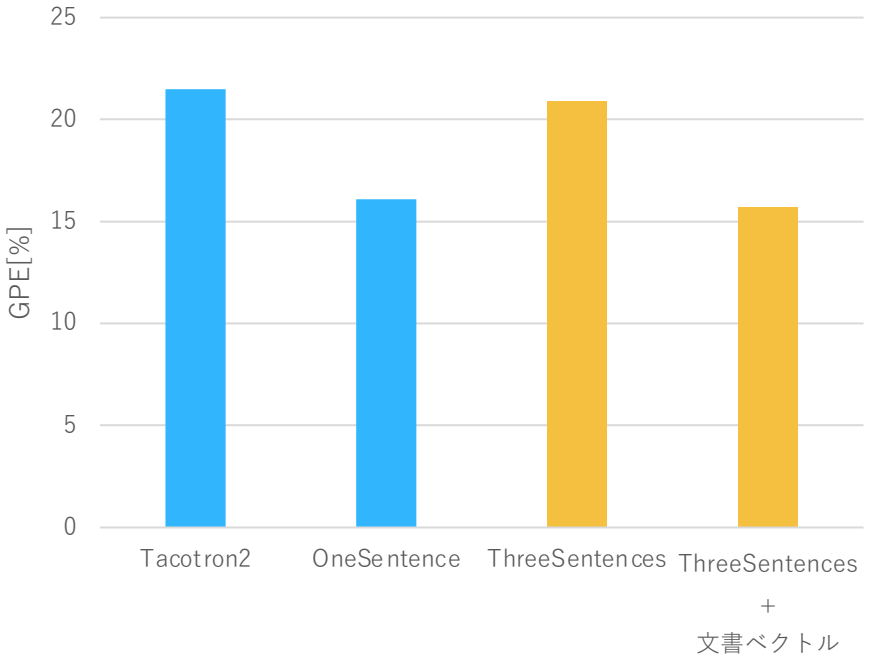
*https://huggingface.co/distilbert-base-uncased

評価結果

平均メルケプストラム距離



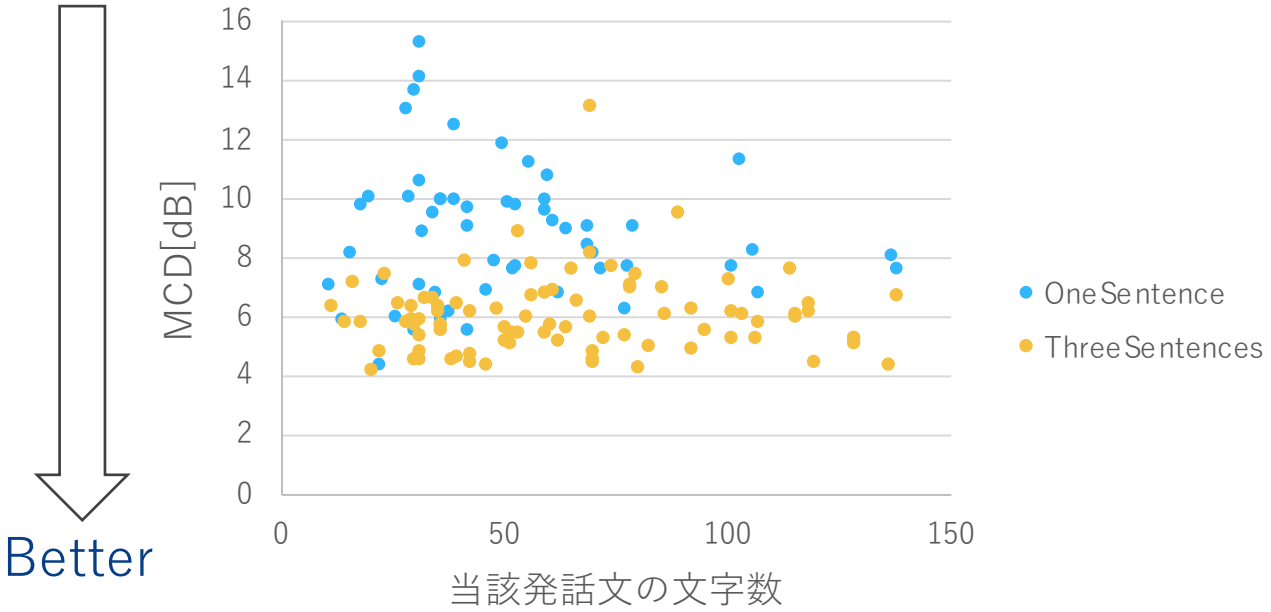
グロスピッチエラー



文横断情報を用いたモデル

文横断情報を用いる事により，合成音声が原音声に近づく

当該発話文とMCDの関係



当該発話文が短い場合でも，文横断情報を用いることにより
平均メルケプストラム距離の変化が少ない

まとめと今後の予定

目的

– オーディオブック音声合成のための長い文脈に対応した韻律の実現

提案法

– 先行する２文を言語モデルに入力することにより得られる文横断情報を用いてTacotron2のエンコーダ出力を拡張

評価結果

– 文横断情報を用いることにより，合成音声の歪が減少

今後の予定

– 合成音声の主観評価
– 英語以外のデータベースでの評価