2021年音響学会秋季研究発表会1-3-8

文横断コンテキストを用いた日本語オーディオブック音声合成の評価

☆中田 亘†, 郡山 知樹†, 高道 慎之介†
井島 勇祐‡, 増村 亮‡, 猿渡 洋†

†東京大学
‡NTT

オーディオブック音声合成

既存のオーディオブック作成方法

- プロの声優の録音音声
- 録音が長時間となるため多大な時間と資金が必要
- **オーディオブック音声合成による負担の軽減**が望まれる

オーディオブック音声合成における課題

- 文横断コンテキストを考慮した韻律の実現（本研究の焦点）
  - e.g.外を眺めた. ハシがある.（複数文の情報を用いてハシの意味が確定）

- キャラ表現の実現
  - 声優はキャラクターに合わせて声を変化させる

文横断コンテキストを考慮した音声合成

我々の先行研究で文横断コンテキストを考慮した音声合成を提案[中田+21]

- 😁 文横断コンテキストを用いることにより合成音声が原音声に近づいた
- 🤔 日本語の音声合成は試されていない
- 😨 主観評価は行われていない

本研究:

- 😊 日本語の音声合成における文横断コンテキストの利用を評価
- 😊 オーディオブック音声としての品質を主観評価

発表概要

➢目的

- 文横断コンテキストを用いた音声合成モデルの
  日本語オーディオブック音声での評価

➢手法

- 先行研究で提案されたOneSentence（文横断コンテキストなし），ThreeSentences
  （文横断コンテキストあり）を日本語オーディオブック音声合成において主観評価，
  客観評価で比較

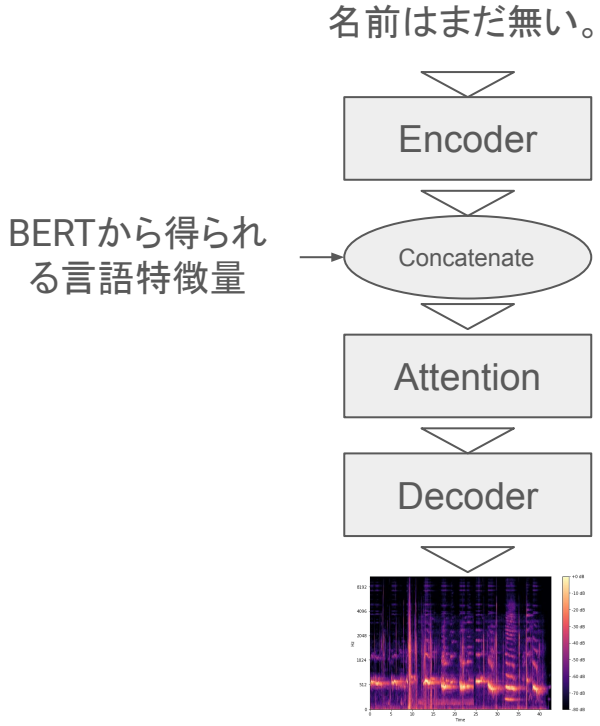
➢結果

- 文横断コンテキストを用いることにより，よりオーディオブックに
  適した音声合成が可能
- BERTをfine-tuningすることで合成音声の品質がより向上

文横断コンテキストを用いた音声合成モデルの概要

OneSentence，ThreeSentences[中田+21]の概要

- Tacotron2[Shen+18] + BERT[Delvin+18]から得られる言語特徴量
- 3つのモデルを評価
  - Tacotron2
  - OneSentence
    - 文横断コンテキストなし
  - ThreeSentences
    - 文横断コンテキストあり

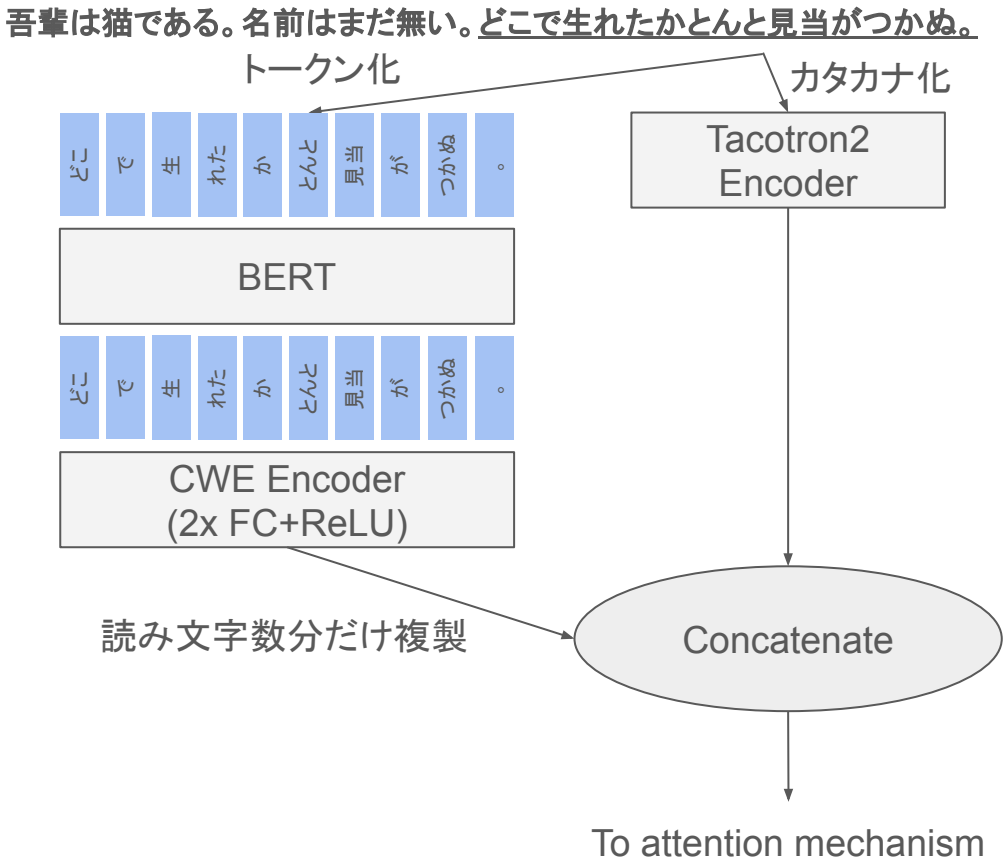


OneSentence

Input: 当該発話文のみ

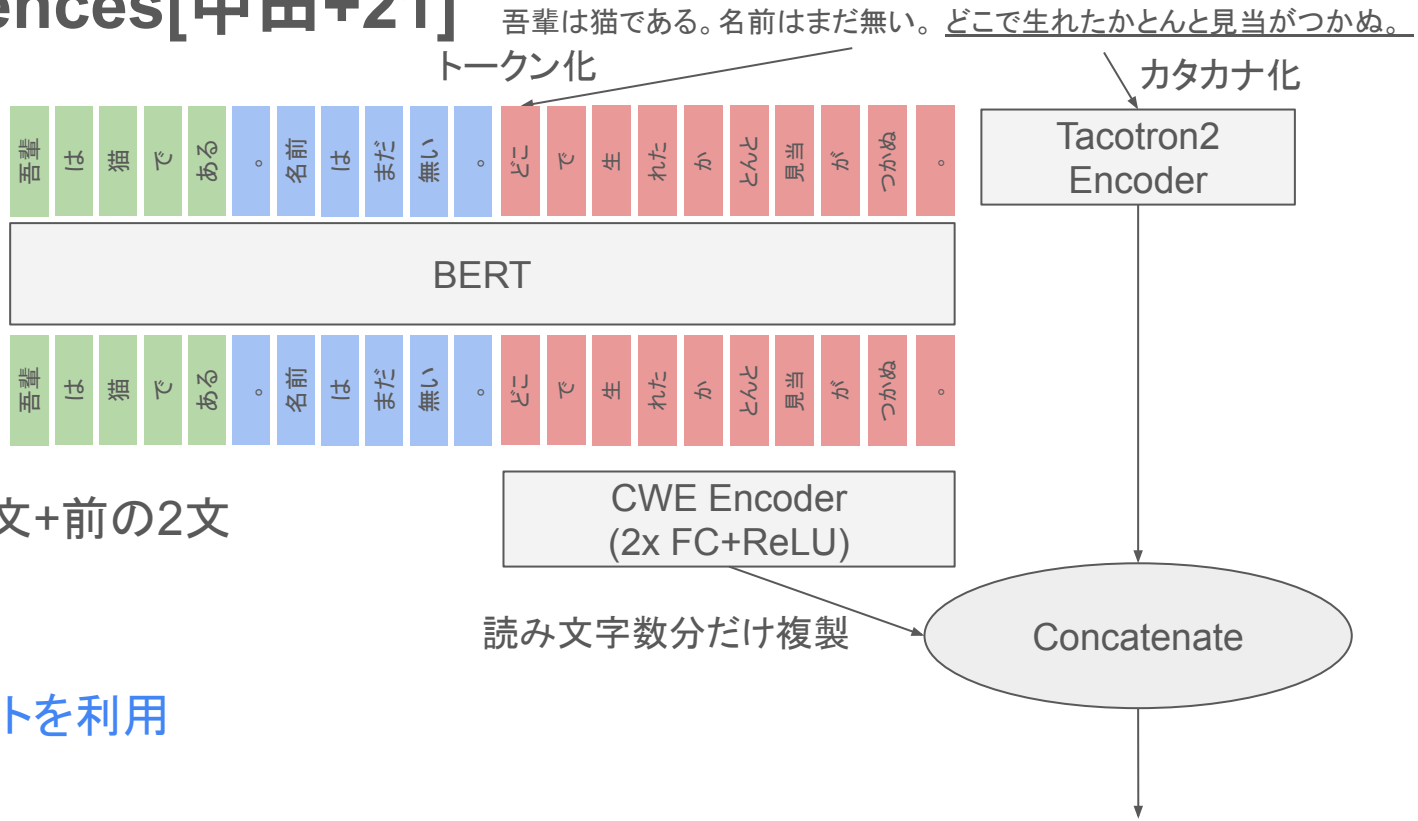
BERTから得られるCWEを利用

文横断コンテキストは用いない



*CWE: 文脈を考慮した単語分散表現

ThreeSentences[中田+21]



Input : 当該発話文+前の2文

CWEを利用

文横断コンテキストを利用

*CWE: 文脈を考慮した単語分散表現

評価を行ったモデル

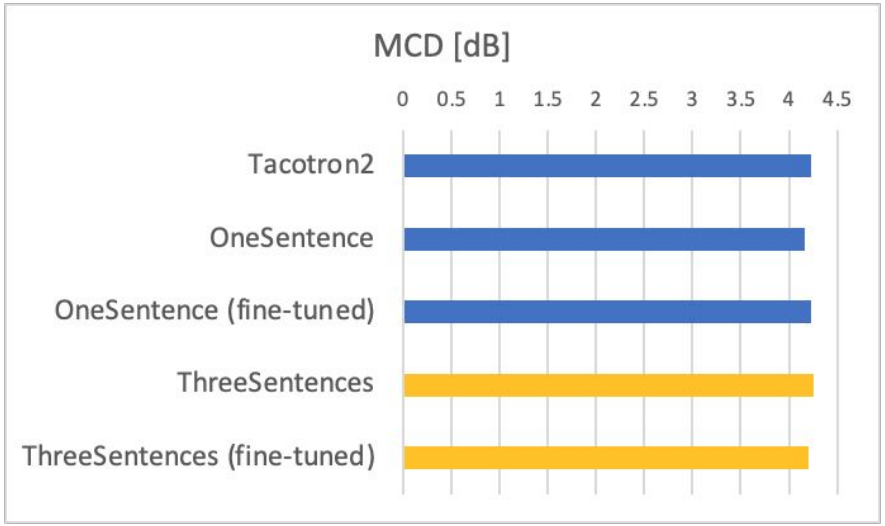
5つのモデルを評価

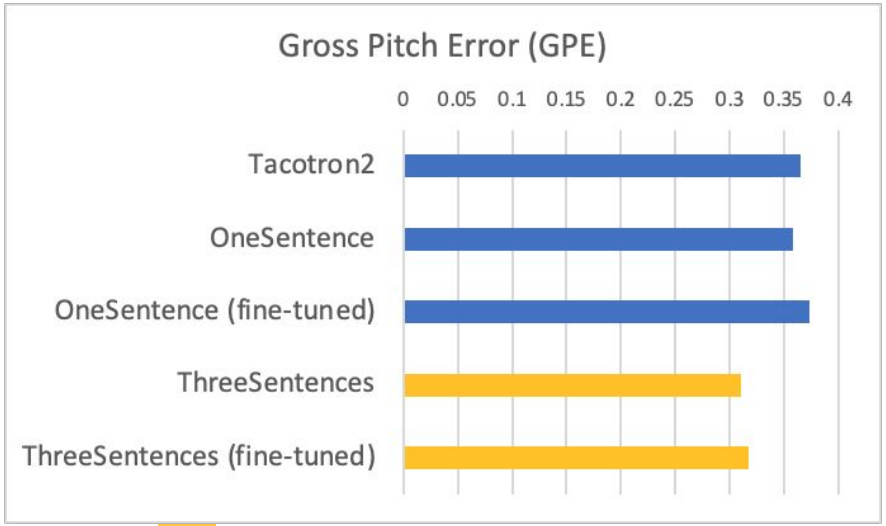
- Tacotron2
- OneSentence
  - BERTのfine-tuning**なし**
- ThreeSentences
  - BERTのfine-tuning**なし**
- OneSentence (fine-tuned)
  - BERTのfine-tuning**あり**
- ThreeSentences (fine-tuned)
  - BERTのfine-tuning**あり**

実験条件

データセット	事前学習: JSUT [Sonobe+17], 10時間 訓練: J-KAC (Japanese Kamishibai and Audiobook Corpus)[高道+21] 男性単独話者によるオーディオブック及び紙芝居音声 6時間
Pretrained BERTモデル	bert-japanese-aozora [akirakubo+20] 日本語wikipedia + 青空文庫で事前学習したモデル
客観評価指標	平均メルケプストラム歪 Gross Pitch Error
主観評価指標	1文での評価 - 自然性MOS 5文での評価 - 自然性MOS - 絵本の読み上げとして適切な方を選ぶABテスト

客観評価結果

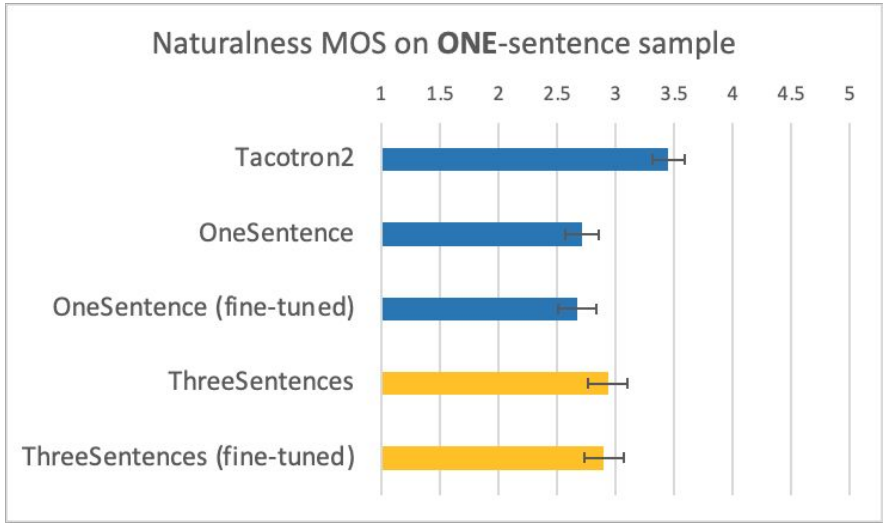


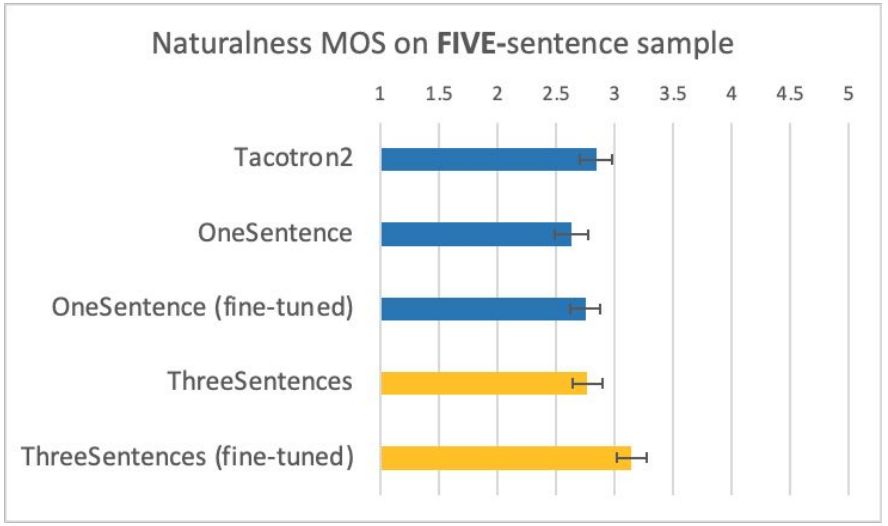


文横断コンテキストを用いたモデル

文横断コンテキストの利用により合成音声が原音声に近づく

主観評価結果 - 自然性MOS





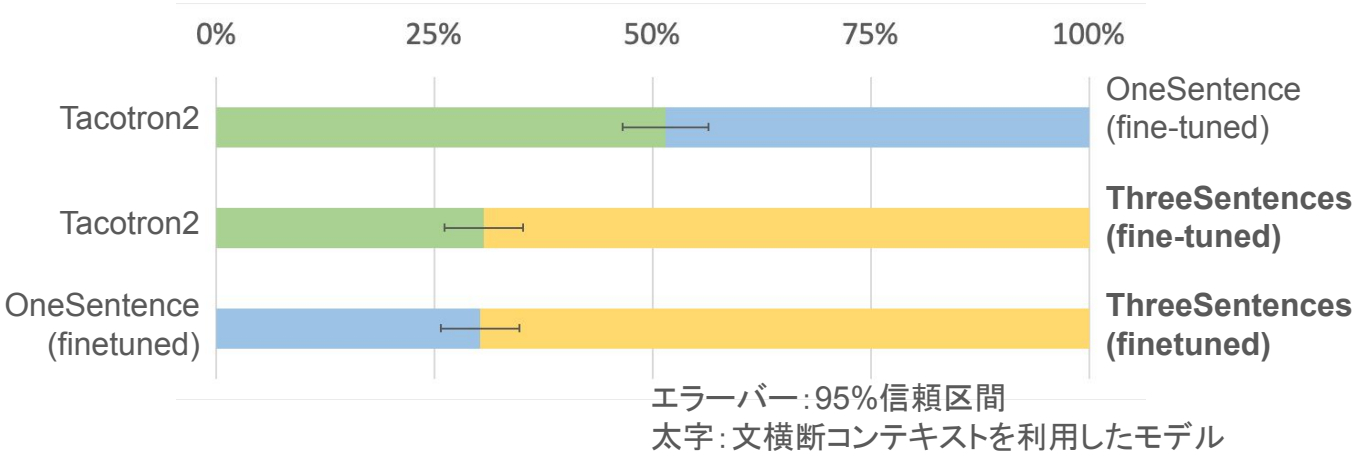
エラーバー：95%信頼区間

文横断コンテキストを用いたモデル

文横断コンテキスト+BERTのfine-tuningにより
5文の合成音声において有意に自然性が向上

主観評価結果 - AB test

どちらがより絵本の読み上げとして適切ですか?



文横断コンテキストを用いたモデルは有意に
より絵本の読み上げとして適切

音声サンプル

会話文
ナレーション

クッキーをほおばりながら、ありくんがいいました。
「このクッキー、おいしいね。」
「とっておきのケーキもあるのよ。」
おいしいおやつをまえに、みんなウキウキ、ルンルンルン。
ところが、きにいらないのはかえるくんです。

原音声	OneSentence (fine-tuned)	ThreeSentences (fine-tuned)

まとめ

目的
- 文横断コンテキストを用いた音声合成モデルの日本語での評価

手法
- OneSentence（文横断コンテキストなし），ThreeSentences（文横断コンテキストあり）を主観評価，客観評価で比較

結果
- 文横断コンテキストの利用により合成音声のオーディオブックとしての品質が向上
- BERTのfine-tuningにより，合成音声の品質がより向上

今後の課題
- より長い文において評価
- 段落や挿絵などの非テキスト情報を考慮した音声合成