

(仮)二次元音声合成
インタフェースを用いた
シナリオ支援アプリの開発

目次

1. はじめに
2. 関連研究
3. 音声合成 Tacotron2+x-vector
4. 話者選択用二次元インタフェース
5. 執筆支援アプリのデモ
6. おわりに



1.はじめに

あるシナリオライターチームでは
書き上げたシナリオを読み上げてみる過程が制作過程に含まれる

読み上げを行うことでシナリオライターは
キャラクターの個性にブレがないか？セリフが自然かなど確認

<https://www.4gamer.net/games/999/G999905/20210720032/>

この書き物の読み上げに音声合成が活かせるのでは？

1. はじめに

音声合成を使用することで

多話者の音声を合成できる音声合成システム
利用することで頭の中にあるキャラクターの声を模索

想像に近い声でシナリオの読み上げをすることで
執筆作業に影響をもたらすのでは？



読み上げで
セリフの自然性
確認できる？



2. 関連研究

・ 戀津,伊藤,他「シナリオ記述支援のための情報管理システムの開発」, 第73回全国大会講演論文集, 2011巻1号605-606p, 2011

シナリオ制作時に発生する情報を, システムを用いて管理, 表示することでシナリオ執筆未経験者でもシナリオ制作ができるシナリオ情報管理システムを開発

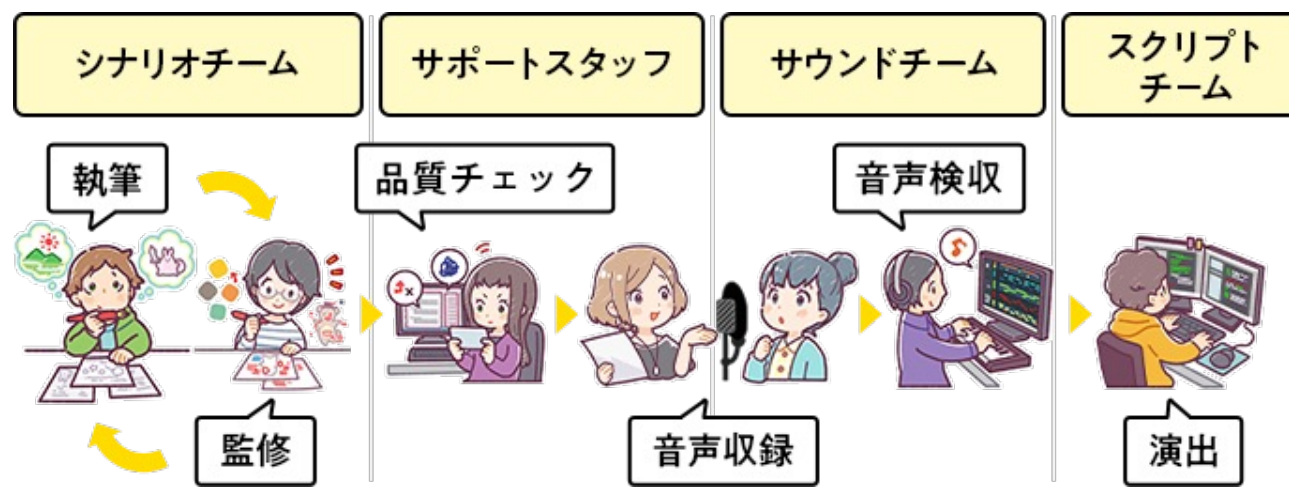
2. 関連研究

シナリオ制作を飛躍的に効率化する「こえぼん」 cygames

執筆サポート・監修・更新履歴とロールバック・検索・品質のチェック・収録台本の作成・音声ファイルの管理と音声検収・シナリオデータの出力 8つの機能を持つ

<https://magazine.cygames.co.jp/archives/19967>

<https://youtu.be/g7uY5zRpz4g>



3. 音声合成 Tacotron2+x-vector

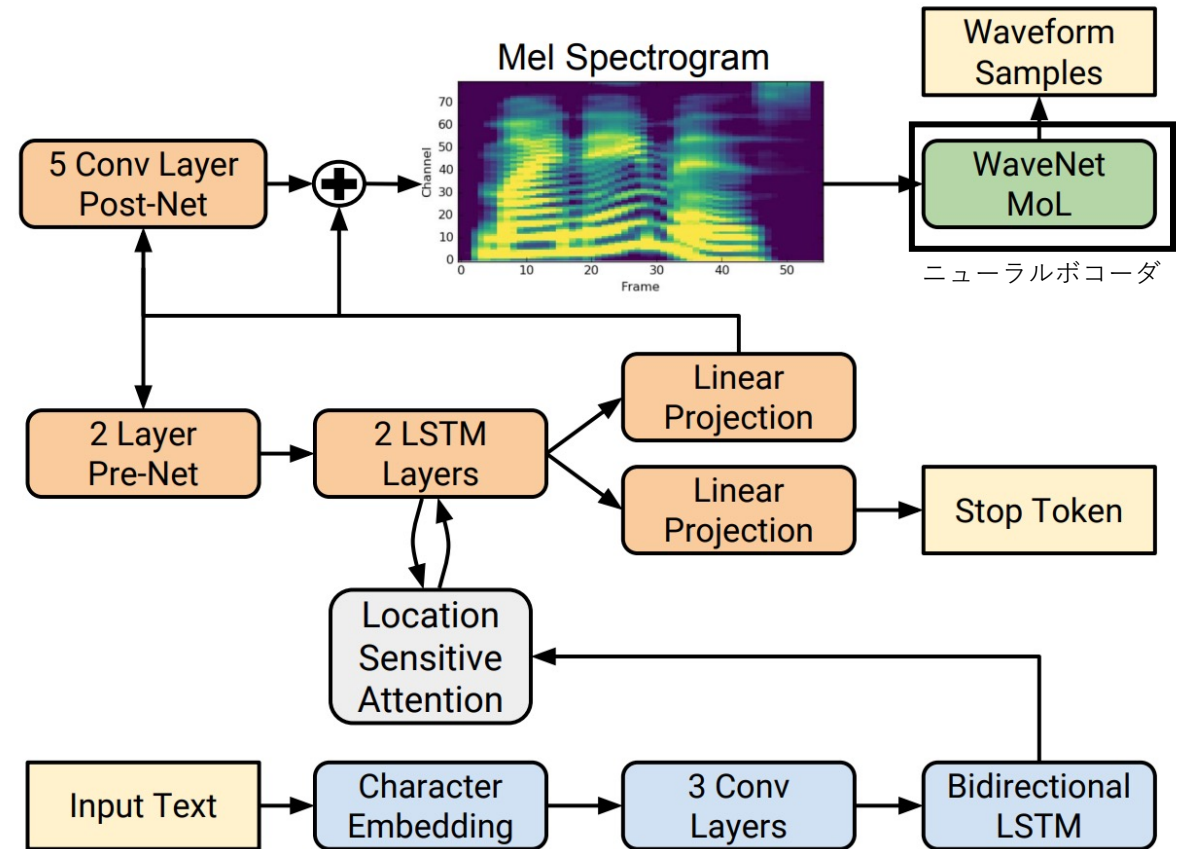
Tacotron2

右記のようなネットワークで
構成される音響モデル
(音素などの文字情報から
メルスペクトログラム(音響特徴量)
変換するためのモデル)

x-vector

話者認証技術
濱田らによって提供される
学習済みモデルを使用

https://github.com/sarulab-speech/xvector_jtubespeech



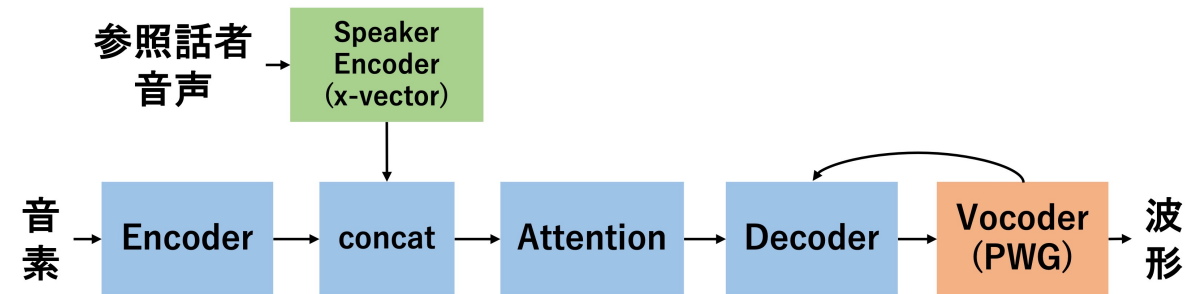
Shen, J., et al.: Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions, ICASSP, pp.4779-4783 (2018)より

3. 音声合成 Tacotron2+x-vector

Jiaらの

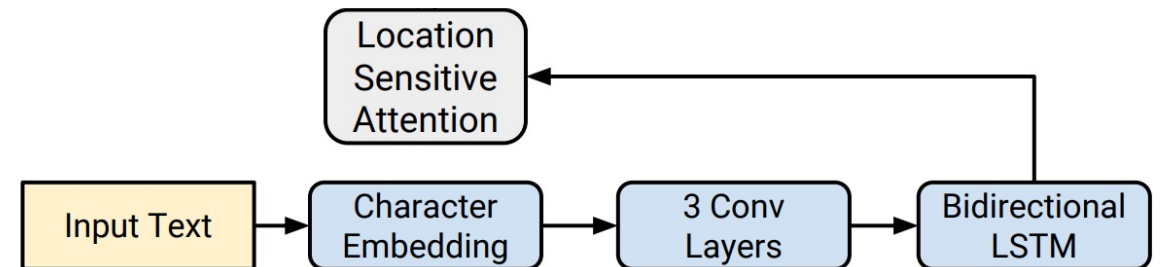
話者特徴量を使用する
音声合成ネットワークを使用

学習時にEricaらの調査を参考に
全結合層によって512次元の
話者特徴量を64次元に削減し
concatを行っている



Jia, Ye, et al. "Transfer learning from speaker verification to multispeaker text-to-speech synthesis." Advances in neural information processing systems 31 (2018).

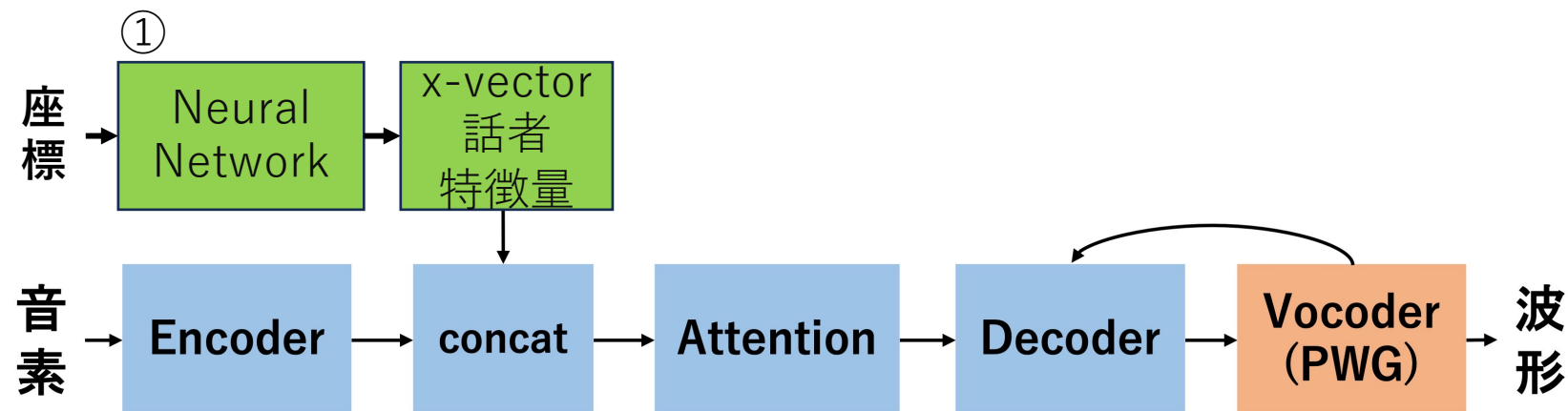
Cooper, Erica, et al. "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.



Shen, J., et al.: Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions, ICASSP, pp.4779-4783 (2018)より

4. 話者選択用二次元インタフェース

- ① ニューラルネット
損失関数: MSELoss
最適化関数: Adam
全結合層(2,16,64,256,512)
BatchSize:256
Epoch:1000
Trainデータ数:11328



ペイントツールで
色を選ぶみたいに
二次元上で話者性を選択して
音声合成できる

女性話者空間 (1, 2) 男性話者空間

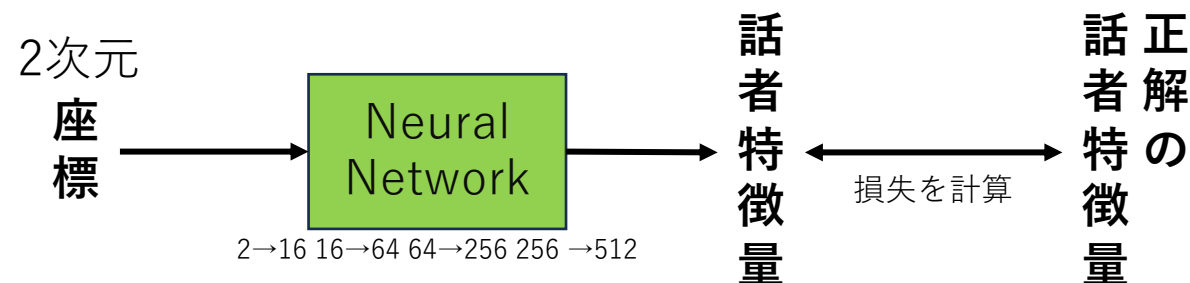
ユーザが自由に座標を選択

座標から話者特徴量への復号方法

最初にUMAPという次元削減手法を用いて話者特徴量(512次元のx-vector)を2次元に削減(学習データ)

2次元の座標を入力として512次元を出力するニューラルネットを訓練(①)

512次元



5. 執筆支援アプリのデモ

