# 合成音声を人間の声に近づけるために

氏名
宮城県仙台第三高等学校　理数科1班

合成音声とは人間の声を人工的に作り出したものである。合成音声はneural networkとmarkov modelの2種類に分けることができ、その差は学習させるデータ量の多さである。そこで本研究ではこの差を可視化させ、何が要因となるのかを調べた。tacotron2を用いneural network、markov modelを作成。人間の声を基準にした上でグラフにおこし、比べる。実験1では人間の声とneural networkを比べ、大きな差異がなかった。実験2では人間の声とmarkov modelを比べ、markov modelにおいて急な周波数の上昇が見られた。実験1，2よりmarkov model特有の確率計算のミスがneural networkとmarkov modelの差だと分かった。

## 1　背景

合成音声は至るところで使用されている。駅構内でのアナウンス、YouTube等動画サービス内での使用、声帯摘出者の使用など数を挙げればきりがないほどに。合成音声には2種類存在する。使用する学習データが多く、音声を作成する際多くの確率計算を試行するneural network。使用する学習データがとても少なく、音声を作成する際最低限度の確率計算しか行わないmarkov modelの以上2種類だ。neural networkは主に声帯摘出者の代わりに声を発する機械や、翻訳機で流れる音声等に使われる。一方markov modelはYouTube内のコンテンツの一つとして使われることが多い。これら2つの音声の相違点を可視化させたいと思い研究するまでに至った。

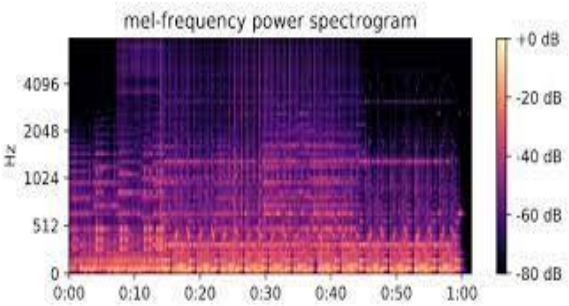
## 2　材料と方法

実験手順

1）tacotron2*を用いた機械学習を通して、markov modelとneural networkの2つの音声を作る。今回は女性のネイティブの声を軸に”Hallo.I am student”と発声させるようにした。

2）人間の声（高校2年生男子）の声を録音

3）作成した2つの音声と人間の声をそれぞれメルスペクトログラムに起こし、差異を調べる。




縦軸が周波数、横軸が時間、値がパワーを表しています。

tacotron2…Text to speechのことです。
テキストをリアルな音声に変換する音声サービス機能を指します。

## 3　仮説

neural networkと markov modelの違いとして確率計算の試行数の差を挙げ、2つの音声の違う大きな要因がこの差であると考えた。

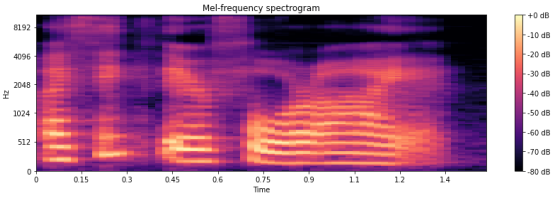
確率のミスマッチがmarkov modelに生じ、グラフ上に急激な音声の上昇・下降が起きると予測する。

4　実験1

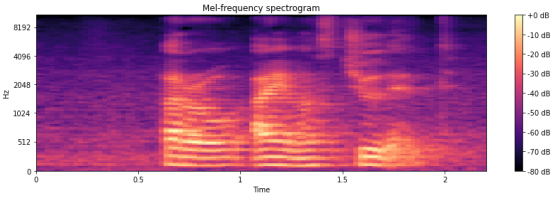
neural networkと人間の声の比較

実際に作り出したグラフがこちら。

neural network




人間の声



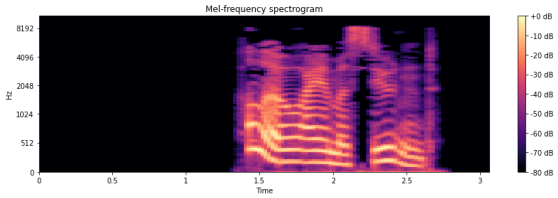

話すスピードが合わなかったこと、人間の声の録音中に無言の間ができてしまったことから時間を合わせることは叶わなかった。それでも初めから終わりまで周波数の差、パワーの大きさの差がそれぞれ微々たるものであった。

5　実験2

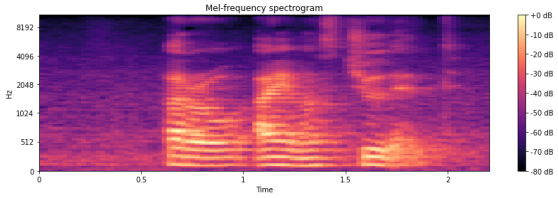
markov model と人間の声の比較

実際に作り出したグラフがこちら

markov model

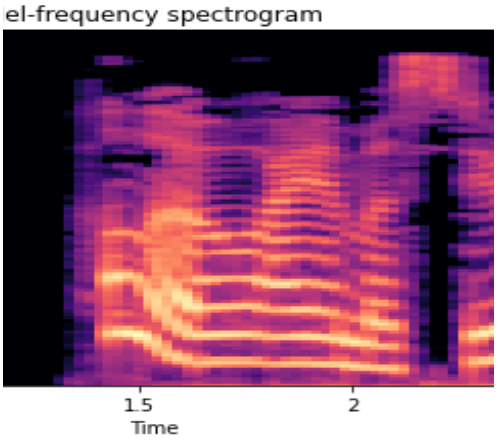



人間の声




markov modelの1．5秒地点にて不可解な周波数の変化が見られた。約400～500ヘルツの急激な増加によりmarkov model特有の確率計算のミスが発生したと考えられる。

また、話者依存性によってmarkov modelと人間の声に差異が生まれたという見解については、事前に日本人男子高校生、日本人女子高校生、男性ネイティブスピーカー、女性ネイティブスピーカーの四人の音声を録音・グラフ化し、比較したところ差が約20～30ヘルツほどしかなかったことから考えられない。




拡大した図がこちら

6　結果・考察

実験1より人間の声とneuraru networkにおいて差異が見つからず、事前調査の通り確率のミスがなかったと言える。一方、人間の声とmarkov model

においては大きな差が一言目に見られた。事前調査において確率のミスが音声に表出するとあった通り約400～500ヘルツの違いが見つかり、隠れマルコフモデルが確率計算のミスにより検出されたと言える。
markov modelはもともと手軽に作られることが良さの一つとされ、確率計算のミスは目をつむられることが多かった。そして今回の結果から一文、それも短い文内で一つ検出されていることから、普段聞いている合成音声に感じる違和感というのはこの隠れマルコフモデルあると考えることが妥当だと考える。

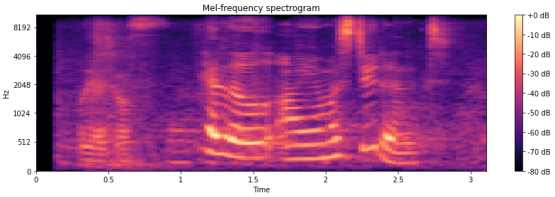
## 7　参考文献


1:https://www.gavo.t.u-tokyo.ac.jp/~mine/japanese/nlp+slp/IPSJ-MGN451003.pdf

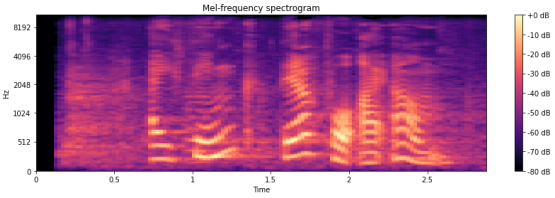
2:音声言語処理と自然言語処理（13年、コロナ社）

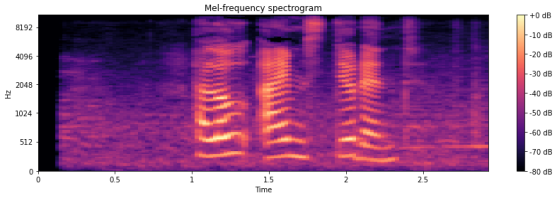
3.河原達也「音声認識技術の展開」
PRMU2015-111

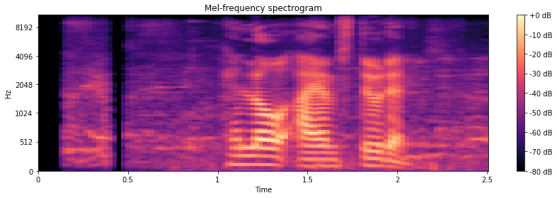
4.https://qiita.com/KojiOhki/items/89cd7b69a8a6239d67ca


## 8　グラフ化した音声





沢山のご協力、誠にありがとうございました