感情表現可能な AI 音声合成システムの開発

西宮市立西宮高等学校地球科学部　2 年 和田遥大　西出皓貴　松本侑大　山崎航生

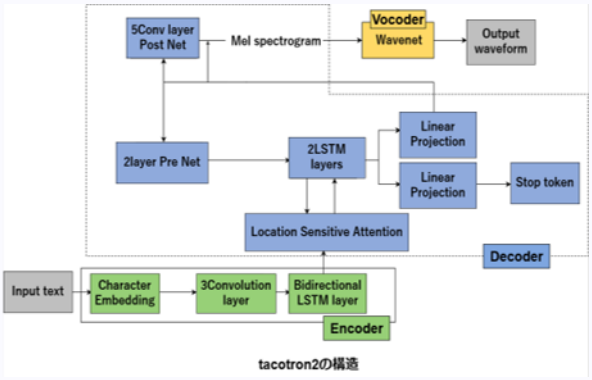
目的

私たちは、病気などで声を失った人がこれからも自分の声で会話するために、誰の声からでも合成音声を作成できるようにすることを目的として研究を行った。

昨年度、一昨年度では音を繋げる方式での音声合成を研究し、最終的に音の素材を切り出し、つなぎ合わせるプログラムを開発した。今年度の研究では、音声に感情を付けることを目的に、AI を用いた音声合成を試みた。

理論

今回は Google の開発した TTS アルゴリズムである tacotron2 を参考に音声合成システムの開発を行った。Tacotron2 の文章から音声が作られる過程は下図のようになる。



tacotron2の構造

大まかな構造としては Encoder, Decoder, Vocoder という 3 つの層から成り立っており、それぞれの役割は以下の通りである。

Encoder

入力された文章を、AI の学習によって感情やアクセントなどの特徴が入ったベクトルへと変換することで文章に感情を加え、機械の読み取れる形式にする。

Decoder

Encoder から出力されたベクトルを入力し、AI によってベクトルを音声特徴量である Mel spectrogram へと変換する。

Vocoder

音声特徴量である Mel spectrogram を wavenet という AI の学習を使い音声波形へと変換する。今回は学習量の削減やより音質のよい音声波形の出力のためデータセットを用いた。

また、開発では、DCGAN（Deep Convolutional Generative Adversarial Network)と呼ばれる構造を用いた。これは生成と判別の二つのネットワークの対立によってそれぞれが学習をし、最終的により良い結果を出力するという方式の学習方法である。

開発

Tacotron2 の複雑な部分は外部の学習済みのデータセットを用いる等で対応した。また、tacotron2 は日本語に対応してないため、日本語に特化した AI にするための前処理をプログラム内に組み込んだ。また、Decoder では理論で述べた DCGAN を用いることで、Tacotron からよりよい構造になるよう工夫をした。

Encoder

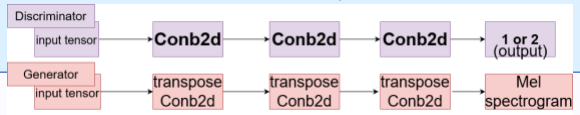
日本語に特化させるため、形態素解析という前処理を含めた。また、文章にベクトルを付けるための学習に多量のデータと時間を必要とするため、今回はデータセットを用いて処理を軽くした。結果として、図のように AI 処理用のテンソル形式でベクトルがついていることを確認した。

[3.1426e-01, -1.3108e+00, -6.2778e-01, 1.5444e-01, -7.3550e-02,2.5071e-01, -3.7715e-01, 1.0569e-01, -1.8187e-01, 9.1854e-01, -2.3662e-01,……]

↑「友達」という単語のベクトル(一部抜粋)

Decoder

この部分は tacotron2 とは全く違い、前述の DCGAN という 2 つのニューラルネットワークを対立させ、互いに学習させる手法を用いることで、Encoder 出力のベクトルを Mel Spectrogram へと変換した。学習データはネット上の音声と部員の声から作成した。また、出力の Mel sp ectrogram は画像を圧縮することで音質は悪くなるが軽量化を行った。出力した結果、データセットに存在した文は自然な声だったものの、存在しない声は不自然に感じられた。



↑Decoder 内で対立させた二つのネットワークの構造(簡

Vocoder

Tacotron2 と同じく wavenet と外部のデータセットを用い Mel spectrogram を音声波形へと変換し、音の完成度や感情の損失を調べた。結果として、もとの音が高いほど再現性がよく、逆に低い音はこもったような音声になった。

結論、今後の展望

今回、音声合成システムの 3 つの層である Encoder, Decoder, Vocoder の開発に成功した。

今後の展望として、外部のデータセットを用いない独自の Encoder の開発や、また他の Vocoder を用いたときの音声の変化について検証することを考えている。