あらゆる人の声を模倣可能なリアルタイム音声変換システムの開発
—誰もが望む声になれるシステム NeuroVoice—

1. 背景

人の声は個人ごとに異なり、自分の声でしか話すことはできない。聞き心地の良い声優や俳優の声、または厳格な印象を与える政治家の声など、場面に応じて自分以外の声で話すことに憧れを抱く人も多いであろう。実際に、多くの小説、漫画または映画にはフィクションとして変声機が登場し、人々の心をおどらせる魅力的なものとして描かれている。

このように、変声機は人々が実現したいと強く願うシステムの一つであり、実現にあたって不特定多数の人が気軽に利用できることが望ましい。しかし、現状では、特定の話者間で限定的に利用できる変声機は存在するものの、誰もが利用できる変声アプリケーションは確立されていない。

2. 目的

本プロジェクトでは、誰もが手軽に利用できる音声変換システムの構築を目的とした。システムの満たすべき要件として、任意話者の音声を入力として利用できること、そして新たな変換対象の拡張が容易であることが求められる。これらの要件を満たすように開発を行いつつ、変声後の音声品質の向上や高速な音声変換にも挑戦した。また構築したシステムを容易に利用できるように、API 環境の構築も行った。

3. 開発の内容

本プロジェクトでは、入力音声を他者の声に変換する変声システムを開発した。構築した変声システムの概要を図 1 に示す。本システムでは、話し方等の時間情報を保存したまま声質とピッチ（声の高さ）のみを対象のものに変換する。入力音声を一度音素列に変換する事によって音声の個人性を排除し、音素認識結果から対象の声質とピッチを推定することによって、任意の入力音声の変換が可能となる。本プロジェクトでは編成システムを構成するモジュールとして、音素認識システム、声質変換システムそしてピッチ変換システムの大きく分けて 3 つの開発を行った。また、構築した変声システムを HTTP 通信によって呼び出せる API 環境を構築した。

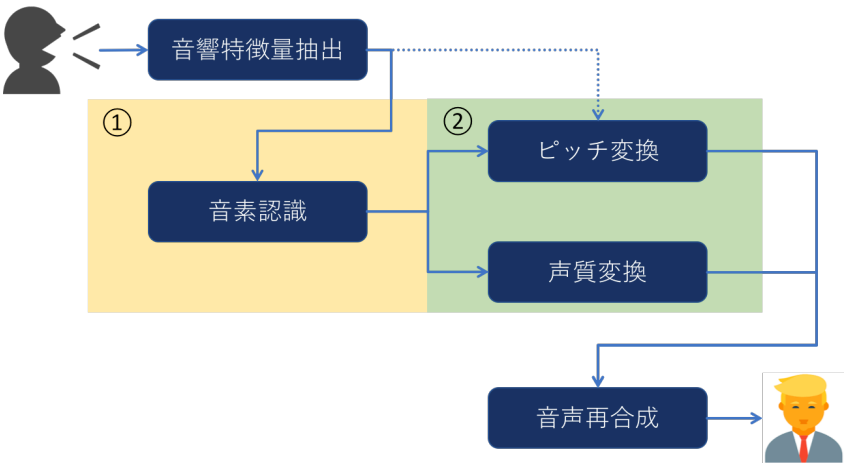

図 1 構築した変声システムのフローの概要図

- 音素認識システム

  本プロジェクトで構築した音素認識システムは、大きく分けて、音声波形から音響特徴量を抽出する前処理、Convolutional Neural Network（CNN）によるフィルター処理そして Recurrent Neural Network（RNN）による時系列処理の 3 つのステップからなる。音響特徴量としては、音声認識に適した MFCC を利用した。抽出した MFCC を、フィルターサイズの異なる幾つかの CNN にかけることによって、様々なスケールで見たときの特徴を抽出し、その結果を RNN に通すことによって時系列的な相関を見る事ができる。この音素認識システムからの出力である音素分布を利用して、対象のピッチと声質の推定を行う。
- 声質変換システム

  各フレームにおける声質特徴は、時間的な相関を持つため、前フレームの予測結果から次フレームの特徴量を順に予測する自己回帰モデルを利用した。本プロジェクトではこの自己回帰モデルとして、WaveNet を参考にシステムを構築した（図 2）。各フレームでの予測にあたって、音素認識システムによって得られた音素分布によって条件付けを行うことで、入力話者の話した内容と生成音声の話す内容を一致させることができる。また、音素分布は時系列性を保持しているため、入力話者が発話したそれぞれの音の長さを保持したまま、声質のみを変換することができる。

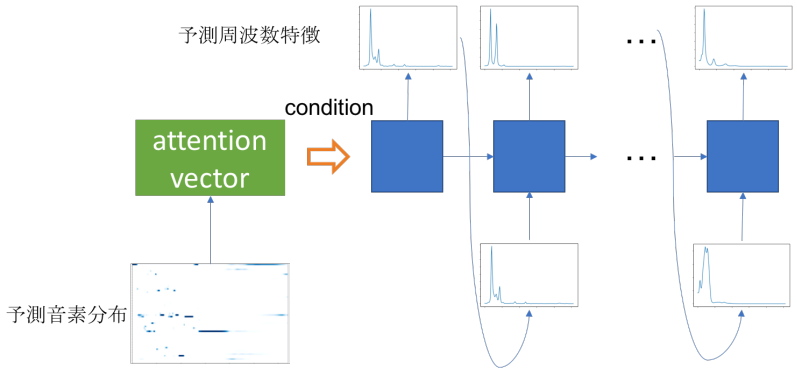

図 2 声質変換システムの概要図

- ピッチ変換システム

  ピッチを変換する際には、入力ピッチの相対変化は保持したまま、絶対値のみを対象の高さに変換することが望ましい。入力の相対変化を残す事によって、入力音声の強調やイントネーションといった情報はそのままに変声を行うことが可能である。従って、ピッチ変換システムは、[0, 1]の範囲に正規化したピッチから対象のピッチ情報を復元するものとして構成した（図 3）。声質変換と同様に、各フレームで音素予測結果による条件付けを行った。
- API 実装

  変声システムを利用可能な環境として HTTP 通信を利用した API を整備した（図 4）。どのようなフロントエンドアプリケーションであっても、HTTP 通信を通じて音声ファイルをアップロードし、変換対象を選択するだけで変声を実行することが可能となる。本プロジェクトでは、Microsoft のクラウドサービスである Azure の Linux サーバを利用して環境構築を行った。

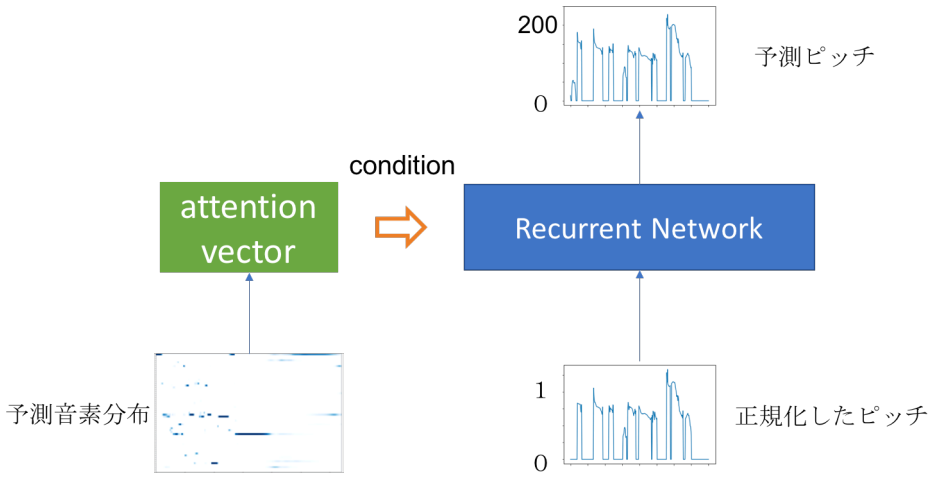

図 3 ピッチ変換システムの概要図

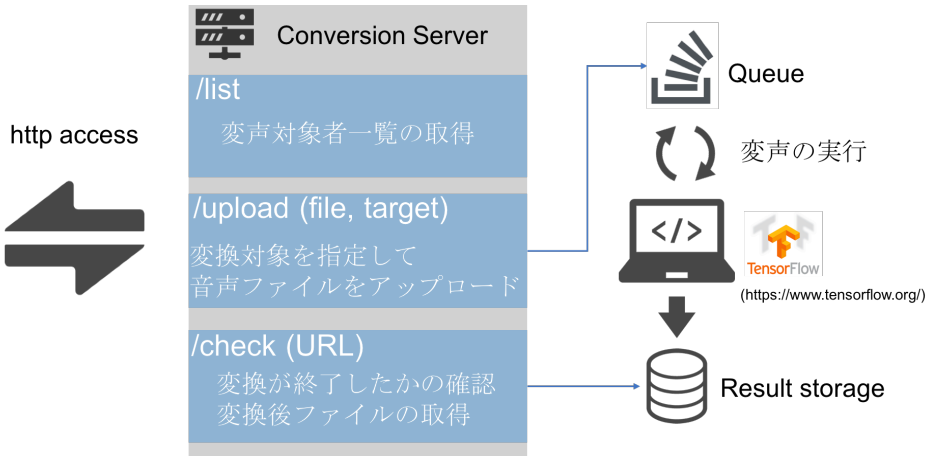

図 4 実装した API の概要図

4. 従来の技術(または機能)との相違

競合する製品として、リアチェン voice が挙げられるが、リアチェン voice ではリアルタイムな変声を実現しているものの、基本的にはユーザは決められた音声を読み上げて自分の声をシステムに学習させる必要がある。ユーザ音声を学習させない場合には変声品質は低下し、変声品質を向上させるためにユーザは話し方を工夫しなければならない。また新しい音声の追加にあたっても、決められた音声を読み上げてもらう必要があり、すでに存在する動画等の音声を学習に利用することはできず、スケーラビリティの点では本手法のアプローチのほうが柔軟である。

また、類似したシステムとして、Lyrebird が存在する。このシステムは、ユーザの音声を数十分程度の音声収録を通じて学習するというものであるが、変声ではなく文字から音声を生成するテキスト読み上げシステムである。従って、本プロジェクトの提案と同様に比較的少数のデータから任意の音声を生成できるものの、話す文章が同一であれば生成される音声は毎回同じである。この点で、本システムの方が話し方まで柔軟に生成することができると言える。

5.　期待される効果

現状で普及している音声合成システムはテキストを入力としたものであるが、本システムの品質が向上し普及することになれば、動画配信者などのクリエータはより自由で柔軟なコンテンツを作ることができるようになる。また SNS などに普及することができれば、多くの人が一度は思い描いたことがあるであろう、自由に他人の声になることができる世界を実現することができる。その他にも、スマートホーム、VR や医療分野など応用先は広い。

6.　普及（または活用）の見通し

今後も変声音声の品質の向上に挑戦しつつ、まずは動画コンテンツ配信者などある程度整えられた環境で音声収録を行うことが保証されたシーンでの普及を目指す。最終的には、テレビ通話などのコミュニケーションツールに載せることや、SNS やコンシューマアプリとして普及させることを目標として取り組んでいきたい。

7.　クリエータ名（所属）

早川 顕生（東京大学大学院情報理工学系研究科）

（参考）関連 URL

http://neurovoice.jp/