IPA

2017 年度 未踏 IT 人材発掘・育成事業 採択案件評価書

1. 担当 PM

プロジェクトマネージャー：藤井 彰人
（KDDI 株式会社 ソリューション事業本部 ソリューション事業企画本部長
兼 クラウドサービス企画部長）

2. 契約者氏名

クリエータ：早川 顕生（東京大学 大学院 情報理工学系研究科）

3. 委託金支払額

2,304,000 円

4. テーマ名

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

5. 関連 Web サイト

http://neurovoice.jp/

6. テーマ概要

本プロジェクトでは、入力音声を任意の選択した人の声に変換する変声システムを開発した。本システムの特徴は、任意の音声を学習無しで入力として利用できること、入力音声の話し方等の時間情報を保存したまま声質のみを対象のものに変換できること、そして録音した音声を数秒程度の時間で高速に変換できることである。本プロジェクトの成果として、多対多の変声システムの構築とそのシステムを任意の端末から HTTP 通信によって実行できる API 環境を整備した。

7. 採択理由

本提案はニューラルネットワーク技術を利用した、音声変換システムの開発提案であった。タイトルの通り、特定の話者の特徴を再現する変声機の実現を目指しており、実現すればそれぞれの話者が持つ、音素のつながりやイントネーションを真似た音声への変換サービスを提供することができる。文字起こしと発

話ではなく、音声そのものを音素認識と声質変換することで音声変換を実現しており、実現する変声品質次第ではあるが、具体的なサービスとしての発展性に大きく期待した。

8. 開発目標

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。

9. 進捗概要

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。

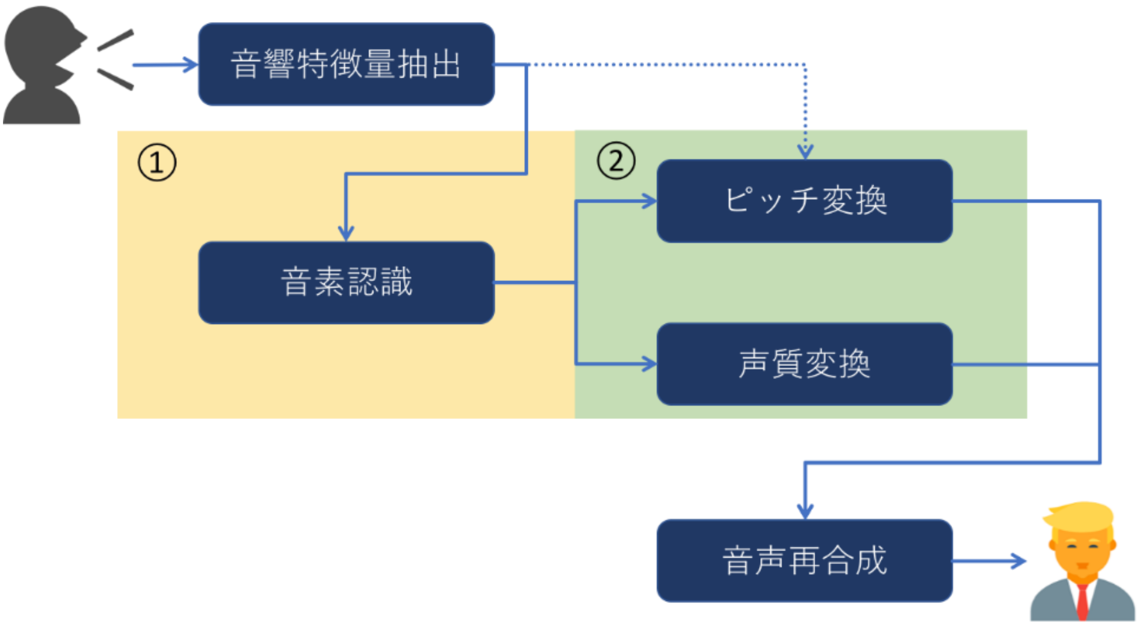

図 1. 構築した変声システムのフローの概要図

- 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。

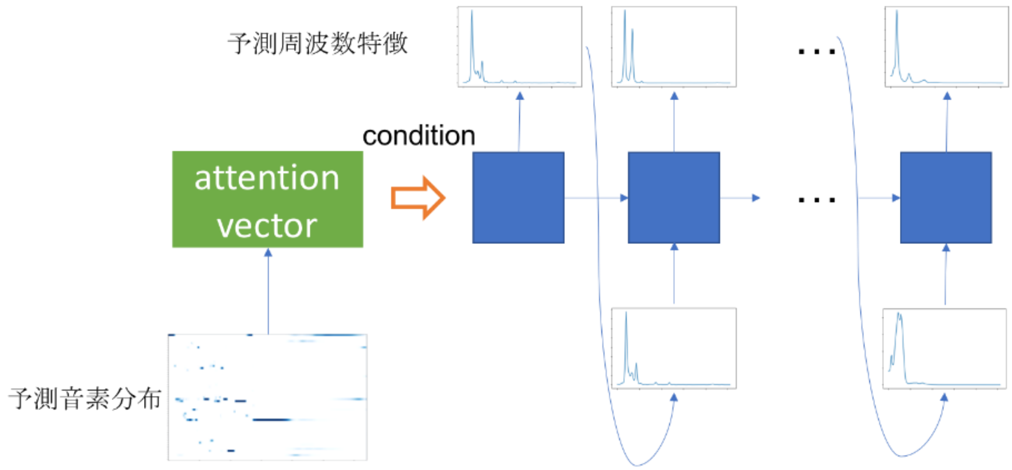

図 2. 声質変換システムの概要図

- 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。

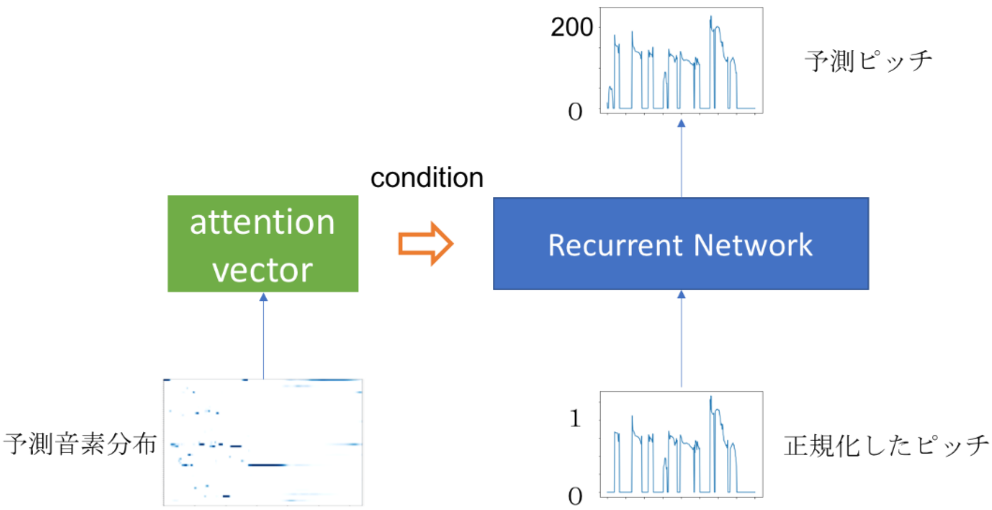

図 3. ピッチ変換システムの概要図

- 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。

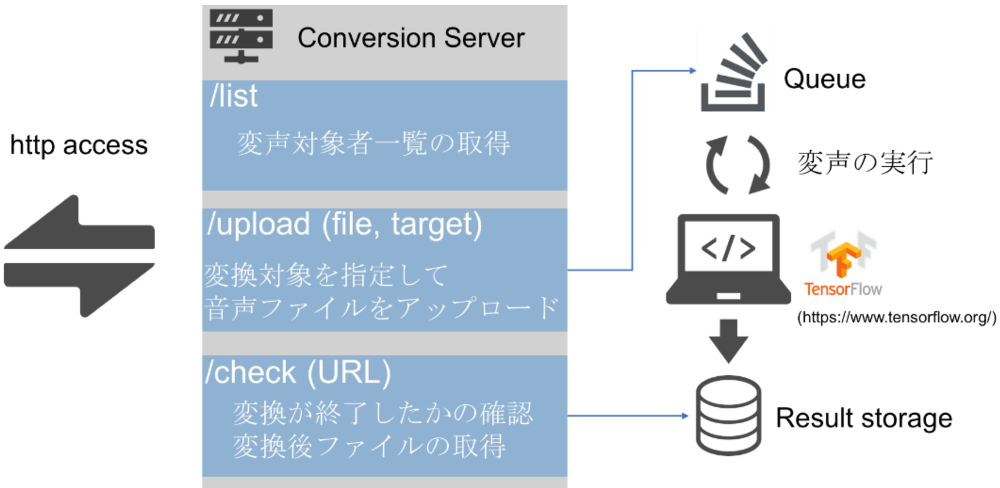

図 4. 実装した API の概要図

10. プロジェクト評価

音素認識と声質変換をベースとして音声変換を実現するという、この野心的なプロジェクトは、誰もが分かりやすい音声変換品質で評価されることもあり、変換品質向上において様々な壁を経験し、粘り強くその改善を行ってきた。本プロジェクト期間中に、当初の音声変換品質を大きく改善し、API 化によるサービ

ス基盤を実装したことは高く評価したい。プロジェクトスタート時に新たな目標として設定した、具体的なサービス化にまでは至らなかったことは残念であるものの、本プロジェクトのチャレンジは未踏そのものと考えている。

11. 今後の課題

　音声変換を実現するという誰もが夢見るゴールに向けて、諦めずにさらなるチャレンジを続けて欲しいと考えている。コア部分の実装改良はもちろん大切ではあるが、ノイズ軽減やサービス実装面での工夫などを行い、ターゲットユーザや適用領域をある程度限定することで、新たな価値提供が可能か、引き続き検討して欲しいことである。