IPA

2018 年度 未踏 IT 人材発掘・育成事業 採択案件評価書

1. 担当 PM

稲見 昌彦（東京大学 先端科学技術研究センター 教授）

2. クリエータ氏名

星野 凌我（豊橋技術科学大学）

3. 委託金支払額

469,600 円

4. テーマ名

独自マイクによる音声サンプリングおよびそれによる声質変換

5. 関連 Web サイト

なし

6. テーマ概要

本プロジェクトでは、ある人の声を別の人の声に変換する声質変換のためのソフトウェアを開発することを目指した。声質変換にはメル周波数ケプストラム係数（MFCC）を用いた声質の特徴量抽出と、動的時間伸縮法（DTW）を用いた時系列の整合という 2 つの前処理を施してから学習する方法がよく用いられる。しかし、この手法にはささやき声や子音に弱かったり、統計的モデルを用いることによる過剰な平滑化が発生したりするという問題があり、声質変換が社会に普及するには未だ至っていない。この 2 つの問題に対して、喉と口元の 2 箇所から音を取る独自手法とディープラーニングを用いることで、従来手法と比べてより似ていて、より自然な声での声質変換を目指した。

本プロジェクトは結果として、どのような収録条件、つまり喉のマイクにどのようなマイクを使用すればよいかの比較実験を行うまでに止まった。使用したマイクは市販の咽喉マイクと、体伝導により微少な音を収録するためのマイクである NAM マイクの 2 つである。この 2 つのマイクに加え、基準用の口元のマイクとして気導マイクを使用した。喉と口元の 2 箇所からサンプリングするという条件のもと 3 種類のマイクを用いて比較した結果、複数のマイクを使用

してサンプリングする際の、子音の回り込みや減衰が問題となることを明らかにした。

7. 採択理由

現在 VR 空間で CG アバタを操作するチャットシステムや VR YouTuber こと VTuber などが話題になっている。そのような背景において、アバタに合わせて演者の声を変換させる声質サンプリングおよび変換技術が注目されている。声質変換自体は各所で研究されているが本提案は発話時の音源を独自のマイクで記録することで、ささやき声や子音などの変換と変換処理の実時間性を確保することを目指している点が興味深い。近年機械学習などの進展により、ソフトウェア上の処理のみで信号処理の諸問題が解決できるようになりつつあるが、人間に関わる技術分野において、ソフトウェアとハードウェアの適切な組み合わせにより劇的に性能が向上することは過去にも多く事例があり、ソフトとハードの匙加減も技術者の腕の見せ所である。星野氏はすでに予備実験で独自マイクの効果の検証を開始しており、実装にあたってのフットワークの軽さと技術バランスのセンス、そして自らが本システムの完成を欲しているという当事者としての熱意を高く評価し、採択するに至った。

8. 開発目標

Mingdi らは「ヒトが話者を認識する際は、母音を手がかりの中心にしている」と主張しており、これは少なくともヒトの知覚という観点では声の個人差が母音に集中していることを意味する。母音は声帯を用いて発声され、子音は主に口元から喉にかけての唇、舌、歯、歯茎、後部歯茎、硬口蓋、軟口蓋、口蓋垂、声門で発声される。各部位の位置を図 1 に示す。

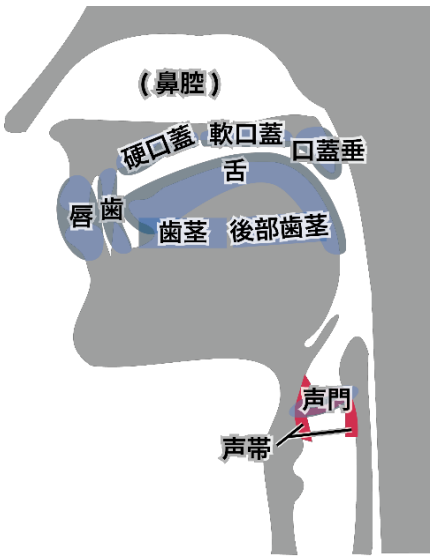

図 1. 子音の発声位置（青）と母音の発声位置（赤）

本プロジェクトでは母音への個人差の集中とヒトの発声機構から着想し、声質変換を行うためのサンプリングの改善を試みた。母音の発声位置に近い喉と子音の発声位置に近い口元という 2 箇所から収録することで母音と子音を分離して収録し、個人差の大きい母音成分を効率的に変換することを目標とした。

9. 進捗概要

本プロジェクトでは、喉のマイクにどのようなマイクを使用すればよいか、マイクを喉につけることが声質変換に有効かを検証するため、実験を行った。使用したマイクは市販の咽喉マイク Retevis JPC9038A と、名古屋大学大学院情報学研究科戸田研究室より借り受けた特注品の NAM マイクである。この 2 つのマイクに加え、基準用の口元のマイクとして気導マイクとして audio-technica AT9904 を使用した。

図 2 の M の位置に気導マイクを、N の位置に NAM マイクを、T の位置に咽喉マイクを取り付けて、50 音を 1 音素ずつ 1 秒間隔で読み上げて収録した。マイクの取り付け位置については、予備実験をもとに SN 比を十分確保できる点を定めた。以降それぞれのマイクをマイク M（Mouth）、マイク N（NAM）、マイク T（Throat）とよぶ。

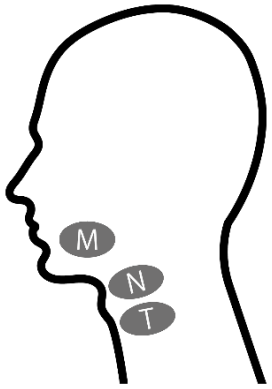

図 2. マイクの取り付け位置

20 代男性話者 1 名で以下を収録した。

あいうえお	*かきくけこ*	*さしすせそ*	*たちつてと*	*なにぬねの*
はひふへほ	*まみむめも*	*やゆよ*	*らりるれろ*	*わをん*
がぎぐげご	*ざじずぜぞ*	*だぢづでど*	*ばびぶべぼ*	*ぱぴぷぺぽ*

5 秒間かけて 1 行を読み上げ、3 秒あけて次の行に移る。読み上げは 2 回連続して行った。

実験結果のスペクトログラムを次の図 3 に示す。それぞれの上段がマイク M、中段がマイク N、下段がマイク T のものである。このスペクトログラムでは、

周波数-時間直交座標系の各点において、当該周波数・当該時刻の強度が高いほど白く、低いほど黒く示されている。

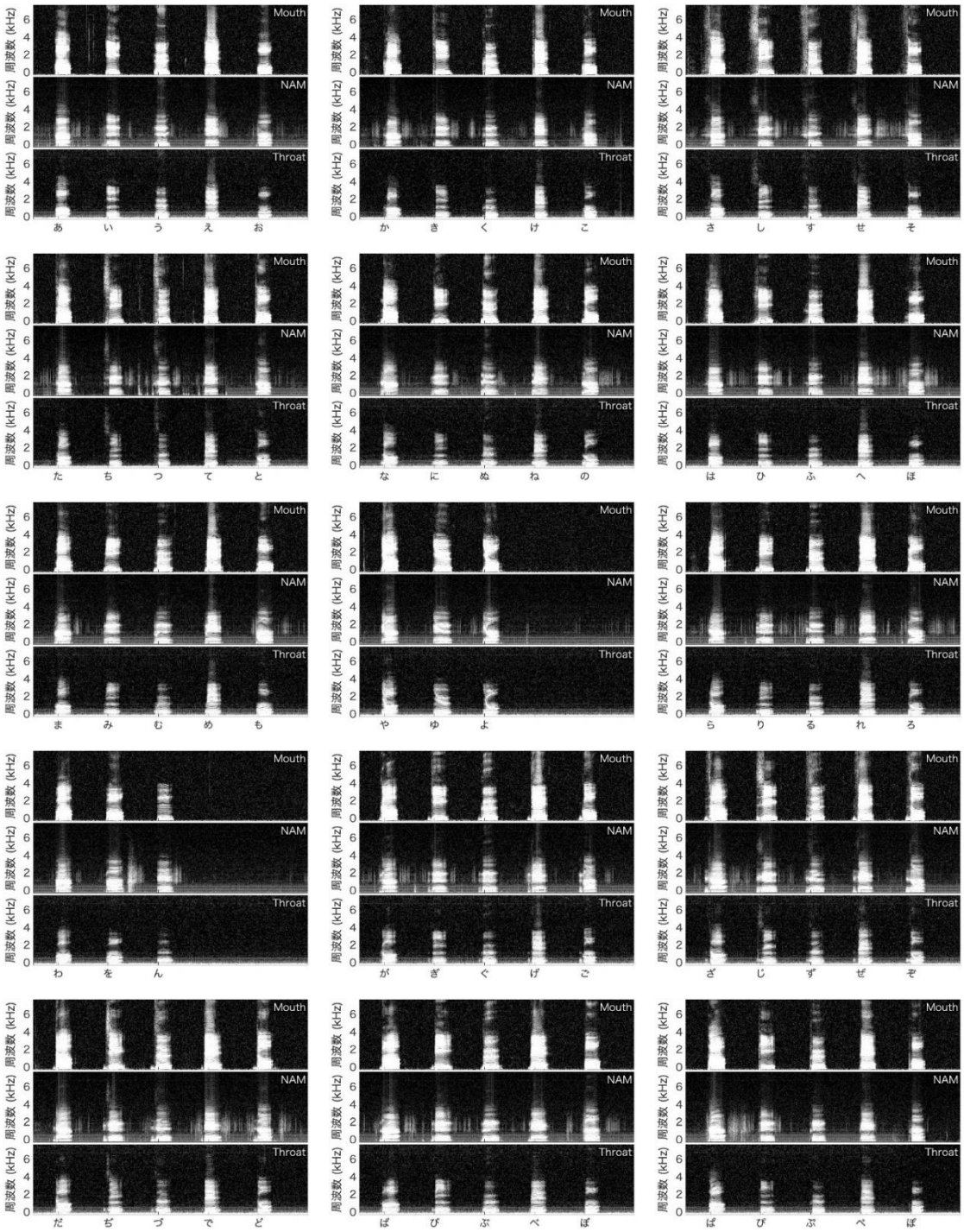

図 3. 3 種類のマイクの比較

先に示した母音と子音を分離して収録することを考えると、マイク T やマイク N は母音の発話中にはマイク M と同一の内容を、子音の発話中には無音となっていることが好ましい。

しかしながら、実際の実験結果はどちらもこの条件を満たしておらず、提案す

るマイクを 2 つ使う手法は高品質な声質変換に適さないことがわかった。これらの原因としては、口元の音声が空気や体内・体表面を伝導して喉に伝達することや、声帯からマイクまで体内を伝導する間に高周波が減衰することが考えられる。

よって本プロジェクトでは喉と口元の 2 箇所からサンプリングするという条件のもと 3 種類のマイクを用いて比較し、複数のマイクを使用してサンプリングする際の、子音の回り込みや減衰が問題となることを明らかにした。

10. プロジェクト評価

本プロジェクトはマイクを咽喉部など複数個所に配置することで、リアルタイム性が高く、精度の高い声質変換を行うことを目標とした。しかしながら前述のように、声帯位置からマイクまでの減衰により、当初の目的を達成しうる性能を得ることが叶わなかった。しなしながら実験によってプロジェクトそのものの難度を提示できた点に一定の価値があると判断する。

11. 今後の課題

複数のマイクというアイデアを活かせるようなアプリケーションの探索、音声マイク以外に声帯周辺部の筋活動を計測する手法の探索など、手法面、アプリケーション面の両輪でプロジェクトを進めていくのが望ましいと考える。