様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 ２１ 日現在

機関番号：３１３０３
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１２０８６

研究課題名（和文）上腕動作をトリガーとして発話の運動指令を制御するMotion-To-Speech

研究課題名（英文）A motor-commend based speech synthesis with the trigger activated by arm motions.

研究代表者
伊藤　仁（Ito, Masashi）

東北工業大学・工学部・教授

研究者番号：００４３６１６４

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：上肢の動作を入力として、音韻やアクセントを即時的に制御するMotion-To-Speech型音声合成システムを用いて/wa,ya/などのわたり音節を出力する際に、入力となる身体動作の違いが操作精度に及ぼす影響を調べた。被験者に目標となる音声を呈示し、動作でそれを再現させる実験を行い、腕全体を用いて音声パラメータを制御した場合と、手首から先だけを用いた場合について比較した。その結果、特に高速性が要求される状況では、手首から先だけの動作の方がオーバーシュートの量が小さくなることが明らかになった。これは動作に関与する部位の質量が重要な要素であることを示唆するものである。

研究成果の学術的意義や社会的意義
Motion-To-Speech型音声合成は、失声者の代替発声装置としての応用が期待されており、誰でも簡単に使えるシステムを実現することが課題となっている。本研究では、この様なシステムの入力として用いる身体動作の動特性を定量化した点に意義があり、これまで個々の研究者が試行錯誤的に行ってきた入力動作の選定に関して、身体工学的な視点からより理に適った設計指針を提供するものであると位置づけられる。本研究で明らかになった身体部位の質量に関する知見を活用することで、より操作性の高いMotion-to-speechシステムが実現できる可能性がある。

研究成果の概要（英文）：Two types of input motions were examined for synthesizing glide /w/ and /y/ with a motion-to-speech synthesizer. Using the synthesizer, subjects were required to reproduce the identical speech signals to the target one which presented beforehand. The reproduced speech parameters showed significant overshoots from the targets. The amounts of the overshoots were smaller when the synthesizer was controlled by hand motions than the case where it was controlled by arm motions. The difference became prominent in the fast transition of the speech parameters. The results indicated the mass of the body to be important factor for precious control in the synthesis.

研究分野：音響信号処理

キーワード：音声合成　生体計測　運動

１．研究開始当初の背景

音声は、肺、顎、舌、口唇などの調音器官の協調的な運動によって生成される。発声に関わる運動は生得的ではなく、自転車の操作等と同様に訓練を通して後天的に学習されたものである。従って、ヒトの発声機構を模擬した人工装置を開発し、その操作に十分に習熟すれば、手足の動作で音声を自在に生成できる可能性がある。この動作を用いてリアルタイムにパラメータを制御する Motion-To-Speech(MTS)型の音声合成は、従来のキーボード等で入力されたテキストを必要とする Text-To-Speech(TTS)型音声合成より、会話の即時性や表現力において有利であると考えられており、失声者の代替発声装置などへの応用が期待されている。

1992 年にトロント大で開発された最初の MTS 型音声合成システムは、手足の動作を入力として流暢な英語音声を出力し得るが、その操作は非常に複雑で一般の使用者が容易に扱えるものとは言い難い。国内では、失声者の発話支援を目的としてペンタブレットに入力された軌跡を用いて音声を合成するシステムが提案されている。また著者らは、手の空間位置で音量、音高、音韻、ビブラートを操作する MTS 型の歌声合成システムを開発した。このシステムは、使用者が両手で二つの歌声を操作することで二部合唱を合成できるが、出力可能な音韻が限定されている点に問題があった。

一般的に MTS 型の音声合成では、入力となる動作と、その動作を用いて制御する音声パラメータの選定が、システムの操作性を決定付ける重要な要因だが、これまでの研究ではこれらは個々の研究者が試行錯誤的に選定しており、より優れたシステムを構築するための有効かつ普遍的な設計指針が呈示されていないという問題があった。

２．研究の目的

本研究では、より使いやすい MTS 型音声合成システムを実現するために、まず入力となる身体動作の特性を定量的に評価する。過去の研究では制御対象を音高だけに限定し、入力動作として手の高さ、肘関節、肩関節を用いた場合の制御精度を比較し、筋骨格系の構造が単純な肘関節の動作が最も正確であることを示したが、本研究ではこのアプローチをさらに進め、2 次元の軌跡を正確に再現する際に有効となる上肢の運動について身体工学に基づく実験を行って評価する。また多様な音韻を効率的に生成するための音声のパラメータについては、発声時の人間の運動指令を模倣することが有効であると考え、この運動指令の実体を把握するための音声の生成実験も行う。上肢の動作により音韻の生成トリガーを発生させ、この運動指令を再現することで、使用者にとって直感的で操作の習熟が容易な MTS 型音声合成システムを構築することを目指す。

３．研究の方法

まず MTS システムへの入力動作の特性を調べるために、被験者を用いた動作計測実験を行う。被験者の右手に図 1 に示すモーションセンサーを取り付け、掌の 3 次元空間座標と向きをリアルタイムで計測する。被験者には、図 2 に示す手続きで目標とする動作を再現させる。まず被験者の正面に設置した液晶ディスプレイに、目標となる動作を表示する。この動作は音声パラメータのうち第 1,2 フォルマント周波数を、ある値から他の値へと線形に変化させるもので、変化速度が速い場合は破裂音（/ba/や/da/など）、中間の場合はわたり音（/wa/や/ya/など）、遅い場合には二連母音（/ua/や/ia/など）が合成される。被験者はフォルマント周波数の目標値だけでなく、変化速度も再現するよう求められる。被験者の動作は、掌の水平・鉛直座標をフォルマント周波数と関連付ける Position 動作と、掌の水平・鉛直の向きをフォルマント周波数と関連付ける Direction 動作の 2 種類とし、それぞれに 8 名の被験者を割り当てて実験を行った。また操作の習熟に要する時間を測定するため、各被験者について 30 回の試行を行った。

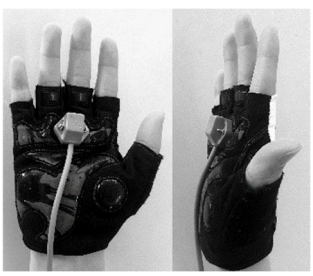

図１．モーションセンサーの設置場所

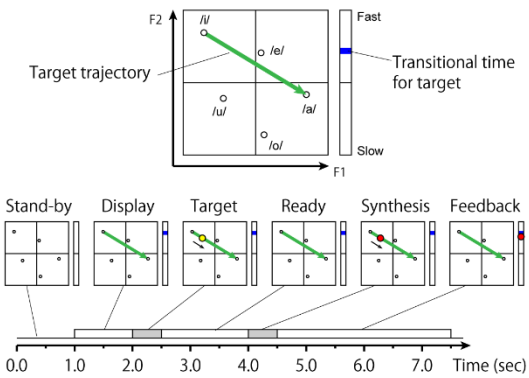

図２．動作計測実験の手続き

４．研究成果

図 3 に被験者の応答の典型例を示す。点線が目標値、丸印が被験者の動作により制御された第 1,2 フォルマント周波数をあらわす。試行回数を重ねるごとに制御値が目標値に近づいていることが分かる。これは被験者が設定された動作と出力される音韻との関係を学習し、MTS 音声合成システムを適切に制御できることを示す結果である。全被験者の目標値と制御値の平均誤差は、約 30 分の操作で一定値に収束する。なお操作の習熟に要する時間は腕全体を用いる Position 動作と、掌だけを用いる Direction 動作で有意な差は見られなかった。

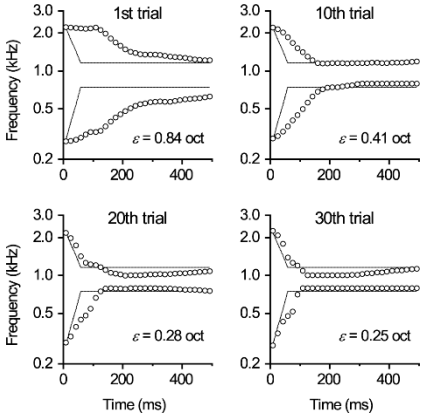

図 3．フォルマント軌跡の例

この実験における被験者の動作を詳細に分析すると、目標となる軌跡と動作による軌跡の間に一貫した傾向が観測された。まず目標値と制御値の誤差の大きさは、Position 動作と Direction 動作で有意差がなかった。動作範囲は Position 動作より Direction 動作の方が広く、モーションセンサーを同じ距離だけ動かした場合のフォルマント周波数の変化は、Position 動作の方が小さくなる。従って、Position 動作の方がより精密な制御が行えると予測されたが、結果結果はこの仮説を支持しなかった。

一方、制御値の時間特性については動作の種類による有意差が観測された。今回の実験では、わたり音を合成するためにフォルマント周波数を 50 ms で変化させる必要があるが、この様に高速ななフォルマント遷移を実現しようとすると、制御値は目標値に達した後でオーバーシュートする傾向がある。このオーバーシュートの量は要求される変化速度が速く、かつ変化量が大きいほど増大する（図 4)。これは、身体動作が質量を持つ物体の運動であり、慣性の法則が関与していることを考えれば自然な結果である。実際、変化速度と変化量が同一の条件では、肩、肘、手首を含む腕全体を動かす Position 動作のオーバーシュート量は、手首から先だけを動かす Direction 動作より有意に大きかった（図 5)。

MTS 型音声合成システムは、例えば日常会話などの状況で即時的に音声を出力する際に有用である。この場合、出力されるのは連続音声であるため、その入力となる身体動作は常に継続することを想定しなければならない。本研究の実験結果は、素早い身体動作が要求される場合は、動作に関与する部位に働く慣性が重要な要素となることを示唆するものであると言える。過去の研究では MTS システムの入力として、本研究でも検討した腕全体や、足を使ったペダルなども提案されてきたが、身体工学的な視点から考慮すると、MTS システムの入力は質量が小さく動きも精密な手指の動作に限定することが理に適っていると考えられる。

また本研究で計画していた、音声生成時の運動指令の解明に関しては、予備実験の段階で被験者間のばらつきが大きく、普遍的な知見を得ることができなかった。同じ言語の音韻を生成する場合でも、脳から調音器官に伝達される運動指令に話者による違いがあることは予測していたが、実際に調べてみると、この個人差が当初の想定を大きく上回るものであった。また発声中に外部から擾乱を加えて、出力される音声の音響的な特徴の変化を比較するという手法では、運動指令の変化だけでなく、話者の調音器官の個人性の違いも同時に扱う必要があるため、解析が難しいことも明らかになった。従って、この問題に対応するためには、脳波や筋電など直接的な計測手法も併用することが必要であると考えられる。

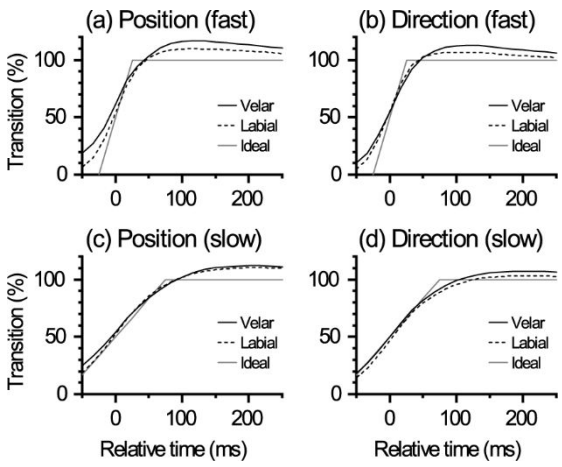

図 4．身体動作のオーバーシュート

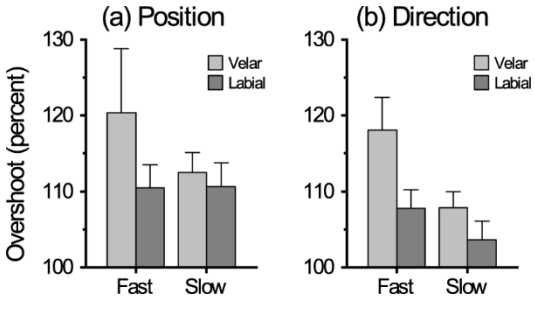

図 5．オーバーシュート量の比較

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 伊藤仁
２．発表標題 腕の動作を用いた母音と わたり音/y,w/のリアルタイム合成
３．学会等名 日本音響学会2021年春季研究発表会
４．発表年 2020年～2021年

１．発表者名 伊藤仁、遠藤慎也、小島銀河
２．発表標題 Motion-to-speech 音声合成におけるわたり音と二連母音の合成に適した 身体動作の検討
３．学会等名 日本音響学会2020春季研究発表会
４．発表年 2020年

１．発表者名 遠藤慎也、小島銀河、伊藤仁
２．発表標題 わたり音と母音のリアルタイム音声合成に適した動作の検討
３．学会等名 2020年東北地区若手研究者研究発表会
４．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関