手の動きを入力としたリアルタイム音声生成系における鼻音の合成に関する検討*

☆國越晶, 喬宇, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

発声器官の制御に障害を持つ構音障害者が会話をする場合、単語を絵や記号で表したコミュニケーションボードや入力した文字を読みあげる VOCA (Voice Output Communication Aids) など、文字や記号の入力を要求する機器を用いることが多い。しかしこれらを使用すると、手話などと比較してリアルタイムに自由な会話をすることが難しく、構音障害者が会話の主導権を握れないといった問題が指摘されている [1]。これは上記した機器の多くが、入力手段として文字や記号を要求するためである。本研究では、構音障害者のリアルタイムで自由な音声コミュニケーションの実現を目標とし、文字や記号を介さない音声生成系として、障害者自身の構音器官以外の身体運動から、直接音声を生成するシステムを検討している。

身体運動から直接音声を生成する研究としては、構音障害者自身によるペンタブレットを使った音声合成器 [2] や、ヒューマンインターフェースの一例として提案された GloveTalkII[3] などが挙げられる。これらは入力機器によってフォルマント周波数、基本周波数、音量などを制御するものである。しかし障害者のコミュニケーションにおける振舞いは、障害の内容などにより極めて多様である [4]。そのため障害者支援機器は個々の障害者に合わせてチューニングされることが多いが、上記の機器において微妙な調整は必ずしも容易ではない。本研究ではこれらを考慮し、身体運動から音声を生成する過程をメディア変換として捉える。近年、与えられた 2 話者間パラレルデータに対して、統計的に空間写像を設計する手法が話者変換の分野で用いられている [5]。この手法を応用し、本研究では、身体運動の特徴量空間から音声の特徴量空間への写像に基づく音声生成系を検討している。

これまでに、手の姿勢（以下ジェスチャーと呼ぶ）を入力とした日本語五母音の連続音声生成において、本手法の有効性を報告した [6]。また「ジェスチャー空間中のジェスチャー群の配置」と「母音空間中の母音群の配置」とが、より等価となるような対応付けによって、より明瞭な音声が生成されることを実験的に検証した [7]。本稿では子音としてまず鼻音に注目し、母音に対して用いた提案手法を子音の合成に拡張する方式について実験的に検討した。

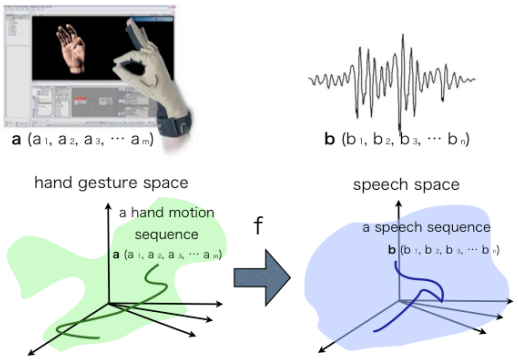


Fig. 1 空間写像に基づくメディア変換の枠組

2 空間写像に基づくメディア変換

2.1 方針

空間写像に基づくメディア変換の枠組を図 1 に示す。ある時刻のジェスチャーが m 次元の特徴量ベクトルで表されるとする。これはジェスチャーを表す m 次元空間（以下ジェスチャー空間と呼ぶ）の中の 1 点に対応する。同様に、ある時刻における音声が n 次元の特徴量ベクトルで表されるとすると、これは n 次元音響特徴量空間の中の 1 点に対応することになる。この 2 つの空間の間の単射な写像関数を求めることで、任意のジェスチャーに対して、対応する音声の特徴量ベクトルを求めることができる。この写像関数は Stylianou らによる手法 [5] を用いて推定することができる。

3 日本語五母音の合成 [6, 7]

3.1 ジェスチャーデザイン

提案するシステムにおいて、日本語五母音による連結母音音声を対象に、手の動きから音声へのメディア変換を実装した [6, 7]。前章で提案した枠組みでは、変換元と変換先の特徴点間の対応がとれたパラレルデータが必要となる。話者変換の場合、DTW などの手法によって 1 対 1 の対応をとることは比較的容易である。本研究の場合，ジェスチャーと音は任意に対応づけることができるため、適切な対応付けを選択することが課題となる。

---

*Nasal sound generation for the real-time hand to speech system. by KUNIKOSHI Aki, QIAO Yu, MINEMATSU Nobuaki, HIROSE Keikichi (The University of Tokyo)

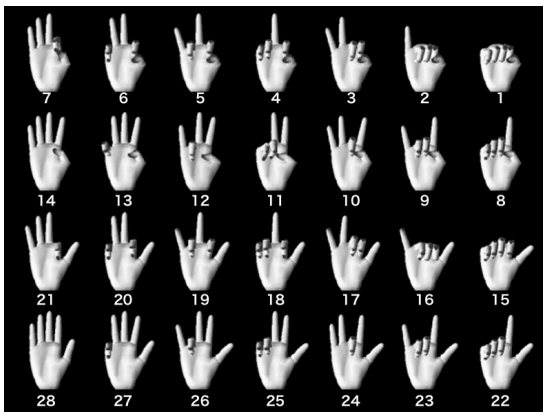


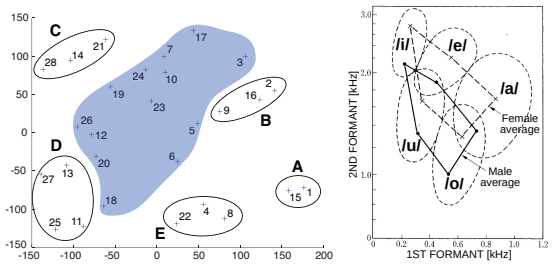
Fig. 2　基本的な 28 種類のジェスチャー

本稿ではジェスチャーの候補として、Wu らが画像認識における論文 [8] の中で用いた、基本的な 28 個のジェスチャーを選んだ。そのジェスチャーを図 2 に示す。これは、五指各々の曲げ延ばしの組み合わせ $2^5 = 32$ 個から、薬指だけを立てるもの、薬指と人差し指を立てるもの、薬指と親指を立てるもの、薬指、人差し指と親指を立てるもの、即ち実現不可能な 4 種類を差し引いたものである。

[7] は、「ジェスチャー空間中のジェスチャー群の配置」と「母音空間中の母音群の配置」とが、より等価となるような対応付けによって、より明瞭な音声が生成されることを示している。「ジェスチャー空間中のジェスチャー群配置」を視覚化するために、この 28 個のジェスチャーを各々 2 回ずつ、計 $2 \times 28 = 56$ 個のデータを、Immersion 製データグローブ CyberGlove で記録し、その全てのデータを用いて PCA を行い、18 次元のデータグローブのデータを 2 次元平面に射影した。結果を図 3(a) に示す。数字は図 2 の各々のジェスチャーを示している。中央の領域は、実現可能であるものの、指に負担がかかったり、同じ姿勢を持続させるのが困難な姿勢などである。それらを除くと、図に示すように凡そ A〜E の 5 つのグループに分類されることが分かる。ジェスチャー空間におけるジェスチャー配置と母音空間における母音配置との等価性を高めることを目的として、図 3(b) に示される F1/F2 図における収録音声の母音群の配置と比較し、A から E をそれぞれ「お」「う」「い」「え」「あ」に対応させた。これらのうち、本実験では日本語五母音に相当するジェスチャーを図 4 のように設定した。

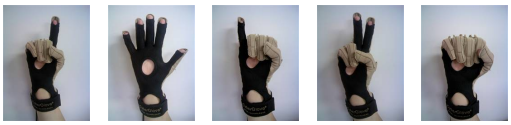
3.2　メディア変換用写像関数の推定

次に学習データとして、データグローブを装着し「あ」「い」「う」「え」「お」および二母音間の遷移 ${}_5P_2 = 20$ 組を各々3 回、計 $(5 + 20) \times 3 = 75$ 個のデータを記録した。センサの数は 18 個, サンプリング周期は 10〜20 ms である[1]。また成人男性 1 名から収録した「あ」「い」「う」「え」「お」および二母音間の遷移 20 組を各々5 回、計 $(5+20) \times 5 = 125$ 個の音声データから、STRAIGHT[9] を用いて分析をおこない、ケプストラム係数 0-17 次を抽出した。フレーム長は 40 ms, フレームシフトは 1 ms とした。そしてデータグローブのデータ 3 セット、音声データ 5 セットから結合ベクトルをつくるために、$3 \times 5 = 15$ 組全ての組み合わせにおいて、データグローブから得られたデータ時系列を、対応するケプストラム時系列の時間長／周期に合わせて線形補完した。得られた結合ベクトルの分布を正規分布でモデル化し（混合数＝ 1)、写像関数を推定した。



(a)　(b)

Fig. 3　(a) PCA 空間における 28 ジェスチャーの位置 (b) 日本語五母音の F1/F2



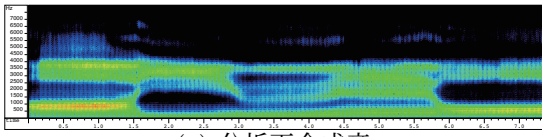
あ　い　う　え　お

Fig. 4　日本語五母音に相当するジェスチャー

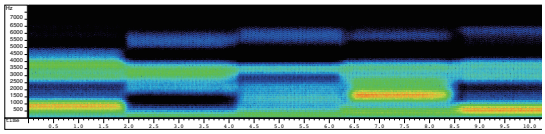
3.3　ジェスチャーを使った音声生成例

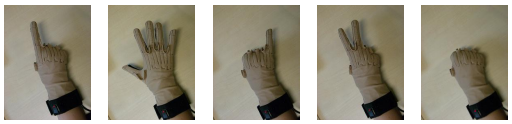
「あいうえお」に対して提案手法により音声を生成した。図 5 に結果を示す。(a) は「あいうえお」の分析再合成音、(b) は明確に動かしたジェスチャーを入力とした場合の合成音、(c) は不明確に動かしたジェスチャーを入力とした場合の合成音の例である。ケプストラム時系列からの再合成には STRAIGHT を用い、F0 はすべて 140 Hz とした。予備的な聴取実験により、日本語五母音による連結母音音声を対象とした合成において、この手法の有効性が示された。また手の動かし方（明確／不明確）によって、発話スタイル（明瞭／不明瞭）も制御できることが分かる。

[1] データグローブからのサンプリング周期は時不変ではない。最終的には、線形補完の形で周期一定となるようデータの再サンプリングを行った。

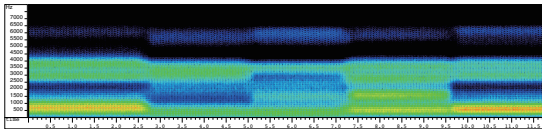


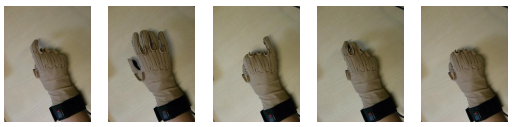
(a) 分析再合成音





(b) 明確に動かしたジェスチャーを入力とした場合の合成音





(c) 不明確に動かしたジェスチャーを入力とした場合の合成音

Fig. 5　「あいうえお」に対する合成音の比較

4　子音の合成

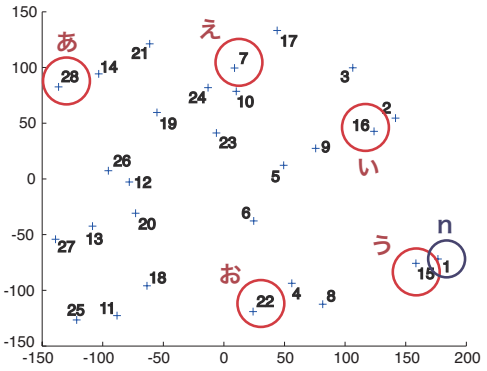
前章では、日本語五母音による連結母音音声を対象とした合成において、提案手法が有効であることが示された。しかし、より自由な発話を目指すためには、さらに子音の追加が必要不可欠である。本章では、本システムにおける子音の合成方式について述べる。

現在、音声の合成方法としては、コーパスベースの波形編集方式や、HMM に基づく分析合成方式などが提案されている。しかし本システムではジェスチャーデータ 1 フレーム毎にリアルタイム変換を行っているため、これらの手法をそのまま応用することは難しい。本稿では子音として鼻音に注目し、母音に対して用いた提案手法を子音の合成に拡張する方式を実験的に検討した。

4.1　子音を考慮したジェスチャーデザイン

本システムに子音を導入するため、子音にもジェスチャーを割り当て、母音に対して用いた提案手法を子音の合成にも拡張する。この方式では、音声にして数 〜数十 ms しか持続しない子音に割り当てるジェスチャーの決定が問題になる。

手の主な働きはものを掴むことであるから、ものを握る動きは最も形成が容易で、速く動かすことができる動きのひとつであると考えられる。そこで本稿ではもの握る動きを子音–母音の遷移に割り当てることとし、図 2 の No.1 のジェスチャーを/n/に割り当てることとした。



ジェスチャー空間における/n/と日本語五母音の位置関係

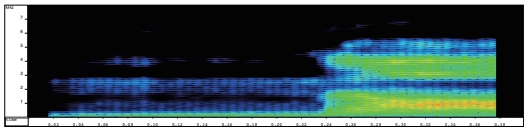
n　あ　い　う　え　お

Fig. 6　/n/を考慮した場合の日本語五母音に相当するジェスチャー

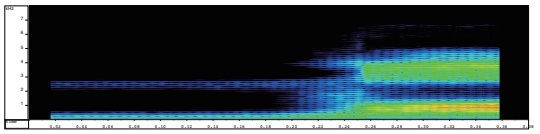
「な」行を発声する際の/n/の口の形は「う」に近いため、図 3(a) の PCA 空間において、No.1 と「う」に割り当てられるジェスチャーは近いことが推測される。これと 3.1 節での議論とから、/n/を考慮した場合のジェスチャーデザインを図 6 のように設定した。

4.2　提案デザインに基づく写像関数の推定

前節で提案したジェスチャーデザインを用いて、提案するシステムにおいて、/n/および日本語五母音による連結母音音声を対象に、手の動きから音声へのメディア変換を実装した。学習データとして、データグローブを装着し「な」「に」「ぬ」「ね」「の」、「あ」「い」「う」「え」「お」および二母音間の遷移 ${}_5P_2 = 20$ 組を各々3 回、計 $(5+5+20)\times 3 = 90$ 個のデータを記録した。また成人男性 1 名から「な」「に」「ぬ」「ね」「の」を各々10 回、「あ」「い」「う」「え」「お」および二母音間の遷移 20 組を各々5 回、計 $5\times 10+(5+20)\times 5 = 175$ 個の音声データを収録した。アライメントをとるため、「な」「に」「ぬ」「ね」「の」のジェスチャーデータおよび音声データは、視察によって、それぞれ子音部分／遷移部分／母音部分に 3 分割した。そして 3.2 節の場合と同様に、ジェスチャーデータの再サンプリング、ケプストラム係数 0-17 次の抽出を行った。その後「な」「に」「ぬ」「ね」「の」においては、データグローブのデータ 3 セットと音声データ 10 セットの全ての組み合わせ $3 \times 10 = 30$ 組、「あ」「い」「う」



(a) 分析再合成音



(b) GMM 混合数 2 の場合の合成音

Fig. 7　「な」に対応する合成音

「え」「お」および二母音間の遷移においては、データグローブのデータ 3 セットと音声データ 5 セットの全ての組み合わせ $3 \times 5 = 15$ 組において、結合ベクトルを作った。この際、それぞれのデータセットにおいて、データグローブのデータ時系列とケプストラム時系列うち、時間長の短い方を長い方に合わせて、線形補完した。計算の便宜のため、このようにして得られた結合ベクトル時系列から、8 データごとに 1 データ抽出したものを学習データとして用いた。図 7 に結果を示す。

4.3　聴取実験

14 人の聴取者を募り、分析再合成音と混合数 2 の GMM によって合成した「な」「に」「ぬ」「ね」「の」、計 $2 \times 5 = 10$ 文字に対し、明瞭度（intelligibility）試験を行った。実験で使用した試料セットには、上記の 10 個の試料の他,「あ」「い」「う」「え」「お」と「ま」「み」「む」「め」「も」の分析再合成音、計 $2 \times 5 = 10$ 個をダミー音声として含めた。回答者には試料がそれぞれ 1 モーラ音声であることを知らせ、ランダムに呈示した 0.3〜0.4 ms 程度の試料それぞれに対し、ローマ字で書き取りを行わせた。結果を表 1 に示す。

提案手法により合成した「な」は分析再合成音の半分以上の明瞭度が得られているのに対し、「に」「ぬ」「ね」は、まったく知覚されていない。間違いの種類としては、「な」では/n/を/m/と知覚した回答が、「に」「の」では/n/を/w/と知覚した回答が、「ぬ」では/n/または/u/という回答が、それぞれの文字における誤回答のおよそ半分を占めていた。また/n/を/w/と置き換える誤りは、分析再合成音に対する回答には見られなかった。

/n/, /m/, /w/は調音点が互いに近接しており [10]、これらの識別には、口唇や鼻腔などの影響が重要である。提案手法では、声道形状のみに注目したために、このような誤りが起こったものと考えられる。/n/と/m/, /n/と/w/の間違いを許容すると、明瞭度は表 2 のように大幅に向上する。

Table 1　明瞭度

合成方式	な	に	ぬ	ね	の	平均
分析再合成	100	93.8	93.8	81.3	100.0	93.8
提案手法	56.3	0.0	0.0	0.0	18.8	15.0

Table 2　/n/と/m/, /n/と/w/の間違いを許容した場合の明瞭度

合成方式	な	に	ぬ	ね	の	平均
分析再合成	100	100	93.8	100	100	98.8
提案手法	87.5	56.3	12.5	18.8	75.0	70.6

5　まとめ

本稿では空間写像に基づいて手の動きを入力とする音声生成系において、母音に対して用いた提案手法を子音の合成に拡張し、鼻音の合成の実験的検討を行った。提案手法による合成音は、明瞭度において分析再合成音を大きく下回っていた。回答の中には/n/を/m/または/w/としたものが多く、それらを許容すると明瞭度は大きく向上したことから、口唇や鼻腔の動きを考慮する必要性が示唆された。今後は、本システムにおいて、調音における声道形状以外のパラメータを制御する方法について検討する予定である。また子音と母音を合わせた音の配置パターンと等価な配置パターンをもつジェスチャーセットについても検討したい。

参考文献

[1] 畠山,「コミュニケーション障害者に対する支援システムの開発と臨床現場への適用に関する研究」シンポジウム予稿集, 12-13 (2007)

[2] 藪他, 信学技報, SP2006-164, 25-30 (2006)

[3] GloveTalkII
http://hct.ece.ubc.ca/research/glovetalk2/index.html

[4] 市川, 手嶋,“福祉と情報技術”, オーム社 (2006)

[5] Stylianou *et.al.*, IEEE Trans. Speech Audio Process., vol.6, 131-142 (1998)

[6] 國越他, 音講論（秋）, 1-Q-23, 375-376 (2007)

[7] A. Kunikoshi *et.al.*, Proc. INTERSPEECH, 308-311 (2009)

[8] Y. Wu *et.al.*, IEEE Trans. Pattern Analysis and Machine Intelligence, vol.27, No.12, 1910-1922 (2005)

[9] 河原, 増田, 信学技報, EA96-28, 9-16 (1996)

[10] 松崎, 河野, “よくわかる音声”, アルク (1998)