ジェスチャー空間と音響空間の写像に基づく リアルタイム音声生成系*

◎國越晶, 喬宇, 鈴木雅之, 峯松信明, 広瀬啓吉 (東大), 坂野秀樹 (名城大)

1 はじめに

現在、発声器官の障害により音声コミュニケーションが困難な構音障害者は、手話や筆談のほか、単語を絵や記号で表したコミュニケーションボード、入力した文字を読みあげる VOCA (Voice Output Communication Aids) などを用いてコミュニケーションを行っている。しかしこれらの方法では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]。 これは構音障害者を支援する機器の多くが、文字や記号を介して情報を伝達しているためと考えられる。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ている。本稿では文字や記号を介さない音声生成として、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、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きわめて多様である [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、話者変換をある話者の音響空間から別の話者への音響空間への写像として捉え、統計的に空間写像を設計する手法が用いられている [5]。筆者らはその手法を応用し、身体運動の特徴量空間から音声の特徴量空間への異メディア間写像を考えることで、音声生成を実現した [6]。

本稿では母音とジェスチャー空間とのより良い対応を求めるために、「ジェスチャー空間におけるジェスチャー群の配置」と「母音空間における母音群の配置」の等価性をより保証できる空間写像の設計し、それをリアルタイム化した。

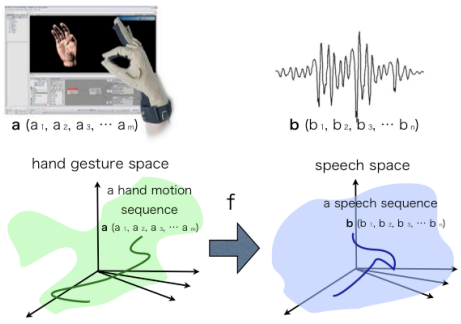


Fig. 1 空間写像に基づくメディア変換の枠組

2 空間写像に基づくメディア変換

空間写像に基づくメディア変換の枠組を図 1 に示す。ある時刻の手の姿勢（以下ジェスチャーと呼ぶ）が m 次元の特徴量ベクトル$\boldsymbol{a}$で表されるとする。これは手の姿勢を表す 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$\boldsymbol{b}$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 予備的検討

3.1 実験

提案するシステムの予備検討として、日本語 5 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。5 母音に相当するジェスチャーは図 2 のように設定した。この際、2 母音遷移途中に他の母音のジェスチャーが混入しないよう配慮した。

次に学習データとして、Immersion 製データグローブ CyberGlove 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。ただし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。また成人男性

*Speech generation from hand motions based on space mapping. by KUNIKOSHI Aki, QIAO Yu, SUZUKI Masayuki, MINEMATSU Nobuaki, HIROSE Keikichi (The University of Tokyo), BANNO Hideki (Meijo University)

Fig. 2　日本語５母音に対応するジェスチャー

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 $(5+20)\times5=125$ 個の音声データから、STRAIGHT[7] 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、$3\times5=15$ 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。

3.2　結果

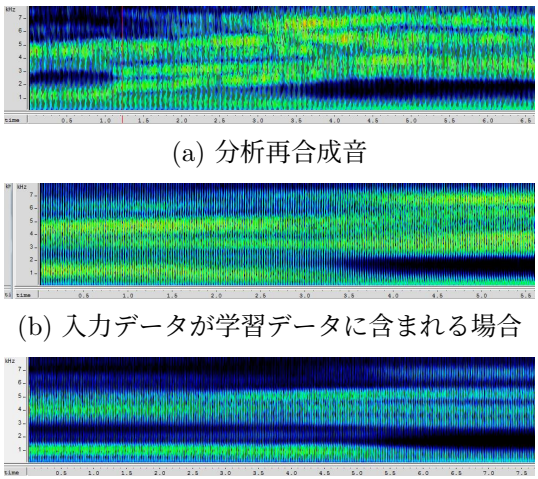
「あ」から「い」への遷移について提案手法により音声を生成した。図 3 に結果を示す。(a) は「あ→い」の遷移の分析再合成音、(b) は「あ」「い」「う」「え」「お」および 2 母音間の遷移 20 組、計 $5+20=25$ 個を学習データとして推定した写像関数に、学習データ内の「あ→い」へのジェスチャー遷移を入力した場合の合成音、(c) は上記のモデルに、学習データに含まれていない「あ→い」のジェスチャー遷移を入力した場合の合成音の例である。ケプストラム時系列からの再合成には STRAIGHT を用い、F0 はすべて 140 Hz とした。予備的な聴取実験によってこの手法の有効性を確認できたが、同時に「い」「う」「お」の不明瞭さが指摘された。

4　ジェスチャーと母音の対応付けの検討

4.1　ジェスチャー空間における５母音の配置

予備実験の結果において、手法の有効性は確認されたものの、合成音において「い」「う」「お」の不明瞭さが指摘された。これは、ケプストラム空間において、これらの母音が近傍に配置されていることを示唆するが、これは逆にジェスチャー空間においても「い」「う」「お」の位置が近接していることが推測される。そこで、ジェスチャー空間における各々のジェスチャーの配置関係を調査した。

Wu らは画像認識における論文 [8] の中で、基本的



(a) 分析再合成音

(b) 入力データが学習データに含まれる場合

(c) 入力データが学習データに含まれていない場合

Fig. 3　「あ→い」の遷移に対応する合成音

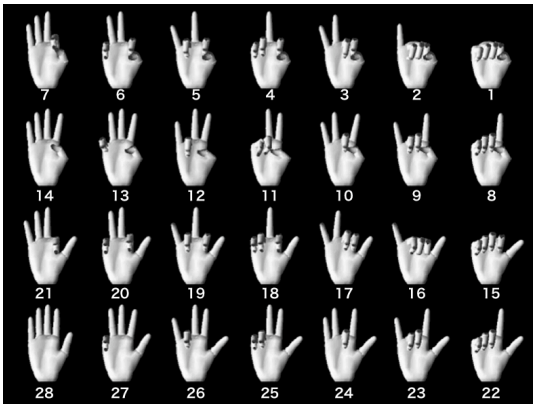


Fig. 4　基本的な 28 種類のジェスチャー

な 28 個のジェスチャーを図 4 のように定めている。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、の実現不可能な 4 種類を差し引いたものである。この 28 個のジェスチャーを各々２回ずつ計 $2\times28=56$ 個のデータをデータグローブで記録し、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。予備検討で用いた「あいうえお」のジェスチャー遷移は、この平面上で、図 5 のようになった。丸で示した部分は各母音に相当するジェスチャーがこの空間で占める凡その位置を、５母音に相当するジェスチャーを各々4 回ずつ計測した計 $4\times5=20$ 個のデータから推定したものである。合成音で不明瞭だった「い」「う」「お」は、ジェスチャー空間においても近傍に配置されていたことが分かった。以上より、合成音において明瞭度を上げるためには、ジェスチャー空間においても五母音の配置を明確化／区別化する必要があると考えられる。

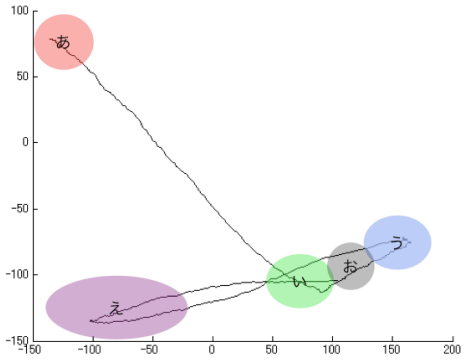


Fig. 5 予備検討モデルによる「あいうえお」の遷移

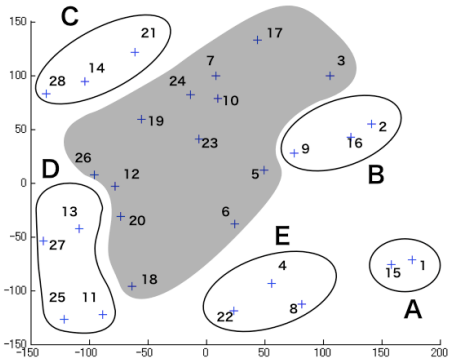


Fig. 6　28 種類のジェスチャーの分類

4.2　5 母音に対するジェスチャーとその最適化

同様の分析を、図 4 に示される 28 種類のジェスチャーに対して行なった。結果を図 6 に示す。数字は図 4 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ジェスチャーなどである。それらを除くと、図 6 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、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。今回は煩雑さを避けるため、A からは基本的な形である No.1 を選び、残りの 4 グループからは、各々、そのジェスチャーの形成が容易だと思われる 2 つを選んだ。また母音図における母音配置を考慮し（但し回転は考えない）、考えられる 16 通りの組み合わせを表 1 に示す。

これらの中から、より最適なジェスチャー・母音対応を決定する。前述したように、障害者支援においては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。そのような場合、チューニングの度に写像関数を推定して実際に音声を合成し、聴取実験によってモデルを評価するのは現実的ではない。そこで本稿では写像関数を推定することなく、ジェスチャー群配置と母音群配置をより簡便に比較することを目的として、構造的表象を用いることとした。

Table 1　16 種類のジェスチャーと母音との対応付け

No.	あ	い	う	え	お	No.	あ	い	う	え	お
1	8	14	2	11	1	9	22	14	2	11	1
2	8	14	2	13	1	10	22	14	2	13	1
3	8	14	16	11	1	11	22	14	16	11	1
4	8	14	16	13	1	12	22	14	16	13	1
5	8	28	2	11	1	13	22	28	2	11	1
6	8	28	2	13	1	14	22	28	2	13	1
7	8	28	16	11	1	15	22	28	16	11	1
8	8	28	16	13	1	16	22	28	16	13	1

5　構造的表象を用いたジェスチャーの決定

5.1　話者不変の音声の構造的表象

音声に不可避的に混入される話者性を捨象して音声を表象する手法として、音声の構造的表象は提案された [9]。音声合成における話者変換技術の多くは、空間写像として実装されるが、如何なる写像・変換に対しても不変な物理量のみで音声を表象できれば、それは話者不変量となる。筆者らの一部はその先行研究において、f-divergence が可逆かつ連続な如何なる変換に対しても不変であることを示し、f-divergence の一つであるバタチャリヤ距離を用いて音声を表象することで話者不変表象を導出した [9, 10]。

音響空間中に時系列として存在する音声ストリームを考える。これを分布系列へと変換し、全ての二分布間距離をバタチャリヤ距離尺度を用いて計測する。得られた距離行列は如何なる変換に対しても不変となるのは自明である。一般に距離行列は一つの幾何学的図形を規定するため、得られた距離行列を音声の構造的表象と呼ぶ。図 7 に一発声を構造として表象する様子を図示する。図 7 における分布数は、例えば 5 つの母音を連結した発声であれば、凡そ 25 個ほどの分布が必要となることが認識実験では示されている。この数を本稿でも使った。例えば「あいうえお」という発声を 25 角形で表現し、また、ジェスチャーの変化の様子も同様に 25 角形で表現し、両者の対応の「尤もらしさ」を検討した。構造的表象は変換不変であるため、本来ならば、各母音をどのジェスチャーに割付けようと、手の運動と音声の両表象は同一となるはずである。しかし、ガウス分布を非線形変換すると一般に非ガウス分布となることから分かるように、例えば分布の形状としてガウス分布のみを許せば、それは、変換不変性を満たす写像関数群に制約を課すことになる。その結果、非線形写像によって構造は歪むことになる。逆に構造間差異が小さいほど、両メディア空間はより線形性の高い写像で変換でき

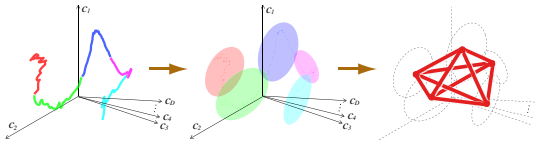


Fig. 7　音のコントラストのみ用いた音声の構造表象

ることを意味する。

本稿では与えられた二構造間の構造間差異は、距離行列をベクトル化し（上三角行列の要素を取り出す)、その二ベクトル間のユークリッド距離で表現する。この値は一方の構造をシフト及び回転して他方に近づけ、対応する 2 事象間距離の和の最小値を近似することが示されている。

5.2　準最適なジェスチャーの設計

表 1 に基づく手の運動と、収録音声の構造ベクトル間距離を求めた結果を図 8 に示す。16 種類の対応付けのうち最も両メディア間の構造ベクトル間距離が小さかったのは No.5、最も大きかったのは No.14 であった。なお図 8 には、16 種類の対応付けによる構造ベクトル間距離の平均と標準偏差についても示す。

各対応付けを用いて写像関数を推定し、生成した「あいうえお」を図 9 に示す。(a) には比較として「あいうえお」の分析再合成音を示してある。視察では、No.5 と No.14 の間には明確な差異が見られなかった。しかしこの二つを予備実験時の対応付けと比較すると、スペクトルの山と谷が強調され、個々の母音の違いがより明確に生成されている様子が分かる。また 10 人の聴取者を募り、分析再合成音と、No.5、No.14、予備実験のモデルを用いて生成した 5 母音の並べ替え 5 単語、計 $4 \times 5 \times 5 = 100$ 文字に対し、了解度試験を行った。結果は分析再合成音で了解度 100 %、No.5 で 99.6 %、No.14 で 99.2 %、予備実験のモデルはこれらを大きく下回る 95.2 %となった。以上より No.5 と No.14 が予備実験のモデルと比べ、より明瞭な音声を生成することがわかった。しかしスペクトログラムおよび聴取実験において No.5 と No.14 の明確な差異が確認できなかった。今回行ったジェスチャーと音声間の構造的比較は、各メディアにおいて全空間をそのまま用いている。[9, 10] などでは適切な部分空間への分割の必要性が示されており、今後の検討事項の 1 つである。

なおこのデザインに基づき、C 言語版 STRAIGHT を用いてシステムをリアルタイム化した。

6　おわりに

本稿では空間写像に基づいて手の動きを入力とするリアルタイムな音声生成系を構築した。その際「ジェスチャー空間におけるジェスチャー群の配置」と「母音空間における母音群の配置」の等価性をより保証できる空間写像を設計することで、より明瞭な音声を生成できることがわかった。しかし構造ベクトル間距離を求めることによる音声の明瞭化については今後より詳しく検討していく必要がある。

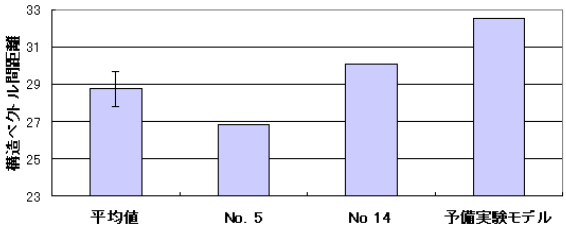
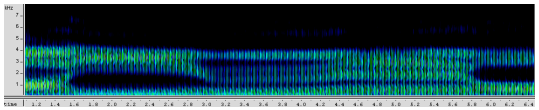
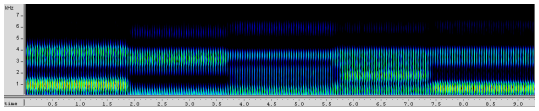


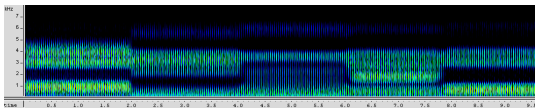
Fig. 8　構造ベクトル間距離



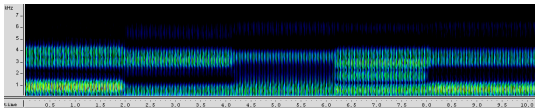
(a) 分析再合成音



(b) No. 5 のモデルから合成した場合



(c) No. 14 のモデルから合成した場合



(d) 予備実験のモデルから合成した場合

Fig. 9　「あいうえお」に対応する合成音の比較

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, pp12-13, 2007

[2] 藪 他, 信学技報 Vol.106 No.613 pp25-30, 2007

[3] 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, pp.131-142, 1998

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

[7] 河原, 増田, 信学技報, EA96-28, pp.9-16, 1996

[8] Ying Wu et.al.,“Analyzing and Capturing Articulated Hand Motion in Image Sequences IEEE Trans. Pattern Analysis and Machine Intelligence, vol.27, No.12, pp.1910-1922, 2005

[9] 峯松他, 音講論（春）, 1-P-12, pp.147-148, 2007

[10] Y. Qiao, et.al., 音講論,（秋） 2-P-2, pp.109-112, 2007