社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

空間写像に基づく手の動きを入力とした音声生成系の構築

國越　晶†　喬　宇††　鈴木　雅之††　峯松　信明††　広瀬　啓吉†††

† 東京大学大学院新領域創成科学研究科　〒 277–8561 千葉県柏市柏の葉 5–1–5
†† 東京大学大学院工学系研究科　〒 113–8656 東京都文京区本郷 7–3–1
††† 東京大学大学院情報理工学系研究科　〒 113–0033 東京都文京区本郷 7–3–1
E-mail: †{kunikoshi,qiao,suzuki,mine,hirose}@gavo.t.u-tokyo.ac.jp

あらまし　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、文字や記号の入力を介して音声を生成する機器を用いることが多い。しかし、リアルタイムに自由な発話をすることが難しく、障害者が会話の主導権を握れない等の問題が指摘されている。そこで本研究では、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の構築を検討する。近年、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。この手法を応用し、本研究では日本語五母音を対象として、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構築する。まず予備的検討として、二母音間遷移中に別の母音が混入しないように母音とジェスチャーとを対応させ、連結母音音声の生成系を構築し、手の運動から音声生成が可能であることを確認した。次に、母音とジェスチャー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を設計した。実験の結果、両メディア間の等価性を考慮した空間写像によって、より明瞭な音声を生成することが可能となった。
キーワード　構音障害、音声生成、手の運動、メディア変換、母音・手姿勢配置、構造的表象

Development of a speech generator from hand motions based on space mapping

A. KUNIKOSHI†, Y. QIAO††, M. SUZUKI††, N. MINEMATSU††, and K. HIROSE†††

†Grad. School of Frontier Sci., The Univ. of Tokyo, 5-1-5 Kashiwano-ha, Kashiwa-shi, Chiba 277-8561, Japan
†† Grad. School of Eng., The Univ. of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
††† Grad. School of Info. Sci. and Tech., The Univ. of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033
E-mail: †{kunikoshi,qiao,suzuki,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract　When individuals with speaking disabilities, dysarthrics, want to communicate using speech, they often use speech synthesizers which require them to type word symbols or sound symbols. This input method often makes realtime operations difficult and dysarthric users fail to control the flow of conversation. In this paper, a new and novel speech synthesizer is proposed where not symbol inputs but hand motions are used to generate speech. In recent years, statistical voice conversion techniques have been proposed based on space mapping. By applying these methods, a hand motion space and a vowel space is mapped and a convertor from hand motions to vowel transitions is developed. In this paper, as a preliminary discussion, the correspondence between Japanese five vowels and five hand gestures is fixed so that a transition between any pair of vowels will not generate a third vowel. Using this correspondence, we develop a converter, which will convince us that the conversion is effective enough. After this preliminary discussion, we make attempts to find a more optimal correspondence between hand gestures and vowels. By considering the equivalence between geometrical features of the gestural arrangement in the gesture space and those of the vowel arrangement in the vowel space, we show that the quasi optimal correspondence can be obtained.
Key words　Dysarthria, speech production, hand motions, media conversion, arrangement of gestures and vowels, structural representation

1. はじめに

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、非音声のコミュニケーション手段として手話や筆談を利用する他、音声メディアの利用に関しても、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) [1], [2] などを用いることで音声対話を行なっている。しかしこれらの機器を使用すると、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 [3]。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最終的な目標とし、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する。

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 [4] 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 [5] 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（前者は F1/F2 平面をペンタブレットに貼り付け、後者は手、腕などの身体姿勢がそのまま音響パラメータに変換される）。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 [6]。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 [7]。本研究で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、音声生成を実現する(注1)。

本稿では、日本語五母音を連結して発声される音声（連結数・順序は任意）を対象として、写像関数を設計する。一般に写像関数の設計は、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、音響空間同士の変換であるから、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。一方本研究の場合、手の姿勢（以下ジェスチャーと呼ぶ）と音は任意に対応づけることができる。手話には日本語の平仮名一つ一つをジェスチャーに対応させた指文字がある。これを採用して空間写像を設計することも可能であるが、この場合、二母音遷移の途中で別の母音が混入するなど（「さ」→「て」を行なうと必ず「え」のジェスチャーを通る）、不適切な対応付けとなってしまう。そこで本稿ではまず、予備実験として、このような不具合が生じないジェスチャーと母音との対応付けを一つとり上げ、写像関数を推定し、手の運動からの音声生成系を構成する。次に、より適切な対応付けの検討を行なう。ここでは「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を実験的に検討する。

(注1)：テルミンという楽器は、手の動きが音高の動きに直接反映されるが、ここでは、音色テルミンを構築する。健常者の舌は、天然の音色テルミンである。

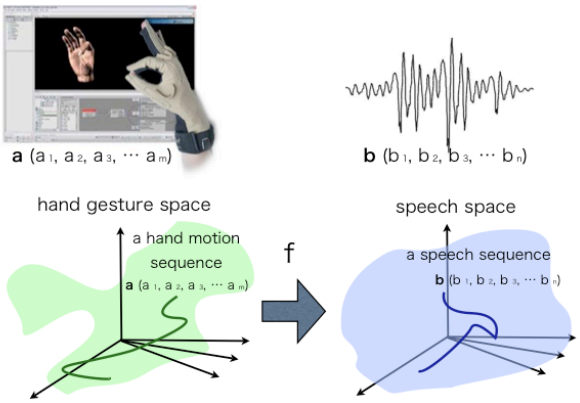


図 1　空間写像に基づくメディア変換の枠組

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。ある時刻の手の姿勢が 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、任意の手の姿勢に対して、対応する音声の特徴量ベクトルを求めることができる。

2.2 写像関数の推定

この写像関数は Stylianou らによる手法 [7] を用いて推定することがができる。その手順を以下に述べる。まず対応関係のわかっているジェスチャー空間の特徴量ベクトル$\boldsymbol{a}$と音響空間のケプストラムベクトル$\boldsymbol{b}$から、結合ベクトル$\boldsymbol{z}$=[$\boldsymbol{a}$, $\boldsymbol{b}$] をつくる。この結合ベクトルの分布を混合正規分布 (Gaussian Mixture Model：GMM) によってモデル化する。

$$p(\boldsymbol{z}) = \sum_{i=1}^{q} \omega_i \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{1}$$

$$\sum_{i=1}^{q} \omega_i = 1, \omega_i \geqq 0 \tag{2}$$

$$\boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_i^A \\ \boldsymbol{\mu}_i^B \end{bmatrix} \tag{3}$$

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i^{AA} & \boldsymbol{\Sigma}_i^{AB} \\ \boldsymbol{\Sigma}_i^{BA} & \boldsymbol{\Sigma}_i^{BB} \end{bmatrix} \tag{4}$$

ここで $\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ は平均 $\boldsymbol{\mu}_i$, 分散 $\boldsymbol{\Sigma}_i$ の正規分布を表し、q は混合数、ω_i は重みを表す。変換関数 $f(\boldsymbol{a})$ は、これらのパラメータを使って、q 個の正規分布で定義された写像関数の重み付け和で表現される。

図 2　日本語 5 母音に対応するジェスチャー

$$f(\boldsymbol{a}) = \sum_{i=1}^{q} h(i|\boldsymbol{a})[\boldsymbol{\mu}_i^B + \boldsymbol{\Sigma}_i^{BA}\boldsymbol{\Sigma}_i^{AA-1}(\boldsymbol{a} - \boldsymbol{\mu}_i^A)] \tag{5}$$

i 番目の正規分布における事後確率 $h(i|\boldsymbol{a})$ は以下で与えられる。

$$h(i|\boldsymbol{a}) = \frac{\omega_i \mathcal{N}(\boldsymbol{a}; \boldsymbol{\mu}_i^A, \boldsymbol{\Sigma}_i^{AA})}{\sum_{j=1}^{q} \omega_j \mathcal{N}(\boldsymbol{a}; \boldsymbol{\mu}_j^A, \boldsymbol{\Sigma}_j^{AA})} \tag{6}$$

ここでジャスチャー空間上のベクトル $\boldsymbol{a}$ に対応する音声の特徴量ベクトル $\boldsymbol{b}$ は、$\boldsymbol{b} = f(\boldsymbol{a})$ として推定される。

3.　予備的検討

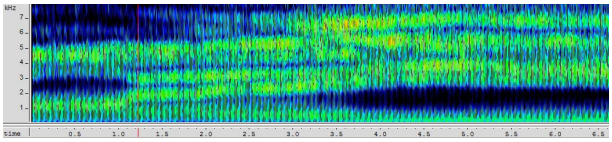
3.1　実　　験

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。五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、既述した様に、二母音遷移途中に他の母音の姿勢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(注2)。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 [8] 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。

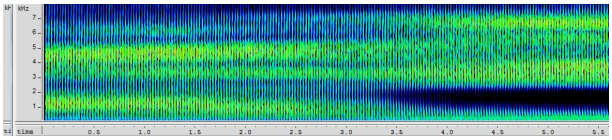
3.2　結　　果

「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」および二母音間の遷移 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),(d) 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。

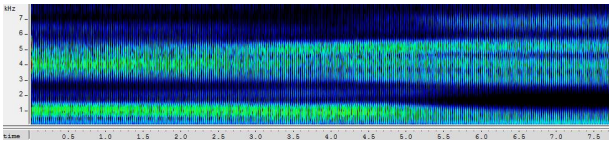
(注2)：データグローブからのサンプリング周期は時不変ではない。最終的には、線形補完の形で周期一定となるようデータの再サンプリングを行った。



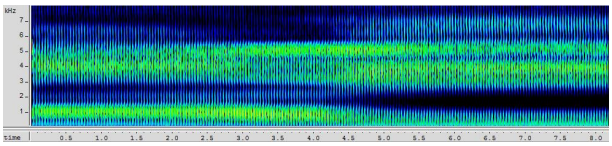
(a) 分析再合成音



(b) 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 1



(d) 入力データが学習データに含まれていない場合 2

図 3　「あ→い」の遷移に対応する合成音

4.　ジェスチャーと母音の対応付けに関する検討

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。予備的検討では、二母音遷移において異なる母音が挿入されないことのみを配慮したが、ジェスチャーと母音の対応付けに対して、より適切な制約条件を導入する必要がある。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 [9] の中で、基本的な 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 $2 \times 28 = 56$ 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

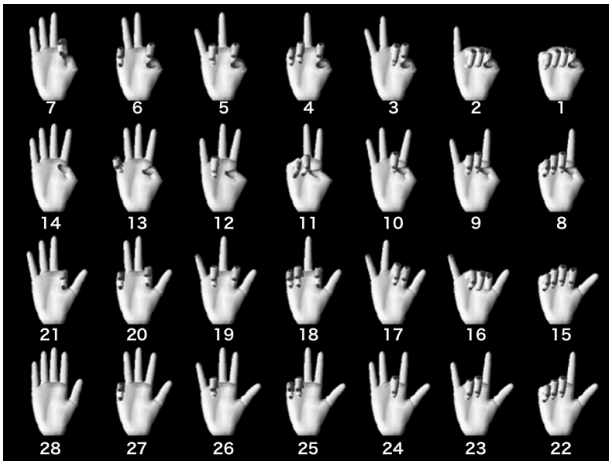


図 4　基本的な 28 種類のジェスチャー

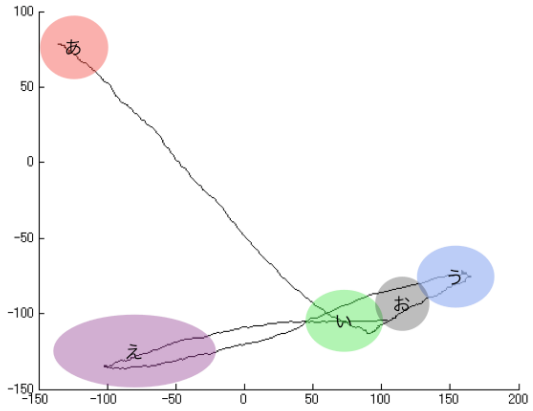


図 5　予備検討で用いたモデルによる「あいうえお」の遷移

間で占める凡その位置を、五母音に相当する手の姿勢を各々4 回ずつ計測した計 4 × 5 = 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。なお、このような分析は、データグローブのセンサーの配置や感度に大きく依存する。センシングの対象やその分解能が異なれば、当然、結果は異なる。

4.2　5 母音に対する 5 種類の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。即ち、28 種類のジェスチャーを PCA により 2 次元平面上に射影した。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 7 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。今回は煩雑さ

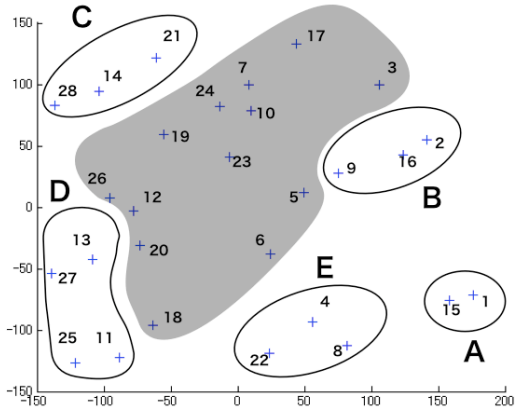


図 6　28 種類のジェスチャーの分類

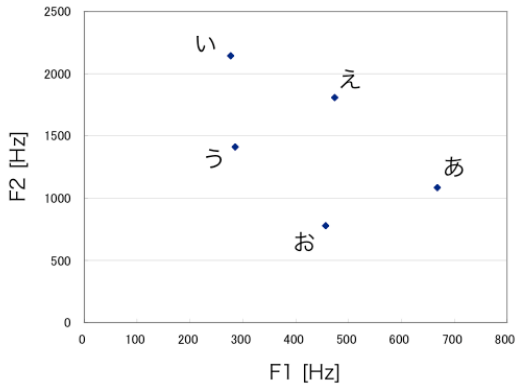


図 7　収録音声の母音図

表 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

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、その姿勢の形成が容易だと思われる 2 つを選んだ。また母音図の回転は考えないこととした。このようにして考えられる 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必要がある。16 通りの候補に対して写像関数を推定し、実際に合成音声を提示・聴取することで各候補の是非を検討することも可能であるが、ここでは写像関数の推定することなく、候補群の中から選定する簡便な方法を考える。

障害者支援の場合、楽器の製作とは異なり、使用者にどのような身体運動能力が残されているのか事前に知ることは困難である。また、最適なジェスチャーと母音の対応付けは、身体運動のセンサーの配置、分解能などにも依存する。これらのことを考慮すると、身体運動と音運動の対応付けを粗く見積もり、

その後、対応付けのより細かい修正については当事者の訓練を通して推定することが予想される。しかしこの場合、実際に合成音声の作成、聴取実験の遂行をその都度行なうことは現実解ではない。

表 1 に示されるジェスチャーと母音の対応付けは、ジェスチャー群の配置と母音群の配置との等価性を高めることを目的としているが、実際の連続発声となった場合、当然母音間遷移区間などは、図 4 や図 7 に示されたジェスチャーや母音のみならず、その中間的な値を呈することになる。これらの中間的な姿勢や母音をも考慮した、ジェスチャー群配置と母音群配置の比較を行なうことを目的として、本稿では構造的表象を用いることとした。

5. 構造的表象を用いた手姿勢の決定

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 [10]〜[12]。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（線形／非線形を問わない）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 [10], [13]。

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 8 に一発声を構造として表象する様子を図示する。構造表象は音声の実体を捨象し、音声の動きのみを話者不変的に表象する方法であり、具体的には事象間のコントラストだけを抽出する方法となっている。既に構造表象を用いた音声認識を検討しており、話者性に対して、極めて高い頑健性が示されている [14]。図 8 における分布数であるが、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うとすれば、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、両者の対応の善し悪しを検討することになる。

与えられた二構造間の類似性評価は、距離行列をベクトル化し（上三角行列の要素を取り出す）、二ベクトル間のユークリッド距離で行なう。この値は図 9 に示す様に、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。即ち、音声とその対応する手の運動の両者を、より細かな多角形で表現し、2 つの多角形を、その「形・配置」のみに着眼して比較することができる。

5.2 メディア普遍の情報の運動的・構造的表象

ここで、音声の構造表象と、手の動きの構造表象を比較することの意味について考察する。話者 A と話者 B の同一言語内容の発声を構造的に表象すれば、両発声の表象は凡そ同一となる。図 1 に示すような異なる物理メディア空間を写像で結びけ、両者における運動を構造的に表象した場合、当然、二つの構造的表象は数学的に同一となる。よって、各母音をどの手の姿勢に割付けようと、変換不変であるため、本来ならば、構造表象は変わらないはずである。しかし、ガウス分布を非線形変換すれば一般に非ガウス分布となるように、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(注3)。その結果、非線形写像によって構造は歪むことになる。本稿の場合、GMM による写像関数推定を混合数＝ 1 で行なえば、(5) 式は線形回帰写像となる。以下では、音声ストリームも手の動きも両方、ガウス分布系列としてモデル化し（図 8 参照）、構造ベクトルを作り、両構造を比較する。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ることを意味する。

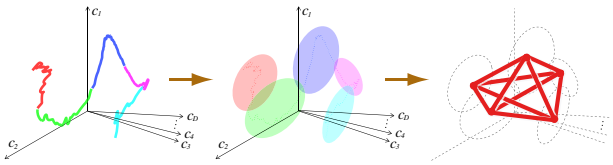


図 8　音と音のコントラストだけを用いた音声の構造的表象

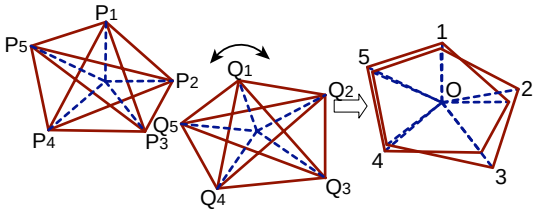


図 9　回転とシフトを通した構造の照合

5.3 準最適な手姿勢の設計

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 10 に示す。手の運動も音声も、ストリームから 25 個のガウス分布を推定して構造化し、構造ベクトルを求めている。なお、構造サイズを揃える処理を導入している（構造ベクトルのノルムを揃える処理に等しい）。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 であった。一方、距離が最も大きかったのは、No.14 であった。なお図 10 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。今回の検討では個々の対応付けによる差はそれほど大きくなかった。

No. 5 と、No.14 及び、予備実験で用いた対応付けとを比較するために、各対応付けを用いて写像関数を推定し、「あいうえお」とは異なるデータとして「いおあうえ」という音声を生成した。結果を図 11 に示す。(a) には比較として「いおあうえ」の分析再合成音を示してある。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

(注3)：音声認識応用の場合も、不変性が強すぎるため、異なる単語が同一視されることがある。これを回避するために、不変性を満たす写像関数に制約をかける必要が生じる。[14] ではストリーム分割を通して、この制約を実装している。

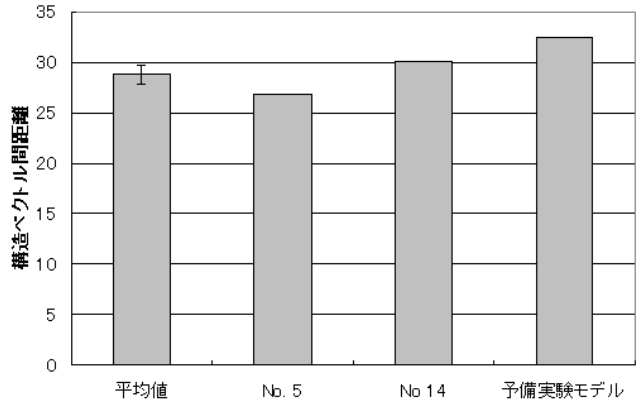
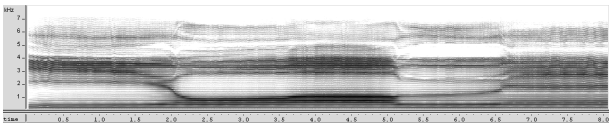
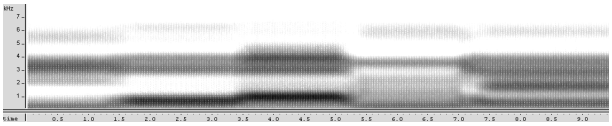


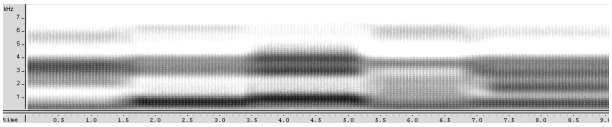
図 10　回転とシフトを通した構造の照合



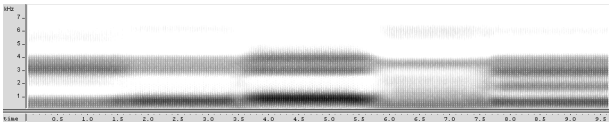
(a) 分析再合成音



(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

図 11 「いおあうえ」の遷移に対応する合成音の比較

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 5 人の聴取者を募り、いずれが最も不明瞭であるかを選ばせたところ、5 人とも予備実験のモデルを選択した。これは構造ベクトル間距離を求めることで導かれる結果と一致する。

以上より「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示された。

6. 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を構築した。その際、ジェスチャーと音の対応の「尤もらしさ」については、峯松らが提案した構造的表象 [10]〜[12] が有効であることを確認し、合成音の音質の改善に成功した。今後は図 4 や 7 に示されたものだけでなく、より多彩なジェスチャーや母音を考慮に入れ、またそれらを計測するセンサについても検討を進めていくつもりである。

文　献

[1] 株式会社ファンコム　携帯用会話補助装置レッツチャット http://www.funcom.co.jp/products/products-fc-lc12-menu.html

[2] 株式会社アルカディア　ボイスエイド http://www.arcadia.co.jp/VOCA

[3] 畠山卓朗, “コミュニケーション支援の現状と課題　ーすべては気づきからー”, 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

[4] 藪 謙一郎 他, “発話障害者支援のための音声生成器ーその研究アプローチと設計概念”, 電子情報通信学会技術研究報告 Vol.106 No.613 pp.25-30, 2007

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

[6] 市川 熹、手嶋 教之,“福祉と情報技術”, pp.169-174, オーム社, 2006

[7] Y. Stylianou *et.al.*, “Continuous probablistic transform for voice conversion,” IEEE Trans. Speech Audio Process., vol.6, pp.131-142, 1998

[8] H. Kawahara *et.al.* “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction,” Speech Commun., 27, 187-207 (1999)

[9] 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

[10] 峯松 信明 他, ”線形・非線形変換不変の構造的情報表象とそれに基づく音声の音響モデリングに関する理論的考察 ”, 日本音響学会春季講演論文集, 1-P-12, pp.147-148, 2007

[11] N. Minematsu, “Mathematical evidence of the acoustic universal structure in speech,” Proc. ICASSP, pp.889-892, 2005.

[12] N. Minematsu, *et.al.* , “Theorem of the invariant structure and its derivation of speech Gestalt,” Proc. SRIV, pp.47-52, 2006

[13] Y. Qiao, *et.al.*, “Structual representation with a general form of invariant divergence”, 日本音響学会秋季講演論文集, 2-P-2, pp.105-108, 2007

[14] 朝川 智 他, “判別分析と構造表象を用いた話者の多様性に超頑健な音声認識”, 日本音響学会秋季講演論文集, 2-P-2, pp.113-116, 2007