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.

手から声のメディア変換モデルと手のジェスチャーモデルの確率的統合に基づく異メディア空間の対応付けの検討

國越　晶† 齋藤　大輔† 喬　宇†† 峯松　信明††† 広瀬　啓吉†††

† 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
†† 中国科学院深セン先進技術研究院　中国広東省深セン市南山区西麗区深セン大学城学苑大道 1068 号
††† 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,dsk_saito,mine,hirose}@gavo.t.u-tokyo.ac.jp, yu.qiao@sub.siat.ac.cn

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している．近年、二話者から与えられた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．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、手姿勢（ジェスチャー）を入力とした日本語五母音の連続音声生成において、本手法が有効であることを報告した．本稿では、母音のみのパラレルデータを用いて音声–ジェスチャー変換システムを構築し、それに子音音声を入力することにより、子音に割り当てるジェスチャーを推定する手法を検討した．

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

Gesture Design of Hand-to-Speech Conversion Based on Probabilistic Integration of Speech-to-Hand Joint Density Model and Hand Gesture Model

A. KUNIKOSHI†, D. SAITO†, Y. QIAO††, N. MINEMATSU†††, and K. HIROSE†††

† Grad. School of Eng., The Univ. of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 113-8656, Japan
†† Shenzhen Inst. of Advanced Tech., 1068 Xueyuan Ave., Shenzhen Univ. Town, Shenzhen, P.R.China
†††Grad. School of Info. Sci. and Tech., The Univ. of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 113-0033,Japan
E-mail: {kunikoshi,dsk_saito,mine,hirose}@gavo.t.u-tokyo.ac.jp, yu.qiao@sub.siat.ac.cn

Abstract　When individuals with speaking disabilities, dysarthrics, try to communicate using speech, they often have to 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 study, we are developing a new and novel speech synthesizer where not symbol inputs but hand motions are used to generate speech. In recent years, statistical mapping techniques have been proposed for voice conversion. Based on these methods, we developed a system to convert hand motions to vowel transitions by finding the mapping between a hand space and a vowel space. We found that the proposed method was effective to generate utterances of Japanese five vowels. In this paper, we discuss how to extend this system for consonant generation. We develop a Speech-to-Hand conversion system trained from parallel data for vowels only to infer the gestures corresponding to consonants.

Key words　Dysarthria, speech production, hand motions, media conversion, arrangement of gestures and vowels

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．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 [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極めて多様である [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、ある話者の音響空間と別話者の音響空間との間の写像を推定し、これを用いて入力音声の話者性を変換する技術が提案されている [7], [8], [9], [10]．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．

これまでに、手姿勢（ジェスチャー）を入力とした日本語五母音の連続音声生成において、本手法が有効であることを報告した [11]．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 [12]．一方、子音に適当な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拡張した場合、合成音において遷移部分が適切に知覚されないなどの問題が指摘されている [13]．その原因として、ジェスチャー空間における母音と子音に対応するジェスチャーの位置関係が、音響空間における母音と子音の位置関係に対応していないこと、また静的な位置関係が適切に対応づけられた場合でも、ジェスチャーと音声の動的な軌跡において適切な対応付けが取れていない可能性があるといった点が挙げられる．それらを回避するため、本稿では、母音のみのパラレルデータを用いて音声–ジェスチャー変換システム（Speech-to-Hand system, 以下 S2H システム）を構築し、それに子音音声を入力することにより、子音に相当するジェスチャーを推定する手法を検討した．

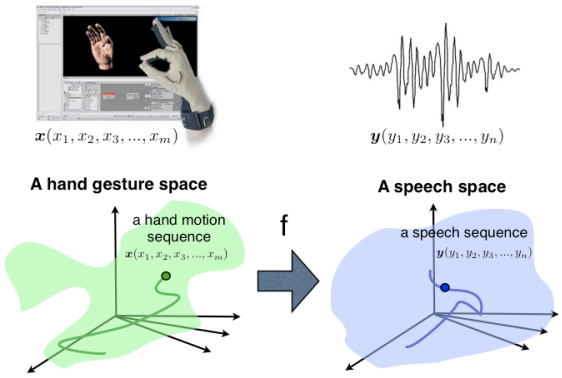


図 1　空間写像に基づくメディア変換の枠組み

次章以降の構成は次の通りである．まず 2 章で、空間写像に基づくメディア変換の枠組みについて述べる．次に 3 章で、提案手法に基づき、日本語五母音連続音声を対象に、手の動きを入力とした音声生成系を構築する．4 章では、提案手法の枠組みにおいて子音に適切なジェスチャーを割り当てるため、確率的統合モデルを導入する方法について述べる．そして 5 章で、4 章で述べた手法を実験的に検証する．最後に 6 章で、まとめと今後の課題を述べる．

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．ある時刻のジェスチャーが m 次元の特徴量ベクトル $\boldsymbol{x}_t$ 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 $\boldsymbol{y}_t$ 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．

[7], [8], [9], [10] は、ある話者の音響特徴量空間から別の話者の音響特徴量空間への空間写像を設計することで、声質変換を実現する手法を提案している．本研究が目指すジェスチャー–音声変換システム（Hand-to-Speech system, 以下 H2S システム）は、それらの手法において、入力話者の音響空間をジェスチャー空間に置き換えたものと考えることができる．そこでジェスチャー空間と音響特徴量空間における空間写像を、Stylianou らの手法 [7] に基づき、次のように推定する．

まず対応関係の判っているジェスチャー空間のベクトル（以下、ジェスチャーベクトル）$\boldsymbol{x}$ と音響特徴量空間のベクトル（以下、音響特徴量ベクトル）$\boldsymbol{y}$ から、フレーム毎に結合ベクトル $\boldsymbol{z} = [\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top$ をつくる．この特徴量系列を用いて、以下の式で表される GMM のパラメータを推定し、$\boldsymbol{z}_t$ の確率密度をモデル化する．

$$P(\boldsymbol{z}|\lambda^{(z)}) = \sum_{m=1}^{M} \omega_m \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}) \tag{1}$$

ここで $\mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)})$ は平均 $\boldsymbol{\mu}_m^{(z)}$, 分散 $\boldsymbol{\Sigma}_m^{(z)}$ の正規分布を表し、M は混合数、ω_m は重みを表す. $\boldsymbol{\lambda}^{(z)}$ は結合ベクトルの GMM のモデルパラメータであり、以下のように表される.

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix} \tag{2}$$

ただし、$\boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)}, \boldsymbol{\mu}_m^{(y)}, \boldsymbol{\Sigma}_m^{(yy)}$ は m 番目の正規分布における、ジェスチャーベクトルおよび音響特徴量ベクトルの平均ベクトルおよび分散共分散行列である. また $\boldsymbol{\Sigma}_m^{(xy)}$ および $\boldsymbol{\Sigma}_m^{(yx)}$ は、入出力空間間の相互共分散行列を表す.

変換写像 $\mathcal{F}(\cdot)$ は入力ベクトル $\boldsymbol{x}_t$ が与えられた場合の $\boldsymbol{y}_t$ の条件付き確率密度に基づいて導出することができる. この確率密度は上述の GMM のモデルパラメータ $\boldsymbol{\lambda}$ によって、以下のように表現される.

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(x)}) = \sum_{m=1}^{M} P(m|\boldsymbol{x}_t, \lambda^{(z)}) P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}) \tag{3}$$

ここで

$$\begin{aligned} P(m|\boldsymbol{x}_t, \lambda^{(z)}) &= \frac{\omega_m \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)})}{\sum_{m=1}^{M} \omega_m \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)})} \\ P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}) &= \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{E}_{m,t}^{(y)}, \boldsymbol{D}_m^{y}) \\ \boldsymbol{E}_{m,t}^{(y)} &= \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}) \\ \boldsymbol{D}_m^{(y)} &= \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} \boldsymbol{\Sigma}_m^{(xy)} \end{aligned} \tag{4}$$

となる. 最小平均二乗誤差基準に基づく変換関数は以下のように表される.

$$\mathcal{F}(\boldsymbol{x}_t) = \sum_{m=1}^{M} P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) \boldsymbol{E}_{m,t}^{(y)} \tag{5}$$

3. 日本語五母音の合成 [11] [12]

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 [11], [12]. 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.

本稿ではジェスチャーの候補には、Wu らが画像認識における論文 [14] の中で用いた、基本的な 28 個のジェスチャーを参考にして選んだ. 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. [12] 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. そこで「ジェスチャー空間中のジェスチャー群配置」を確かめることを目的として、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. 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、図に示すようにおよそ 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.

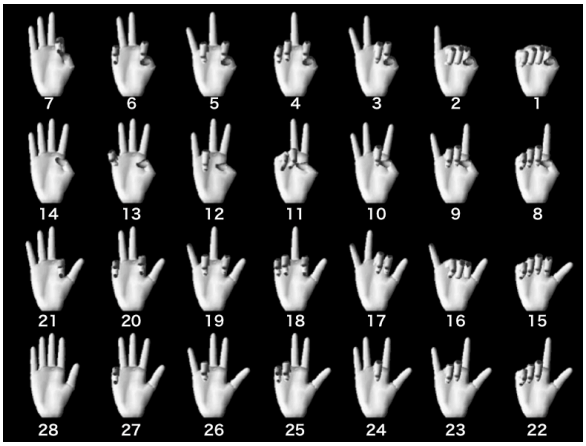
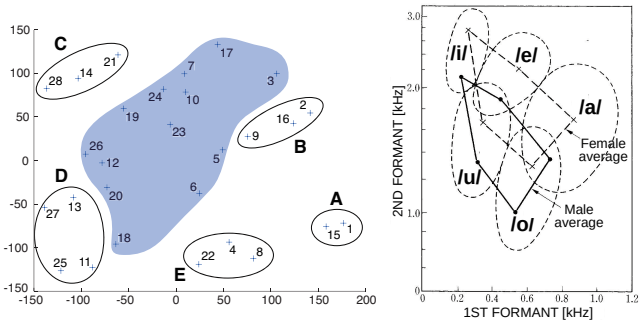


図 2　基本的な 28 種類のジェスチャー



(a)　(b)

図 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

図 4　日本語五母音に相当するジェスチャー

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、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．データグローブは、各関節に取り付けられたセンサーにより、第 2〜4 指の DIP 関節を除く 18 個の関節の曲げ角度を、それぞれ 8bit の値として出力するものである．サンプリング周期は 10〜20 ms である(注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 [15] 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．

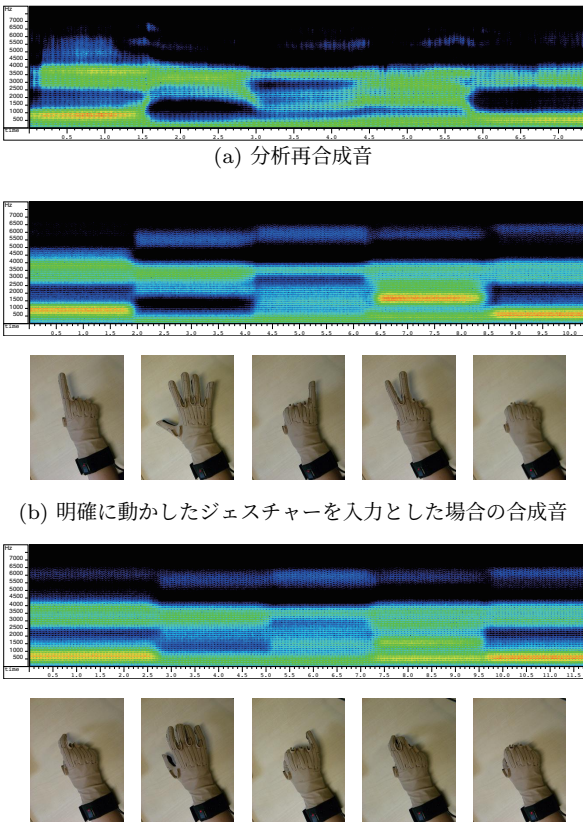
「あいうえお」に対して提案手法により生成した結果を図 5 に示す．(a) は「あいうえお」の分析再合成音、(b) は明確に動かしたジェスチャーを入力とした場合の合成音、(c) は不明確に動かしたジェスチャーを入力とした場合の合成音の例である．ケプストラム時系列からの再合成には STRAIGHT を用い、F0 はすべて 140 Hz とした．予備的な聴取実験により、日本語五母音による連結母音音声を対象とした合成において、この手法の有効性が示された．また手の動かし方（明確／不明確）によって、発話スタイル（明瞭／不明瞭）も制御できることが分かる．

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を示した．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．

本システムに子音を導入する場合、子音に割り当てるジェスチャーの決定が問題になる．子音に適当な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拡張した場合、合成音において遷移部分が適切に知覚されないなどの問題が指摘されている [13]．その原因として、ジェスチャー空間における母音と子音に対応するジェスチャーの位置関係が、音響空間における母音と子音の位置関係に対応していないこと、また静的な位置関係が適切に対応づけられた場合でも、ジェスチャーと音声の動的な軌跡において適切な対応付けが取れていない可能性があるといった点が挙げられる．本稿ではそれらを回避するため、母音のみのパラレルデータを用いて S2H システムを構築し、それに子音音声を入力することにより、子音に対応するジェスチャーを推定する手法を検討した．

(注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

図 5 「あいうえお」に対する合成音の比較

確率的な変換モデル $P(\boldsymbol{y}|\boldsymbol{x})$ の、y に関する最大化問題に対して、ベイズの法則より $P(\boldsymbol{y}|\boldsymbol{x})$ を $P(\boldsymbol{x}|\boldsymbol{y})P(\boldsymbol{y})$ に変換し、これを最大化する問題として解くことが、統計翻訳の世界で広く行われている [16]．$\boldsymbol{x}$ = 日本語、$\boldsymbol{y}$ = 英語として, $P(\boldsymbol{y}|\boldsymbol{x})$ を直接モデル化、最大化するためには大量のパラレルデータが必要となるが、これを、$P(\boldsymbol{x}|\boldsymbol{y})P(\boldsymbol{y})$ とすれば、$P(\boldsymbol{x}|\boldsymbol{y})$ 推定用のパラレルデータがたとえ十分になくても、大量な英語コーパスより得られる精度の高い $P(\boldsymbol{y})$ により、結果的に、品質の高い翻訳が可能になっている．この枠組みは、声質変換のタスクにおいても適用され、少量のパラレルデータから推定される変換モデルと、大量の目標話者データより構成される目標話者モデルを用いた声質変換が検討されている [17]．

我々の目的は、どの音声をどのジェスチャーに対応させるかを求めることにある．そこで、本来の目的であるジェスチャー $(\boldsymbol{x})$ から音声 $(\boldsymbol{y})$ への統計的変換モデル $P(\boldsymbol{y}|\boldsymbol{x})$ ではなく、その逆の変換モデル $P(\boldsymbol{x}|\boldsymbol{y})$、S2H を考え、これにベイズ則を適用することを考える．すなわち、$\underset{x}{\operatorname{argmax}}\, P(\boldsymbol{x}|\boldsymbol{y}) = \underset{x}{\operatorname{argmax}}\, P(\boldsymbol{y}|\boldsymbol{x})P(\boldsymbol{x})$ より、音声 $\boldsymbol{y}$ に対するジェスチャーを求めることを考える．$P(\boldsymbol{y}|\boldsymbol{x})$ は母音のみからなるパラレルデータより構成された変換モデル（先行研究で構築済み）であり、$P(\boldsymbol{x})$ は大量のジェスチャーデータから推定されるジェスチャーの統計モデルである．このようにして得られた統計モデルに対して、子音を入力した場合に得られるジェスチャーを検討し、音声とジェスチャーとの対応を母音以外にも拡張する．$P(\boldsymbol{y}|\boldsymbol{x})$ のみでジェスチャーを検討すれば、不自然なジェスチャーが推定されることが予想さ

れるが、$P(\boldsymbol{x})$ により、ジェスチャーとしての自然性が考慮され、より適切なジェスチャーが得られると期待される．

5. 実　　験

5.1 音声からジェスチャーへの変換

S2H システムにおいて、パラレルデータセットに含まれていない音に相当するジェスチャーも適切に推定されることを確認するため、以下の通り実験を行った．まず母音に相当するジェスチャーデザインを設定し、母音に相当するパラレルデータとジェスチャーモデルのみを用いて、S2H システムを構築した．そのシステムに、パラレルデータにない子音音声を入力することにより、子音に対応するジェスチャーを推定した．

まず学習データとして、データグローブを装着し、図 3(a) の中央の領域を除いた 15 種類のジェスチャーに、中央の領域で比較的容易であった No.7 を加えた合計 16 種類のジェスチャー、及びそのうち二ジェスチャー間の遷移、$16 + {}_{16}P_2 = 256$ 個のジェスチャーを記録し、これを用いてジェスチャーモデル $P(\boldsymbol{x})$（混合数 64）を構築した．

次に日本語五母音に対応するジェスチャーの候補を考え、$P(\boldsymbol{x}|\boldsymbol{y})$ を構築する．計算の便宜上、「あ」は No.28 とした．「い」「う」「え」「お」に対応するジェスチャーの組み合わせのうち、3 章の議論から母音図との等価性に配慮し、ジェスチャー空間内のユークリッド距離において、あ–い間の距離が、あ–え間よりも短いもの、及び、あ–う間の距離が、あ–お間の距離よりも短いものは候補外とした．これにより五母音に相当するジェスチャーデザインの候補は、8190 通りとなった．各ジェスチャーデザインに対し、以下のように音声からジェスチャーへのメディア変換を実装した．

学習データとして、上記のジェスチャーデータセットから、それぞれのジェスチャーデザインにおいて「あ」「い」「う」「え」「お」および二母音間の遷移に相当する ${}_5P_2 = 20$ 個のデータを抽出し、学習データに用いた．また成人男性 1 名から「あ」「い」「う」「え」「お」および二母音間の遷移 ${}_5P_2 = 20$ 組、計 $5 + 20 = 25$ 個の音声データを収録した．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 次の抽出を行い結合ベクトルを作った．これらの結合ベクトルを用いて変換モデル（混合数 8）を学習した．このようにして作られた 8190 通りの S2H システムに、子音として、な行、ま行、ら行、ば行、ぱ行の音を入力し、/n/, /m/, /r/, /p/, /b/に対応するジェスチャーを推定した．すなわち計 8190 通りの/a/, /i/, /u/, /e/, /o/, /n/, /m/, /r/, /p/, /b/ のジェスチャーデザインが定義されることになる．得られたジェスチャーデザインの例を図 6, 7 に示す．

5.2 ジェスチャーから音声への変換

次に、前節で得られた 8190 通りのジェスチャーデザインの比較および有効性の検証を目的に、以下の通り実験を行った．まずそれぞれのジェスチャーデザインにおいて、4 章で述べた手法を用いて H2S システムを構築した．すなわち、$\underset{y}{\operatorname{argmax}}\, P(\boldsymbol{y}|\boldsymbol{x}) = \underset{y}{\operatorname{argmax}}\, P(\boldsymbol{x}|\boldsymbol{y})P(\boldsymbol{y})$ により、ジェスチャー

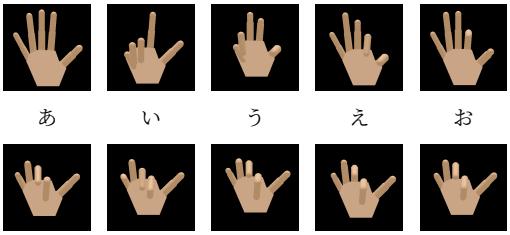


図 6　S2H システムを構築した日本語五母音に相当するジェスチャーと、それによって推定された子音に相当するジェスチャーの例 1

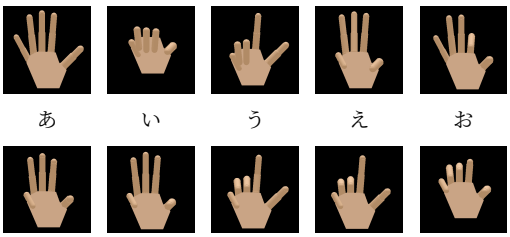


図 7　S2H システムを構築した日本語五母音に相当するジェスチャーと、それによって推定された子音に相当するジェスチャーの例 2

$\boldsymbol{x}$ に対する音声を求めるシステムである．変換モデル $P(\boldsymbol{x}|\boldsymbol{y})$ は、前節で構築した S2H システムと同様の学習データ／混合数で学習した．話者モデル $P(\boldsymbol{y})$ は、同一話者から収録した ATR 音素バランス 503 文の A セット 50 文で学習した．混合数はジェスチャーモデルと同様に 64 とした．

このようにして構築された H2S システムに、前節で構築した S2H システムによって推定されたジェスチャーを入力する．構築された両システムが理想的ならば、S2H システムに入力した音声と、S2H システムの出力ジェスチャーから H2S システムによって推定される音声は、同一のものとなるはずである．しかし実際には、変換モデル $P(\boldsymbol{x}|\boldsymbol{y})$ や $P(\boldsymbol{y}|\boldsymbol{x})$ が 2 つの空間の特徴量を完全に対応づけていないことなどから歪みが生じる．本稿ではこの歪みを、ジェスチャーデザインの評価指標とした．すなわち、S2H システムに入力した音声と、その音声から S2H システムによって推定されるジェスチャーを、H2S システムに入力した場合に推定される音声とのケプストラム平均自乗距離が近いものほど、よりよい変換を実現するジェスチャーデザインと判断した．

8190 通りのジェスチャーデザインにおけるケプストラム平均自乗誤差の平均と標準偏差を図 8 に示す．計算の都合上、子音として/n/のみに注目した．S2H システムに入力した音声は、学習データ内の「あ」「い」「う」「え」「お」、および同一話者から録音した「な」「に」「ぬ」「ね」「の」の再合成音、合計 $5 + 5 = 10$ 個である．8192 通りのジェスチャーデザインそれぞれにおいて、各モーラごと 10 個のケプストラム平均自乗誤差が求められることになる．ここで「な」「に」「ぬ」「ね」「の」各モーラごとに 8190 通りのジェスチャーデザインに順位をつける．この 5 つの順位の合計が最も小さかった準最適なデザインは、「あ」が No.28、「い」が No.22,、「う」が No.11、「え」が No.7、「お」が No.21 の場合であった．図 8 には、準最適な

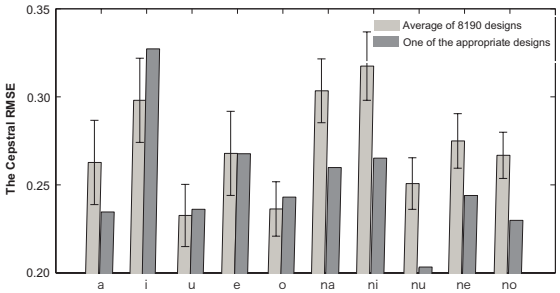
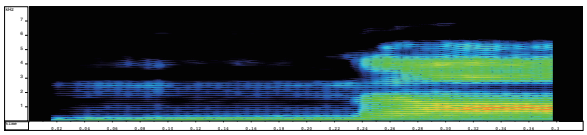
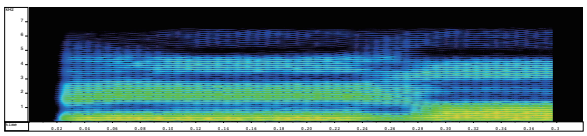


図 8　S2H システムに入力した音声と、その音声から S2H システムによって推定されるジェスチャーを、H2S システムに入力した場合に推定される合成音のケプストラム平均自乗誤差



(a) S2H システムに入力した分析再合成音



(b) S2H システムの出力を、H2S システムに入力して得られた合成音

図 9 「な」に対応する音声

デザインにおけるケプストラム平均自乗距離も示す．提案手法によって構築された S2H および H2S システムでは、文字ごとにケプストラム平均自乗誤差が異なる傾向が見られた．日本語五母音のうち、最もケプストラム平均自乗誤差の小さかったものは「う」であり、最も大きかったものは「い」であった．8190 通り全体の平均では、学習データに含まれる母音に比べ、学習データに含まれていない子音ではケプストラム平均自乗誤差は大きくなる傾向がある．一方、準最適なデザインでは、学習データに含まれていない子音が学習データに含まれている母音とほぼ同程度の音質を達成していることがわかる．準最適なデザインにおける「な」に対応する分析再合成音と、S2H および H2S システムによって生成された子音音声の例を図 9 に示す(注2)．[13] は、適当なジェスチャーを子音に割り当てた際に、子音から母音への遷移部分が正しく知覚されない点を指摘している．一方、本手法によって合成された子音音声は、予備的な聴取実験において、子音を含んだ 1 モーラ音声として知覚されることが確認された．しかしながら、図 6, 7 に示されるように、/n/と/m/, /p/と/b/など、異なる子音に対し互いに類似したジェスチャーが推定されることがあった．これは本システムにおいて、口唇や鼻腔の動きを考慮していないためと考えられる．今後、これらのパラメータを本システムに導入する方法について検討する予定である．

(注2)：スペクトログラムは、推定されたケプストラム系列に平滑化処理を施してから可視化を行っている．

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において、子音に相当するジェスチャーを理論的に推定する一つの枠組みを示した．提案手法で用いた、母音音声のみのパラレルデータを用いて構築した S2H モデルは、パラレルデータに含まれない子音に相当するジェスチャーを生成するのに有効である可能性が示唆された．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．

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] Glove-Talk II 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] A. Kain and M.W.Macon, “Spectral voice conversion for text-to-speech synthesis,” Proc. ICASSP1998, vol.1, pp.285–288, 1998.
[9] H. Zen *et al.*, “Reformulating the HMM as a Trajectory Model by Imposing Explicit Relationships between static and Kynamic Feature Vector Squnces,” Computer Speech & Language, vol. 21, no. 1, pp.153–173, 2007.
[10] T. Toda *et al.*, “Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory,” IEEE Trans. Audio Speech Language Proc., vol.15, pp.2222–2235, Nov. 2007.
[11] 國越 晶 他, “空間写像に基づく手の動きを入力とした音声生成系 ”, 日本音響学会春季講演論文集, 1–Q–23, pp.375–376, 2008
[12] A. Kunikoshi *et al.*, “Speech generation from hand gestures based on space mapping,” Proc. INTERSPEECH, pp.308–311, 2009
[13] 國越 晶他, “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 ”, 日本音響学会春季講演論文集, 1–P–8, pp.419–422, 2010
[14] Y. Wu *et al.*, “Analyzing and Capturing Articulated Hand Motion in Image Sequences,” IEEE Trans. PAMI., vol.27, No.12, pp.1910–1922, 2005
[15] 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
[16] P. Brown *et al.*, “A statistical approach to machine translation,” Computational Linguistics Vol.16, No.2, pp.79–85, 1990
[17] 齋藤 大輔他, “変換モデルと話者モデルの確率的統合に基づく声質変換法の検討”, 日本音響学会秋季講演論文集, 3–P–4, pp.335–338, 2010