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*

☆國越晶, 齋藤大輔 (東大), 喬宇 (深セン先進院), 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、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 [1]. 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.

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、与えられた二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）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ている [6, 7]. 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 [8]. 本稿では、学習データに子音を含まない、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、子音に割り当てるジェスチャーを推定する手法を検討した.

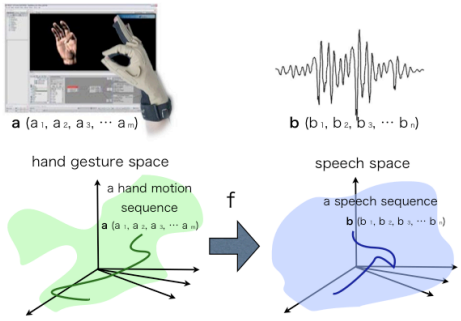


Fig. 1 空間写像に基づくメディア変換の枠組

2 空間写像に基づくメディア変換

空間写像に基づくメディア変換の枠組を Fig.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.

Stylianou らは、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 [5]. 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、Stylianou 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. そして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、[5] の手法に基づき、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 $\boldsymbol{x}$ と音響特徴料空間のベクトル $\boldsymbol{y}$ 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。

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

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。

---

*Gesture design of Hand-to-Speech Conversion based on Probabilistic Integration of Speech-to-Hand Joint Density Model and Hand Gesture model. by KUNIKOSHI Aki, SAITO Daisuke (The University of Tokyo), QIAO Yu (Shenzhen Institutes of Advanced Technology), MINEMATSU Nobuaki, HIROSE Keikichi (The University of Tokyo)

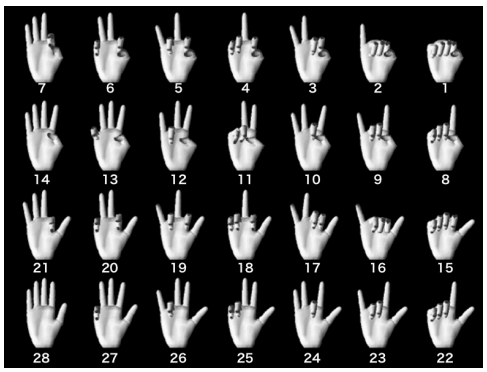


Fig. 2　基本的な 28 種類のジェスチャー [9]

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。

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t,\boldsymbol{\lambda}^{(x)})=\sum_{m=1}^{M}P(m|\boldsymbol{x}_t,\lambda^{(z)})P(\boldsymbol{y}_t|\boldsymbol{x}_t,m,\boldsymbol{\lambda}^{(z)}) \quad (2)$$

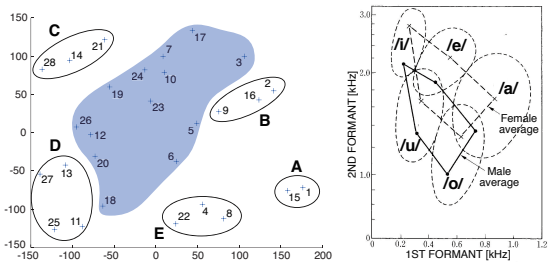
写像 $\mathcal{F}(\cdot)$ は最小平均二乗誤差基準に基づいて求める。

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. 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の対応付けが考えられるため、その中から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として、Wu らが画像認識における論文 [9] 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 Fig.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、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 Fig.3(a) に示す. 数字は Fig.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、Fig.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〜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「あ」「い」「う」「え」「お」に対応するジェスチャーを、それぞれ、No.8, No.28, No.2, No.11, No.1 とした.



(a)　(b)

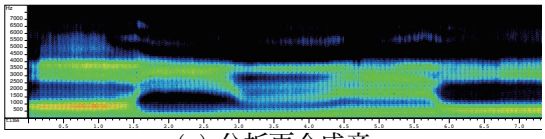
Fig. 3　(a) PCA 空間における 28 ジェスチャーの位置 (b) 日本語五母音の F1/F2

3.2　メディア変換用写像関数の推定

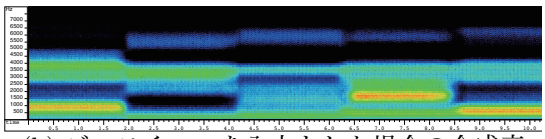
次に学習データとして、データグローブを装着し「あ」「い」「う」「え」「お」および二母音間の遷移 ${}_5P_2 = 20$ 組を各々3 回、計 $(5+20)\times 3 = 75$ 個のデータを記録した. センサの数は 18 個, サンプリング周期は 10〜20 ms である. また成人男性 1 名から収録した「あ」「い」「う」「え」「お」および二母音間の遷移 ${}_5P_2 = 20$ 組を各々5 回、計 $(5+20)\times 5 = 125$ 個の音声データから、STRAIGHT[10] を用いて分析をおこない、ケプストラム係数 0-17 次を抽出した. フレーム長は 40 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 Fig.4 に示す。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.

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



(a) 分析再合成音



(b) ジェスチャーを入力とした場合の合成音

Fig. 4　「あいうえお」に対する合成音の比較

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 [8]．その原因としては、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、学習においてジェスチャーの遷移部分と音声の遷移部分を適切に対応づけることが難しいことなどが挙げられる．本稿ではそれらを回避するため、学習データに子音を含まない、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、子音に割り当てるジェスチャーを推定する手法を検討した．

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 [11]．$\boldsymbol{x}$ = 日本語、$\boldsymbol{y}$ = 英語として, $P(\boldsymbol{y}|\boldsymbol{x})$ を直接モデル化、最大化すると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、$P(\boldsymbol{x}|\boldsymbol{y})P(\boldsymbol{y})$ 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 [12]．

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、$P(\boldsymbol{x})$ は大量の手形データから推定される手形の統計モデルである．この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手形を検討し、音声と手形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．

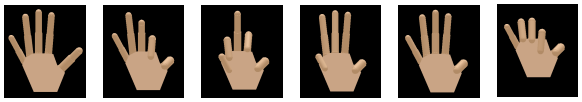
5　実験

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、パラレルデータにないことが予想される子音に相当するジェスチャーを推定した．

まず学習データとして、Fig.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る．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．これらのジェスチャーデザインに対し、以下のようにそれぞれ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、線形補完を行って結合ベクトルを作った．これを用いて変換モデル（混合数 8）を学習した．



あ　い　う　え　お　/n/

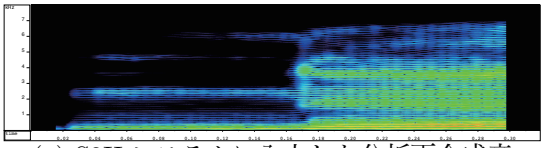
Fig. 5　有効なジェスチャーデザインとそれによって生成された/n/に相当するジェスチャー

このようにして作られた 8190 通りの S2H システムに、「な」「に」「ぬ」「ね」「の」を入力してジェスチャーを推定した．データグローブのデータは、各関節に取り付けられたセンサーにより、第 2〜4 指 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。グローブのはめ直しや機器の再起動などにより、第 1 指 MP 関節および IP 関節、第 2〜4 指 PIP 関節および DIP 関節それぞれにおいて、測定値に ±10 程度の誤差が現れることが確認されている．そこで、S2H システムで得られたジェスチャーデータのうち、学習に用いたジェスチャーデータのレンジ ±20 の範囲に入らないものは、形成不可能と判断した．ただし形成不可能なフレーム数が、全フレーム数の 5%以内であるデータは許容した．「な」「に」「ぬ」「ね」「の」に対応するすべてのジェスチャーが形成可能であったのは、8190 個中 3 個のジェスチャーデザインだけであった．有効であったジェスチャーデザインの一例を Fig.5 に示す．S2H モデルにより推定された、/n/に相当するジェスチャーも示した．

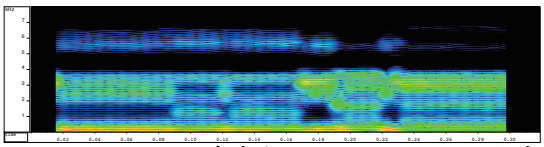
5.2　ジェスチャーから音声への変換

前節で得られた 3 つのジェスチャーデザインに対し、母音のみのパラレルデータを用いて、従来手法により H2S システムを構築した．学習データとして、データグローブを装着して「あ」「い」「う」「え」「お」および二母音間の遷移に相当するデータを 2 セット、合計 $(5 + {}_5P_2) \times 2 = 50$ 個を記録した．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を 2 セット、計 $(5 + {}_5P_2) \times 2 = 50$ 個の音声データを収録した．そして 5.1 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、および線形補完を行い、結合ベクトルを作った．この際、各結合ベクトルにおいて前フレームからの変位が一定値以下になるものは除去し、全体のフレーム数が元の結合ベクトル全フレーム数の 5 分の 1 程度になるようにした．これを用いて、変換モデル（混合数 32）を学習した．

このようにして構築された H2S システムに、S2H システムによって子音音声から推定されたジェスチャーを入力する．H2S システムに入力した音声は、学習に用いたのと同一話者から録音した、な行、ま行、ら行、ぱ行、ぱ行の各文字の再合成音、合計 $5 \times 5 = 25$ 個である．生成された子音音声の例を Fig.6 に示す．



(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

Fig. 6　「ぬ」に対応する音声

[8] 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、ま行／な行／ら行の音、ぱ行／ば行の音の識別が難しいことも指摘された．これは、/m/, /n/, /r/に相当するジェスチャー、/p/, /b/に相当するジェスチャーが互いに近似しているためと思われる．このことから、今後、これらの識別に重要な、口唇や鼻腔の動きを考慮する必要性が示唆された．

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．

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, pp.12–13 (2007)

[2] 藪他, 信学技報, SP2006–164, pp.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, pp.131-142 (1998)

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

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

[8] 國越他, 音講論（春）, 1–P–8, pp.419–422 (2010)

[9] Y. Wu et al., IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.27, No.12, pp.1910–1922 (2005)

[10] 河原, 増田, 信学技報, EA96–28, pp.9–16 (1996)

[11] P. Brown et al., Computational Linguistics Vol.16, No.2, pp.79–85 (1990)

[12] 齋藤他, 音講論（秋）, 3–P–4, pp.335–338 (2010)