手の動きを入力とした音声生成系における有声／無声の明確化に関する検討*

☆國越晶 (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

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 [1] 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 [2] 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 [3]．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 [4]．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 [5, 6]．また、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（目的とするジェスチャー → 音声変換システムの逆システム）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提案した [7]．子音を含むジェスチャーデザインにおいて、提案手法は従来手法と比べ、より自然な音声を生成することが主観評価実験によって 確認されている．一方で、提案手法では、/b/と/p/などに類似したジェスチャーが割り当てられてしまう点も指摘されていた．これは音声の特徴量として有声度が考慮されていないため、GMM の学習において有声／無声が明確に区別されていないためであると考えられる．本稿では音声の特徴量として、ケプストラムのみならず、F0 および正規化相関係数を考慮することで、有声／無声に相当するジェスチャーの明確化を検討した．

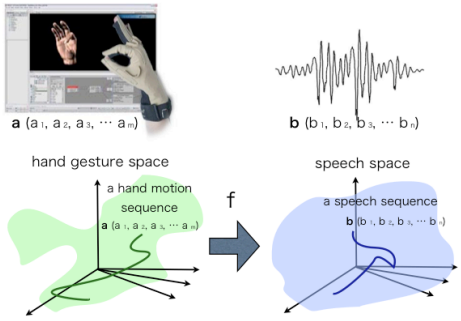


Fig. 1　空間写像に基づくメディア変換の枠組

2　空間写像に基づくメディア変換

2.1　方針

空間写像に基づくメディア変換の枠組を Fig.1 に示す．時刻 t 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、時刻 t 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 [4]．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られる．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、[4] 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．

*Differentiation of Voiced/Unvoiced Gestures for Hand-to-Speech System. by KUNIKOSHI Aki (The University of Tokyo), QIAO Yu (Shenzhen Institutes of Advanced Technology), MINEMATSU Nobuaki, HIROSE Keikichi (The University of Tokyo)

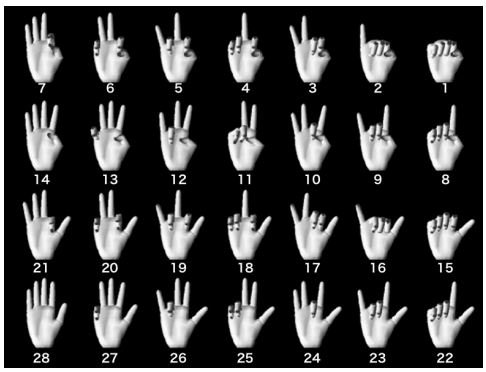


Fig. 2　基本的な 28 種類のジェスチャー [8]

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．

2.2　ジェスチャー候補

前節で述べた枠組みでは、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．

Wu らは画像認識における論文 [8] の中で、基本的な 28 個のジェスチャーを挙げている（Fig.2）．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．本稿ではこのうち、操作者である成人女性にとって比較的形成の容易であった No.1, 2, 4, 7, 8, 9, 11, 13, 14, 15, 16, 21, 22, 25, 27, 28 のジェスチャー計 16 個を、ジェスチャー候補として用いる．

3　子音の合成 [7]

3.1　手から声のメディア変換モデルとジェスチャーモデルの確率的統合に基づく子音ジェスチャーの推定

これまでに、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 [5, 6]．その際、[6] 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．

一方、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、2 章で述べた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 [9]．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、母音のみのパラレルデータを用いて S2H システム（目的とする H2S システムの逆システム）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提案している．

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．

3.2　音声からジェスチャーへの変換

前節で述べた手法を応用することにより、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実験的に検証した．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、成人女性 1 名が Immersion 製データグローブ CyberGlove を装着し、2.2 で設定した 16 種類のジェスチャー候補、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た．サンプリング周期は 10〜20 ms である[1]．各関節に取り付けられたセンサーは、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

[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．

8bit の値として出力する．指の曲げ角度はそれぞれ独立ではないから（たとえば第 5 指の PIP 関節を曲げると、MP 関節も曲がる）、データグローブによって記録された 18 次元データの各次元は互いに高い相関がある．そこで直交性の高いケプストラム空間との等価性を高めることを目的に、記録した全てのデータを用いて各データに対し主成分分析（以下 PCA）を行った．PCA を行った後の 18 次元データをジェスチャーの特徴量とし、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設定し、$P(\boldsymbol{x}|\boldsymbol{y})$ を構築する．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、「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、STRAIGHT[10] を用いて分析をおこない、ケプストラム係数 1-18 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．結合ベクトルを作る際、ジェスチャーデータ時系列はケプストラム時系列の時間長／周期に合わせて、線形補間した．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 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．母音に対応するジェスチャーデザインと提案手法によって得られた子音のジェスチャーデザインの例を Fig.3 に示す．

[7] 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、提案手法によって生成されたジェスチャーデザインは、1 モーラ音声として知覚される子音音声を合成することが確認された．しかしながら、いくつかの子音の識別が難しいことも指摘された．これは、それらの子音に相当するジェスチャー、たとえば/p/と/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考えられる．この原因の一つとして、音声の特徴量として有声度が考慮されていないため、学習および推定フェーズにおいて有声／無声が明確に区別されていないことが挙げられる．

4　F0 を考慮したシステム

4.1　F0 と NCC

S2H システムにおいて、有声音／無声音に相当するジェスチャーを明確化することは、声から声への変換における有声／無声誤りの改善に相当する．近年、TTS や声質変換の分野において、特徴量として F0 と共に NCC（Normalized Correlation Coefficient: 正規化相関係数）を考慮することにより、従来の F0 推定手法と比べ、有声／無声誤りが改善されることが報告されている [11, 12]．NCC は次式で表される [13]．

$$\phi_{i,k} = \frac{\sum_{j=m}^{m+n-1} s_j s_{j+k}}{\sqrt{c_m c_{m+k}}} \tag{3}$$

ただし

$$c_m = \sum_{l=m}^{m+n-1} s_l^2 \tag{4}$$

ここで s_j は音声信号、$i = 0, 1, ..., M-1$ はフレーム番号である．$k = 0, 1, ..., K-1$ は遅れ、n はフレーム長 、z はフレームシフト幅をそれぞれサンプル数で表したものである．ただし $m = iz$ である．本稿では F0 および NCC を音響特徴量に付加することにより、有声音／無声音に相当するジェスチャーの明確化を目指す．

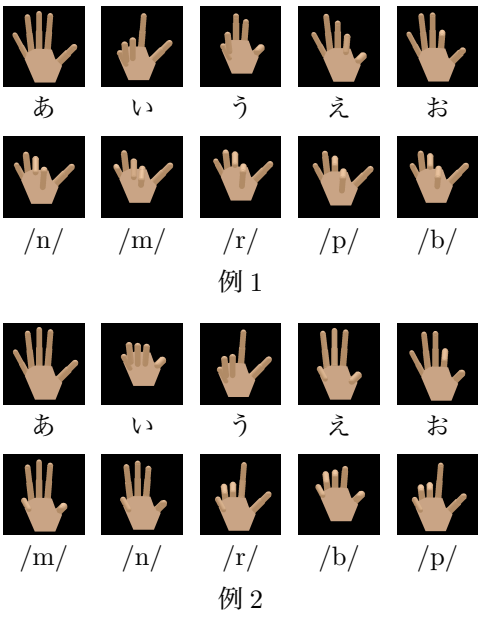


Fig. 3　S2H システムを構築した日本語五母音に相当するジェスチャーと、それによって推定された子音に相当するジェスチャーの例

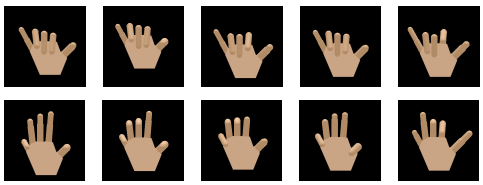
4.2　F0 を考慮した S2H モデル

前節で述べた手法が、有声音／無声音に相当するジェスチャーの明確化に有効であることを実験的に検証した．まず 3.2 と同様に、母音に相当するジェスチャーデザインのみから 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した．そして、このようにして得られたジェスチャーを、3.2 の手法によって得られたジェスチャーと比較した．

まず日本語五母音に相当するジェスチャーとして、Fig.4 のジェスチャーを選んだ．次に 3.2 と同様に、ジェスチャー／音声データの収録、ジェスチャーデータの再サンプリングおよび PCA、ジェスチャーモデルの

あ　い　う　え　お

Fig. 4　日本語五母音に相当するジェスチャー



/m/　/n/　/r/　/b/　/p/

Fig. 5　S2H によって推定された子音に相当するジェスチャー（上）従来手法（下）提案手法

構築を行う．音声データに対しては、STRAIGHT[10] を用いて分析をおこない、ケプストラム係数 1-16 次および F0 を抽出した．F0 値が 0 となる無声区間は、スプライン関数を用いて補間した。NCC は (3) 式から計算した．STRAIGHT の F0 抽出プログラムと同じ条件になるよう、フレーム長は 40 ms, フレームシフトは 1 ms とした．これら、ケプストラム 1-16 次、F0 および NCC を連結した 18 次元のベクトルを、音声の特徴量ベクトルとした．

3.2 で述べたように、ジェスチャー空間は PCA によって直交性を高めている．空間同士の対応付けを考慮し、音声の各特徴量ベクトルに対しても PCA を行った．PCA に用いたのは、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 文である．これらのジェスチャーの特徴量ベクトル（18 次元）と音の特徴量ベクトル（18 次元）から、3.2 と同様に結合ベクトルを作り、それを用いて変換モデル（混合数 16）を学習した．

このようにして構築された S2H システムによって推定された子音のジェスチャーデザインを Fig.5 に示す．有声音である/m/, /n/, /r/においては、従来手法と提案手法ともに、ジェスチャーが類似している．一方、/b/と/p/に相当するジェスチャにおいては、従来手法のデザインでは第 2 指の曲げ角度以外に大きな差異が見られなかったのに対し、提案手法では第 1〜3 指に差異が見られる．提案手法によって推定された子音のジェスチャーは、従来手法のものと比べ、有声音と無声音のジェスチャーデザインの差異がより明確になっていることが分かる．

なお同様の手法を用いて、F0 を考慮した H2S システムを構築すれば、手の動きを入力とした F0 生成が可能である．提案手法によって構築された H2S システムに、S2H システムで推定されたジェスチャーを入力することによって得られた F0 および NCC の一例を Fig.6 および Fig.7 に示す．NCC が 1 に近いフレームほど有声である可能性が高い．合成フェーズにおいて、ここでは仮に NCC< 0.75 となるフレームは無声と仮定し、F0 値を 0 としている．

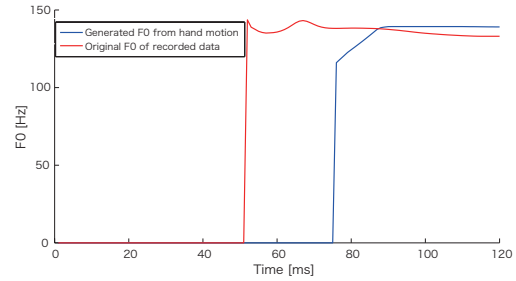


Fig. 6　/pi/に相当する F0

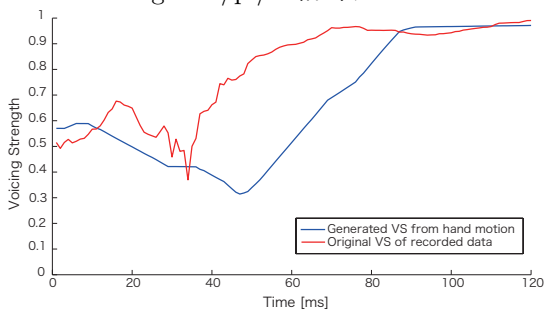


Fig. 7　/pi/に相当する NCC

5　まとめ

本稿では、S2H システムにおいて、F0 および NCC を音響特徴量に付加することにより、有声音／無声音に相当するジェスチャーを明確化できることを示した．また H2S システムにおいて同様の特徴量を用いることにより、手の動きからの F0 生成の可能性が示唆された．今後はさらに適切なジェスチャーデザインの推定とより自然な F0 の生成を目指し、ジェスチャー／音声の特徴量において、Δ特徴を考慮することを考えている．

参考文献

[1] 藪他, 信学技報, SP2006–164, pp.25–30 (2006)

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

[3] 市川, 手嶋,“福祉と情報技術”, オーム社 (2006)

[4] Stylianou et al., IEEE Trans. Speech Audio Process., Vol.6, pp.131-142 (1998)

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

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

[7] 國越他, 信学技報, SP2010–127, pp.73–78 (2011)

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

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

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

[11] Q. Zhang et al., Proc. ICASSP, pp.4606–4609 (2010)

[12] A. Kunikoshi et al., Proc. ICASSP, pp.4568–4571 (2011)

[13] D. Talkin et al., Eds. Elsevier, pp.495–518 (1995)