空間写像に基づく手の動きを入力とした音声生成系*

◎國越晶, 喬宇, 峯松信明, 広瀬啓吉 (東大)

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や [2]、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。そのため個々の障害者への適応は容易ではない。本研究では身体運動から音声を生成する過程を異メディア間の情報変換としてとらえる。すなわち身体運動の特徴量空間から音声の特徴量空間への写像を考えることで、音声生成を実現する。本稿では近年声質変換の分野で広く用いられるようになった、統計的手法を用いた空間写像の構築 [4] を応用し、手の動きから音声を生成する手法について述べる。

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。ある時刻の手の形が m 次元の特徴量ベクトル$\boldsymbol{a}$で表されるとする。これは手の形を表す 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、任意の手の形に対して、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　写像関数の計算

この写像関数は Stylianou らによって提案された手法 [4] によって求めることができる。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。

結合ベクトルの確率密度関数は GMM によって以下のように定義される。

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

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

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

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

ここで $N(\mathbf{z}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ 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、各正規分布で定義された写像関数の重み付け和で表される。

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

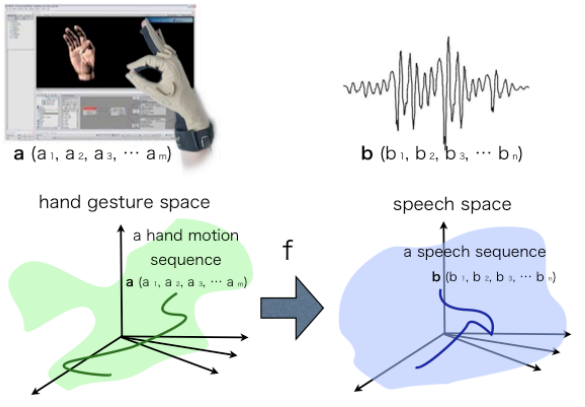


Fig. 1　空間写像に基づくメディア変換の枠組

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

Fig. 2　日本語５母音に対応する手の形

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

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

時刻 t におけるジャスチャー空間上のベクトル $\boldsymbol{a}(t)$ に対応する音声の特徴量ベクトル $\boldsymbol{b}(t)$ は、この写像関数 $f(\boldsymbol{a}(t))$ を使って $\boldsymbol{b}(t) = f(\boldsymbol{a}(t))$ と近似される。

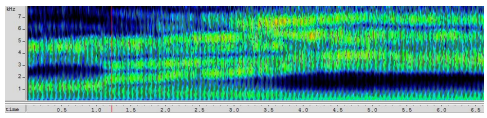
3　実装

3.1　学習

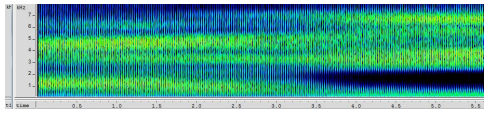
今回は提案するシステムの予備検討として、日本語５母音に対する連続音声を対象に手の動きから音声へのメディア変換を実装した。５母音に相当する手の形は Fig.2 のように設定した。この際、別の母音への遷移途中に形が重複しないように配慮した。次に学習データとして、Immersion 製データグローブ CyberGlove を装着し「あ」「い」「う」「え」「お」および２母音間の遷移 ${}_5P_2 = 20$ 組、計 $5 + 20 = 25$ 個のデータを記録した。センサの数は 18 個, サンプリング周期は 10〜20 ms である。また成人女性１名から収録した「あ」「い」「う」「え」「お」および２母音間の遷移 20 組、計 $5 + 20 = 25$ 個の音声データから、STRAIGHT[5] 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２種類のデータから結合ベクトルをつくるために、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に合わせて線形的に伸縮した後、ケプストラム時系列に対応するようにデータを線形補完した。こうして得られた結合ベクトルの分布を正規分布でモデル化した。

3.2　結果

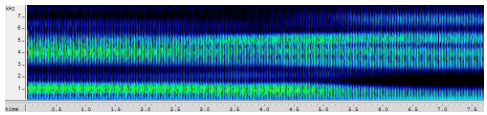
「あ」から「い」への遷移について提案手法により音声を生成した。Fig.3 に結果を示す。(a) は「あ」から「い」への遷移の分析再合成音、(b) は「あ」「い」「う」「え」「お」および２母音間の遷移 20 組、計 $5+20 = 25$ 個を学習データとして用いたモデルに、学習データ内の「あ」から「い」への手の形の遷移を入力した場合の合成音、(c),(d) は上記のモデルに、学習データに含まれていない「あ」から「い」への手の形の遷移を入力した場合の合成音の例である。ケプストラム



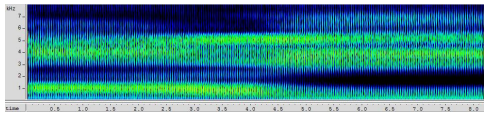
(a) 分析再合成音



(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

Fig. 3　「あ」から「い」への遷移に対応する合成音

時系列からの再合成には STRAIGHT を用い、F0 はすべて 205 Hz とした。また予備的な聴取実験によってこの手法の有効性を確認した。

4　おわりに

本稿では空間写像に基づく統計的手法 [4] を使い、手の動きを入力として母音を生成するシステムを検討した。今後の課題として、より自由な発話を実現するために加速度センサを用いた F0 制御、より多くの音素に対応するための空間写像のデザインについて検討していくつもりである。

謝辞　リアルタイムのシステム実装にあたり、名城大学の坂野秀樹先生が書かれた herium のソースコードから多くを引用させていただいた。またデータグローブは東京大学の中村仁彦先生からお貸しいただいた。両先生に心からお礼を申し上げたい。

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, 電子情報通信学会技術研究報告 Vol.106 No.613 pp25-30, 2007
[3] http://hct.ece.ubc.ca/research/glovetalk2/index.html
[4] Stylianou *et.al.*, IEEE Trans. Speech Audio Process., vol.6, pp.131-142, 1998
[5] 河原, 増田, 信学技報, EA96-28, pp.9-16, 1996