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とピッチ制御に関する検討

國越　晶†　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
†† 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,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している。近年、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。さらに自由な発話の実現を目指し、本稿では本システムにおける鼻子音の合成とピッチ制御に関して実験的検討を行った。本システムへの子音の導入方法としては、子音の開始時刻を前腕の姿勢角などで指定し、子音の波形と、ジェスチャーから得られる母音波形を接続する方式や、母音に対して用いた提案手法を子音の合成に拡張する方式などが考えられる。本稿では子音としてまず鼻音に注目し、この２通りの合成方法について実験的検討を行った。また磁気センサを導入し、前腕の姿勢角によるピッチ制御を実現したので報告する。

**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

# Nasal sound generation and pitch control for the real-time hand to speech system

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

† 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,mine,hirose}@gavo.t.u-tokyo.ac.jp

**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 voice conversion techniques have been proposed based on space mapping between given parallel data sequences. By applying these methods, a hand space and a vowel space is mapped and a converter from hand motions to vowel transitions is developed. It has been reported that the proposed method is effective enough to generate Japanese five vowels. In this paper, we discuss expansion of this system to consonant generation and pitch control. For the former, two methods are examined: waveform concatenation and space mapping for consonant sounds are discussed. For the latter, pitch control is realized using posture of the arm measured by a magnetic sensor.

**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指定し、これを求め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]。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)。

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 [8]。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 [9], [10]。

このシステムの実際の使用を考えた場合、子音の追加やピッチやパワーなどの制御が必要不可欠である。しかし本システムではジェスチャーデータ 1 フレーム毎にリアルタイム変換を行っているため、コーパスや HMM に基づいた通常の音声合成方法を応用するのは難しい。そこで本システムへの子音の導入方法として、1) 子音の開始時刻のみを与え、ジェスチャーから得られる母音波形に子音の波形を接続する波形接続方式、2) 母音に対して用いた提案手法を子音の合成に拡張した空間写像方式を考える。本稿では子音としてまず鼻音に注目し、この 2 通りの合成方法について実験的検討を行う。また磁気センサを導入し、前腕の姿勢角によるピッチの制御についても試みる。

(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。

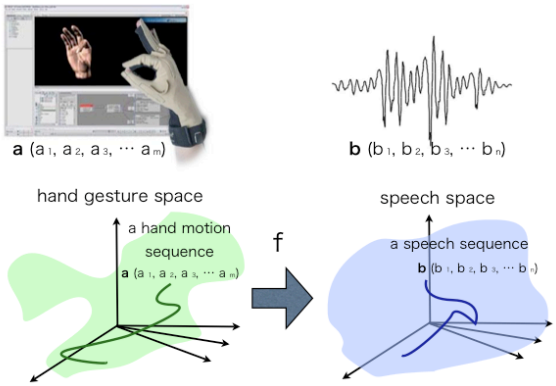



図 1　空間写像に基づくメディア変換の枠組

## 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。

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

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

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

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

ここで $\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ は平均 $\boldsymbol{\mu}_i$, 分散 $\boldsymbol{\Sigma}_i$ の正規分布を表し、$q$ は混合数、$\omega_i$ 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、次のように表現される。

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

$i$ 番目の回帰式に対する事後確率 $h(i|\boldsymbol{a})$ は以下で与えられる。

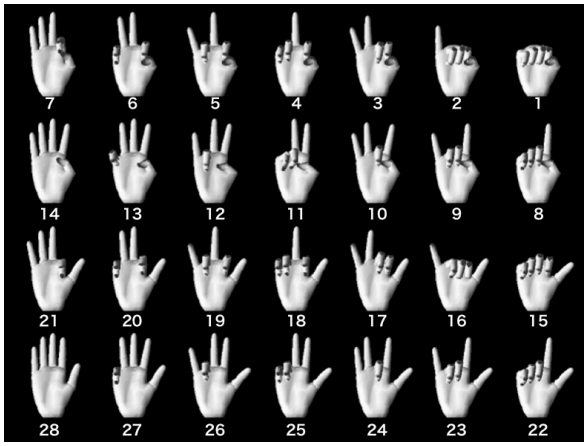



図 2　基本的な 28 種類のジェスチャー

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

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。

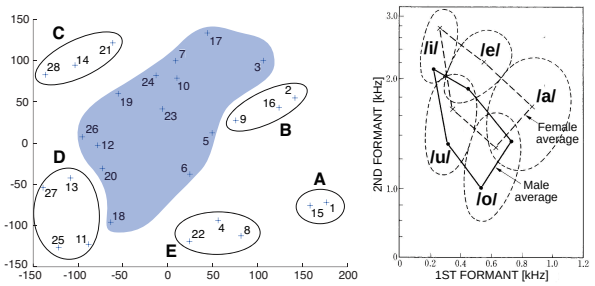
## 3.　日本語五母音の合成

### 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 [8]〜[10]。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。

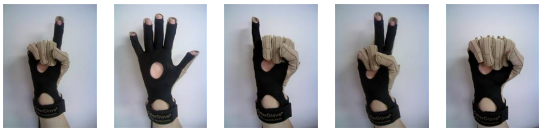
本稿ではジェスチャーの候補には、Wu らが画像認識における論文 [11] の中で用いた、基本的な 28 個のジェスチャーを参考にして選んだ。そのジェスチャーを図 2 に示す。これは、五指各々の曲げ延ばしの組み合わせ $2^5 = 32$ 個から、薬指だけを立てるもの、薬指と人差し指を立てるもの、薬指と親指を立てるもの、薬指、人差し指と親指を立てるもの、即ち実現不可能な 4 種類を差し引いたものである。

[9], [10] 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、




(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

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。

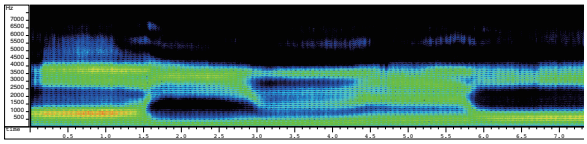
### 3.2　学習データ

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 [12] 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。

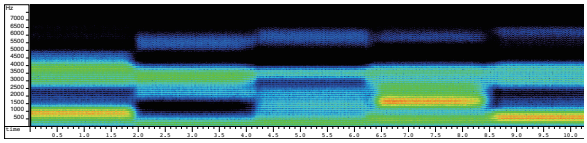
### 3.3　結　　果

「あいうえお」に対して提案手法により音声を生成した。図 5 に結果を示す。(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、

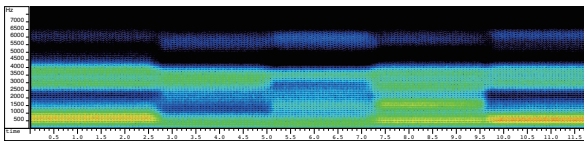
(注2)：データグローブからのサンプリング周期は時不変ではない。最終的には、線形補完の形で周期一定となるようデータの再サンプリングを行った。

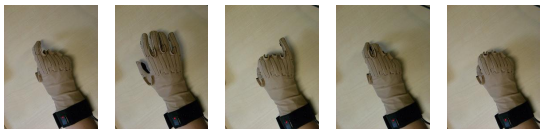


(a) 分析再合成音



(b) 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

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。しかし、より自由な発話を目指すためには、子音の追加やピッチ・パワーなどの制御が必要不可欠である。以降の章では、本システムにおける子音の合成方式やピッチの制御方法について述べる。

## 4. 子音の合成

現在、音声の合成方法としては、コーパスベースの音声合成方式や、HMM に基づく音声合成方式などが提案されている。しかし本システムではジェスチャーデータ 1 フレーム毎にリアルタイム変換を行っているため、これらの手法をそのまま応用することは難しい。本システムへの子音の導入方法としては、1) 子音の開始時刻のみを前腕の姿勢角などで特定し、ジェスチャーから得られる母音波形に子音の波形を接続する波形接続方式、2) 母音に対して用いた提案手法を子音の合成に拡張した空間写像方式などが考えられる。本稿では子音として鼻音に注目し、波形接続方式および空間写像方式の 2 通りについて実験的検討を行った。

### 4.1 波形接続方式

母音と異なり、子音継続長の発声速度に関する影響は小さい。そこで腕姿勢などで特定した子音の開始時刻から、あらかじめ収録音声から切り出した子音の波形を出力し、それに続けてジェスチャーから生成した母音の波形と接続することによって子音を含んだ音声を合成する(注3)。

(注3)：現在のシステムには、子音の開始時刻を特定する機能は搭載していない。本実験では、子音の開始時刻が適切に特定できるものと仮定している。

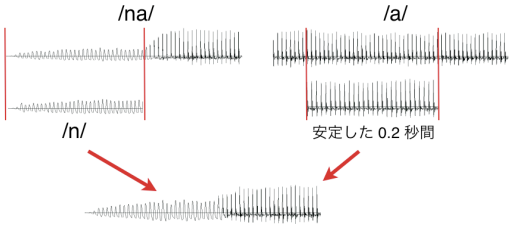
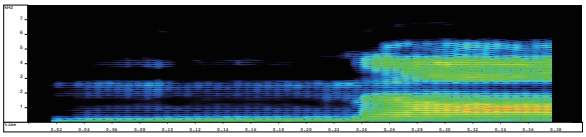
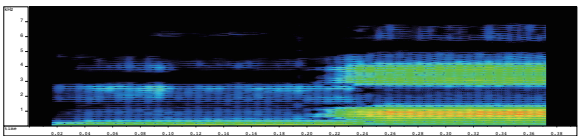



図 6 /n/と/a/の切り出し



(a) 分析再合成音



(a) 波形接続による合成音

図 7 「な」に対応する合成音

今回は予備的検討として、成人男性 1 名から「あ」「い」「う」「え」「お」および「な」をそれぞれ 3 回ずつ、合計 18 個のデータを収録した。次にそれぞれに対して STRAIGHT 分析を行い、ケプストラム係数 0-17 次、フレーム長 40 ms, フレームシフト 1 ms で再合成した。そして図 6 のように、母音の再合成音からはそれぞれ安定した 0.2 秒間、「な」からは視認によって/n/に相当するおよそ 0.24 秒を切り出した。なお、テキスト合成と異なり、/n/をコンテキスト依存で用意することはしていない。リアルタイム変換の場合、コンテキストをジェスチャーから指定するのは困難である。

/n/の波形の後半 $d$ [ms] にはブラックマン窓の後半部を、母音部の波形の前半 $e$ [ms] にはブラックマン窓の前半部をそれぞれ掛け、/n/の後半部と母音の前半部を $m$ [ms] だけ重ね合わせた。$d, e, m$ の値は話者によって変化する。本稿では実験的に、$d = 30$ [ms], $e = 60$ [ms], $m = 60$ [ms] の値を用いた。図 7 に結果を示す。

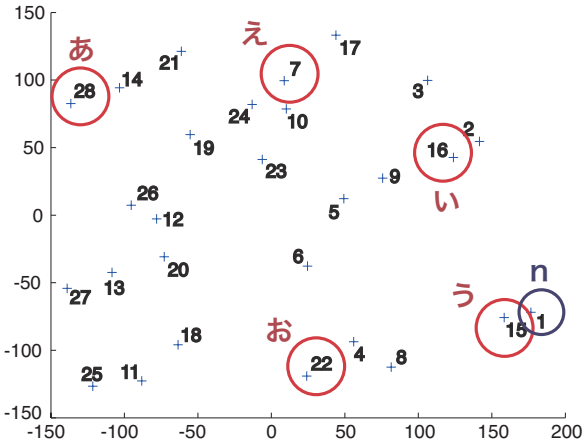
### 4.2 空間写像方式

#### 4.2.1 子音を考慮したジェスチャーデザイン

もう一つの方法として、子音にも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も拡張する方法が考えられる。この方式では子音と母音のつなぎ目は滑らかになるが、その一方で、音声にして数 〜数十 [ms] にしか持続しない子音に割り当てるジェスチャーの決定が問題になる。

手の主な働きはものを掴むことであるから、ものを握る動きは最も形成が容易で、速く動かすことができる動きのひとつであると考えられる。そこで本稿ではもの握る動きを子音–母音の遷移に割り当てることとし、図 2 の No.1 のジェスチャーを/n/に割り当てることとした。

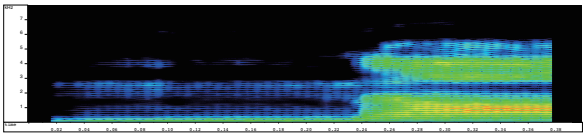
/n/だけを発音する際の口の形は「う」に近いため、図 3(a) の PCA 空間において、No.1 と「う」に割り当てられるジェスチャーは近いことが推測される。これと 3.1 節での議論とから、/n/を考慮した場合のジェスチャーデザインを図 8 のように設定した。



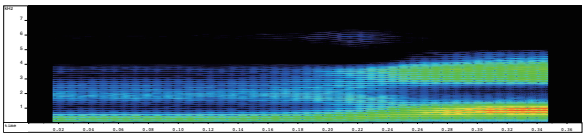

ジェスチャー空間における/n/と日本語五母音の位置関係

n　あ　い　う　え　お

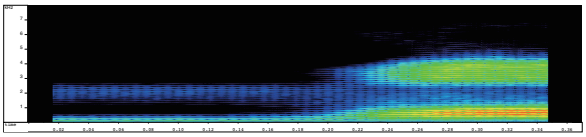
図 8　/n/を考慮した場合の日本語五母音に相当するジェスチャー



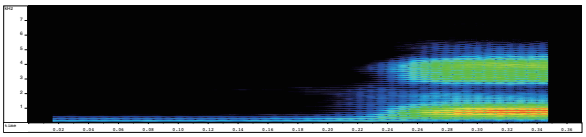
(a) 分析再合成音



(b) GMM 混合数 1 の場合の合成音



(c) GMM 混合数 4 の場合の合成音



(d) GMM 混合数 32 の場合の合成音

図 9　「な」に対応する合成音

表 1　明　瞭　度

| 合成方式 | な | に | ぬ | ね | の | 平均 |
|---|---|---|---|---|---|---|
| 分析再合成 | 85.7 | 85.7 | 71.4 | 64.3 | 85.7 | 78.6 |
| 波形接続 | 14.3 | 50.0 | 28.6 | 71.4 | 14.3 | 35.7 |
| 空間写像 | 7.14 | 78.6 | 14.3 | 14.3 | 21.4 | 27.1 |

表 2　/n/と/m/の間違い（/n/と/nu/, /m/と/nu/の間違いを含む）を許容した場合の明瞭度

| 合成方式 | な | に | ぬ | ね | の | 平均 |
|---|---|---|---|---|---|---|
| 分析再合成 | 92.9 | 92.9 | 85.7 | 100 | 92.9 | 92.9 |
| 波形接続 | 71.4 | 100 | 71.4 | 85.7 | 100 | 85.7 |
| 空間写像 | 7.14 | 85.7 | 78.6 | 21.4 | 42.9 | 47.1 |

#### 4.2.2　提案デザインに基づく写像関数の推定

前節で提案したジェスチャーデザインを用いて、提案するシステムにおいて、/n/および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。学習データとして、データグローブを装着し「な」「に」「ぬ」「ね」「の」、「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 $(10 + 20) \times 3 = 90$ 個のデータを記録した。また成人男性 1 名から「な」「に」「ぬ」「ね」「の」、「あ」「い」「う」「え」「お」および二母音間の遷移 20 組を各々 5 回、計 $(10 + 20) \times 5 = 150$ 個の音声データを収録した。アライメントをとるため、「な」「に」「ぬ」「ね」「の」のジェスチャーデータおよび音声データは、視察によって、それぞれ子音部分／遷移部分／母音部分に 3 分割した。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 次の抽出を行った。その後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。図 9 に結果を示す。

### 4.3　聴 取 実 験

14 人の聴取者を募り、分析再合成音と 2 つの手法で合成した「な」「に」「ぬ」「ね」「の」、計 $3 \times 5 = 15$ 文字に対し、明瞭度（intelligibility）試験を行った。空間写像方式では、混合数 4 の GMM を用いている。実験で使用した試料セットには、上記の 15 個の試料の他,「あ」「い」「う」「え」「お」の分析再合成音 2 セット、合成音と波形接続方式で合成した「ぱ」「ぴ」「ぷ」「ぺ」「ぽ」($d = 20$ [ms], $e = 20$ [ms], $m = 10$ [ms])、計 $2 \times 5 + 2 \times 5 = 20$ 個の試料をダミー音声として含めた。回答者には試料がそれぞれ 1 モーラ音声であることを知らせ、ランダムに呈示した試料に対し、ローマ字で書き取りを行わせた。結果を表 1 に示す。

波形接続方式の「ね」、空間写像方式の「に」を除き，両方式ともに分析再合成音の半分以下の明瞭度しか得られていないことが分かる。文字ごとに比較すると、「な」「ぬ」「ね」では波形接続方式の方が空間写像方式よりも明瞭度が高いが、「に」「の」では空間写像方式の明瞭度が波形接続方式を上回っている。合成方式ごとに比較すると、波形接続方式では「ね」が最も明瞭度が高く、「な」「の」は非常に低い。空間写像方式では「に」が、分析再合成音の明瞭度にせまる非常に高い数値を示しているのに対し、「な」「ぬ」「ね」はほとんど知覚されていない。

間違いの種類としては、波形接続方式では、/n/を/m/と知覚した回答が最も多かった。一方、空間写像方式では文字ごとに間違いの傾向が異なっており、「な」「ね」では/n/を知覚していない回答が多く、「ぬ」では/n/を/m/と知覚したり、母音を落として/n/または/m/とした回答が多かった。

/n/と/m/の間違い（/n/と/nu/, /m/と/nu/の間違いを含む）を許容すると、明瞭度は図 2 のように大幅に向上する。

### 4.4 考　　察

提案した 2 方式はともに、明瞭度において分析再合成音を大きく下回っていた。

波形接続方式では、/n/を/m/と知覚した回答が最も多かった。フォルマントの遷移部には子音の知覚に影響を与える何らかの音響特性があると考えられている [13]。提案手法ではその遷移部分が十分に考慮されていないために、子音の知覚間違いが起こったと考えられる。今後、遷移部分を考慮に入れるための手法を検討する必要がある。

空間写像方式では、明瞭度、間違いの傾向ともに、文字ごとのばらつきが大きかった。これは/n/から母音に遷移する際の遷移部分が、母音ごとに適切に学習されていないためと考えられる。子音から母音への遷移部分を学習するため、提案方式ではジェスチャーデータと収録音声を、子音部分／遷移部分／母音部分に分割してから、結合ベクトルを作っている。しかし音声データの遷移部分は、3 周期程度であり、子音部の 5 分の 1 程度である。そのため GMM を学習する際、遷移部分が無視されやすくなっていると考えられる。遷移部分を適切に学習するために、GMM の混合数の増加や、遷移部分の学習データの増加などが必要であると考えられる。

また現在のシステムは、ジェスチャーデータ 1 フレーム毎に変換を行っているため，音声の特徴量としてケプストラムを用いている。しかし動きの情報を取り入れるためには、Δケプストラムなどの特徴量を考慮に入れることが望ましい。今後は前後のフレームに配慮してシステムを検討したい。

## 5. ピッチの制御

本システムでは、ジェスチャーによって口の形に相当するパラメータを制御している。そこでジェスチャー以外のパラメータとして、手の向きを導入した。手の方向の取得には 3 軸地磁気・3 軸加速度・気圧測定用センサモジュールキット TDS01V を用いた。TDS01V における 3 つの角度の定義を図 10 に示す。加速度に比べ、角度は安定した制御が可能なため、手の腕を軸とした回転（Roll）を合成音の F0 に対応させた。Roll は、手のひらを真下に向けた時を 0° として、−90°〜+90° まで変化する。

本システムでは $Roll \times 1/3 + const.$ を F0 として、合成音の F0 が 120〜160 [Hz] 程度になるようにした。これによって、トーンを表現できるようになったが、ジェスチャーと手の向きの両方を同時に制御するためには、訓練が必要であることが指摘された。今後パラメータを増加するにあたっては、他のパラメータとの関係を考慮し、ユーザーの操作性に配慮する必要がある。

## 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、波形接続方式、空間写像方式の 2 方式で鼻音の合成の実験的検討を行った。提案した 2 方式ともに、明瞭度において、分析再合成音を大きく下回ることがわかった。さらにその結果は文字ごとに大きく異なっていた。これはジェスチャー

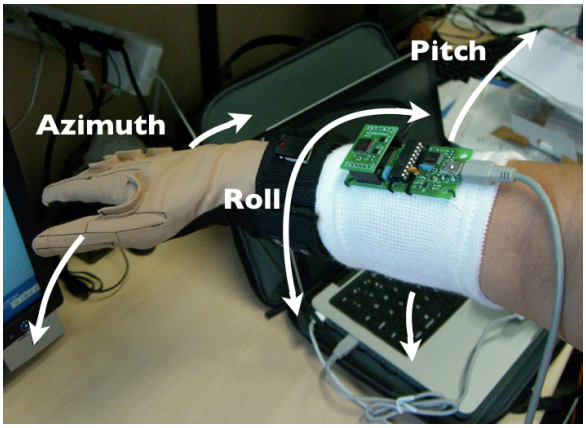



図 10　TDS01V における Azimuth（方位角）, Pitch（傾斜）, Roll（回転）

の遷移部分／音声の調音結合部分が適切に考慮されていないためと考えられる。今後は遷移部分やジェスチャー／音の動きに配慮した手法を検討する必要がある。

また前腕の姿勢角を用いて、ピッチの制御を行った。これによって、トーンを表現できるようになったが、操作の困難さが指摘された。今後パラメータを増加するにあたっては、他のパラメータとの関係を考慮し、ユーザーの操作性に配慮する必要があることがわかった。

## 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] 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
[9] 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の構築”, 電子情報通信学会音声研究会, SP2008-78, pp.45-50, 2008
[10] A. Kunikoshi *et.al.*,“Speech generation from hand gestures based on space mapping,” Proc. INTERSPEECH, pp.308-311, 2009
[11] Y. 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
[12] 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
[13] ジャック・ライアルズ,“音声知覚の基礎”, pp.31-38, 海文堂, 2003