情報処理学会論文誌 Vol.53 No.9 2291–2301 (Sep. 2012)

空間写像に基づく母音と鼻子音を対象としたジェスチャ–音声変換システム

國越 晶[1,a)] 喬 宇[2] 齋藤 大輔[1] 峯松 信明[1] 広瀬 啓吉[1]

受付日 2012年1月17日, 採録日 2012年6月1日

概要：調音音声合成に代表される文字や記号を介さない合成方式は，運動の連続性に基づく滑らかな合成音の生成やその話速制御などにおいて有効性が注目されている．しかしそのアプリケーションのほとんどは，入力機器の特性を活用して設計されているため，その方法論を他のメディアや機器に応用することは容易ではない．本研究では身体運動から音声を生成するプロセスを，特定の身体部位に限定せずに一般化してとらえ，音声以外のメディア情報の動きを入力として音声を出力するプロセス，異メディア間写像の問題としてとらえる．そして近年声質変換の分野で広く用いられている統計的空間写像構築法を応用した，メディア非依存の方法論を提案する．本稿ではその一例として手の運動からの音声出力を考える．この手法においては，どの手の姿勢（以下ジェスチャ）をどの音に割り当てるかが課題となる．これまでに，ジェスチャを入力とした日本語 5 母音の連続音声生成において，本手法の有効性および適切なジェスチャ選択手法を報告している．本稿では，子音として鼻子音に注目し，母音に関して，ジェスチャと音が時間同期されたデータを用いて構築した音声→ジェスチャ変換システム（目的とするシステムの逆システム）に鼻子音音声を入力することにより，鼻子音に割り当てるジェスチャを推定する手法を提案する．聴取実験の結果，音声→ジェスチャ変換システムによって推定されたジェスチャは，ジェスチャ候補から選ばれた準最適なジェスチャと比較して，より自然な音声を出力するジェスチャ→音声システムを構築することが示された．

キーワード：音声生成，手の運動，メディア変換，ジェスチャと母音の配置

A Speech-to-Hand Conversion System for Vowels and Nasals Based on Space Mapping

Aki Kunikoshi[1,a)] Yu Qiao[2] Daisuke Saito[1]
Nobuaki Minematsu[1] Keikichi Hirose[1]

Received: January 17, 2012, Accepted: June 1, 2012

Abstract: Synthesis methods which do not require symbol inputs, such as articulatory synthesis, are useful in continuous speech synthesis and pitch control based on dynamic body motion, in which there are no inherent symbols. Conventional applications based on these methods, however, are strongly dependent on their input media because those applications are designed to make use of their specific characteristics. Once an application is constructed for one media therefore, its methodology is difficult to apply to another media. Considering this point, we treat speech generation from body motion as a mapping problem between different media, non-acoustic media to speech, and propose a media-independent methodology. As one example of our methodology, media conversion from hand motion to speech is discussed. In recent years, the GMM-based statistical mapping techniques have become widely used for voice conversion. Using similar techniques, we have developed a speech generation system which maps gesture space to vowel space and converts hand motions to vowel transitions. In this paper, we expand the system to nasal sound generation. In order to derive the gestures for nasals, a Speech-to-Hand conversion system was developed using the parallel data for vowels. Subjective evaluations showed that our proposed method is effective to generate more natural speech than the quasi-optimal design among a given gesture candidate set.

Keywords: speech production, hand motions, media conversion, arrangement of gestures and vowels

1. はじめに

音声合成技術は，TTS（Text-to-Speech）に代表される文字や記号を入力とする合成方式と，調音音声合成に代表される文字や記号を介さない合成方式に大別される．合成

[1] 東京大学
The University of Tokyo, Bunkyo, Tokyo 113–8656, Japan
[2] 中国科学院深セン先進技術研究院
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen University Town, Shenzhen, P.R. China
[a)] kunikoshi@gavo.t.u-tokyo.ac.jp

音の明瞭性や操作の容易性などから広く実用化されている前者と比較し，後者は運動の連続性に基づく滑らかな合成音の生成や話速制御などにおいて有効性が注目されている [1]．これらの利点を生かした様々なアプリケーションが，芸術的歌声生成 [2], [3]，教育応用 [4], [5]，障害者支援 [1], [6], [7] などの分野で提案されている．その一例として，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 [6] や，身体運動からの歌声生成器 GloveTalk II [3] があげられる．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が（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），入力メディアに依存する各種特性を活用して設計されているため，その方法論を他のメディアに応用することは必ずしも容易ではない．たとえば，ペンタブレットを使った構音障害者用音声合成器は，ペン先の運動の連続性や，ペン先の運動と筆圧を同時に制御できる点などを利用している．そのためペンの使用が困難なユーザに対しては（構音障害者の中には手先の制御に困難をかかえる障害者もいる），別の入力メディアを選択し，その特性に基づいて機器を再設計しなくてはならない．

機器のメディア依存性は，障害者支援やアートなどの分野において重要になる．障害者支援技術においては，ユーザの身体能力や技能などに合わせて，適切なメディアが選択される．たとえば，先天的視覚障害者ならば点字メディアが利用可能な場合が多いが，後天的な障害者の多くは点字が苦手であり，音声メディアを利用した支援機器が望まれる [8]．肢体不自由者支援機器として広く使われている電動車椅子も，障害者の残された能力に応じて，頭部ジェスチャや音声，力覚や筋電を入力としたものが開発されている [9]．またアートの世界では，しばしば表現の可能性を広げるために，様々なメディアが検討される．空間中の手の位置によって音程と音量を調節する電子楽器テルミンを開発したテルミン博士は，そのバリエーションとして，ダンサの身体の動きによって音高が変化する楽器テルプシトンを開発した．テルプシトンはダンサが演奏者にもなりうるという，表現者の意識改革を生み出すきっかけとなったといわれている [10]．このように，開発者ではなく，使用者のニーズに応じて入力メディアが選択されることが多い分野において，メディア選択の自由度は重要な意味を持つ．メディアに依存しない方法論を構築することができれば，その応用先は多岐にわたるものと考えられる．これをふまえ，本研究では身体運動から音声を生成する技術として，音声以外のメディアを入力として音声を出力する異メディア間写像の問題としてとらえ，特定のメディアに限定されない方法論を構築する．

近年，話者変換をある話者の音響空間から別の話者への音響空間への写像としてとらえ，統計的に空間写像を設計する手法が用いられている [11], [12], [13]．これを応用し，本研究では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とらえ，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構築することで，音声生成を実現する．この手法に基づいた音声合成として，これまでに調音運動からの音声合成 [14] や顔面筋電からの音声合成 [15] などが報告されている．我々はこの手法を，音声との対応付けが明確でない入力メディアに拡張する．本稿ではその一例として手の運動からの音声出力を考える．

手の運動からの音声生成系の構築にあたっては，どの手の姿勢（以下ジェスチャ）をどの音に割り当てるかが課題となる．これまでに，ジェスチャを入力とした日本語 5 母音の連続音声生成において，「ジェスチャ空間におけるジェスチャ群の配置」と「母音空間における母音群の配置」の等価性を，より保証できる空間写像を設計した場合，より明瞭な音声を生成できることを確認している [16], [17].

本稿では，この枠組みを鼻子音の合成に拡張する方法について検討する．まず母音に対して，ジェスチャと音が時間同期されたデータ（以下パラレルデータ）を用いて学習された変換モデルと，変換モデルとは別に用意されたジェスチャモデルを用い，音声→ジェスチャ変換システム（Speech-to-Hand system，以下 S2H システム，本研究の目的とするジェスチャ→音声変換システムの逆変換）を構築する．それに鼻子音音声を入力することにより，鼻子音に割り当てるジェスチャを推定する手法を提案する．

次章以降の構成は以下のとおりである．まず 2 章で，提案する H2S システムの枠組みについて述べる．3 章では，その手法を鼻子音の合成に拡張する．その際，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，それに鼻子音音声を入力することにより，鼻子音に割り当てるジェスチャを推定する手法を提案する．4 章では提案手法を実験的に検証する．5 章では提案したジェスチャデザインをもとに H2S システムを構築し，聴取実験によってその性能を評価する．最後に 6 章で本稿をまとめる．

2. 空間写像に基づくメディア変換

時刻 t のジェスチャが m 次元の特徴量ベクトル $\boldsymbol{h}_t = (h_1, h_2, \ldots, h_m)$（以下ジェスチャベクトル）で表されるとする．これはジェスチャを表す m 次元空間（以下ジェスチャ空間）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 $\boldsymbol{s}_t = (s_1, s_2, \ldots, s_n)$（以下，音響特徴量ベクトル）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 $\mathcal{F}$ を求めることで，任意のジェスチャに対して，対応する音声の特徴量ベクトルを求めることができる[*1]．

[*1] 便宜上，本研究では $m = n$ としている．

近年，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が提案されている [11], [12], [13]．本研究が目指す H2S システムは，それらの声質変換手法において，入力話者の音響空間をジェスチャ空間に置き換えたものと考えることができる．本稿ではジェスチャ空間と音響特徴量空間における空間写像を，Kain らの手法 [12] に基づき，次のように推定する．

まず対応関係の分かっているジェスチャベクトル $\boldsymbol{h}$ と音響特徴量ベクトル $\boldsymbol{s}$ から，フレームごとに結合ベクトル $\boldsymbol{z} = [\boldsymbol{h}^\top, \boldsymbol{s}^\top]^\top$ 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，$\boldsymbol{z}$ の確率密度をモデル化する．

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

ここで $\mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)})$ は平均 $\boldsymbol{\mu}_m^{(z)}$，分散 $\boldsymbol{\Sigma}_m^{(z)}$ の正規分布を表す．M は混合数，m は混合インデックス，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．

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(h)} \\ \boldsymbol{\mu}_m^{(s)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(hh)} & \boldsymbol{\Sigma}_m^{(hs)} \\ \boldsymbol{\Sigma}_m^{(sh)} & \boldsymbol{\Sigma}_m^{(ss)} \end{bmatrix} \tag{2}$$

ただし，$\boldsymbol{\mu}_m^{(h)}$，$\boldsymbol{\Sigma}_m^{(hh)}$，$\boldsymbol{\mu}_m^{(s)}$，$\boldsymbol{\Sigma}_m^{(ss)}$ は m 番目の正規分布における，ジェスチャベクトルと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 $\boldsymbol{\Sigma}_m^{(hs)}$，$\boldsymbol{\Sigma}_m^{(sh)}$ は，入出力空間間の相互共分散行列を表す．

Kain らは写像関数 $\mathcal{F}(\cdot)$ を，重み付け線形和として以下のように近似している [12]．

$$\hat{\boldsymbol{s}} = \mathcal{F}(\boldsymbol{h}) = \sum_{m=1}^{M} P(m|\boldsymbol{h}, \boldsymbol{\lambda}^{(z)})(\boldsymbol{A}_m \boldsymbol{h} + \boldsymbol{b}_m) \tag{3}$$

ここで $P(m|\boldsymbol{h}, \boldsymbol{\lambda}^{(z)})$，$\boldsymbol{A}_m$ および $\boldsymbol{b}_m$ は，m 番目の正規分布における事後確率，変換行列およびバイアスベクトルであり，以下のように表現される．

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

$$\boldsymbol{A}_m = \boldsymbol{\Sigma}_m^{(sh)} \boldsymbol{\Sigma}_m^{(hh)-1}$$

$$\boldsymbol{b}_m = \boldsymbol{\mu}_m^{(s)} - \boldsymbol{\Sigma}_m^{(sh)} \boldsymbol{\Sigma}_m^{(hh)-1} \boldsymbol{\mu}_m^{(h)}$$

この枠組みでは，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，動的計画法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，ジェスチャと音は任意に対応付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．これまでに，ジェスチャを入力とした日本語 5 母音の連続母音音声生成において，「ジェスチャ空間中のジェスチャ群の配置」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（ジェスチャ群と母音群の形態的等価性）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 [16], [17]．次章では，本システムにおける子音の合成方式について述べる．

3. 子音の合成

3.1 日本語子音の分類

日本語における子音の分類を**表 1** に示す．

本研究では，これらを半母音，摩擦音/破擦音/破裂音，鼻音および弾き音，の 3 グループに分類し，それぞれのグループにおいて次のような合成方式を考える．

半母音は，母音音声の遷移によって特徴付けられる．たとえば/wa/は，/u/から/a/への遷移によって表現される．薮らは母音のフォルマント遷移のみで，これらの子音を含んだ単語を知覚させることが可能であることを報告している [6]．これを参考に，本システムでも母音ジェスチャの遷移およびその音量を連続的に制御することによってこれらの子音を含んだ音声を合成する．

摩擦音/破擦音/破裂音は，発話速度や後続母音による継続長への影響が小さい [19]．すなわち，本システムにおいてこれらの子音を合成する場合，空間写像に基づいて合成する母音や半母音とは異なり，ユーザが身体の運動速度や後続母音によって，これらの子音に相当する波形（継続長）を調整する必要性は低い．システムには，あらかじめ収録音声から切り出した波形をプリセットしておくことが可能であると考えられる．その一方で，声帯振動が始まる有声開始時間（VOT：Voice Onset Time）がこのグループの子音の知覚に与える影響は大きい．VOT の違いによって，/t/が/d/に，/p/が/b/に知覚されることなどが知られている [20] ほか，薮らは，/sa/の波形が/tsa/，/ta/に知覚

表 1 調音点・調音法による日本語子音の分類 [18]
Table 1 Japanese consonants classification [18].

		調音点												S2H システムにおける合成方式
		両唇		歯茎		歯茎硬口蓋		硬口蓋		軟口蓋		声門		
調音方式	摩擦音	ɸ		s	z	ɕ	ʑ	ç				h		波形接続方式
	破擦音			ts	dz	tɕ	dʑ							
	破裂音	p	b	t	d				k	g				
	鼻音	m		n		ɲ		ŋ						母音音声に対して用いた S2H システムの拡張
	弾き音			ɾ										
	半母音							j		w				母音音声の遷移

されることを報告している [21]．これらを考慮し，本システムでは，腕姿勢などで VOT を制御することでプリセットした子音の波形を出力し，それに続けてジェスチャから生成した母音波形と接続する方式（波形接続方式）により，これらの子音を含んだ CV 音声を合成する．

表 1 において鼻音および弾き音として分類される 5 つの単音と，日本語音素との対応を書くと，[m] が/m/，[ɲ]，[ŋ]，[n] が/n/や/N/，[r] が/r/となる．これら 4 つの音素を，本システムでは鼻音および弾き音（以下，鼻子音と呼ぶ）として検討する．鼻子音は母音同様，共鳴および反共鳴特性によりその特性が記述されるため，本システムでは母音に対して用いた手法をこれらの子音音声の合成にも応用することを考える．すなわち，これらの子音にもジェスチャを割り当て，空間写像に基づく音声合成方式を提案する．

上記の 3 つのグループのうち，半母音の合成は，2 章で述べた母音合成の枠組みをそのまま用いる．摩擦音/破擦音/破裂音は，時間構造が母音と異なり，空間写像に基づく合成法とは異なる議論が必要となるため，本稿では取り上げない．以降の章では，鼻子音の合成のみに焦点を置き，母音に対して用いた枠組みを拡張する手法について議論する．

3.2 母音のパラレルデータを用いた鼻子音ジェスチャの推定

本システムにおいて，鼻子音にもジェスチャを割り当て，母音に対して用いた提案手法を拡張することにより．これらを含んだ音声を合成する．この合成方法では，これらの鼻子音に割り当てるジェスチャの決定が問題になる．我々は予備実験として，日本語 5 母音に対する/n/の配置に配慮したジェスチャデザインに基づき，H2S システムを構築した．すなわちケプストラム空間における/n/と 5 母音間それぞれの距離において/n/–「う」間の距離が最小であることに配慮し，ジェスチャ空間において/n/–「う」間の距離が，/n/と他の 4 母音との距離よりも小さくなるように，/n/および日本語 5 母音のジェスチャを選択した．このジェスチャデザインを用い，2 章の手法に基づいて，母音および/n/を含んだ音声を合成したところ，合成音において/n/が/m/や/w/などに知覚される問題などが指摘された [22]．この原因として 2 つの理由が考えられる．1 つ目は，ジェスチャデザインの問題である．上記のデザインでは，ジェスチャ空間における母音と/n/に対応するジェスチャの位置関係が，音響空間における母音と/n/の位置関係に適切に対応していない可能性がある．そのため，/n/に割り当てられたジェスチャを入力しても/n/相当の音声が生成されなかったことが考えられる．もう 1 つは，ジェスチャデータと音声データの動的な軌跡の対応付けの問題である．子音から母音への遷移部分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 [20]．そのため，子音が適切に知覚されない原因として，遷移部分が適切に合成されていない可能性があげられる．すなわち上記のデザインにおいて静的な位置関係が適切に対応付けられていた場合でも，ジェスチャと音声の動的な軌跡において適切な対応付けがとれていないことが考えられる．これらの問題を回避するため，本稿では，母音のみのパラレルデータを用いて S2H システム（目的とする H2S システムの逆のシステム）を構築し，それに鼻子音音声を入力することにより，鼻子音に対応するジェスチャを推定する．

確率的な変換モデル $P(\boldsymbol{y}|\boldsymbol{x})$ の，$\boldsymbol{y}$ 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 [23]．$\boldsymbol{x}$ = 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十分になくても，大量の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による高品質な変換法が提案されている [24]．

我々の目的は，どの音声をどのジェスチャ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 ($\boldsymbol{h}$) から音声 ($\boldsymbol{s}$) への統計的変換モデル $P(\boldsymbol{s}|\boldsymbol{h})$ ではなく，その逆の変換モデル $P(\boldsymbol{h}|\boldsymbol{s})$ を考え，これにベイズの法則を適用することを考える．すなわち，音声 $\boldsymbol{s}$ が与えられた場合のジェスチャ推定問題を，以下の式で与えられる $P(\boldsymbol{h}|\boldsymbol{s})$ の最大化問題として扱う．

$$\begin{aligned}\hat{\boldsymbol{h}}(s) &= \operatorname*{argmax}_h P(\boldsymbol{h}|\boldsymbol{s}) = \operatorname*{argmax}_h \frac{P(\boldsymbol{s}|\boldsymbol{h})P(\boldsymbol{h})}{P(\boldsymbol{s})} \\ &= \operatorname*{argmax}_h P(\boldsymbol{s}|\boldsymbol{h})P(\boldsymbol{h}) \end{aligned} \quad (4)$$

ここで，$P(\boldsymbol{s}|\boldsymbol{h})$ は母音のみからなるパラレルデータにより構成された変換モデル，$P(\boldsymbol{h})$ は子音も含む大量のジェスチャデータから推定されるジェスチャの統計モデルである．

声質変換のタスクにおいて，齋藤らはこの問題を解くため，式 (4) に基いた以下のような尤度関数を導入した [24]．

$$\mathcal{L}(\boldsymbol{h}_t; \boldsymbol{s}_t, \boldsymbol{\lambda}^{(z)}, \boldsymbol{\lambda}^{(g)}) \triangleq P(\boldsymbol{s}_t|\boldsymbol{h}_t, \boldsymbol{\lambda}^{(z)})P(\boldsymbol{h}_t|\boldsymbol{\lambda}^{(g)})^{\alpha} \quad (5)$$

$\boldsymbol{\lambda}^{(g)}$ はジェスチャモデルのモデルパラメータである．声質変換において，右辺第 1 項は入力発話と出力発話における内容の同一性を保証する変換モデルであり，第 2 項は出力音声における話者性を表現する話者モデルに相当する．そしてそれらのバランスをとるため，齋藤らは話者モデルを α 乗している．S2H システムでは，α はジェスチャモデルの重みに相当する．α が小さいと，ジェスチャの自然性が考慮されず，形成困難なジェスチャが推定される可能性がある．しかしジェスチャモデルは音声とは独立にモデル

化されるため，α が大きくなると，入力音声の変化によるジェスチャの変化が過小評価されることにつながる．本研究では，ジェスチャモデルは混合正規分布でモデル化されているため，より平均的なジェスチャが推定されることになる．その結果，動きのある入力音声を与えられたとしても，より平均的でかつ動きの少ないジェスチャ遷移となることが予想される．予備的な検討から，本システムでは α を 1 に設定した．

式 (5) で表される尤度関数は，その対数の $\boldsymbol{h}_t$ に対する微分を 0 としても，陽な解が得られない．そこで齋藤らは，EM アルゴリズムによって逐次的に最適解を求める手法を提案している [24]．本研究でもその手法に従い，以下の式 (6) および式 (7) を交互に計算することで，最適解 $\hat{\boldsymbol{h}}_t$ を得る．

$$\gamma_{m,t} = P(m|\boldsymbol{h}_t, \boldsymbol{\lambda}^{(z)}),\ \gamma_{n,t} = P(n|\boldsymbol{h}_t, \boldsymbol{\lambda}^{(g)}) \tag{6}$$

$$\hat{\boldsymbol{h}}_t = \Big(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m'^{(h)-1} + \alpha \sum_{n=1}^{N} \gamma_{n,t} \Sigma_n^{-1}\Big) \times \Big(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m'^{(h)-1} \boldsymbol{E}_{m,t}'^{(h)} + \alpha \sum_{n=1}^{N} \gamma_{n,t} \Sigma_n^{-1} \boldsymbol{\mu}_n\Big) \tag{7}$$

ここで $\boldsymbol{\mu}_n$ および $\boldsymbol{\Sigma}_n$ はジェスチャモデル GMM の n 番目の分布の平均ベクトルおよび分散共分散行列である．また $\boldsymbol{E}_{m,t}'^{(h)}$，$\boldsymbol{D}_m'^{(h)-1}$ は以下のように表される．

$$\boldsymbol{E}_{m,t}'^{(h)} = \boldsymbol{\mu}_m^{(h)} + \boldsymbol{\Sigma}_m^{(hh)} \boldsymbol{\Sigma}_m^{(sh)+} (\boldsymbol{s}_t - \boldsymbol{\mu}_m^{(s)}) \tag{8}$$

$$\boldsymbol{D}_m'^{(h)-1} = \big[\boldsymbol{\Sigma}_m^{(hh)} - \boldsymbol{\Sigma}_m^{(hs)} \boldsymbol{\Sigma}_m^{(ss)-1} \boldsymbol{\Sigma}_m^{(sh)}\big]^{-1} - \boldsymbol{\Sigma}_m^{(hh)-1} \tag{9}$$

ただし，$(\cdot)^+$ は一般化逆行列を表す．式 (6) の初期値については，パラレルデータに対して式 (3) より直接 S2H システムを構築し，入力音声 $\boldsymbol{s}_t$ を変換して初期値 $\boldsymbol{h}_t = \mathcal{G}(\boldsymbol{s}_t)$ を求め，これを式 (6) に代入する．このようにして構築された S2H システムに，鼻子音音声を入力することで，鼻子音音声に相当するジェスチャを推定する．次章では，この手法の有効性を実験的に検証する．

4. 実験

4.1 実験の流れ

鼻子音に相当するジェスチャの推定手法として，提案手法が有効であることを確認するため，以下のとおり実験を行った．実験手順を**図 1** に示す．まず母音のパラレルデータによって学習した変換モデル $P(\boldsymbol{s}|\boldsymbol{h})$ と，ジェスチャモデル $P(\boldsymbol{h})$ を用いて，前節の手法によって S2H システムを構築する．その S2H システムに，鼻子音音声を入力することで，鼻子音音声に相当するジェスチャベクトル時系列を得る．このジェスチャベクトル時系列が H2S システムにおいて適切な音声を推定することを確認するため，S2H システムと同様の枠組み・母音のパラレルデータを用いて H2S システムを構築し，それに S2H システムによって推定されたジェスチャベクトル時系列を入力することで，提案するジェスチャデザインの有効性を検証する．

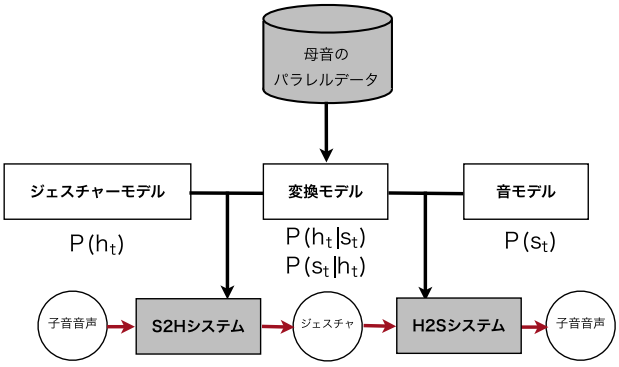


図 1 実験手順
Fig. 1 The procedure of the experiments.

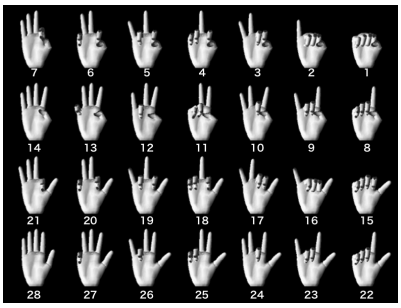


図 2 基本的な 28 種類のジェスチャ
Fig. 2 The 28 basic hand gestures.

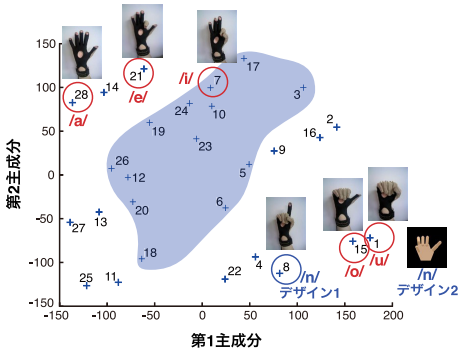
4.2 音声からジェスチャへの変換

まず母音に相当するジェスチャデザインを設定し，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を推定した．S2H システム構築に用いたジェスチャモデルおよび変換モデル用の学習データを以下に示す．

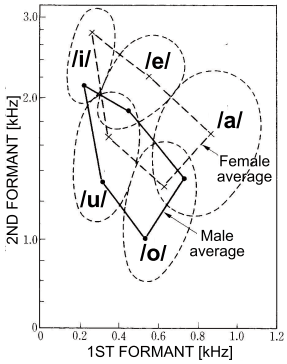
4.2.1 ジェスチャモデル用学習データ

Wu らが画像認識における論文 [25] の中で使用した基本的な 28 個のジェスチャを**図 2** に示す．これは，5 指各々の曲げ延ばしの組合せ $2^5 = 32$ 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，すなわち実現不可能な 4 種類を差し引いたものである．このうち，本システムのユーザである成人女性にとって形成の比較的容易であった，No.1，2，4，7，8，9，11，13，14，15，16，21，22，25，27，28 の計 16 個を，日本語 5 母音に相当するジェスチャの候補とした．

次にジェスチャデータセットとして，これら合計 16 種



(a)



(b)

図 3 (a) PCA 空間における 28 ジェスチャの位置，(b) 日本語 5 母音の F1/F2

Fig. 3 (a) The location of the 28 gestures in the PCA space. (b) The 5 Japanese vowels in the F1-F2 plane.

類のジェスチャ，およびそのうち 2 ジェスチャ間の遷移，計 $16 + {}_{16}P_2 = 256$ 個のジェスチャを，Immersion 製データグローブ CyberGlove を使用して記録した．CyberGlove は，各関節に取り付けられたセンサにより，人差し指，中指，薬指，小指の第 1 関節を除く 18 個の関節の曲げ角度を，それぞれ 8 bit の値として出力するものである．指の曲げ角度はそれぞれ独立ではないから（たとえば小指の第 2 関節を曲げると，薬指の第 2 関節も曲がる），これら 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，記録したすべての 18 次元データを用い，各データに対し主成分分析（以下 PCA）を行った．PCA 後の 18 次元データをジェスチャの特徴量とし，これを用いてジェスチャモデル $P(\boldsymbol{h})$（混合数 64）を構築した．

4.2.2 変換モデル用学習データ

ジェスチャモデル構築に用いた 16 ジェスチャの中から日本語 5 母音に対応するジェスチャの候補を設定し，それを用いて変換モデル $P(\boldsymbol{s}|\boldsymbol{h})$ を構築する．計算の便宜上，「あ」は No.28 とした．我々は先行研究で，「ジェスチャ空間中のジェスチャ群の配置」と「母音空間中の母音群の配置」（図 3）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 [17]．そこで母音図との等価性に配慮し，「い」「う」「え」「お」

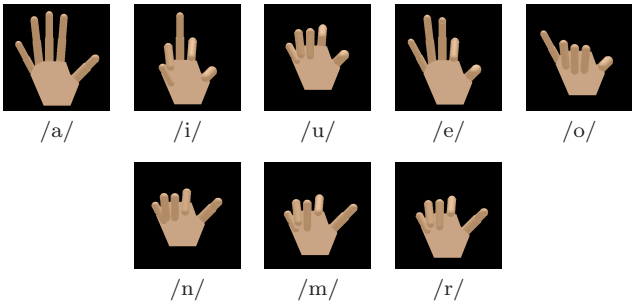


図 4 S2H システムを構築した日本語 5 母音に相当するジェスチャと，それによって推定された子音に相当するジェスチャの例

Fig. 4 An example of a gesture design for vowels which was used to train an S2H system and a derived gesture design for consonants.

に対応するジェスチャの組合せのうち，ジェスチャ空間内のユークリッド距離において，$d_h(a,i) < d_h(a,e)$ および $d_h(a,u) < d_h(a,o)$ となるデザインは候補外とした．ただし $d_h(x,y)$ は音素/x/と音素/y/に相当するジェスチャベクトル間のユークリッド距離を表す．これにより 5 母音に相当するジェスチャデザインの候補は，8,190 通りとなった．

各ジェスチャデザインに対し，以下のように音声からジェスチャへのメディア変換を実装した．変換モデル $P(\boldsymbol{s}|\boldsymbol{h})$ を構築するための学習データとして，上記のジェスチャデータセットから，それぞれのジェスチャデザインにおいて「あ」「い」「う」「え」「お」および 2 母音間の遷移に相当する ${}_5P_2 = 20$ 個のデータを抽出した．サンプリング周期は 10〜20 ms である[*2]．また成人男性 1 名から収録した，「あ」「い」「う」「え」「お」および 2 母音間の遷移 ${}_5P_2 = 20$ 組，計 $5 + 20 = 25$ 個の音声データから，STRAIGHT [26] を用いて分析を行い，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 8 ms とした．そしてジェスチャデータと音声データから結合ベクトルをつくるために，ジェスチャデータ時系列を，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．これらの結合ベクトルを用いて変換モデル $P(\boldsymbol{s}|\boldsymbol{h})$（混合数 8）を構築した．

これらを用い，$P(\boldsymbol{s}|\boldsymbol{h})P(\boldsymbol{h})$ として構築された 8,190 通りの S2H システムに，子音として，な行，ま行，ら行の音を入力し，/n/，/m/，/r/に対応するジェスチャを推定した．すなわち計 8,190 通りの/a/，/i/，/u/，/e/，/o/，/n/，/m/，/r/のジェスチャデザインが定義されることになる．得られたジェスチャデザインの一例を図 4 に示す．

/n/，/m/，/r/のジェスチャが互いに類似していることが分かる．8,190 通りを通して，同様の傾向が見られた．これは音響特徴量空間内で近接しているこれらの音が，ジェスチャ空間内でも近接する位置に変換されたためと考えら

[*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．

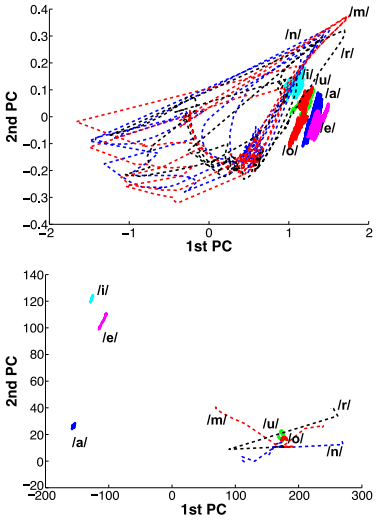


図 5 ケプストラム空間における子音の位置（上）と推定されたジェスチャのジェスチャ空間における位置（下）

Fig. 5 The location of consonants in the cepstral space (top) and that of derived gestures in the gesture space (bottom).

れる（図 5）．これらの差異を明確にする特徴量の選択は，今後の課題である．

一方で，調音運動の習得が不十分な日本語学習者には，これらの音を区別して発音することが難しいこと [27] や，韓国語などの言語においては，後続する音によって，これらの音が互いに変化する現象 [28] などが見られることから，/n/，/m/，/r/は，ジェスチャだけでなく，調音運動においても区別が難しい音であることが予想される．そのため，音からジェスチャへ理想的に変換された場合でも，操作者の（ジェスチャ言語に対する）習熟度によってはこれらの音の差異が正しく知覚されない可能性がある．すなわち，本システムを実用化する際は，ジェスチャデザインに加え，習熟度の影響を考慮する必要があると考えられる．

本稿では，ジェスチャデザインの設計のみに主眼を置き，習熟度が影響すると予想されるこれらの音の区別化については取り上げない．以降の章では，鼻音として/n/のみに注目し，推定されたジェスチャが有効であることを確認する．

4.3 ジェスチャから音声への変換

前節で得られた 8,190 通りのジェスチャデザインの比較および有効性の検証を目的に，以下のとおり実験を行った．まずそれぞれのジェスチャデザインにおいて，3.2 節で述べた手法を用いて H2S システムを構築した．すなわち，$\underset{\boldsymbol{s}}{\operatorname{argmax}} P(\boldsymbol{s}|\boldsymbol{h}) = \underset{\boldsymbol{s}}{\operatorname{argmax}} P(\boldsymbol{h}|\boldsymbol{s})P(\boldsymbol{s})$ により，ジェスチャ $\boldsymbol{h}$ に対する音声を求めるシステムである．変換モデル $P(\boldsymbol{h}|\boldsymbol{s})$ 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 $P(\boldsymbol{s})$ 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 $P(\boldsymbol{s}|\boldsymbol{h})$ や $P(\boldsymbol{h}|\boldsymbol{s})$ が 2 つの空間の特徴量を完全に対応付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，より良い変換を実現するジェスチャデザインと判断した．

8,190 通りのジェスチャ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 6 に示す．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．8,192 通りのジェスチャ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」の各モーラごとに 8,190 通りのジェスチャ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.7，「う」が No.1，「え」が No.21，「お」が No.15 の場合であった．図 6 には，準最適な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 5 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．8,190 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分かる．

また準最適なデザインにおける「ね」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 7[*3]に，子音部分/遷移部分/母音部分における両合成音の周波数特性を図 8 に示す．合成音は分析再合成音と比較して，スペクトルが平坦になっていることが分かる．統計的な声質変換法 [12] においては，変換音声のスペクトルが平坦になるという問題が指摘されている [13]．この手法を異メディア間の変換に応用した本システムにおいても，同様の問題が現れたと考えられる．この影響を軽

[*3] 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．

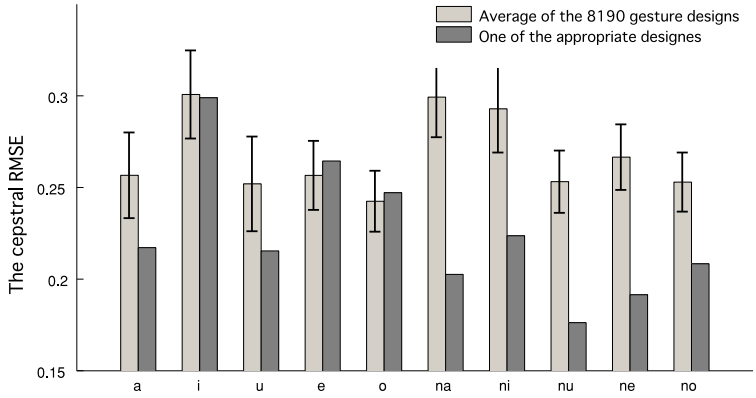
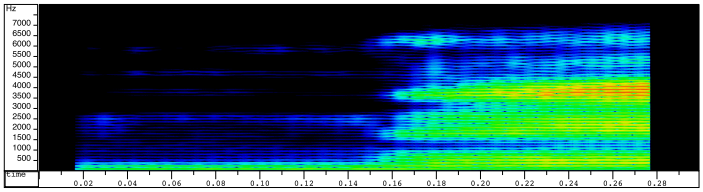


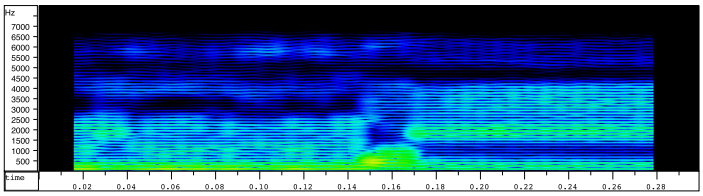
図 6 S2H システムに入力した音声と，その音声から S2H システムによって推定されるジェスチャを，H2S システムに入力した場合に推定される合成音のケプストラム平均自乗誤差

Fig. 6 The cepstral RMSE between input and output speech.



(a) S2H システムに入力した分析再合成音

(a) Re-synthesized speech used as input for an S2H system.



(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

(b) The output of the S2H–H2S combined system.

図 7 「ね」に対応する音声

Fig. 7 Synthesized speech for /ne/.

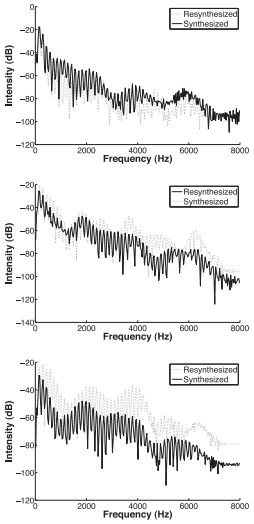


図 8 子音部分（上），遷移部分（中），母音部分（下）における分析再合成音と合成音の比較

Fig. 8 The consonant part (top), the transition part (middle) and the vowel part (bottom) of the synthesized sounds.

減するためには，変換後のスペクトルに対し補正を行うなどする必要がある．子音部および遷移部においては，フォルマントは比較的再現されている一方，2,000 kHz 付近のアンチフォルマントが再現されていない．鼻音は，口腔を閉鎖し鼻腔から音を放射するため，そのスペクトルには，声道が二又になることに起因するアンチフォルマントが現れる [29]．今回使用した音響特徴量では，鼻腔の開閉の影響を十分に表現できておらず，たとえば/n/と/u/など，鼻腔の開閉によって区別されるべき鼻音と母音が，音響特徴量空間内で近接した位置になっている（図 5 参照）．その結果，それらの音が近接したジェスチャに対応付けられてしまい，同一の GMM で近似されてしまったため，合成音のスペクトルも類似してしまったものと考えられる．また合成音のエネルギーは，子音部においては，分析再合成音より強く，母音部においては分析再合成音より弱くなっていることが分かる．本実験では，学習および変換において，音声のエネルギーに相当するケプストラム 0 次項を一定としている．すなわち，エネルギーが高い部分も低い部分も，

同じ1つのGMMで近似している．そのため変換された音声においてはエネルギーが平滑化され，その結果，本来エネルギーが小さい子音部が強められ，逆にエネルギーが大きい母音部が弱められたものと考えられる．以上のことから，特徴量として，スペクトル項だけでなく鼻腔の影響やエネルギー項を考慮する必要があることが分かる．

5. 聴取実験

提案するジェスチャデザインの有効性を検証するため，(1) ジェスチャモデルに用いたジェスチャ候補から選択されたジェスチャ（デザイン1）と (2) S2Hシステムに鼻子音音声を入力して得られたジェスチャ（デザイン2：提案手法），それぞれを用いてH2Sシステムを構築し，出力音声を聴取実験によって比較した．

まず「あ」「い」「う」「え」「お」は前節で得られた準最適なデザインを用い，それぞれ，No.28，No.7，No.1，No.21，No.15とした．これに基づき，4.2.2項と同様に，母音に相当する結合ベクトルを作った．これにデザイン1，2それぞれにおいて/n/に相当する結合ベクトルを加え，変換モデル（混合数64）を学習した．デザイン1，2それぞれにおいて，/n/に相当する結合ベクトルを得る手順を以下に示す．

5.1 デザイン1

4.2.2項で用いた音声データ「あ」「い」「う」「え」「お」および「な」「に」「ぬ」「ね」「の」を用いて，ケプストラム空間内の/n/の位置を推定し，「ジェスチャ空間内のジェスチャ配置」と「ケプストラム空間内の配置」とがより等価となるように，/n/に相当するジェスチャを選択した．

まず「な」「に」「ぬ」「ね」「の」5サンプルを視察によって，子音部分/遷移部分/母音部分に分割した．この子音部分5サンプルを/n/相当の音声と考える．この/n/と，4.2.2項で用いた音声データ「あ」「い」「う」「え」「お」との，ケプストラム空間における5つのユークリッド距離のうち，最小であったのは $d_s(n,u)$ であり，また $d_s(n,a) < d_s(n,i)$，$d_s(n,e) < d_s(n,o)$ の関係も見られた．ただし $d_s(x,y)$ は音素/x/と音素/y/の平均ケプストラムベクトル（エネルギーは一定とした）のユークリッド距離を表す．これから，ジェスチャ空間内のユークリッド距離 $d_h(n,a)$，$d_h(n,i)$，$d_h(n,u)$，$d_h(n,e)$，$d_h(n,o)$ において，$d_h(n,u)$ が最小となり，かつ $d_h(n,a) < d_h(n,i)$ および $d_h(n,e) < d_h(n,o)$ を満たすジェスチャを，/n/に相当するジェスチャとした．ジェスチャモデル構成に用いた16ジェスチャから，5母音相当のジェスチャを除いた11ジェスチャのうち，上記の条件を満たすものはNo.8のみであったため，No.8を/n/に相当するジェスチャとした．

このデザインに基づき，データグローブを装着し「な」「に」「ぬ」「ね」「の」を各々1回，計5個のデータを収録した．また成人男性1名から「な」「に」「ぬ」「ね」「の」を各々10回，計 $5 \times 10 = 50$ 個のデータを収録した．アライメントをとるため，ジェスチャ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/母音部分に分割し，4.2.1項と同様に，ジェスチャデータの再サンプリング，PCA，ケプストラム係数0–17次の抽出を行った．そしてジェスチャデータ1セット，音声データ10セットから結合ベクトルを作るために，10組すべての組合せにおいて，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．

5.2 デザイン2

4.2節で構築したS2Hシステムに，上記した「な」「に」「ぬ」「ね」「の」を入力して，それらに相当するジェスチャ遷移を得た．この変換はフレームごとに行われているため，入力した音声と出力されたジェスチャベクトルとをフレームごとに結合することで，対応付けのとれた子音のパレルデータが得られる．

5.3 結果

前述の2デザインに基づいて作られた子音のパラレルデータを，母音のパラレルデータに加え，変換モデル $P(\boldsymbol{s}|\boldsymbol{h})$（混合数32）を学習した．2デザインにおけるジェスチャの，PCA空間における位置を，図3 (a) に示す．数字は図2のジェスチャ番号に対応している．このようにして構築された2つのH2Sシステムにそれぞれ「あ」「い」「う」「え」「お」「な」「に」「ぬ」「ね」「の」に相当するジェスチャを入力して得られた音声，10組20サンプルに対し，ABプレファレンステストを行った．被験者は日本語母語話者15名である．サンプルはランダムに提示し，被験者にはより自然に聞こえたものを「A」「B」「どちらも同じ」の中から選択するよう指示した．プレファレンススコアは，デザイン1が24%，デザイン2（提案手法）が48%，どちらも同じが24%であった．これによって，S2Hシステムによって推定されたジェスチャは，ジェスチャ候補から選ばれた準最適なジェスチャと比較して，より自然な音声を出力するH2Sシステムを構築することが示された．

6. まとめ

本稿では，文字や記号を介さない音声合成方式として，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新しい音声生成系を提案した．そして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を推定する手法を提案した．聴取実験の結果，S2Hシステムによって推定されたジェスチャは，ジェスチャ候補から選ばれた準最適なジェスチャと比較して，より自然な音声を

出力する H2S システムを構築することが示された．一方で，類似したジェスチャが推定されるという問題や，変換されたスペクトルやエネルギーが平坦になるという問題，アンチフォルマントが正しく表現されないという問題なども指摘された．適切な特徴量の選択，変換後のスペクトルの補正，鼻腔の影響やエネルギー項の検討は今後の課題である．また本稿では，システムのジェスチャデザインの観点から，提案手法の有効性を検証した．今後は，使用者の習熟度による影響も考慮し，/m/や/r/が合成可能であることを検証する必要がある．さらに摩擦音/破擦音/破裂音の生成，ジェスチャ以外の入力方式などについても検討したい．

参考文献

[1] 緒方公一，山下健太郎，掛谷拓史，広瀬　賢，中島邦久：声道形状マッピングインタフェースのコンバータとしての応用，日本音響学会春季講演論文集，2-7-3, pp.283–286 (2011).

[2] d'Alessandro, N. and Dutoi, T.: Handsketch: Bi-manual control of voice quality dimensions and longterm practice issue, *QPSR of the Numediart Research Program*, Vol.2, No.2 (2009).

[3] Nordstrom, K., Fels, S., Hassall, C.D. and Pritchard, B.: Developing vowel mappings for an interactive voice synthesis system controlled by hand motions, *Journal of Acoustic Society of America*, Vol.127, No.3, p.2021 (2010).

[4] 荒井隆行：声道形状を単純化した模型による音声の音響教育，電子情報通信学会技術研究報告，Vol.109, No.10, pp.7–12 (2009).

[5] 坂田　聡，佐伯勇哉，柴田　航，上田裕市：母音発声のリアルタイム視聴覚フィードバックのための正規化構音空間の検討とその応用，電子情報通信学会技術研究報告，Vol.111, No.225, pp.55–60 (2011).

[6] 藪謙一郎，伊福部達，青村　茂：ポインティングデバイスを利用した音声生成方式：発話障害者のための支援機器として，日本保健科学学会誌，Vol.12, No.1, pp.49–57 (2009).

[7] 藪謙一郎，伊福部達：構音機能障害者のための音声生成器の抑揚制御方式に関する基礎的検討，電子情報通信学会技術研究報告，Vol.111, No.225, pp.43–48 (2011).

[8] 市川　熹，手嶋教之：福祉と情報技術，オーム社 (2006).

[9] 井上剛伸：重度障害者の自立移動を支援する技術の開発，第 23 回国立身体障害者リハビリテーションセンター業績発表会 (2006).

[10] 竹内正実：テルミン—エーテル音楽と 20 世紀ロシアを生きた男，岳陽舎 (2000).

[11] Stylianou, Y., Cappe, O. and Moulines, E.: Continuous probablistic transform for voice conversion, *IEEE Trans. Speech Audio Process.*, Vol.6, pp.131–142 (1998).

[12] Kain, A. and Macon, M.W.: Spectral voice conversion for text-to-speech synthesis, *Proc. ICASSP*, Vol.1, pp.285–288 (1998).

[13] Toda, T., Black, A.W. and Tokuda, K.: Voice conversion based on maximum likelihood estimation of spectral parameter trajectory, *IEEE Trans. Audio, Speech and Language Processing*, Vol.15, No.8, pp.2222–2235 (2007).

[14] Toda, T. and Tokuda, K.: Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory HMM, *Proc. ICASSP*, pp.3925–3928 (2008).

[15] Nakamura, K., Janke, M., Wand, M. and Schultz, T.: Estimation of fundamental frequency from surface electromyographic data: EMG-to-F0, *Proc. ICASSP*, pp.573–576 (2011).

[16] 國越　晶，喬　宇，峯松信明，広瀬啓吉：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).

[17] Kunikoshi, A., Qiao, Y., Minematsu, N. and Hirose, K.: Speech generation from hand gestures based on space mapping, *Proc. INTERSPEECH*, pp.308–311 (2009).

[18] 町田　健（編），猪塚　元，猪塚恵美子（著）：日本語音声学のしくみ，研究社 (2003).

[19] 匂坂芳典，東倉洋一：規則による音声合成のための音韻時間長制御，電子情報通信学会論文誌 A，Vol.J67-A, No.7, pp.629–636 (1984).

[20] ジャック・ライアルズ：音声知覚の基礎，海文堂 (2003).

[21] 藪謙一郎，伊福部達，青村　茂：発話障害者支援のための音声合成器の基礎的設計，電子情報通信学会技術研究報告，Vol.105, No.686, pp.59–64 (2006).

[22] 國越　晶，喬　宇，峯松信明，広瀬啓吉：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，電子情報通信学会技術研究報告，Vol.109, No.260, pp.43–48 (2009).

[23] Brown, P.F., Cocke, J., Pietra, S.A.D., Pietra, V.J.D., Jelinek, F., Lafferty, J.D, Mercer, R.L. and Roosin, P.S.: A statistical approach to machine translation, *Computational Linguistics*, Vol.16, No.2, pp.79–85 (1990).

[24] 齋藤大輔，渡部晋治，中村　篤，峯松信明：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).

[25] Wu, Y., Lin, J. and Huang, T.S.: Analyzing and Capturing Articulated Hand Motion in Image Sequences, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.27, No.12, pp.1910–1922 (2005).

[26] Kawahara, H., Masuda-Katsuse, I. and de Cheveigne, A.: Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction, *Speech Commun.*, Vol.27, pp.187–207 (1999).

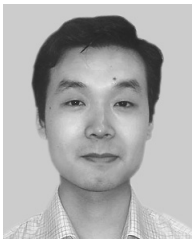
[27] 国際交流基金：音声を教える，ひつじ書房 (2009).

[28] 金裕　鴻：しっかり学ぶ韓国語：文法と練習問題，ベレ出版 (1999).

[29] 中尾睦彦，稲川千津，福原　始：白色ガウス雑音を用いた日本語音声の合成，明石工業高等専門学校研究紀要，pp.6–13 (2006).

國越 晶　（学生会員）

2009 年東京大学大学院新領域創成科学研究科基盤情報学専攻修士課程修了．修士（科学）．現在，同工学系研究科博士後期課程に在籍．メディア変換に関する研究に従事．IEEE，ISCA，電子情報通信学会，日本音響学会各会員．



喬 宇

2006 年電気通信大学大学院情報システム学研究科博士課程修了．博士（工学）．現在，中国科学院深セン先進技術研究院准教授．画像処理，コンピュータビジョン，音声工学，統計学習に関する研究に従事．

齋藤 大輔 （正会員）

2006 年東京大学工学部電子情報工学科卒業．2011 年同大学大学院工学系研究科電気系工学専攻博士課程修了．博士（工学）．2010 年から 2011 年まで日本学術振興会特別研究員（DC2）．現在，東京大学大学院情報理工学系研究科システム情報学専攻助教．音声合成，声質変換技術を中心として，広く音響分析，話者認識，音声認識等の音声言語情報処理の研究に従事．IEEE，ISCA，日本音響学会，電子情報通信学会，人工知能学会，映像情報メディア学会各会員．

峯松 信明 （正会員）

1995 年東京大学大学院工学系研究科博士課程修了．博士（工学）．同年豊橋技術科学大学情報工学系助手．2000 年東京大学大学院工学系研究科助教授．2002 年スウェーデン国王立工科大学客員研究員．現在，東京大学大学院工学系研究科教授．音声科学から音声工学に至るまで幅広く音声コミュニケーションに関する研究に従事．IEEE，ISCA，IPA，CALICO，電子情報通信学会，日本音響学会，人工知能学会，日本音声学会，日本音声言語医学会，外国語教育メディア学会等各会員．

広瀬 啓吉 （正会員）

1972 年東京大学工学部電気工学科卒業．1977 年同大学大学院博士課程修了．工学博士．同年東京大学工学部電気工学科講師．1994 年同電子工学科教授．1996 年東京大学大学院工学系研究科電子情報工学専攻教授．1999 年同新領域創成科学研究科教授．2004 年 10 月より同情報理工学系研究科教授．1987 年米国 MIT 客員研究員．音声言語情報処理分野一般についての研究開発に従事，特に韻律に着目した研究．IEEE，米国音響学会，ISCA（Board メンバ），電子情報通信学会（フェロー），日本音響学会，人工知能学会，言語処理学会，信号処理学会各会員．