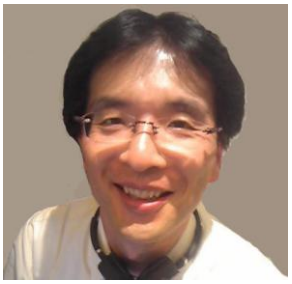
構造不変性に基づく音声の構造的表象とそれを用いた話者性に頑健な音声処理系の構築

Proposal of a new speech representation based on structural invariance and its application to robust speech processing systems against speaker variability



峯松 信明 (Nobuaki MINEMATSU, Dr. Eng.)
東京大学大学院工学系研究科教授
(Graduate School of Engineering, Professor, The University of Tokyo)

IEEE ISCA IPA 電子情報通信学会 情報処理学会 日本音響学会 人工知能学会 音声学会 音声言語医学会 発達心理学会 外国語教育メディア学会　会員

受賞：2007 年 Best Paper AwardThe Research Institute of Signal Processing　2007 年 全国大会優秀賞 人工知能学会

著書：音声言語処理と自然言語処理（共著）コロナ者 2013 年発刊予定　韻律と音声言語情報処理（共著）丸善 2006 年 人と共存するコンピュータ／ロボット学（共著）オーム社 2004 年 音声認識システム（共著） オーム社　2001 年

研究専門分野：音声コミュニケーション

あらまし　音声言語を聴取すると通常，誰が、何を、どのように話したのか、などの様々な情報が話し手から聞き手に伝わる。空気の粗密波でしかない音声信号のどの部分に、これら情報が符号化されているのか、を見極めることは音声技術の高度化には不可欠である。例えば音声認識を例にとれば「音声の知覚過程は位相の変化に鈍感である」「音韻的特徴は韻律的特徴(音の高さなど）とは独立である」事実から、音声信号から位相成分、調波成分を除去してスペクトル包絡を抽出することが一般的である。しかしながら、スペクトル包絡は話者の体格、年齢、性別によって大きく変動する。本研究では位相・調波同様、話者成分を除去した上で音声（の言語的情報のみ）を表象する技術を提案し、それを音声認識、合成、発音評価などに応用することを目的とする。研究を遂行する中で、「音声から話者成分を除去して言語情報のみに着眼する情報処理能力」が、発達心理学的、及び、進化心理学的に非常に重要な意味を持つ、凡そ人間特有の能力であることに気付かされ、この能力が言語獲得や言語発生において不可欠な能力であると考えるに至った。

１．序論

音声認識にしろ、音声合成にしろ、スペクトル包絡が基本的な音声特徴量として広く使われている。しかしスペクトル包絡は話者の違いによっても大きく変動する。音声認識において話者の違いに頑健なシステムを構築する場合、1)多くの話者から音声を集め、話者性を隠れ変数として扱い音声の統計モデルを構築する、2)入力音声の話者性を常に正規化して扱う（特徴量正規化）、3)音響モデルを常時入力話者に合わせる（モデル適応）などを行い対処する。位相成分や調波成分は音響的に除去するものの、話者成分に関しては除去せずに対処しているのが現状である。音声合成に目を向ければ、合成音声の評価が、（日本語としての）自然性のみならず、学習データ提供者の個人性の再現性も評価対象となっていることから分るように、音声の言語的な情報（何を話したのか）と、非言語的な情報（例えば誰が話したのか）を音響的に分離することなく技術構築が行なわれている。

ここで幼児の言語発達を考える。子供は親の発話を真似て言葉を獲得するが（音声模倣）、この時声帯模写はしない。話者性までは真似ない。話者成分には鈍感な模倣をする。人の音声知覚が位相成分に鈍感なように、子供の言語獲得（音声模倣）は、話者成分には鈍感である。子供は親の声の何を真似ているのだろうか？当然、音韻を正確に把握することは困難であるため、音声を音韻系列に変換し、それを読み上げるような過程を考えることは不適切である[(1),(2)]。結局、ある発声から話者成分をそぎ落とし、発話内容（メッセージ）に相当する「音のある側面」だけに着目できる情報処理能力を仮定することになる。

ちなみに、他者の発話を真似る際に声帯模写が基本となるケースは、重度の自閉症者に見られる。しかしこの場合、音声言語の獲得は一般に困難となる[(3),(4)]。動物にまで目を向けると、他個体の発声を真似る（音声模倣）行為は幾つかの種で観測されているが、それは音響的模倣である[(5)]。いずれも、模倣対象は音であ

構造不変性に基づく音声の構造的表象とそれを用いた話者性に頑健な音声処理系の構築

Proposal of a new speech representation based on structural invariance and its application to robust speech processing systems against speaker variability

って、音が運ぶメッセージではないのだろう。発達や進化というコンテキストをふまえて（人間の）音声コミュニケーション能力を計算機上に実現することを目的とした場合、体格や年齢による声色・音色のバイアスを取り去り、メッセージに直接対応する「音の側面」をあぶり出す音声表象（特徴抽出）技術が不可欠であると考えるに至った。現在の音声合成技術は、実現されている機能を考えれば声帯模写技術と呼ぶべき技術である。それをヒューマノイドに搭載したとしても、言語障害者や動物を模擬したシステムとして考えた方が適切であると考察することもできる。

本稿では、1)話者の違いに不変な音声特徴量の導出方法、2)音声認識や発音評価などに応用した場合の効果、3)発達心理学や進化心理学的の観点から考える本研究の意味付け、について述べる。詳細については参考文献を参照して戴きたい。

２．話者不変量の理論的導出とその数学的導出[6]-[8]

スペクトル包絡に代表される音声の音響的特徴は、音声の時系列性を表現するために、数十 msec ほどの時間幅に対して包絡特性を抽出し（フレーム）、音声をフレームの時系列として表すことが多い。従来提案された話者の違いに不変な音声特徴量は全て、「個々のフレーム特徴量を如何にして話者不変量として定義し直すか」に関する検討である[9],[10]。話者の差異による声色の変化に対して簡素な変換関数を仮定し、この変換関数に対して不変な特徴量を導出している。本研究で提案する話者不変量は、従来研究とは発想の着眼点（理論的導出）も技術的実装も大きく異なる。

幼児の音声模倣・言語獲得を再度考える。幼児は親の発話の話者性までは真似ないが、方言性は真似てくる。親の声が太いからといって太い声を出そうとはしないが、地方訛りの様子は真似される。日本語では方言性はアクセントに出やすいが、より一般的には（外国語まで考慮すれば）地方訛りは母音の音色変化として議論されることが多い。母音体系の変化である。**図１**に米語の幾つかの母音が地方訛りによってどのように変化するのかを第一・第二フォルマント平面で示している[11]。話者の違により各母音の絶対座標は変化するが、各話者の母音体系は（同一方言内の話者間であれば）、およそ類似した母音配置を示す。即ち、子供は親の発声の「音」を真似るのではなく、「音の体系」を真似ると解釈できる。類似した例は音階の地方性にも現れる。例えば**図２**は西洋音楽の長音階とアラビア音階を示している。アラビア音階でメロディーを奏でれば、旋律に「アラブらしさ」が加わるが、この「アラブらしさ」は**図２**の音の体系がもたらす効果である。音楽の世界では、調（ハ長調・ト長調など）に不変に

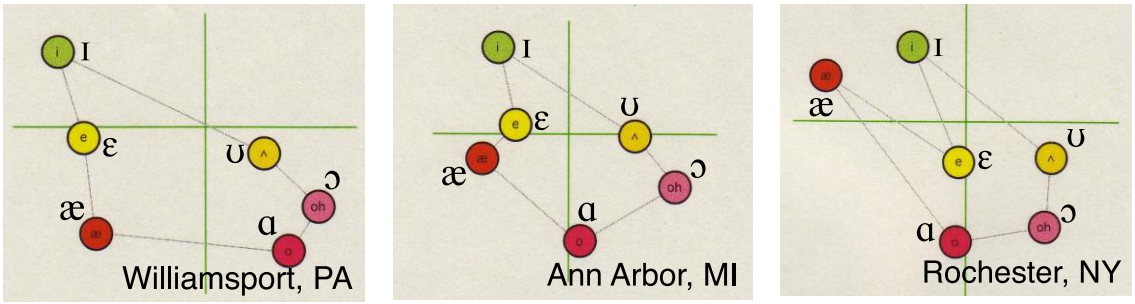


図１　幾つかの米語方言における幾つかの母音（正規化された F1/F2 を用いた散布図）

構造不変性に基づく音声の構造的表象とそれを用いた話者性に頑健な音声処理系の構築

Proposal of a new speech representation based on structural invariance and its application to robust speech processing systems against speaker variability

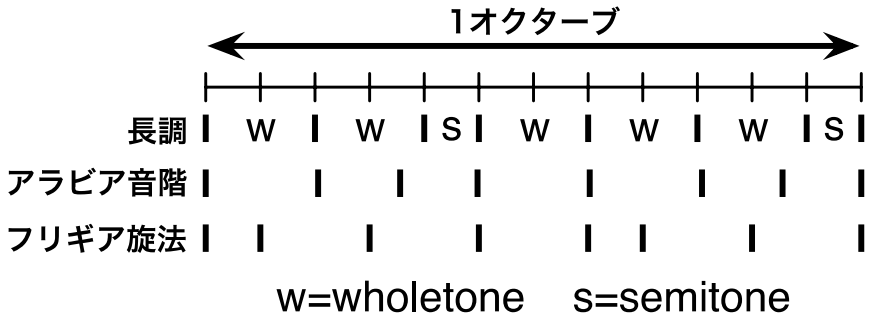


図２　西洋の現代音楽における長音階とアラビア音階

メロディーを記録するために階名がある。キーを上げ下げしても変わらない個々の音の命名法であり、これは音と音の相対的な差に基づいて命名される（相対音感）。本研究で目指した話者不変量は（音高ではなく）音色・声色の相対音感の技術的実装である。個々の音の絶対的属性ではなく、音の体系に基づく相対的属性を用いて音声の情報処理を行なう枠組みの構築である。理論的には、個々の音への着眼は音響音声学に基づく音の捉え方であり、音の体系への着眼は構造的音韻論[12]に基づく音の捉え方である。構造的音韻論は（非常に古い）言語学の一分野であるが、本研究は、それを物理的・音響的に再定義する試みであると言える。

音高は基本周波数という一次元物理量で定義できるため、音高の相対音感は音と音との基本周波数差をそのまま計測すれば良い。一方音色は多次元の特徴量であるため音高の相対音感とは異なる扱いが必要になる。まず「複数の事象群から成る体系」を数学的にどのように定義するのか、を考える。多次元ユークリッド空間の N 点からなる多角形は全ての対角線の長さを規定すればその多角形の形態が定義できる。即ち距離行列が多角形（体系）の形態定義となる。問題は、事象間距離を話者不変に定義できるのか、である。話者の違いを空間写像として考えれば、話者不変量は写像不変量となる。本研究では、f-divergence（以下、f-div. と略す）が可逆な任意の連続写像に対して不変であること、更に、可逆な任意の連続写像に対して不変な計量は f-div. のみであることを証明した[6]。二つの分布 $p(x)$、$q(x)$ 間の f-div. は下記で定義される。$g(x)$ は汎関数である。

$$f-div(p,q)=\oint q(x)g(\frac{p(x)}{q(x)})dx$$

ここで全ての事象を分布として扱い、分布間距離を f-div. で計測して距離行列を求めると、その写像不変性が保証される（**図３**参照）。一つの発声を話者不変・写像不変に表象することを考えると、**図４**に示すように、発声を一旦分布の系列として捉え、時間的に離れた事象間も含め、全ての事象間（分布間）距離を求め、距離行列（＝形態）として音声を捉える。この形態が不変情報となる。構造的音韻論を提唱したヤコブソンの言葉を借りれば、“the sound shape of language” を数式的に導出していることになる。本研究ではこれを、音声の構造的表象、略して音声構造と呼んでいる。

構造不変性に基づく音声の構造的表象とそれを用いた話者性に頑健な音声処理系の構築

Proposal of a new speech representation based on structural invariance and its application to robust speech processing systems against speaker variability

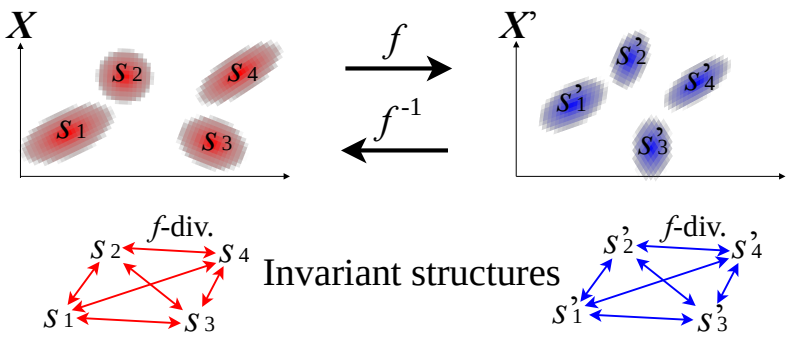


図３　不変計量である f-div. を用いた構造的な情報表象

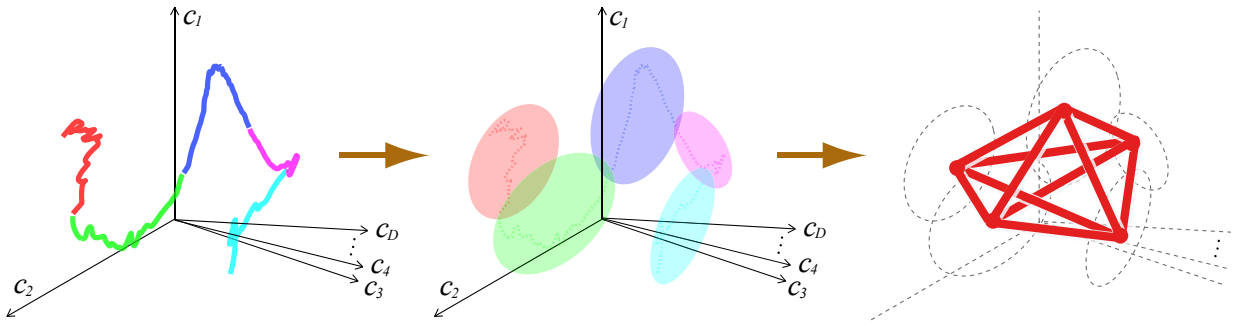


図４　一発声から構成される構造的表象

３．音声認識への応用 [6),(7),(13),(14)]

入力された発声から音声構造を抽出し、これを特徴量として用いた音声認識系を実装した。二つの方法を試みた。一方は、音声の構造的表象のみを用いた音声認識の実装であり、他方は、従来技術に対して音声の構造的表象を組み込むことで検討した精度向上である。音声の構造的表象は、スペクトルの時系列として求まる従来の音声特徴とは異なる音声特徴である。そのため、従来の仮説探索（デコーディング）技術との相性は非常に悪い。音声認識系は音響モデル、言語モデル、デコーダの三つのモジュールから構成されるが、音響モデルを構造的なモデルに置き換えても、これらモジュールをシステムとして構築することは難しい。そこで前者の検討としては、簡素な孤立単語認識をタスクとして認識系を構築し（この場合複雑なデコーダは不要）、構造的表象の効果を検討した。一方後者では大語彙連続音声認識をタスクとし、従来の音声認識系を用いて仮説を複数出力し、その仮説を再評価するリランキングの枠組みに音声の構造的表象を導入した。リランキングでは、デコーディング処理において直接利用することが難しい音声特徴・言語特徴も利用できるため、構造表象を導入する実用的な方策として検討した。

構築した構造的な孤立単語認識系を**図５**に、実験結果を**図６**に示す。語彙としては日本語の五母音「あいうえお」の母音順序を変えて定義される 120 語を語彙とした場合と、（子音も含む）音素バランス 212 語を

構造不変性に基づく音声の構造的表象とそれを用いた話者性に頑健な音声処理系の構築

Proposal of a new speech representation based on structural invariance and its application to robust speech processing systems against speaker variability

語彙とした場合を示す。比較対象としては、スペクトル時系列を隠れマルコフモデル（HMM）でモデル化した孤立単語認識系を採択した。話者の違いに対する頑健性を見るため、入力話者の体格を（人工的に）操作して、巨人や小人の音声を模擬してシステムに入力した場合の結果も示している。図６の横軸が身長に相当し、0 が通常の成人男女、負になるほど身長が伸び、正になるほど縮む。従来法では、話者正規化や話者適

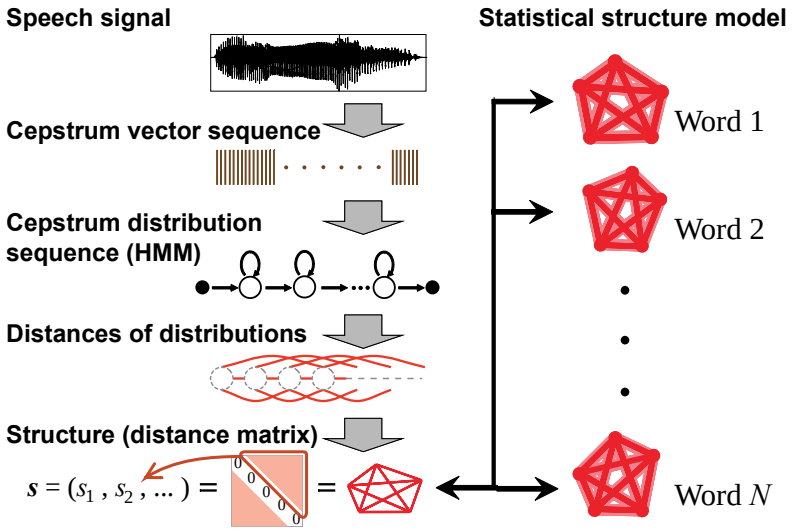


図５　構造的な孤立単語認識系

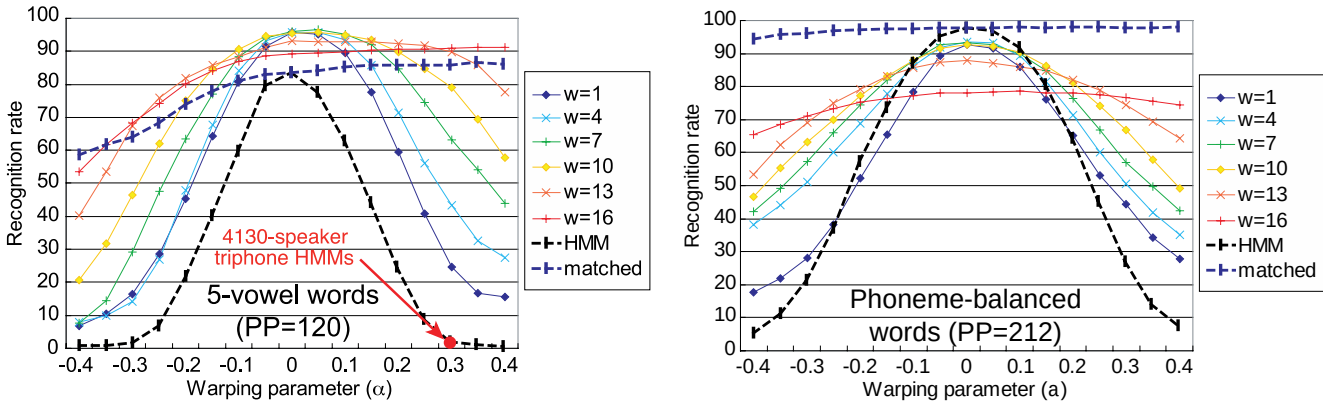


図６　母音単語（左）及び音素バランス単語（右）を対象とした孤立単語音声認識結果

応を行なわなければ、巨人や小人の音声には追従できず精度は容易に下落する。その一方で、構造表象に基づく単語認識系は、入力話者の体格によらず安定した認識性能を示している。音素バランス単語では構造表象の優位性が母音単語と比べて低減するが、話者の違いに頑健な性能は十分示されている。

リランキング処理に導入した場合のシステム構成を**図７**に、結果を**図８**に示す。この場合のタスクは（よ

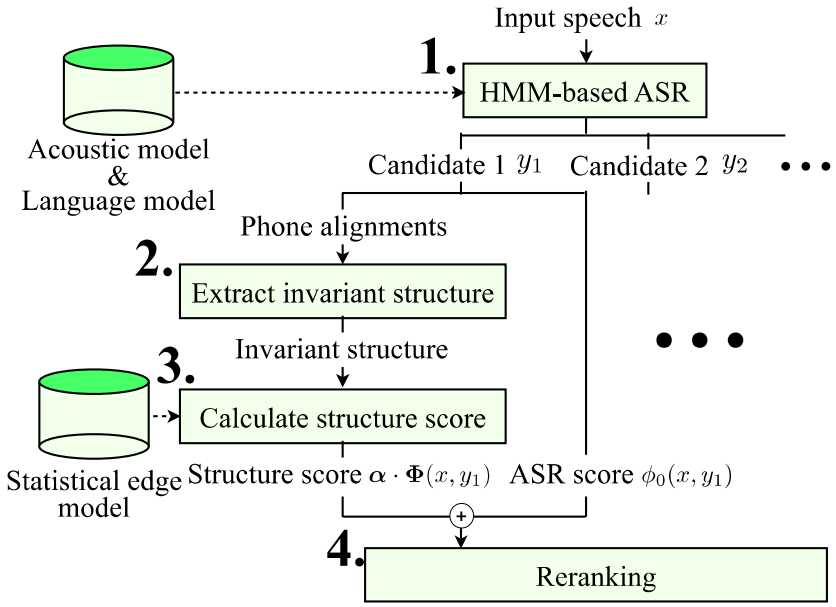


図７　リランキング処理を導入した大語彙連続音声認識系

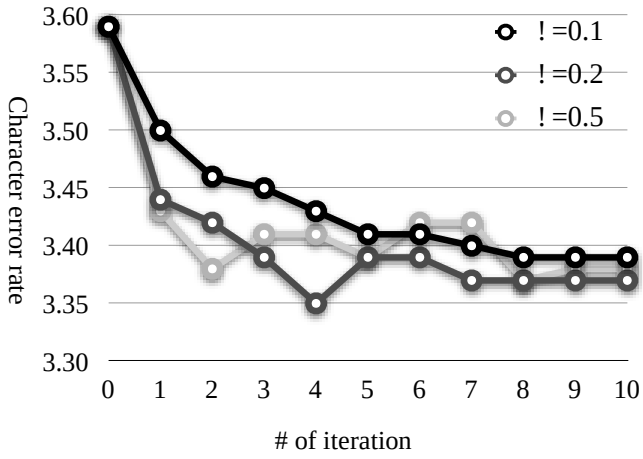


図８　構造的リランキングによる認識精度向上

り現実的な）大語彙連続音声認識である。HMM を用いた大語彙連続音声認識系より得られる複数の出力に対し、各仮説を構造的特徴に基づいて再評価し、最終的な結果を決定する。**図８**より 6.7 %の誤り削減率が得られ、本手法の有効性を示すことができた。

４．発音評価への応用 [15)-18)]

ある外国語を学ぶ学習者の発音の善し悪しを判定する場合、当然教師の音声と比較する必要がある。この場合、両者の間でスペクトル比較を行なえば、それは発音の善し悪しの評価ではなく、声帯模写の善し悪しを測ることになる。本研究では学習者や教師の音声から、体格、年齢、性別に相当する音の成分を除去して発音構造を抽出し、これを両者間で比較することで発音評価を行なう枠組みを検討した。**図９**に示す「比較したい先生を選ぶ」インタフェースは、本技術によって初めて可能になったインタフェースである。従来、教師・学習者間の話者性のミスマッチ（教師：女性、学習者：男性など）の解消手段としては、多人数の教師音声を用いて話者性を分布の中に隠す形で構築された母語話者モデルを用いることが多く、**図９**のインタフェース構築は困難であった。

発音評価及び学習者分類という二つのタスクに対して構造的表象を適用した。前者では、従来方法である音素の事後確率を用いたスコアリング技術である GOP（Goodness Of Pronunciation）手法との融合を検討した。また、音声認識実験同様、学習者の体格を人工的に操作した場合の頑健性についても検討した。結果を**図10**に示す。縦軸は教師によるマニュアル評定結果と自動評定結果との相関係数であり、これが高いほど良い技術である。GOP と構造表象とを比較すると、単体では GOP とほぼ同等の精度を示しており、両者を組み合わせることで GOP を超える精度を示している。また、巨人や小人の学習者音声を（人工的に）作成してシステムに入力すると、音声認識実験同様、GOP の精度は極端に下落する。その一方で、構造表象

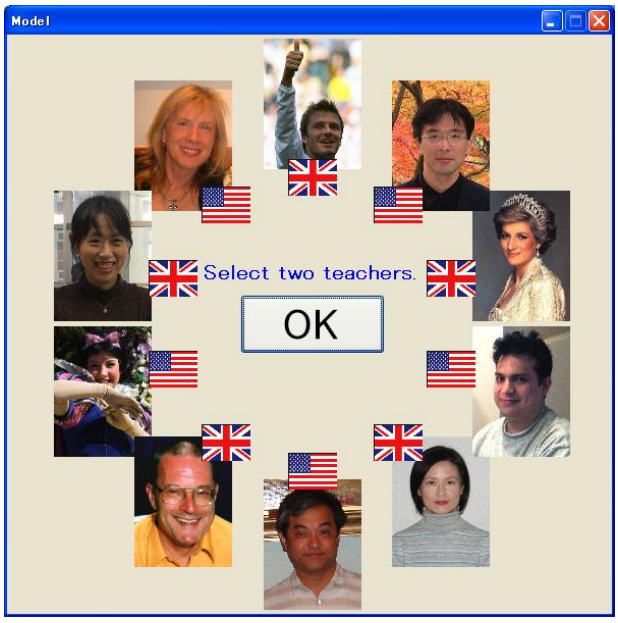


図９　比較したい教師を選ぶインタフェース

を用いた場合は極めて安定した精度を示している。

図10は学習者と教師の間の構造的差異を用いて発音習熟度を予測しているが、任意の二学習者間で両者の発音差異を定量化できれば、N 人の学習者に対して、N×N の発音距離行列が得られ、これを用いれば N 人の学習者を発音に基づいて分類する事が可能になる。12 人の帰国子女（話者 A～L）から米語発声と日本語発声を収録した。一部の米語母音を日本語母音で置き換えるなどして、一人の帰国子女から（完全な）米語発音構造、及び 7 種類の日本語訛りが一部混入した米語発音構造（合計 8 種類の発音。発音 1～発音 8）を得た。つまり、12×8=96 通りの発音構造を得た。発音間差異＝二話者間のスペクトル距離として得られた 96×96 の距離行列、発音間差異＝発音構造差異として得られた 96×96 の距離行列を用いて、96 発音の分類を行なった。**図11**にスペクトル距離に基づく分類、**図12**に構造間距離に基づく分類を示す。前者は完全な話者分類（A～L の分類）、後者は若干雑音が観測されるが、発音分類（1～8 の分類）となっていることが分る。スペクトルという音声の絶対的特性のみに着眼すれば話者が分類され、構造的特徴、即ち相対的特性のみに着眼すれば発音が分類されることとなった。

５．発達心理学や進化心理学的の観点から考える本研究の意味付け [19)-21)]

本研究では、音声から話者成分を除去して音声の言語的側面だけを直接的に表象する技術の構築を目指し、不変量の導出や音声認識・発音評価への適用を試みてきた。技術的検討を行なう中で、提案手法が古い言語学の物理的実装に相当することに気付かされ、また、発達心理学や進化心理学的に非常に興味深い検討を行なっていることに気付いた。健常者の音声言語発達と、先天的な障害のために音声言語発達に困難を示す自閉症者との音情報処理の差異、更には、人間の音情報処理と動物のそれとの差異に関する論文・図書の調査の結果、音の関係性に基づいて音声を処理する能力は、音声言語活動には必須の能力であると考えている。誌面の都合でこれらの考察についての詳細は参考文献に委ねるが、言語の起源を理論的・実験的に議論する国際会議にて発表するなどの学術活動を行なうに至っている。

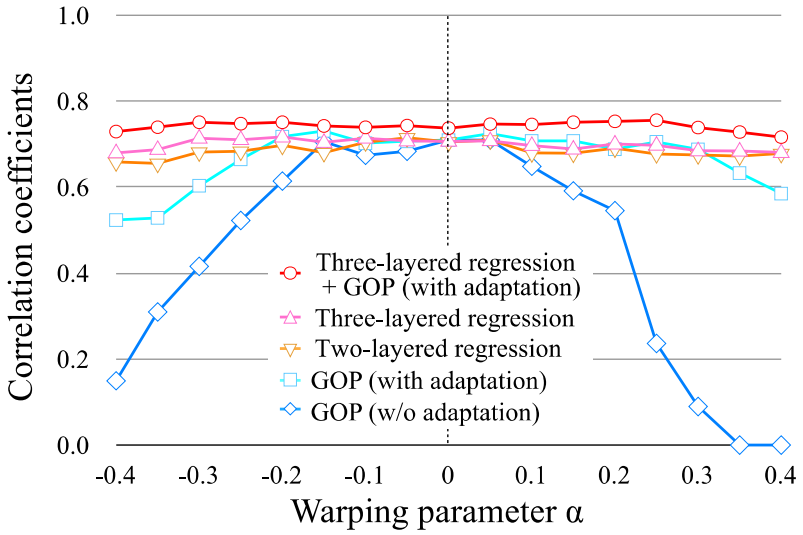


図10　構造表象と GOP を用いた発音習熟度推定結果

構造不変性に基づく音声の構造的表象とそれを用いた話者性に頑健な音声処理系の構築

Proposal of a new speech representation based on structural invariance and its application to robust speech processing systems against speaker variability

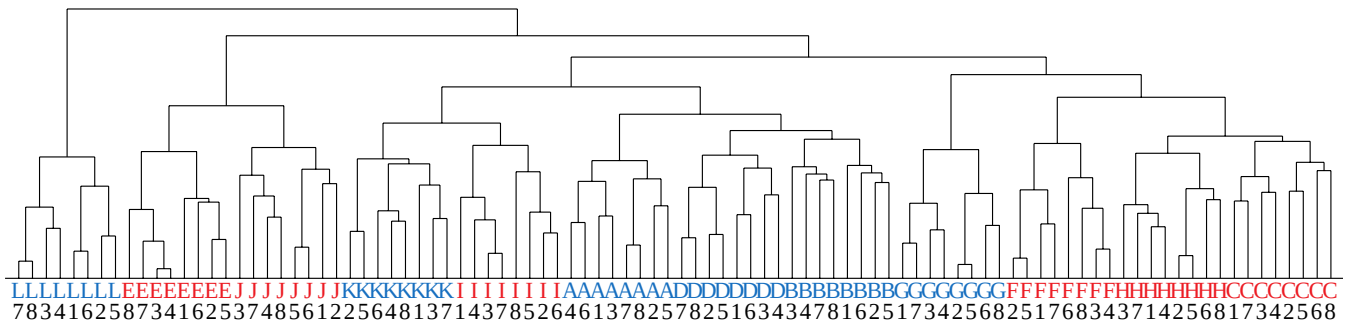


図11　発音のスペクトル比較に基づく学習者分類（A～L：話者 ID）

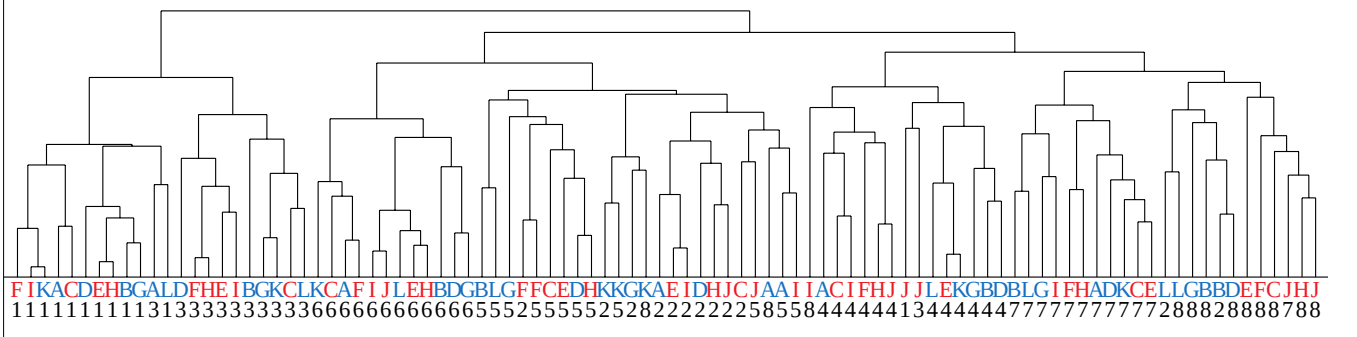


図12　発音の構造比較に基づく学習者分類（１～８：発音 ID）

６．まとめ

19 世紀のドイツに「計算できる馬」がいた。サーカスでの見せ物になっていた馬である。「3+5」と書かれた紙を見せると、蹄で地面を 8 回蹴った。調査の結果「トリック無し」との判定を得たこともあったが、精密調査の結果、正解の回数だけ蹄を蹴った時の聴衆の雰囲気（息を飲んだ雰囲気）を察知して「蹴り」を止めていたことが明らかとなり、「計算できる馬」ではなく「計算しているように見せかけることができる馬」であることが判明した。言葉を話し、聞くロボットが市場を賑わしつつある。可愛い人型ロボット（子供ロボット）に音声認識や音声合成技術を搭載して作られたロボットである。彼らが本当に「話し、聞く」ロボットなのか、「話し、聞いているように見せかけることに長けた」ロボットなのか、研究者・技術者はどこまで真剣に考える必要があるのだろうか？本研究を通して一番考えさせられたのはこの点である。「話す、聞く」ロボットとして流通させれば、子供たちは「話す、聞く」能力の技術的実装は完了したと考え、本当の意味で「話す、聞く」ことができるロボットの開発を志す人材を減らすことになるだろう。構築した技術を多角的に評価し、何が出来るのか、そして何がまだ出来ないのかを正しく、隠さず後輩に伝えることも必要なのではないだろうか。

構造不変性に基づく音声の構造的表象とそれを用いた話者性に頑健な音声処理系の構築

Proposal of a new speech representation based on structural invariance and its application to robust speech processing systems against speaker variability

参考文献

(1) 早川勝廣、“言語獲得と育児語”、月刊言語、**Vol.35、No.9**、pp.62-67、(2006)。

(2) 原恵子、“子どもの音韻障害と音韻意識”、コミュニケーション障害学、**Vol.20, No.2**, pp.98- 102, (2003)。

(3) U. Frith（著）、冨田真紀、清水康夫（訳）、自閉症の謎を解き明かす、東京書籍、(1991)。

(4) ニキリンコ、スルーできない脳-自閉は情報の便秘です-、生活書院、(2008)。

(5) 岡ノ谷一夫、“小鳥の歌と言語：共通する進化メカニズム”、音響春季講論集、**1-7-15**, pp.1555-1556, (2008)。

(6) Y. Qiao, N. Minematsu, “A study on invariance of f-divergence and its application to speech recognition,” *IEEE Trans.* on Signal Processing, **Vol.58, No.7,** pp.3884-3890, (2010).

(7) N. Minematsu, Y. Qiao, S. Asakawa, M. Suzuki, “Speech structure and its application to robust speech processing,” J*ournal of New Generation Computing,* **Vol.28, No.3**, pp.299-319,(2010).

(8) 峯松信明、櫻庭京子、西村多寿子、喬宇、朝川智、鈴木雅之、齋藤大輔、“音声に含まれる言語的情報を非言語的情報から音響的に分離して抽出する手法の提案～人間らしい音声情報処理の実現に向けた一検討～”、電子情報通信学会論文誌、**Vol.J94-D, No.1,** pp.12-26, (2011、招待論文)。

(9) T. Irino, R. D. Patterson, “Segregating information about the size and shape of the vocal tract using a time-domain auditory model: the stabilized wavelet-Mellin transform,” Speech Communication, **Vol.36,** pp.181-203, (2002).

(10) A. Mertins and J. Rademacher, “Vocal tract length invariant features for automatic speech recognition,” *Proc. IEEE Workshop on Automatic Speech Recognition & Understanding,* pp.308-312, (2005).

(11) W. Labov, S. Ash, C. Boberg, Atlas of North American English, Mouton and Gruyter, (2005)

(12) R. Jakobson, L. Waugh, 言語音形論、岩波書店、(1986)。

(13) N. Minematsu, S. Asakawa, Y. Qiao, D. Saito, and T. Nishimura, “Implementation of robust speech recognition by simulating infants' speech perception based on the invariant sound shape embedded in utterances”, *Proc. Speech and Computer,* pp.35-40, (2009).

(14) M. Suzuki, G. Kurata, M. Nishimura, N. Minematsu, “Discriminative reranking for LVCSR leveraging invariant structure, ” *Proc.* INTERSPEECH, (2012)

(15) 朝川智、峯松信明、広瀬啓吉、“音声の構造的表象 に基づく英語学習者発音の音響的分析”、電子情報通信学会論文誌、**Vol.J90-D, No.5**, pp.1249-1262, (2007)。

(16) 峯松信明、“グローバル時代における英語発音とその科学的な分析方法"、大学英語教育学会関東支部学会誌、**No.7**, pp.5-14, (2011)。

(17) 鈴木雅之、峯松信明、広瀬啓吉、“音声の構造的表象と多段階の重回帰を用いた外国語発音評価”、情報処理学会論文誌、**Vol.52, No.5**, pp.1899-1909, (2011)。

(18) 峯松信明、鎌田圭、朝川哲、牧野武彦、西村多寿子、広瀬啓吉、“音声の構造的表象に基づく学習者分類の検証と発音矯正度推定の高精度化、情報処理学会論文誌、**Vol.52, No.12**, pp.3671-3681, (2011)。

(19) 峯松信明、“「あ」という声を聞いて母音「あ」と同定する能力は音声言語運用に必要か？～音声認識研究からの一つの提言～”、日本語学 4 月号、p.187-197, 明治書院, (2008)。

(20) 最相葉月、“ビヨンド最相葉月、"ビヨンド・エジ

ソン～12 人の博士が見つめる未来～（第六章「言葉の不思議を探求する ～音声工学者・峯松信明と動物科学者テンプル・グランディンの自閉症報告～」）"、pp.119-139, ポプラ社, (2009).

(21) N. Minematsu, “A modulation-demodulation model for speech communication and its emergence," *In The Evolution of Language*, edited by T. C. Scott-Phillips, M. Tamariz, E. A. Cartmill, and J. R. Hurford, pp.234-241, World Scientific, (2012).

この研究は、平成２０年度ＳＣＡＴ研究助成の対象として採用され、平成21年度～23年度に実施されたものです。