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——

峯松 信明[†a)] 櫻庭 京子[††] 西村多寿子[†††] 喬 宇[†]
朝川 智[††††] 鈴木 雅之[†††††] 齋藤 大輔[†††††]

Proposal of a Method to Extract the Linguistic Information in Speech Based on Acoustic Separation of the Linguistic and Extra-Linguistic Aspects of Speech
——An Attempt toward Realizing Human-Like Speech Processing on Machines——

Nobuaki MINEMATSU[†a)], Kyoko SAKURABA[††], Tazuko NISHIMURA[†††], Yu QIAO[†], Satoshi ASAKAWA[††††], Masayuki SUZUKI[†††††], and Daisuke SAITO[†††††]

あらまし 近年の計算機性能の飛躍的な向上により，大規模語彙を対象とした音声認識は実用段階を迎えている．音声合成においても話者性や発話スタイルを制御できる合成方式など，種々の応用場面を念頭においた技術開発が行われている．その一方で，音声工学研究の目的を「人間に匹敵するような」音声言語情報処理能力の計算機実装と考えた場合，人間と機械との間には，今なお，大きな溝があることも指摘されている．本研究ではまず，現在の音声認識・音声合成相当の情報処理を行う人間が現に存在した場合，その人間の挙動は，音声言語の獲得に困難を示す重度自閉症者の挙動と類似するであろうことを指摘する．その上で，(定型発達を遂げた）人間らしい音声情報処理の実現に向けて，現在の音声技術に欠けている基礎技術は何であるのかを幅広い視点から考え，欠損技術の一つとして「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技術」を主張する．と同時に，その実現に向けて一つの技術的提案を行い，いくつかの実験結果を述べる．

キーワード 音響モデリング，情報分離，音声模倣，自閉症，知覚の恒常性，変換不変量，音声の構造的表象

1. まえがき

近年の計算機性能の飛躍的な向上により，音声認識・音声合成ともに技術的精度が改善され，各種の実用アプリケーションが開発されるに至っている [1]．これら音声技術の中核の一つは，「音声のどの音響的側面をどのように表現・表象すべきか」という問いに対する技術的回答である，音響モデリング技術である．隠れマルコフモデル (Hidden Markov Model, HMM) が音声認識に導入されて以来，ゆう度最大化基準，あるいは，事後確率最大化基準に基づくパラメータ推定，識別学習など，数理統計的な機械学習に基づく，より精度，識別力の高いモデル学習方法が構築されてきた．この枠組みは音声合成にも導入され，HMM に基づく音声合成は現在主流の音声合成方式の一つである．

「音声言語を操れる機械を構築する場合に，人間のアルゴリズムを真似る必要は必ずしもない」という言葉は頻繁に聞かれる [2]．しかし，採択する方法論がどうであれ，音声工学研究の究極の目的が「人間に匹敵するような音声言語情報処理を計算機に実装すること」であることは多くの研究者が同意するものと考える．それを裏づけるように，人間と機械による音声言語運

† 東京大学大学院情報理工学系研究科，東京都
Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, 113–0033 Japan
†† 獨協医科大学越谷病院，越谷市
Dokkyo Medical University Koshigaya Hospital, Koshigaya-shi, 343–8555 Japan
††† 東京大学大学院医学系研究科，東京都
Graduate School of Medicine, The University of Tokyo, Tokyo, 113–0033 Japan
†††† 東京大学大学院新領域創成科学研究科，柏市
Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa-shi, 277–8561 Japan
††††† 東京大学大学院工学系研究科，東京都
Graduate School of Engineering, The University of Tokyo, Tokyo, 113–0033 Japan
a) E-mail: mine@gavo.t.u-tokyo.ac.jp

用能力の差異に対して，これまで様々な報告が行われてきた [3]〜[5]．いずれの報告においても共通していることは「両者の間に大きな溝があることは否めない」という事実である．最近では半世紀以上にわたる音声認識研究史を踏まえた上で「何かが足りない」という言葉を残した古井による講演が記憶に新しい [5], [6]．特に [5] では，機械は人間と比較して音響的な汎化能力が非常に乏しいことを指摘している．多様に変形する声に対して頑健に動作する技術が求められている．

人間に匹敵する汎化能力を計算機実装することを考えた場合，現在の音声工学の技術体系の中に，基礎技術として何が足りないのだろうか？　音声は，時間も振幅も連続的な値を有する一次元信号（波形）として観測される．それを標本化・量子化して整数値列とし，計算機上で各種の処理が行われる．計算機にとって音声とは単なる数値列でしかないが，この数値列の中に様々な情報が埋め込まれて（符号化されて）いる．そして人間はその情報をいとも簡単に解読して（復号化して）しまう．多様な情報を適切に反映しつつ数値列を導出するのが音声合成であり，その数値列から多様な情報を的確に抽出するのが音声認識・理解である．

これらの技術を構築する場合，数値列のある側面を切り落とし，処理の効率化を図っている．例えば人間の聴覚は音声信号の位相成分には鈍感であるとの知見から，パワースペクトルのみを特徴量として使用する場合が多い．更に，音声生成を「声帯による音源生成」と「声道による共鳴」との 2 段階に分け（ソース・フィルタモデル），両者による音響特性を関数の積で表現することで，後者による音響効果のみに着眼することも頻繁に行われている．現在の音声認識技術が好例であり，音源の音響特性を切り落としたスペクトル包絡特性を基本的な音響特徴量として用いている（図 1 参照）．2 段階の分離を通して得られる包絡特性であるが，なお，様々な情報源がこの音響量を変形させる．

音声に含まれる情報は大きく言語情報，非言語情報に分類される（言語情報は文字面情報だけに限定した言語情報と，文字面では表現困難なパラ言語情報に細分化される）．スペクトル包絡（共鳴特性）は声道形状を直接反映した音響量であるが，言語，パラ言語，非言語情報のいずれによっても容易に変形を被る．

不特定話者単語音声認識，テキスト非依存話者認識を例として考える．包絡特性 o は，単語 w（言語情報），話者 s（非言語情報），いずれにも依存する．統計的音響モデルの構築を考えた場合，単語認識の場合

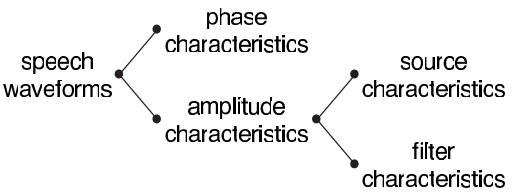


図 1　2 段階の分離に基づく特徴量抽出
Fig. 1　Feature extraction based on 2-step separation.

は $P(o|w)$ を，話者認識の場合は $P(o|s)$ を推定することになる．ここで，認識対象とは独立な要因を期待値（周辺化）操作で消失させることが広く行われている．

$$P(o|w)=\sum_{s} P(o|w,s)P(s|w)\approx\sum_{s} P(o|w,s)P(s)$$

$$P(o|s)=\sum_{w} P(o|w,s)P(w|s)\approx\sum_{w} P(o|w,s)P(w)$$

しかし言語情報（単語）と非言語情報（話者）は，そもそも独立した情報である．にもかかわらず，それらを運ぶ音響的対象物（特徴量）として，各々に対応した特徴量（o_w や o_s）を求めずに，共通項 $P(o|s,w)$ に対する期待値操作で各音響モデルを導出する．音声工学では常套手段となっているが，機械学習を専門とし，音声・話者認識は一応用としてとらえている研究者には，これを不思議な方法論と考える者もいる [7]．

人間のように汎化能力の高い音声言語情報処理の計算機実装を目的とした場合，基礎技術として何が欠けているのだろうか？　本論文は,「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技術」の欠損を主張し，これを可能にする一つの技術的提案を行う．本主張に至るには，幅広い言語・人間研究の知見を踏まえた上で現在の技術体系を見直す必要があるが，本論文の主題は，上記主張に至るまでの経緯を読者に提示することにある．欠損部分に対応する技術構築については論文後半で概説するに止める．詳細は，筆者らの他論文を参照して頂きたい．

「ある情報処理体が音声言語がない状態から，ある状態へと遷移する」場合，どのような情報処理を新たに獲得することが必要なのだろうか？　これは正答困難な問いであるが，本研究では，以下の検討を通して回答への指針を得，それに基づいて高い音響的汎化能力の技術的実現を試みる．すなわち，先天的欠損により音声言語の獲得に困難を呈する障害者と健常者の間に観測される，音情報処理のビヘービア（行動パターン）

レベルの差異（発達的側面から考察する音声言語）[8]，及び，動物と人間の間に観測される，音情報処理のビヘービアレベルの差異（進化的側面から考察する音声言語）の検討を通して，回答への指針を得る．

2. 発達的・進化的側面から考察する音響モデリングの技術的欠損

2.1 発達的側面から考察する技術的欠損

幼児の言語獲得は「音声模倣・学習」を基本とする [9], [10]．他個体の発声を積極的に模倣する行為である．ここで注意すべきは，彼らの模倣行為の音響的対象物である．幼児の音声模倣は，音響的模倣（声帯模写）ではない．音響的には彼らは何を真似ているのか？

「親の声をシンボル（音韻，平仮名）列に変換し，個々のシンボルを自らの口で生成する」という説明は不適切である．彼らは音韻意識が未熟であり，「しり取り」も困難な状況にある [11]．発達心理学の文献を調査すると，各種用語でこの模倣対象を説明している．[12] では「幼児は単語全体の語形・音形を獲得し，その後，個々の分節音を獲得する」と説明し，[13] では「語形の全体ゲシュタルトを認知する」と述べ，[14] では「related spectral pattern」と呼んでいる．これらは同一対象に対して異なる名称を用いていると解釈できるため，本論文では以下，「語ゲシュタルト」と呼ぶ．

語ゲシュタルトに話者情報が含有されていれば，幼児は音響的模倣を試みることになり，現実とは合致しない．つまり，この語ゲシュタルトは音声から話者情報が切り離された音響パターンとなる．筆頭著者は国内外の発達心理学研究者に，「語ゲシュタルト」の物理的定義の提示を促したが，明確な回答はなかった．

さて「幼児の聞く声の大半は両親の声であり，また，自らが話せるようになると，その子の聞く声の約半分は自らの声である」という記述を否定することは困難である．すなわち，人が聴取する音声の話者性は極めて偏りが大きい．そして，この話者的に偏った音声の聴取を通して，人は頑健な情報処理を獲得する．話者情報を切り落として言語的情報のみを音響パターンとして抽出する能力があれば，当然の帰結である．その一方，話者情報の分離技術が確立せず，集めることで $P(o|w)$ を推定する枠組みで不特定話者音声認識を実装すれば，話者バランスがとれた音声サンプルが必要になる．かつて IBM 社が自社製の音声認識エンジンの宣伝に用いた「集めた話者数」は 35 万人であった．

音声模倣が音響的模倣になる場合があるのだろうか？　そのような事例は（重度）自閉症者に見られる．七色の声をもつと呼ばれる声優の中村メイ子の声をそっくりまねる例 [15]，外国語発音練習やカラオケにおいて，音響的模倣以外のまね方が難しい例 [16]，相手そっくりの声を模倣する例 [17]〜[19]，音声に限らず車や列車の音など，様々な音響音を模倣する例は，自閉症関連図書において頻出する．刺激音をそのまま記憶し（そのため音響的汎化能力も低下すると考えられる），再生しようとする情報処理が主体となっているわけだが，重度自閉症者の場合「音声コミュニケーションが困難となる場合が多い」という事実は注目に値する．中には，母親の音声は正しく認識・理解できるが，母親以外の音声への対応が難しい例もある [20]．電話越しであれば母親でも難しくなるようである．

ある話者の音声を学習データとして音声合成システムを構築すれば，その話者の声が出力される．成人話者を多数集めて構築した音響モデルで子供の音声を認識すれば，認識率は下落する．音声合成，認識ともに，言語情報と非言語情報が同居したままスペクトル包絡特性の統計モデルを構築する点では同じである．つまり，音そのものを記憶・モデル化対象としている．その意味において，現在の音声認識・合成システムと自閉症者の挙動（情報処理）は類似していると考察できる．

発達的側面から考察したが，健常者の音声模倣行為と，重度自閉症者のそれとの差異を考えれば，「ある情報処理体を音声言語がない状態から有る状態へと遷移させる際に必要となる音情報処理」として考えられる回答の一つは「音声に含まれる言語情報を非言語情報から音響的に分離して抽出する処理」であると考える．

2.2 進化的側面から考察する技術的欠損

動物を対象とした場合，音声模倣はまれな行為と位置づけられている．例えば霊長類では，人のみが行う行為であると考えられている [21]．動物種の範囲を広げた場合，音声模倣を行う動物種は鳥，クジラ，イルカなどで確認されているが [22]，動物の音声模倣は音響的模倣が基本となっている [22]．また，進化人類学の実験研究によれば，人以外の霊長類は相対音感が非常に乏しく，移調前後のメロディーの同一性判定が困難であることが示されている [23], [24]（注1）．すなわち，人以外の霊長類は極端な絶対音感を有している（注2）．

自閉症（アスペルガー症候群）者として世界で初め

（注1）：ただし，1 オクターブずらすと同一性が分かるとのことである．
（注2）：彼らがメロディーを音名で記述できたり，採譜できるわけではない．違う音は違う音，と認識しているだけである．

て書籍を出版した [17] グランディンは動物学の教授であるが，彼女は，自閉症者と動物の情報処理における類似性を指摘している [25]．いずれも，入力刺激の詳細な様子をそのまま記憶・保持する傾向が強い．入力された情報を無意識的に取捨選択できず，汎化能力に乏しく，情報過多の渦に巻き込まれる様子は多くの自閉症関連図書に散見される [16], [19], [26], [27] (注3)．自閉症者の多くは絶対音感保有者である [28]．

以上，人間と動物の音情報処理の差異に関して，筆者らの文献調査の結果を述べた．音を用いた情報伝達を行う場合，情報の同一性を保証するために，音響的同一性が必要とされるのか否か，が問うべき焦点であると考える．必要であれば，音響的に同一の音を自らが生成したり，他者に要求することになる．重度自閉症者や動物の音声模倣，更には，動物における移調前後のメロディー同一性の欠損などはその良い例である．

音声認識における音響的照合とは，ある発声と別の発声（あるいは音響モデル）の言語的な同一性検証を，音響的な同一性検証を通して行う技術である．言い換えれば，2 種類の同一性を置換可能と仮定して，初めて成立する技術である．この仮定は正しいのだろうか? 身長が 2.5 m 近い世界一の巨人と 1.0 m に満たない世界一の小人が難なく会話する様子がテレビで報道されることがある．世界一の音色・声色の音響的差異をもつ両者は，それを全く気にせず会話を楽しむのである．

筆者らはこの 2 種類の同一性は置換可能なものではないと考える．(言語的) 情報の同一性を保証する場合に，音響的同一性を必要としなくなったのが人間である．進化的側面からの考察を行ったが，本節においても「音声言語がない状態からある状態へと遷移させる際に必要となる音情報処理」に対する回答の一つは「非言語情報に非依存な音響パターンを通して言語情報の同一性を検証する技術」の構築であると考える．

本論文では，以下，非言語的情報が分離された語ゲシュタルトの数学的導出を試みるが，その前に，どのような形式で導出すべきか，に関して検討を行う．求めるべきは，年齢，性別，体格といった話者特性，更には収録や伝送に用いた機器の音響特性に対して独立・不変な音響パターンであるが，このような刺激の多様性に対する認知の不変性は，心理学の世界では広く「知覚の恒常性」として知られる現象である．筆者らは音声を他の物理メディアに対して特別視すべきではないと考えており，ここでは，色やメロディーの知覚恒常性と対比しながら音声の知覚恒常性を考え，その後に，語ゲシュタルトの数学的導出を試みる．

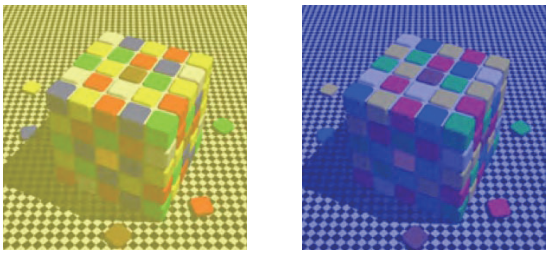


図 2　異なる色眼鏡を通して見たルービックキューブ [32]
Fig. 2　Rubik's cube seen through differently colored glasses [32].

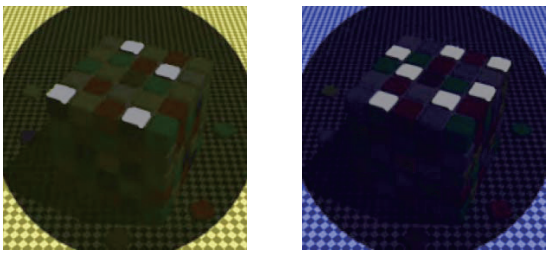


図 3　コンテクストを隠した場合の色知覚 [32]
Fig. 3　Perception of color without context [32].

3.　刺激の物理的多様性とその認知的不変性

3.1　色，メロディーに見る知覚の恒常性

同一の情報であっても，環境要因により，異なった物理量として感覚器に入力されることは頻繁に起きるが，通常，情報の同一性認知は容易である [29]〜[31]．

図 2 は，同一のルービックキューブを黄眼鏡，青眼鏡で覗いた場合の「見え」を表現している．左側が黄眼鏡，右側が青眼鏡による像であることは容易に認識できる．これは，両図において対応する各部位は，観測者の網膜に異なる波長を届けることを意味する．しかし，通常，各部位に同一の色シンボルを振り，最終的に両キューブの同一性を認識する．つまり，両図の違いを認識しつつ，同時に，同一性を認識している．

更に，左キューブ上面には四つの青部位を，右キューブ上面には七つの黄部位を認める．しかしコンテクスト情報を消失させ，対象部位を単独で観察すれば，同一色（同一波長）であることが分かる（図 3 参照）．我々は，異なる色を同一と判断し，同一色を異なると

(注3)：ある当事者は，自閉症とは「情報の便秘」である，と述べている [19]．同様に，自閉症を（人工知能の世界でいう）「フレーム問題」が解けない症状として関連づける書籍もある [27]．

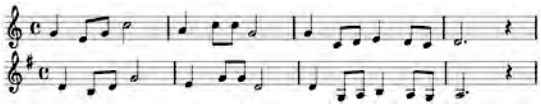


図 4　ハ長調（上）とト長調（下）の同一メロディー
Fig. 4　The same melody with different majors: C major (upper) and G major (lower).

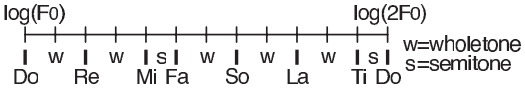


図 5　長調におけるオクターブ内の音配置
Fig. 5　Tone arrangement of a major key.

判断する．我々の認知が，個々の要素刺激の物理特性のみで行われていないことを示す好例である．

同様の認知は，音高においても観測される．図 4 に示す二つの音系列は，同一メロディーのハ長調（上）とト長調（下）であるが，両者の同一性認知は通常容易である．両メロディーをドレミで書き起こすことを要求した場合，絶対音感者であれば，個々の音を，その基本周波数に基づいて書き起こすため，前者は「ソミソド‥」となり後者は「レシレソ‥」となる．彼らにとってドレミとは音名である．その一方，相対音感者であれば，個々の音を，そのメロディー・音階における機能に基づいて書き起こすため，前者も後者も等しく「ソミソド‥」となる．すなわち，提示されたメロディーの調に対して非依存に，メロディーを書き起こす(注4)．彼らにとってドレミとは階名である．

上曲の最初の音と，下曲の最初の音の基本周波数は異なるにもかかわらず，彼らは同じ音（ソ）と判断する．更に，上曲の最初の音と，下曲の四番目の音の基本周波数は同一であるにもかかわらず，彼らは異なる音（ソとド）と主張する．色知覚と同様，異なる音高を同一と判断し，同一音高を異なると判断する．コンテクスト情報を消失させ孤立音として提示すれば，機能を知覚できず，階名として同定できない．これも色知覚と同様である．我々の認知が，個々の要素刺激の物理特性のみで行われていないことを示している．

心理学研究によれば，これら知覚恒常性は，刺激群のコントラスト（インターバル）情報を用いた処理が寄与していると考えられている [29]〜[31]．各要素刺激の物理量は容易に変形するが，対象刺激と周辺の刺激群との関係性は不変である．図 5 に長調のオクターブ内音配置を示す.「全全半全全全半」という音配置は調に対して不変であり，メロディー中の 2 音（時間的に離れていてもよい）が 3 全音の音高差をもつ場合，それらは（階名でいえば）「ファとシ」のいずれかとなる [33]．このように，調に独立な関係性を制約条件として，相対音感者はメロディーを階名で書き起こす．

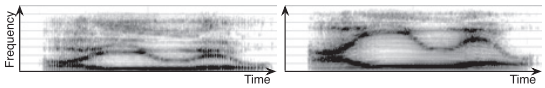


図 6　長身の話者と短身の話者による同一内容の発話
Fig. 6　Utterances of tall and short speakers.

なお，色の恒常的知覚は蝶や蜂でも観測されているが [34], [35]，音高に対する恒常的知覚は，**2. 2** に示したように，霊長類であっても人間以外では難しい．

3. 2　音声における音色知覚の恒常性

図 4 を女性と男性のハミングだとすれば，両者の違いは，声帯の長さ・重さの個人差に由来する．一方，声道の長さ・形状の個人差は，音声の音色・声色（スペクトル包絡）を大きく変形する（図 6 参照）．

色の知覚恒常性に対して「蝶や蜂は数千の色眼鏡の試着を通して各色の統計モデルを個別に構築する」と主張する仮説を筆者らは知らない．そもそも，キューブの各部位に色シンボルを振るという作業すら，両キューブの同一性認知には，本来必要ない．しかし従来の音声認識研究では，音声を音素列（シンボル列）を通して眺め，各音シンボルに対応するスペクトル包絡を数千の喉形状を通して観測し，得られた観測量を統計的に，かつ，個別にモデル化する方法（音韻の統計的音響モデル）が標準技術となっている．生態学的，進化論的，発達心理学的に考えた場合，この方法論は非常に不自然である．色や音高の知覚を参考にすれば，音色知覚の恒常性は，各音とそのコンテクストが形成する関係性に基盤を置くべきであると考える．

しかし，音声の場合メロディーとは異なり，孤立音の同定は容易である．例えば，孤立母音を同定させるタスクは容易である．しかし母音数が多い英語の場合，孤立母音同定率は 57%という報告もある [36]．このタスクは，母音カテゴリーを獲得する前の幼児であれば当然困難であるが，幼児のように音韻意識が未発達なまま成人となる例は海外では広く観測されている [37]．音韻性 Dyslexia と呼ばれ，この場合，音韻操作・文字言語使用に困難を示すが，音声言語使用には大きな困難を示さない．そもそも音声を音韻列として表現したり，それを操作する能力は文字言語使用には不可欠で

(注4)：なお，ドレミ書き起こしが困難な相対音感者もいる．

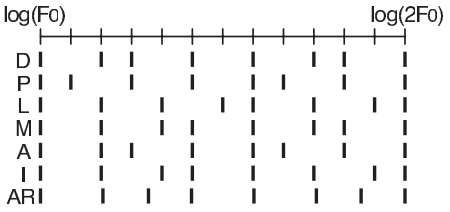
あるが，音声言語使用には必ずしも必要ではない [38].

なお，日本語であっても巨人や小人による発声を技術的に模擬した音声の場合，孤立母音の同定は難しくなる [39]．しかし，無意味モーラ列として連続提示すれば，母音同定率は有意に向上する [40]．意味や統語の情報がなくても，前後コンテクストが存在すれば同定率は向上する．筆者らはこの結果を，メロディーの階名書き起こしと類似した情報処理の結果として解釈している．すなわち音声の場合も，各音がそのコンテクストとの間に有する関係性に基づいた情報処理を実装することで，頑健な処理系構築が期待できる．その上で，音の関係性に基づいた情報処理と音の絶対的な特性に基づいた情報処理とを組み合わせることで，人間の様々な音声処理能力に対応した情報処理系を構築できると考える．次章で音群の不変な関係性に基づく「語ゲシュタルト」の数式的な定義を述べ，その後，特に音の関係性のみに基づいた情報処理系と，その高い頑健性について実験的に検討する．しかしその前に，これまでの議論を簡単な思考実験を通して総括する．

4. 思考実験を通して考察する情報の分離

4.1 音高の配置パターンと音色の配置パターン

図 5 に長調における音階の音配置を図示した．この音配置には，時代・民族に依存する形で多様な配置パターンが存在する．一例として中世の教会音楽で使用された音階とアラビア音階を図 7 に示す．D~I が教会音階であり，I と A が現代音楽でいう長調と短調である．また，AR がアラビア音階である．アラビア音階を用いて西洋音楽の曲を演奏すると，調律のずれたピアノを用いた演奏として聞こえる(注5)．しかしアラブ人に聴取させると「なじみのあるメロディー」として受け止める．つまり彼らにとって，この音配置が本来の配置である．各音の基本周波数ではなく，音の配置パターンを獲得していることは言うまでもない．



D=Dorian, P=Phrygian, L=Lydian, M=Mixolydian
A=Aeolian(短調), I=Ionian(長調), AR=Arabic

図 7　6 種類の古典的教会音階とアラビア音階
Fig. 7　Six scales of Medieval church music and Arabic scale.

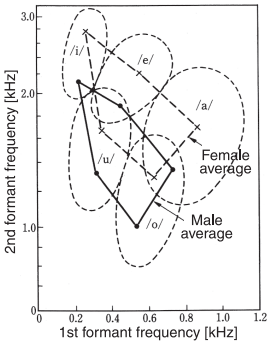


図 8　日本語五母音の第 1，2 フォルマント周波数 [41]
Fig. 8　The first and second formant frequencies of five Japanese vowels [41].

日本語における五母音の音色の配置を図 8 に示す．各母音の第 1，第 2 フォルマント周波数の（多数話者における）分布の様子と，成人男性・女性の平均値が示されている．音楽の場合，音階の各音の基本周波数の対数値は，移調によって，等しい値だけ上下するが，音声の場合，例えば性差によってフォルマント周波数は図 8 のように移動する．メロディーの階名同定と音声の音韻同定とを類似した情報処理としてとらえることは，前者が音高に対する相対音感に基づく情報処理であるように，後者を音色に対する相対音感に基づく情報処理として考えることに相当する．個々の音の物理特性ではなく，他音との関係性に基づいて（例えば）母音同定のメカニズムを検討することは，音声科学の分野では古くから議論されている [42]〜[44].

さて，性差や話者差によっておよそ不変な母音配置を，図 7 のように多様に変形させるような要因を考えれば，それが方言であることは周知の事実である．図 9 に米語方言の例を示す．声道長正規化後のいくつかの単母音を第 1，2 フォルマント周波数平面に配置している [45]．各地方で生まれた場合，両親の母音のフォルマント周波数をまねるのではなく，この母音配置を獲得する．以上の事実を踏まえ，思考実験を行う．

4.2 一卵性双生児の言語獲得に関する思考実験

「出産直後に両親が離婚した一卵性双生児」に関して思考実験を行う．離婚後，父親，母親が一人ずつ別々に育てた場合，10 年後，彼らがどのような発音を獲得

(注5)：「子犬のワルツ」をアラビア音階で演奏した WAV ファイルを下記にアップロードしている．一度聴取することを強く勧める．
http://www.gavo.t.u-tokyo.ac.jp/~mine/material/western.wav
http://www.gavo.t.u-tokyo.ac.jp/~mine/material/arabic.wav

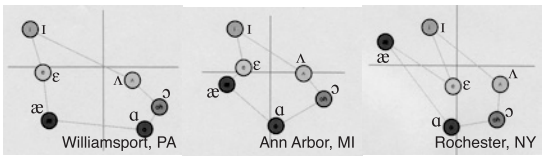


図 9　米国方言における母音配置（ただし一部）の差異 [45]
Fig. 9　Vowel arrangement (in part) of several American English dialects [45].

しているのかを考える．父親の音声を模倣して言語を獲得するとより太い声を有し，母親の音声を模倣して言語を獲得するとより細い声を有するようになることはない．しかし，父親，母親が異なる方言話者であった場合，両者の発音は全く異なったものとなる．

性差（声道の長さ・形状における差）に起因する音声の音響的差異はスペクトル包絡の差異である．同様に，方言差に起因する音響的差異もスペクトル包絡の差異である．しかし人間が音声言語を獲得する場合，前者は影響せず，後者は決定的に影響を与える．つまり，与えられた音声試料から非言語的情報を分離し，言語的（及びパラ言語的）情報を獲得する．より音響的に考察すれば，各音の物理特性をそのまま獲得するのではなく，音群がなす関係性を獲得するといえる．

父親の地方訛りの英語を学習データとした音声合成システムは，父親の声色の地方訛り英語を話すようになる．一見言葉を話しているように見えるシステムであるが，筆者らは，これらシステムと人間の間には，情報処理的に，大きな隔たりがあると考えている．人間らしい音声情報処理の実現には，音声に含まれる非言語情報を分離し，音の体系として言語情報を抽出し，獲得・学習する技術が必要であると考える．

4.3　体系としてとらえる言語音と古典的音韻論

音の体系として言語音群をとらえる方法論は音韻論では古典的な議論である．研究者が音声を波形やスペクトルとして観測できるようになる以前の時代から，体系としての音声言語が議論されてきた．近代言語学の祖と呼ばれるソシュールは，“What defines a linguistic element, conceptual or phonic, is the relation in which it stands to the other elements in the linguistic system.” と述べている [46]．また，ソシュールに啓蒙されたヤコブソンは関係的・体系的不変性に基づいて，言語音群を幾何学的に描画し [47], [48]，“Physiologically identical sounds may possess different values in conformity with the whole sound system, i.e. with their relations to the other sounds,” “We have to put aside the accidental properties of individual sounds and substitute a general expression that is the common denominator of these variables.” などの言葉を通して，言語音群の相対的関係性に言及している．著者らが現在の音響モデリング技術の欠損を議論し始めた当初は，これら古典的議論を既知とするものではなかった．しかし最終的に，類似した議論を重ねていたことは，筆者らにとって非常に興味深い．

以下，語ゲシュタルトの数学的解釈について述べる．

5.　完全変換不変量と音声の構造的表象

5.1　可逆な変換に対する完全変換不変性

入力音声の話者情報だけを変換する話者変換技術では(注6)，話者変換を空間写像として扱っている．話者 A の声空間(注7)と話者 B の声空間との間に写像を張る．話者 A の発声が軌跡として与えられると，対応する話者 B の発声（軌跡）が写像によって得られる．同様に，収録機器特性や伝送特性などの音響特性も空間写像となるため，非言語的要因による音声の変形はすべて空間写像として考えられる．すなわち，非言語的情報から言語的情報を分離させ，前者に対して不変・非依存な形式で後者を表象する技術は，音声を変換・写像不変な音響量のみを用いて表象することで得られる．

変換不変量で音声を表象する試みは先行研究にも見られるが，すべてが周波数 f の線形変換 ($\hat{f}=\alpha f$) に対する不変表象であり，話者変換技術で用いられる一般的な写像関数 ($\hat{f}=\beta(f)$) を対象としていない．更に，音の関係性ではなく，個々の音を不変的に扱う方法に終始している [49], [50]．この場合，時間軸に沿って話者性が（例えば音素単位に）変化する合成音声に対しても各音を不変的に扱うことができるが，これを人間に聴取させると，話者性の変化（スペクトル包絡の変化）を音韻の変化として知覚する例が報告されている [43], [51]．本研究では，話者性というのは静的・時不変な特徴であるとの前提に立ち，各音を不変的に扱う枠組みではなく，音と音との関係性（以下に示すように距離）を不変に表象することを考える．

筆者らは [52] において，二事象間距離尺度の一つである f-divergence [53] が，微分可能かつ可逆ないかなる変換に対して不変であること（十分性，図 10 参

(注6)：ただし，言語情報を非言語情報から分離する技術ではない．
(注7)：具体的にはケプストラム空間となるが，スペクトル（包絡）空間でもよい．両者は線形写像（FFT）で変換されているだけである．

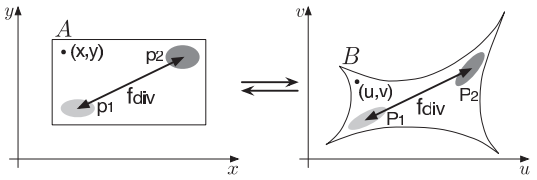


図 10　連続かつ可逆な変形に対して不変な f-divergence
Fig. 10　f-divergence is invariant with any kind of differentiable and invertible transform.

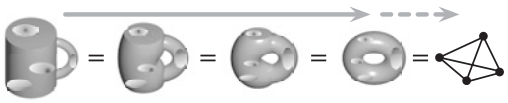


図 11　f-divergence に基づく形態的不変性
Fig. 11　Topological invariance based on f-divergence.

照），更に，不変となる事象間距離は f-div. しかないこと（必要性）を証明した．f-div. は $t>0$ において凸な関数 $g(t)$ に対して，下式で定義される．

$$f_{div}(p_1,p_2)=\oint p_2(\boldsymbol{x})g\left(\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\right)d\boldsymbol{x}$$

ここで，$p_i(\boldsymbol{x})$ は i 番目の事象である．事象は点ではなく，確率密度分布として表象される．$g(t)$ を換えることで様々な f-div. が定義可能であるが，$g(t)=t\log(t)$ とすれば，f-div. は KL-div. となり，$g(t)=\sqrt{t}$ とすれば(注8)，$-\log(f\text{-div.})$ はバタチャリヤ距離になる．つまり，これらの分布間距離尺度は変換不変量である．

例えば図 11 に示すような，連続かつ可逆な空間写像による形状の変形を考える．各々の変形された形状の表面（及び内部）に分布としての事象群が存在している場合，任意の二事象間の f-div. は如何なる写像によっても変化しないため，f-div. より構成される距離行列は一切不変であることが導かれる．

5.2　一発声の構造化に基づく音声の構造的表象

ケプストラム空間などの特徴量空間にてフレーム系列として表象された一発声（一軌跡）を分布系列化し，任意の二分布間距離を（時間的に離れた事象間を含め）f-div. で計測し，距離行列を求める [54]．一般に $N\times N$ の距離行列は，N 個の事象によって構成される N 角形に対してその幾何学的形態を規定するため，この変換不変な距離行列を，音声の構造的表象と呼ぶ（図 12 参照）．筆者らはこの構造的表象を語ゲシュタルトの数学的解釈であると考えている．入力音声を構造化し，その後は構造表象のみを用いた処理を行えば，個々の音の音響特性はすべて捨象することとなる．音群の関係性のみを利用し，各音のスペクトル包絡特性は直接的には参照しない処理となる．

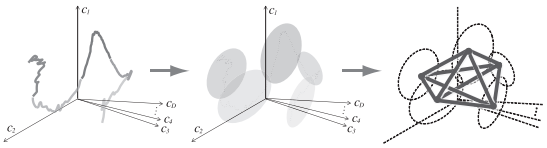


図 12　f-divergence を用いた一発声の構造化
Fig. 12　Structuralization of a single utterance using f-divergences.

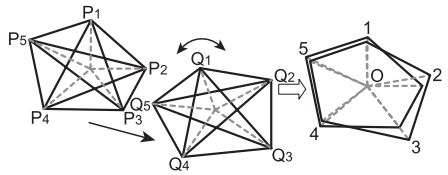


図 13　回転及びシフトによる構造の重ね合わせ
Fig. 13　Overlapping two structures through rotation and shift.

二発声から各々計算された，等事象数の音声構造（距離行列）に対して，行列をベクトルとしてみなして（構造ベクトル）計算されるユークリッド距離は，二つの構造をシフト及び回転して重ねた後に算出される，対応する二点間距離の総和の最小値に近似的に比例する [55], [56]．図 13 は二つの構造の重ね合わせを示している．ここでケプストラム空間を考えれば（次節で詳述するが），シフトは音響機器特性の違い，回転は声道長の違いを相殺する演算に対応する．言い換えれば，構造ベクトル間距離は，適応処理を通して話者や音響機器特性をそろえた後に計算される音響照合距離と，およそ比例関係にある [55], [56]．すなわち，適応処理を施した後の音響照合距離は，構造表象を用いれば，明示的に適応処理を行わずに推定できる．

環境変化に頑健な情報処理モデルを考える場合 [57]，一般に，環境変化に対する逐次適応に基づくモデルと環境変化に対する不変量に基づくモデルが考えられる．しかし，不変量として事象間距離（コントラスト）を採択すれば，この二つの考え方は相反するものではなく，後者は，「明示的適応を行うことなく前者と等価な効果をもたらす情報処理」として位置づけられる．その意味で [54] では，前者を explicit adaptation，後者を implicit adaptation として説明している．

(注8)：[53] の定義式では $g(1)\equiv 0$ であるが，$g(t)=\sqrt{t}$ はこの条件を満たさない．しかし，f-div. 不変性は $g(1)\equiv 0$ を要求しないため，ここでは $g(1)\neq 0$ である $g(t)$ を用いた場合でも f-div. と呼んでいる．

5.3 部分空間を用いた不変性の制御

音声の構造的表象は連続かつ可逆な一切の変換に不変である．異なる 2 単語が空間写像で対応づけられる場合を仮定すると，構造表象はこの 2 単語を区別できなくなる（強すぎる不変性）．つまり，非言語的な変換のみに対して不変性を有するように不変性を制御する必要が生じる．例えば，ケプストラム空間における一次変換 $c'=Ac$ を考える（ケプストラムの一次変換は，一般に，周波数軸変換では非線形変換となる）．任意の行列 A に対する不変性ではなく，下記で示される帯行列のみに不変性を有する構造表象の計算法を考える．

$$\begin{pmatrix} c'_1 \\ c'_2 \\ \vdots \\ c'_n \end{pmatrix} = \begin{pmatrix} & & 0 \\ & \iddots & \\ 0 & & \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}$$

上式より明らかなように，変換後の成分 c'_i は変換前の自成分及びその近隣の成分 $c_{i-d},..,c_i,..,c_{i+d}$ とのみ関係する．言い換えれば，変換前後で，離れた次元間では成分は独立性をもつ．そこで，連続した w 次元の幅のみを用いて部分空間を構成し，各部分空間で構造化，及び，構造照合を行う [54]．w の大・小は，「不変性」「識別力」の「高・低」「低・高」に対応する．図 14 は三次元空間を二つの二次元空間 (c_1, c_2)，(c_2, c_3) に分割し，各々で構造化する様子を示している．

5.4 声道長変換行列の幾何学的意味

音声認識研究において，声道長適応（変換）を一次の全域通過ディジタルフィルタの周波数変換特性で近似することが広く行われている [58]．この変換はケプストラム空間では，下記の行列 B を用いた変換 $c'=Bc$ となる [58]．α は $|\alpha|<1.0$ を満たす定数である．

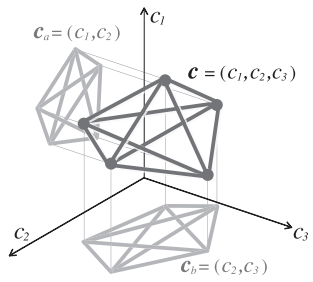


図 14 部分空間を用いた不変性の制御
Fig. 14 Control of invariance using sub-spaces.

$$B=\begin{pmatrix} 1-\alpha^2 & 2\alpha-2\alpha^3 & \cdots & \cdots \\ -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

$$B_{ij}=\frac{1}{(j-1)!}\sum_{m=m_0}^{j}\binom{j}{m}\frac{(m+i-1)!}{(m+i-j)!}(-1)^m\alpha^{(2m+i-j)}$$

ただし $m_0=\max(0, j-i)$ である．B はおよそ帯行列となるが，筆者らは B が多次元回転行列で近似できることを導いている [59]．収録機器，伝送機器の音響特性は，ケプストラム空間では定ベクトルを足す演算であるため，シフトによって音響特性の差異が，回転によって声道長の差異が相殺される（図 13 参照）．

6. 音声の構造的表象の応用と情報分離

提案する音声の構造的表象の効果的応用例として，孤立単語認識，外国語発音習熟度推定，及び，音声合成を取り上げ，その結果について概説する．なお，詳細は [54], [62], [65] などを参照して頂きたい．

6.1 孤立単語音声認識実験

日本語五母音を並び換えて定義される語彙数 120 の単語セット，及び，子音を含む音素バランスのとれた語彙数 212 の単語セット [60] を用いた孤立単語認識実験を行った [54]．認識実験の枠組みを図 15 に示す．図が煩雑となるため，部分空間化は省いてある．また，パラメータ次元数を抑えるため線形判別分析も導入している．母音単語の場合，各単語を 20 個の分布系列へ，バランス単語の場合は 25 個の分布系列へ変換

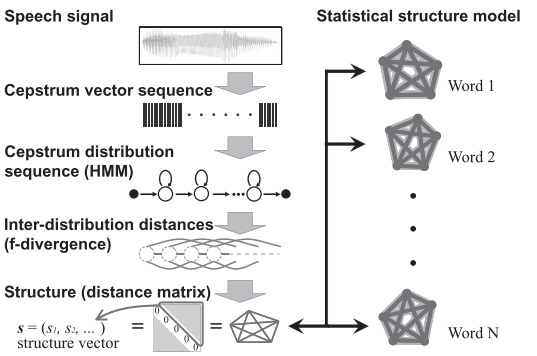


図 15 構造表象を用いた孤立単語音声認識の基本的枠組み
Fig. 15 Basic framework of structure-based spoken word recognition.

し構造化している．この構造化は，各発声から MAP（Maximum A Posteriori，事後確率最大化）推定に基づき HMM を学習することで行っている（図 15 参照）．学習データは母音単語の場合，成人男女 4 名による 5 回ずつの発声であり（各単語 40 発声），バランス単語の場合，成人男女 15 名による 1 回ずつの発声である（各単語 30 発声）．これらより，構造ベクトルのガウス分布として各単語の統計構造モデルを構築した．評価データは同規模の他話者発声であるが，行列 B を用いて，声道長変換データも用意した．

比較実験として，同一学習データより構築した単語 HMM を用いた孤立単語認識実験も行った．なお，音声対話技術コンソーシアムより配布されている 4,130 人の話者より構築されたトライフォンモデル [63] を用いた孤立単語認識システムによる実験も一部行った．

結果を図 16，図 17 に示す．α の負（正）は，声道長の長（短）に対応し，$|\alpha|$=0.4 で約倍，半分になる．図中，HMM とは単語 HMM の性能であり，matched とは，α の各値に対応した学習データを用いた（学習・評価間の不一致を事前に手動で解消）単語 HMM の性能である．w は部分空間化における次元幅である．

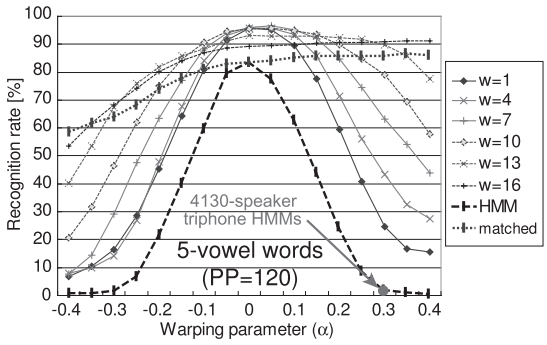


図 16　五母音単語セットに対する認識率
Fig. 16　Recognition rates with 120 words of five Japanese vowels.

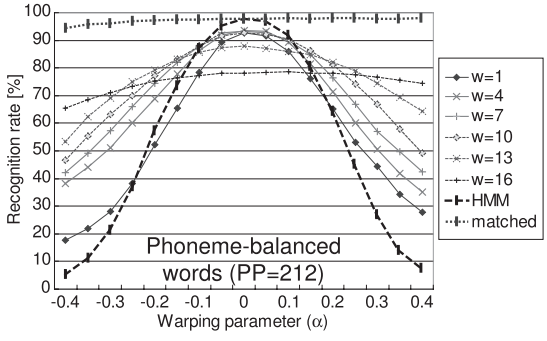


図 17　バランス単語セットに対する認識率
Fig. 17　Recognition rates with 212 words of balanced phonemes.

母音単語の場合，適切な w を設定することで，提案手法は単語 HMM より極めて高い頑健性を示している．なお，トライフォンの結果より，話者数を増やすだけでは対処できない学習・評価条件の不一致にも提案手法は十分に対応できている．バランス単語においても高い頑健性は示されているが（w=10, 13 など），matched には及んでいない．一部の子音（無声破裂音，摩擦音など）は話者差による変形が母音に比べて小さいため，音と音の相対的な関係だけでは十分に対応できていないと解釈される．**3. 2** でも述べたように，音と音の関係性に基づく情報処理と，音の絶対的な特性に基づく情報処理の融合についても検討を始めており [61]，興味のある読者は参照して頂きたい．

6. 2　外国語発音評定実験

日本人学習者による英語発音の自動評定実験を行った [62]．発音評定は，同一内容の教師発声と学習者発声の比較が基本となるが，両者を音響的に比較すれば，それは発音の善しあしではなく，声帯模写のそれを定量化することになる．発音評価に必要な音響的側面のみを抽出，表象する手段として構造的表象を導入した．

英語教師による発音習熟度が付与された 26 名の英語学習者の音声資料（約 60 文）に対して，その習熟度を推定する．従来手法としては，母語話者音声より学習した不特定話者音素 HMM を用いて計算される GOP (Goodness Of Pronunciation) スコアを採択した [64]．読み上げ音声を対象としており，学習者が意図したテキスト（すなわち音韻列）を既知として，これを習熟度推定時に用いることができる．意図した音韻列の，観測量 o に対する事後確率が GOP である．

構造表象を用いて発音習熟度を推定する場合，一発声を構造化するのではなく，音素 HMM（3 状態）を約 60 文の音声データより学習者ごとに学習し，状態単位での構造を構成した．各学習者の発音構造と，教師 1 名の発音構造とのユークリッド距離を求める（図 13 参照）．このとき，評価話者以外の音声を用いて，構造間差異と習熟度との相関が最大化するよう，不要な状態対を準貪欲探索により削除し，選択された状態対のみで構造間差異を定義した．手順を図 18 に示す．

図 19 に結果を示す．横軸は図 16，図 17 同様，評価データの声道長変形の度合いである．縦軸が教師による手動評定値と計算機による自動評定値との相関で

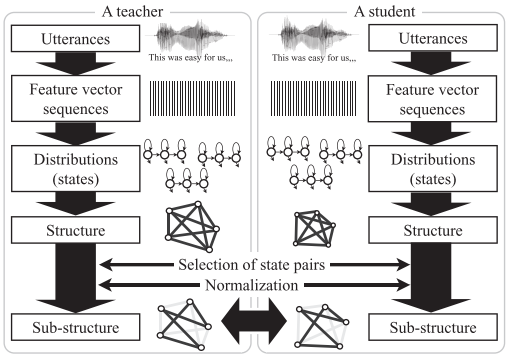


図 18　選択された状態対を用いた発音構造比較
Fig. 18　Structure comparison using selected states.

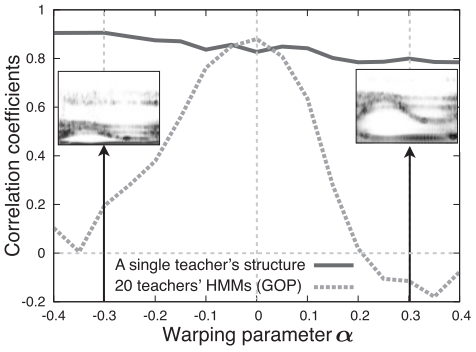


図 19　学習者音声を用いた習熟度推定結果
Fig. 19　Results of proficiency estimation of learners.

ある．GOP と比較して発音構造による自動評定は，極めて高い頑健性を有している．GOP 計算時に用いる不特定話者 HMM を適宜話者適応すれば，同様の性能は導出可能である．しかし，HMM を学習者の声色に合わせて適宜修正するということは，これは発音評定ではなく，声帯模写評定技術として考えるべきである．声帯模写評定技術を発音評定に直接応用するには，話者適応が常時必須となる．このような実装は技術的には可能であるが，**2.**〜**4.** での議論を既知とする立場から考えれば，不自然な技術構築となる．

図 19 は教師 1 名と各学習者との発音構造間差異であるが，任意の学習者間の構造間差異も同様に求めることができ，26 名全体で行えば，学習者間距離行列が得られる．この距離行列を用いて学習者群を樹形図化する．ここでは巨人化 (α=−0.3)，小人化 (α=0.3) した音声も含め，合計 78 名の学習者の樹形図を求めた．Ward 法による結果を図 20 に示す．アルファベットが学習者 ID であり，$\overline{\mathrm{X}}$ は巨人化した学習者 X を，$\underline{\mathrm{X}}$ は

図 22　音声 − 声道の長さ・形状 = 語ゲシュタルト
Fig. 22　Speech − length and shape of the vocal tract = word Gestalt.

図 23　音声の構造的表象 + 声道の長さ・形状 = 音声
Fig. 23　Speech structure + length and shape of the vocal tract = speech.

小人化したそれを表す．字体の違いは性別である．身長差が全く捨象され，26 名の学習者の発音分類となっている．これに対して，学習者 i と学習者 j の学習者間距離を，i の HMM と j の HMM の対応する状態間距離（バタチャリヤ距離）の和で定義した場合の樹形図を図 21 に示す．こちらは原身長，小人，巨人（すなわち体格）でまず分かれ，各サブツリーでは性別で分類されている．言い換えれば，図 20 は言語情報のみに着目した分類であり，図 21 は非言語情報のみに基づいた分類である．情報分離が実現されている．

6.3　構造からの音声合成

音声認識と発音評定では，多様な非言語情報を分離し，言語情報を話者不変に表象する応用であったが，話者不変表象に対して非言語情報を再度加味することで，多様性を生成する応用を考える．図 22 に示すように，語ゲシュタルトは音声からその話者の声道の長さ・形状を消失させた表象である．これに対して図 23 に示すように，構造表象（語ゲシュタルト）に別話者の声道（別話者の身体特性）を戻すことで，その話者の声を生成することを検討している [65]．いうなれば，幼児の音声模倣のシミュレーションに相当する．紙面の制約のため具体的なアルゴリズムや結果の提示は省略するが，興味ある読者は [65] を参照して頂きたい．

7.　む　す　び

従来より音声の技術構築においては，種々の知見に基づき，情報を適宜捨象する形で特徴抽出，音響モデリングを試みてきた．位相スペクトルの切落し，調波構造の切落しがそれに相当する．しかし，人の聴覚が位相スペクトルに鈍感なように，人は言語を獲得するときに，親の音声の非言語的情報には鈍感である．確

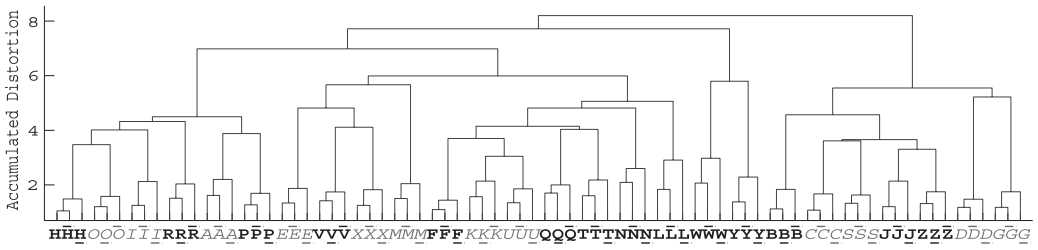


図 20　発音構造間差異に基づく 78 名の学習者分類結果
Fig. 20　Learner clustering based on structure-based pronunciation comparison.

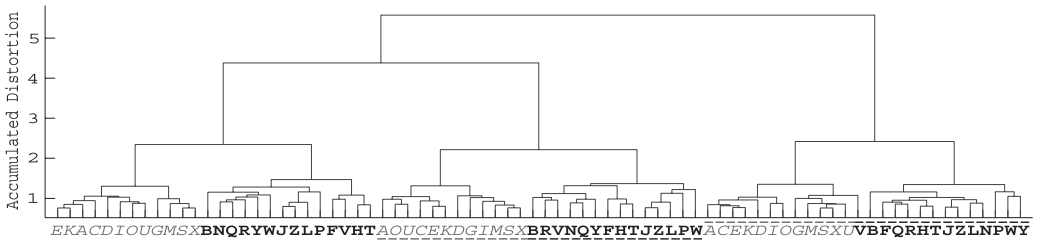


図 21　学習者の音素 HMM 間距離に基づく 78 名の学習者分類結果
Fig. 21　Learner clustering based on spectrum-based pronunciation comparison.

かに，人は音声の非言語的情報を認知し，それに基づいた行動をとることができる．その意味では敏感である．しかし，言語獲得（音声模倣）においては，自らの発声にその情報を反映しようとしない．その意味で鈍感である．本論文ではこの鈍感さに着目し，言語的情報を非言語的情報から分離し，多様な音響特性をもつ声に対する高い汎化能力の実現を試みた．

脳科学では，感覚器から入力された情報がいったん分離される情報処理モデルが広く受け入れられている．視覚情報の場合は，第 1 次視覚野からの情報が腹側経路と背側経路とに分かれ，各々が，what の情報，where（あるいは how）の情報を表象する [66]．聴覚情報の場合でもこれに倣い，言語情報，非言語情報の分離モデルが検討されている [67], [68]．これらの研究動向から見ても（実装方式の是非は問えないが），情報を分離する技術を構築することは，より人間らしい情報処理の実装に近づいていると筆者らは考えている．

積極的な分離を行わない場合，言語的同一性と音響的同一性とを等価なものと考えたり，発音評価を声帯模写評価として扱うことになる．情報分離を行わなくても，着目する情報カテゴリーと共起する観測量を大量に集め，期待値操作にて無関係の要因を消去すれば，あるいは，明示的適応を常時行えば，ある程度動作する機械は構成できる．しかしその場合人間と機械を比較すると，「何かが足りない」という言葉を再度繰り返すことになると考えている．統計的機械学習や明示的適応は強力なツールであるが，自然な技術構築を目的とするのであれば，実装している情報処理を接地（グラウンディング）させ，何を実装しているのかを十分に吟味しながら技術構築する必要があるだろう．

なお，本研究では多様な音響特性をもつ人間の声に対する高い汎化能力の実現を試みたのみであり，これは「人間らしい音声情報処理」に対する必要条件でしかない．特に意味や記憶の情報処理に関しては，本論文は直接的には何も言及していない．今後，様々な人間研究の成果を考慮しつつ，検討していきたい．

文　献

[1] 古井貞熙，田中穂積他，“特集：音声情報処理技術の最先端，” 情報処理，vol.45, no.10, pp.1002–1049, 2004.

[2] M.J.F. Gales, “Acoustic modelling for speech recognition: Hidden Markov Models and beyond?,” Proc. IEEE Workshop on Automatic Speech Recognition & Understanding, p.44, 2009.

[3] R. Lippmann, “Speech recognition by machines and humans,” Speech Commun., vol.22, pp.1–15, 1997.

[4] R.K. Moore, “A comparison of the data requirements of automatic speech recognition systems and human listeners,” Proc. EUROSPEECH, pp.2581–2584, 2003.

[5] 古井貞熙，“何かが欠けている音声認識研究，” 信学技報 SP2009-80, 2009.

[6] S. Furui, “Generalization problem in ASR acoustic model training and adaptation,” Proc. IEEE Work-

shop on Automatic Speech Recognition & Understanding, pp.1–10, 2009.
[7] N.D. Lawrence and J. Barker, “Dealing with high dimensional data with dimensionality reduction,” Tutorial of INTERSPEECH, 2009.
[8] 峯松信明他, “小特集：言語障害を通して再考する音声言語情報処理,” 音響誌, vol.63, no.7, pp.363–398, 2007.
[9] P.W. Jusczyk, The discovery of spoken language, The MIT Press, 2000.
[10] P.K. Kuhl, “Early language acquisition: Cracking the speech code,” Nature Reviews Neuroscience, vol.5, pp.831–843, 2004.
[11] 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.
[12] 加藤正子, “特集「音韻発達とその障害」にあたって,” コミュニケーション障害学, vol.20, no.2, pp.84–85, 2003.
[13] 早川勝廣, “言語獲得と育児語,” 月刊言語, vol.35, no.9, pp.62–67, 2006.
[14] P. Lieberman, “On the development of vowel production in young children,” in Child Phonology vol.1, ed. G.H. Yeni-Komshian, J.F. Kavanagh, and C.A. Ferguson, Academic Press, 1980.
[15] 深見　憲, ひろしくんの本 (V), 中川書店, 2006.
[16] 綾屋紗月, 熊谷晋一郎, 発達障害当事者研究, 医学書院, 2008.
[17] T. Grandin, M.M. Scariano（著）, カニングハム久子（訳）, 我, 自閉症に生まれて, 学研, 1994.
[18] L.H. Willey（著）, ニキリンコ（訳）, アスペルガー的人生, 東京書籍, 2002.
[19] 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.
[20] 東田直樹, 東田美紀, この地球にすんでいる僕の仲間たちへ, エスコアール, 2005.
[21] W. Gruhn, “The audio-vocal system in sound perception and learning of language and music,” Proc. Int. Conf. language and music as cognitive systems, 2006.
[22] 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.
[23] A.A. Write, J.J. Rivera, S.H. Hulse, M. Shyan, and J.J. Neiworth, “Music perception and octave generalization in rhesus monkeys,” J. Exp. Psychol. Gen., vol.129, pp.291–307, 2000.
[24] M.D. Hauser and J. McDermott, “The evolution of the music faculty: A comparative perspective,” Nature neurosciences, vol.6, pp.663–668, 2003.
[25] T. Grandin, C. Johnson（著）, 中尾ゆかり（訳）, 動物感覚〜アニマル・マインドを読み解く, 日本放送出版協会, 2006.
[26] 泉　流星, 僕の妻はエイリアン, 新潮社, 2005.
[27] 藤井　学, 神谷栄治, 自閉症, 新曜社, 2007.
[28] 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.
[29] R.B. Lotto and D. Purves, “An empirical explanation of color contrast,” Proc. National Academy of Science USA, vol.97, pp.12834–12839, 2000.
[30] R.B. Lotto and D. Purves, “The effects of color on brightness,” Nature neuroscience, vol.2, no.11, pp.1010–1014, 1999.
[31] 谷口高士, 音は心の中で音楽になる, 北大路書房, 2003.
[32] http://www.lottolab.org/illusiondemos/Demo%2012.html
[33] 東川清一, 読譜力ー「移動ド」教育システムに学ぶ, 春秋社, 2005.
[34] C. Neumeyer, “Chromatic adaptation in the honeybee: Successive color contrast and color constancy,” J. Comparative Physiology, vol.144, pp.543–553, 1981.
[35] A. Balkenius and A. Kelber, “Colour constancy in diurnal and nocturnal hawkmoths,” J. Experimental Biology, vol.207, pp.3307–3316, 2004.
[36] W. Strange, R. Verbrugge, D. Shankweiler, and T. Edman, “Consonant environment specifies vowel identity,” J. Acoust. Soc. Am., vol.60, pp.213–224, 1976.
[37] S. Shaywitz（著）, 藤田あきよ, 加藤醇子（訳）, 読み書き障害（ディスレクシア）のすべて〜頭はいいのに本が読めない〜, PHP 研究所, 2006.
[38] R. Port, “How are words stored in memory? Beyond phones and phonemes,” New Ideas in Psychology, vol.25, pp.143–170, 2007.
[39] 青木美和, 入野俊夫, R.D. Patterson, 河原英紀, “スケール変形した日本語 5 母音の知覚特性,” 音響秋季講論集, 2-P-6, pp.373–374, 2004.
[40] 林　芳恵, 入野俊夫, R.D. Patterson, 河原英紀, “話者の寸法を変化させた時の母音と単語の知覚特性の比較,” 音響春季講論集, 2-Q-27, pp.473–474, 2007.
[41] R.K. Potter and J.C. Steinberg, “Toward the specification of speech,” J. Acoust. Soc. Am., vol.22, no.6, p.807, 1950.
[42] L. Gerstman, “Classification of self-normalized vowels,” IEEE Trans. Audio Electroacoust., vol.AU-16, no.1, pp.78–80, 1968.
[43] P. Ladefoged and D. Broadbent, “Information conveyed by vowels,” J. Acoust. Soc. Am., vol.29, pp.98–104, 1957.
[44] W. Ainsworth, “Intrinsic and extrinsic factors in vowel judgments,” Auditory Analysis and Perception of Speech, ed. G. Fant and M. Tatham, pp.103–113, Academic, London, 1975.
[45] W. Labov, S. Ash, and C. Boberg, Atlas of North American English, Mouton and Gruyter, 2005.
[46] F.D. Saussure（著）, 小林英夫（訳）, 一般言語学講義, 岩波書店, 1940.
[47] R. Jakobson and J. Lutz, Notes on the French phonemic pattern, Hunter, N.Y. 1949.
[48] R. Jakobson and L. Waugh（著）, 松本克己（訳）, 言語音形論, 岩波書店, 1986.
[49] T. Irino and R.D. Patterson, “Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised

wavelet-Mellin transform," Speech Commun., vol.36, pp.181–203, 2002.
[50] 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.
[51] N. Minematsu, T. Nishimura, K. Nishinari, and K. Sakuraba, "Theorem of the invariant structure and its derivation of speech Gestalt," Proc. Int. Workshop on Speech Recognition and Intrinsic Variations, pp.47–52, 2006.
[52] Y. Qiao and N. Minematsu, "A study on invariance of f-divergence and its application to speech recognition," IEEE Trans. Signal Process., vol.58, no.7, pp.3884–3890, 2010.
[53] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," Studia Scientiarum Mathematicarum Hungarica, vol.2, pp.299–318, 1967.
[54] N. Minematsu, S. Asakawa, M. Suzuki, and Y. Qiao, "Speech structure and its application to robust speech processing," J. New Generation Computing, vol.28, no.3, pp.299–319, 2010.
[55] 峯松信明，"音声の音響的普遍構造の歪みに着眼した外国語発音の自動評定，" 信学技報，SP2003-180, 2004.
[56] 峯松信明，志甫　淳，村上隆夫，丸山和孝，広瀬啓吉，"音声の構造的表象とその距離尺度，" 信学技報，SP2005-13, 2005.
[57] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," Speech Commun., vol.49, pp.763–786, 2007.
[58] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," IEEE Trans. Speech Audio Process., vol.13, no.5, pp.930–944, 2005.
[59] D. Saito, R. Matsuura, S. Asakawa, N. Minematsu, and K. Hirose, "Directional dependency of cepstrum on vocal tract length," Proc. ICASSP, pp.4485–4488, 2008.
[60] Tohoku univ. – Matsushita isolated Word database http://research.nii.ac.jp/src/eng/list/detail.html#TMW
[61] Y. Qiao, M. Suzuki, and N. Minematsu, "A study of Hidden Structure Model and its application of labeling sequences," Proc. IEEE Workshop on Automatic Speech Recognition & Understanding, pp.118–123, 2009.
[62] M. Suzuki, N. Minematsu, D. Luo, and K. Hirose, "Sub-structure-based estimation of pronunciation proficiency and classification of learners," Proc. Int. Workshop on Automatic Speech Recognition and Understanding, pp.574–579, 2009.
[63] T. Kawahara, A. Lee, K. Takeda, K. Itou, and K. Shikano, "Recent progress of open-source LVCSR engine Julius and Japanese model repository," Proc. INTERSPEECH, pp.3069–3072, 2004.
[64] S.M. Witt and S.J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," Speech Commun., vol.30, pp.95–108, 2000.
[65] D. Saito, Y. Qiao, N. Minematsu, and K. Hirose, "Optimal event search using a structural cost function – Improvement of structure to speech conversion," Proc. INTERSPEECH, pp.2047–2050, 2009.
[66] L.G. Ungerleider, "Two cortical visual systems," in Analysis of Visual Behavior, ed. David J. Ingle, pp.549–586, MIT Press, 1982.
[67] S.K. Scott and I.S. Johnsrude, "The neuroanatomical and functional organization of speech perception," Trends in Neurosciences, vol.26, no.2, pp.100–107, 2003.
[68] P. Belin and R.J. Zatorre, "'What', 'where' and 'how' in auditory cortex," Nature Neuroscience, vol.3, no.10, pp.965–966, 2000.

（平成 22 年 8 月 24 日受付，9 月 28 日再受付）

峯松　信明　（正員）

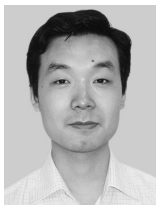
1995 東京大学大学院工学系研究科博士課程了．博士（工学）．現在，同大学院情報理工学系研究科准教授．音声科学から音声工学に至るまで，幅広く音声コミュニケーションに関する研究に従事．

櫻庭　京子

2003 名古屋大学大学院人間情報学研究科満期退学．博士（医学）．現在，獨協医科大学越谷病院で言語聴覚士として勤務．専門は自閉症者の認知，コミュニケーション，感情のコントロール等の臨床及び研究．

西村多寿子

1997 東京大学大学院医学系研究科国際保健学専攻修士課程了．同研究科公共健康医学専攻客員研究員．看護師，保健師の実務経験後，医療翻訳者として独立．現在，医学サイトの論文紹介記事等を執筆．



喬　　宇

2006 電気通信大学大学院情報システム学研究科博士課程了．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．

朝川　　智　（正員）

2008 東京大学大学院新領域創成科学研究科博士課程了．博士（科学）．2006〜2008 日本学術振興会特別研究員 DC1．現在，ソニー（株）勤務．音声信号処理，パターン認識に関する研究に従事．

鈴木　雅之　（学生員）

2010 東京大学大学院工学系研究科修士課程了．修士（工学）．現在，同大学院工学系研究科博士後期課程に在籍．音声認識，音声分析，音声強調に関する研究に従事．

齋藤　大輔　（学生員）

2008 東京大学大学院新領域創成科学研究科修士課程了．修士（科学）．現在，同大学院工学系研究科博士後期課程に在籍．音声合成，音声変換，音声分析，音声認識に関する研究に従事．