情報の分離と音響モデリング ～人間らしい音響モデリング～*

○峯松 信明 (東大情理)

1　はじめに

音声の生成過程を，二つの過程（音源の生成と声道による共鳴）に分離するソース・フィルタモデルが音声認識，合成の分野で広く用いられている。しかし，声道の共鳴特性は語彙（言語的情報）によっても，話者（非言語的情報）によっても変形する。音声認識システムは，音声から言語的情報を抽出することを目的とするが，音響特徴量 o としては，声道の共鳴特性（スペクトル包絡）が使われることが多い。そのため，話者独立なシステムを構築する場合，音響モデル $P(o|w)$ を多数話者から構築される統計モデルとして構築したり，あるいは，入力話者の声質に対して逐一モデル適応（修正）を行うことが一般的である。

本稿ではまず，定型発達を遂げた人間，音声コミュニケーションの獲得に困難を示す重度自閉症者，更には，動物の音（声）情報処理の様子を概観する。そして，定型発達を遂げた人間が示す「声質の違いに影響を受けない柔軟な音声認識の実現」のためには，統計モデルや適応モデルとは異なり，ソース・フィルタモデルによって得られるスペクトル包絡特性に対して，更に，言語的情報と非言語的情報へと情報を分離する技術の構築が不可欠であるとの主張を行う [1]。

2　集める・合わせることは問題解決なのか?

音声は一次元信号（波形，数値列）として観測されるが，その中に，言語・パラ言語・非言語的な様々な情報が符号化されている。人間はそれをいとも簡単に復号化する。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，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，位相情報を切り離し，パワースペクトルに着眼する場合が多い。更に，ソース・フィルタモデルに基づき，調波構造を切り離し，包絡特性のみに着眼することも多い（Fig. 1）。音声認識技術が好例である。

話者独立単語音声認識，テキスト独立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），何れにも依存する。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。しかし，期待値操作は大量のサンプルを必要とする。

$$P(o|w) = \sum_s P(o|w,s)P(s|w) \approx \sum_s P(o|w,s)P(s)$$

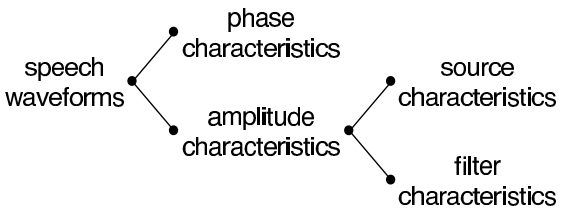

Fig. 1　Feature extraction based on separation

$$P(o|s) = \sum_w P(o|w,s)P(w|s) \approx \sum_w P(o|w,s)P(w)$$

ここで，言語情報と非言語情報は，そもそも独立した情報である。にも拘わらず，それらを運ぶ音響的対象物（特徴量）として，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々の音響モデルを導出する。期待値操作を行えば，数式上，$P(o|w)$ や $P(o|s)$ は確かに求まる。

調波構造や位相も明示的分離に依らずに，期待値操作で消失可能である。調波を h，位相を p とすれば，

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

は，単語 w に対する「話者・調波・位相に独立」な音声波形 o の統計的音響モデルとなる。

期待値操作を行わず，その話者・調波・位相に対する波形モデル $P(o|w,s_0,h_0,p_0)$ を推定すれば，これは話者・調波・位相適応した波形モデルとなる。

筆者は，波形ベースの統計モデルや，適応モデルを用いた研究例を知らないが，これは，精度の高低以前に，抽出すべき言語的情報に対して凡そ独立な要因である，調波や位相の情報を分離する常套手段的技術が存在する（Fig. 1）からであると考える。

ある観測量が複数の情報を伝達する場合に，特定の情報のみを抽出することを考える。この時，1) 各種情報に対応する形で観測量を分解するのか，2) 分解せずに無関係な要因に関して期待値操作で統計モデルを構築するのか，3) 分解せずに無関係な要因に対して目下の状況に即した値を推定して与えた適応モデルを構築するのかは，技術的には，精度が高くなる手法を選択すれば良い。しかし，以下の節で主張するように，技術構築の目的を「定型発達を遂げた人間が行うような」柔軟な情報処理の構築を目的とした場合，上記の選択は慎重に行うべきであると考える。

3　音声言語獲得に関する発達的・進化的考察

3.1　音声コミュニケーションに関する発達的考察

幼児の言語獲得は「音声模倣・学習」を基本とする [2]。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，彼らの模倣の音響的対象物であ

*Acoustic modeling of speech based on information separation – Toward human-like acoustic modeling –.
by MINEMATSU, Nobuaki (The University of Tokyo)

る。幼児の音声模倣は，音響的模倣（声帯模写）では無い。音響的に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 [3]。文献を調査すると，模倣対象に各種用語を当てている。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 [4]。幼児は語形の全体ゲシュタルトを認知する [5]。幼児は related spectral pattern を模倣する [6]。本稿では以降,「語ゲシュタルト」と呼ぶこととする。

語ゲシュタルトに話者情報が含まれていれば，幼児は音響的模倣を試みることになる。しかし彼らは自らの声を親の声に合わせることはせず，両者の音響的ずれに鈍感な模倣を行っている。語ゲシュタルトは音声から話者情報が分離された音響パターンと言える。

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。即ち，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，頑健な情報処理を獲得する。話者情報を音響的に分離する情報処理能力があれば，当然の帰結である。分離能力（技術）が無ければ，言語的・非言語的情報が共存した状態で音声を扱うこととなり，統計モデルや，適応モデルを採択することになる。

音声認識における従来の音響モデリング同様，音声合成においても，言語・非言語情報が共存した音響モデリングが広く行われている。話者 A の音声資料と対応するテキストを話者 B に与え，更に新規テキストを話者 B に与える。話者 B がこれを，話者 A そっくりに読み上げた場合，話者 B は物まね芸人と呼ばれる。これを計算機に実装したのがテキスト音声合成である。Blizzard Challenge では合成音声の話者性と学習話者のそれとの等価性も評価対象である [7]。

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。七色の声を持つと呼ばれる声優の中村メイ子の声をそっくり真似る例 [8]，外国語発音練習やカラオケにおいて，音響的模倣以外の真似方が難しい例 [9]，音声に限らず，様々な音響音を模倣する例は，関連図書において頻出する [10–12]。刺激音を丸暗記し，再生する処理が主体となっている訳だが（故に汎化能力も低い），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 [13]。

音声認識にしろ，音声合成にしろ，従来，言語・非言語情報の分離を積極的に行わず，集めたり合わせたりすることで対処する技術開発を進めてきた。果たしてこの戦略で，人間が示す柔軟な情報処理は実装できるのだろうか？「人間的であること」を技術開発の目的とする必要は必ずしも無い。しかし，これらの技術をヒューマノイドに搭載しているのも事実である。

3.2 音声コミュニケーションに関する進化的考察

動物を対象とした場合，音声模倣は稀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 [14]。霊長類以外で音声模倣を行う種は鳥，クジラ，イルカなどが確認されているが [15]，動物の音声模倣は音響的模倣が基本となっている [15]。また，人以外の霊長類は相対音感が乏しく，移調前後のメロディーの同一性判定が困難であることも示されている [16][1]。即ち，人以外の霊長類は極端な絶対音感を有している[2]。

アスペルガー症候群（自閉症の一種）を患う者として世界で初めて書籍を出版した [10] 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 [17]。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 [9, 12] 自閉症者の多くは絶対音感保有者である [18]。

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，それとも，音のある側面だけに限定された同一性で十分なのか，が問うべき焦点である。前者が必要であれば，音響的同一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，動物におけるメロディー同一性の欠損は，良い例である。

前節，本節と（重度）自閉症者や動物の音情報処理について文献調査の結果を述べたが，これらをまとめるに，筆者は「（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」種が人間である，と考えている。換言すれば「音のある側面が同一であれば，（その側面が伝達する）情報の同一性を認知できるようになった」種が人間である。以降，この「ある側面」をどのように抽出するのか，非言語的情報を如何に分離するのか，について筆者が検討している技術構築について紹介する。

4 写像不変量に基づく音声の構造的表象

声質変換・話者変換に代表されるメディアモーフィングは，特徴量空間の写像として一般化できる。非言語的情報の違い（例えば話者の違い）による特徴量変形は，ある写像となる。よって，非言語的情報の違いに独立な音響パターンとは，写像不変量のみで構成されたパターンとして導出できる。ここでは単純な例として，調独立なメロディーの階名同定を例にとり，その一般化として，音声の構造的表象を説明する。

4.1 音高（メロディー）の相対音感と階名同定

相対音感に基づいてメロディーを階名で書き起こす場合，調を上下させても（移調），書き起こされる階名列（ドレミ列）は変わらない。調に独立な音同定

[1]但し，1 オクターブずらすと同一性が分かるとのことである。
[2]彼らがメロディーを音名で記述できたり，採譜できる訳では無い。違う音は違う音，と認識しているだけである。

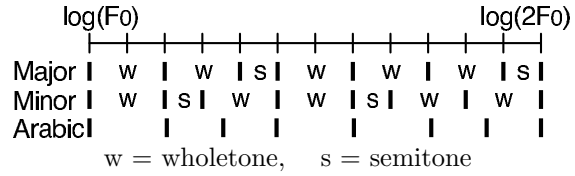

Fig. 2　Three scales of Major, Minor, and Arabic

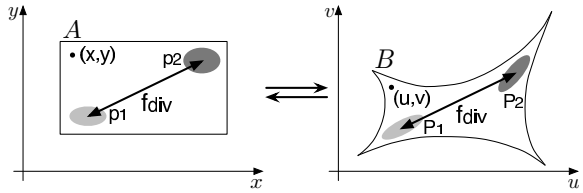

Fig. 3　Transform-invariance of f-divergence

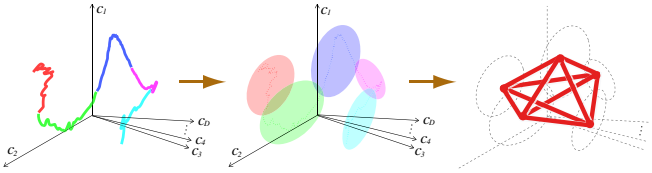

Fig. 4　Utterance to structure conversion

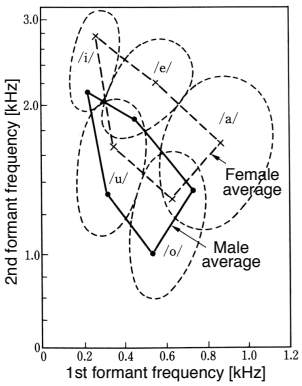

Fig. 5　Distribution of the five Japanese vowels

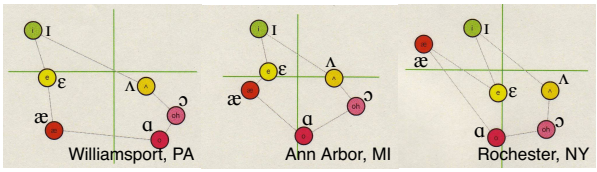

Fig. 6　Regionally accented American English[22]

が行われる。移調とは，全ての音の基本周波数（F_0）を定数倍する写像であるため，任意の二音間の（対数軸上での）音高差（音程）は移調前後で不変である。この音程の不変性が調独立な階名同定の必要十分条件であり [19]，例えば長音階を（移調後），アラビア音階にして奏でると（Fig. 2），階名同定は一部，困難になる。音程が一部，変わるからである。

4.2　音色（スペクトル包絡）の相対音感と音韻同定

対象とする特徴量を，一次元の音高から多次元の音色に拡張する。また写像も一般化する（$\hat{\boldsymbol{x}} = f(\boldsymbol{x})$）。写像 f が連続かつ可逆であれば，二分布間の距離尺度である f-divergence が写像不変となる（Fig. 3）[20]。

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

また，上記条件を満たす全ての写像に不変な二事象間距離は f-div. しか存在しない [20]。音高の場合は音程の方向性（上がるのか下がるのか）も含めて不変量であるが，多次元空間の場合，写像が回転性を持つ場合がある。結局，ある事象から別の事象へ向かう方向性は非不変量となるため，Fig. 3 に示すように，二事象間の距離（スカラー量）のみが不変量となる。

音高の相対音感を一般化して得られた，多次元特徴量の相対的不変量に基づいて音声の不変表象を考える。特徴量空間にて発声（軌跡）を分布系列に変換し，全ての二事象間の f-div. を計測して得られる距離行列がそれに相当する（Fig. 4）。一般に距離行列は一つの幾何学的形態を規定するため，この不変表象を，音声の構造的表象と呼んでいる。筆者はこれを「語ゲシュタルト」の数学的解釈であると考えている。

なお，声道長変化による音色変化（周波数ウォーピング）を，一次の全域通過ディジタルフィルタの周波数特性で近似することが広く行われているが，この時の写像は，回転性の非常に強い写像となる [21]。

Fig. 5 に日本語の五母音の F1/F2 母音図を示す。男女間で母音の体系が保持されている様子が分かる。なお，この体系が崩れると，それは方言となる。Fig. 6 に米語方言の幾つかを示す。子供が言語を獲得する時に，通常，親の声は真似ないが，親の声の中の母音体系はしっかり真似，立派な方言話者になる。

4.3　韻律的特徴としての音声の構造的表象

第 4.1 節, 第 4.2 節に示したように，構造的表象は，音高の相対音感の自然な次元拡張でしかない。言い換えれば，幼児が親の発声のイントネーションパターンを真似る際の模倣対象物と，親の発声の音色の動きパターンを真似る際の模倣対象物を，共通の枠組みで記述する試みでしかない。前者は韻律的特徴（の一部）として扱われるが，筆者は後者も韻律の一部であると考えている。音色は多次元の物理量であるので，その動きの様子を視覚的に把握することは困難である[3]。音高は一次元であるため，容易である。この違いが構造表象の韻律として解釈を困難にしていると思われるが，その導出過程を考えれば，韻律としての解釈は妥当であると考えている。

身体の成長による声道の伸長を線形伸長と仮定し（周波数軸の線形縮小），声道長の伸縮に不変な音声表象が提案されている [23–25]。これらは，各フレームの包絡特性を不変的に表象する試みであり，時間軸に沿って音声の話者性を変形しても，不変性が保たれる。しかしそのような（合成）音声を提示すると，話者性の変化を音韻性の変化として知覚する様子が示されている [26]。本研究では，非言語的情報は時不変

[3]様々な共鳴音を連続的に発声してみて，その動きに対する視覚的把握を試みて欲しい。恐らく困難なはずだ。

の情報であると仮定し，個々の音を不変にするのではなく，音の動きに対する不変項を導出している。

また，構造表象は話者正規化を意図して提案された音声表象ではない。Fig. 1 に示す，スペクトル平滑化による調波構造の分離は F_0 正規化ではないように，話者情報を分離する構造表象は話者正規化ではない。F_0 情報が存在しない音声表象としてスペクトル包絡があるように，構造表象は話者情報を保有しない（話者独立な）音声表象として提案された。音声は，ある話者の調音器官を通して生成される以上，話者の情報は不可避的に混入する。その情報が無い以上，構造表象のみを用いて音色（スペクトル包絡）を復元することは，原理的に不可能である。復元には話者の身体（声道形状・長さ）の情報が必要になる [27][4]。

4.4 音声の構造的表象の応用例

音声の構造的表象は，元来，二話者（教師，学習者）間での発音比較を目的として提案された。音響的な比較をすれば発音の善し悪しではなく，声帯模写の善し悪しを定量化することになる。一方の音声を話者変換し，他方の話者性に合わせて比較すれば発音評価は技術的には可能である。しかしこれは，発音評価を声帯模写評価の枠組みで実装しているに過ぎない。「人間らしい」発音評価を実装するには，発声から非言語情報を分離した上で得られる音声表象を用いて二話者を比較する必要があると考え，構造的表象は提案された [28]。その後，孤立単語音声認識，音声合成，方言性に基づく話者分類に応用されている [1, 29]。

4.5 音声の構造的表象の技術的問題点

音声から非言語情報（様々な時不変バイアス項）を除去するために提案された音声の構造的表象であるが，技術的問題点も少なくない。

分布分割のずれ 構造推定には軌跡の分布系列化が必要である（Fig. 4）。二発声を構造化して照合する場合，Fig. 7 に示すように，二形態比較を行うことになる。この場合，対応する事象間の適切な対応が必要となる。同一単語を複数回発声して構造化した場合，n 番目の分布が異なる音素に対応することがあり，これが精度劣化に繋がる。

強すぎる不変性 異なる二単語が写像で結ばれていれば，同一単語と判断されてしまう。つまり，ある部分写像群に対してのみ不変性を有するように，制約をかける必要がある。帯行列の乗算で表現される変換群に関しては，部分空間での構造照合という解決策を提案している [1] が，それ以外の部分写像群に関する検討は行われていない。

子音に対する処理 二話者の音響空間が一つの「連続かつ可逆な」写像で対応づけられると仮定している。当然，母音などの共鳴音と，無声子音とでは写像関数は異なることが予想される。現実的な解決策として，話者性による音響変化が小さい子音群については，従来の絶対的な特徴量を用い，構造的な処理結果との統合を検討している [30]。

構造デコーダの不在 構造表象は時間的に離れた二時刻間の差異を特徴量としている。従来のデコーディングは，発声が，各時刻における特徴量 o_t の時系列であることを前提としているものが多い。構造的特徴に対応した構造デコーディング（仮説探索）に関する初期的な検討を行った [31] が，十分な性能は得られていない。

[4]「話者が聞こえない音声表象」という構造表象の抽象性も，構造表象の理解を困難にしているのかもしれない。

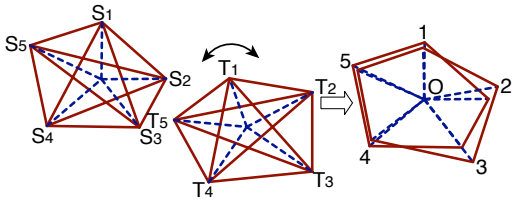

Fig. 7 Structure matching

音素列表象との整合性 音声から非言語情報を捨象した表象として，通常は音素列表象が使われる。任意の音素列に対応する構造を生成する場合，どのようなエッジを用いるべきか，という問題が生じる。構造表象は発声全体を表現するため，新規単語や任意音素列から構造を生成することは難しくなる。現在は母音列を対象として検討している [32]。

5 おわりに

定型発達を遂げた人間が示す，非常に高い音響的汎化能力を有する音声認識の実現に向け，筆者が提案する，非言語的情報を音響的に分離する一手法を紹介し，理論的及び技術的な話題提供を行った。

参考文献

[1] 峯松他, “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～（招待論文）”, 電子情報通信学会論文誌, J94-D, 1, 12-26, 2011.
[2] P.K. Kuhl, *Nature Reviews Neuroscience*, 5, 831–843, 2004.
[3] 原, コミュニケーション障害学, 20, 2, 98–102, 2003.
[4] 加藤, コミュニケーション障害学, 20, 2, 84–85, 2003.
[5] 早川, 月刊言語, 35, 9, 62–67, 2006.
[6] P. Lieberman, *Child Phonology*, vol.1, chap.7, Academic Press, 1981.
[7] S. King *et al.*, “The Blizzard Challenge 2009”, *Proc. Blizzard Challenge 2009 Workshop*, 2009.
[8] 深見, ひろしくんの本 (V), 中川書店, 2006.
[9] 綾屋他, 発達障害当事者研究, 医学書院, 2008（但し, 著者らの対談を含む）.
[10] T. Grandin, 我, 自閉症に生まれて, 学研, 1994.
[11] L.H. Willey, アスペルガー的人生, 東京書籍, 2002.
[12] 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.
[13] 東田他, この地球にすんでいる僕の仲間たちへ, エスコアール, 2005.
[14] W. Gruhn, “The audio-vocal system in sound perception and learning of language and music,” *Proc. Int. Conf. on language and music as cognitive systems*, 2006.
[15] 岡ノ谷, 春音講論, 1-7-15, 1555–1556, 2008（但し, 質疑応答を含む）.
[16] M.D. Hauser *et al.*, *Nature neurosciences*, 6, 663–668, 2003.
[17] T. Grandin, 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.
[18] U. Frith, 自閉症の謎を解き明かす, 東京書籍, 1991.
[19] 東川, 読譜力－「移動ド」教育システムに学ぶ, 春秋社, 2005.
[20] Y. Qiao *et al.*, *IEEE Transactions on Signal Processing*, 58, 7, 3884–3890, 2010.
[21] D. Saito *et al.*, *Proc. ICASSP*, 4485–4488, 2008.
[22] W. Labov *et al.*, *Atlas of North American English*, Mouton and Gruyter, 2005.
[23] T. Irino *et al.*, *Speech Commu.*, 36, 181–203, 2002.
[24] A. Mertins *et al.*, *Proc. ASRU*, 308–312, 2005.
[25] 益子, 秋音講論, 3-7-2, 105–106, 2005.
[26] N. Minematsu *et al.*, *Proc. Int. Workshop on Speech Recognition and Intrinsic Variations*, 47–52, 2006.
[27] D. Saito *et al.*, *Proc. INTERSPEECH*, 2047–2050, 2009.
[28] 峯松, 信学技報, SP2003-179, 25–30, 2004.
[29] X. Ma *et al.*, *Proc. SPECOM*, 350–355, 2009.
[30] M. Suzuki *et al.*, *Proc. INTERSPEECH*, 586–589, 2010.
[31] Y. Qiao *et al.*, *Proc. ASRU*, 118–123, 2009.
[32] 齋藤他, 信学技報, SP2009-77, 7–12, 2009.