線形・非線形変換不変の構造的情報表象とそれに基づく音声の音響モデリングに関する理論的考察*

○峯松信明，朝川智（東大・新領域），広瀬啓吉（東大・情報理工）

1　はじめに

年齢，性別，個人性，音響機器など，音声には不可避的に非言語的情報が付随し，これが音声の音響的特徴を変形させ，音声記号としては同一音であったとしても，物理的には異なる特性を持った音響現象として観測されることになる。従来の音声工学では，数千・万の話者の音声から同一記号音を収集して統計的にモデル化することで多様性問題の解決を図って来た。

幼児は両親という非常に限られた音環境で，電話越しのお婆ちゃんの音声にも対応できるようになる。限られた音環境での音声活動は一生続く。人の聞く声の半分は自分の声であり（speech chain），話者バランスのとれた音声コーパスの聴取など，不可能である。

幼児は両親との音声コミュニケーションを通して言語を獲得する。発達心理学では音声模倣と呼ばれるが，音を模倣する幼児はいない。父親の太い声を出そうと努力する幼児はいない。音韻意識が未熟な幼児は，聴取音声を個々のモーラに分割し，各モーラを自らの口で生成する，という技は不可能である。幼児と異なり，音を模倣するのが九官鳥である。優秀な九官鳥は聞けば飼い主が分かるが，優秀な幼児を聞いても飼い主の同定は全くもって不可能である。幼児は親の声の何を真似ているのだろうか?筆者らはこの問いを多くの発達心理学者に問いかけたが，その回答は得られなかった。その一方で,「語全体の音形，語ゲシュタルトを獲得する」と説明する。即ち，この語ゲシュタルトの物理的定義を求めないまま，その存在を仮定した議論を重ねているのが実情である。

相対音感者がメロディーをドレミで書き起こす場合，そのメロディーを移調したとしても，書き起こされるドレミ（階名）は変わらない。相対音感者は個々の音の同定を，その音の絶対的な物理特性ではなく，他の音群（曲全体）に対する相対的特性，即ち，音の機能的価値を感覚して行なう。全体把握から要素を同定し，長調の主音を「ド」と同定する。物理的に異なる音に対して「ド，と聞こえる」と答える。

異なる母音の音色の違いは，その母音を生成する際の口腔形状の違い，即ち，定常波の共鳴周波数の違いに起因する。これは，異なる形状の楽器が異なる音色を生成するのと物理的に等価であり，日本語には五つの母音楽器が存在する。音楽の音ストリームは音高の変化によって構成されるが，音声（特に母音）の音ストリームは，音色の連続変化によって構成される。本稿は，音声聴取を「音色変化に対する相対音感」として解釈し，こうすることで，頑健なる話者不変表象，語ゲシュタルト，が数学的に存在することを示す。

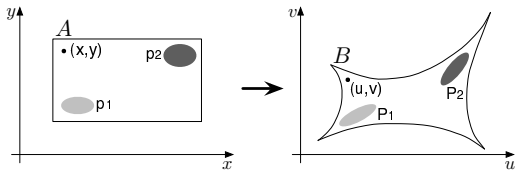

Fig. 1　一対一対応関係を有する二つの空間 A と B

2　頑健に話者不変な音声の音響的表象

図 1 に示す空間 A と B を考える。両者には一対一の対応関係があり，空間 A のある点は空間 B の対応点へ写像され，逆もまた成立する。但し，その写像関数は与えられていない。以下，一般性を失わない範囲で 2 次元空間を用いて説明する。空間 A，B の対応する二点を (x, y)，(u, v) とし，両空間の対応付け（変数変換）を一般的に下記の様に示す。

$$x = x(u, v), \quad y = y(u, v)$$

空間 A における事象 p を考える。但し，全ての事象は点ではなく，確率密度分布関数として存在する。

$$1.0 = \iint p(x, y)dxdy$$

空間 A における積分演算は，変数変換によって空間 B における演算へと変換可能である。

$$\begin{aligned}\iint f(x, y)dxdy &= \iint f(x(u, v), y(u, v))|J(u, v)|dudv \\ &= \iint g(u, v)|J(u, v)|dudv\end{aligned}$$

$g(u, v) \equiv f(x(u, v), y(u, v))$ であり，$J(u, v)$ はヤコビアンである。分布関数も同様に写像可能である。

$$\begin{aligned}1.0 &= \iint p(x, y)dxdy \\ &= \iint p(x(u, v), y(u, v))|J(u, v)|dudv \\ &= \iint q(u, v)|J(u, v)|dudv \\ &= \iint P(u, v)dudv \qquad p() \text{ in A} \rightarrow P() \text{ in B}\end{aligned}$$

$q(u, v) \equiv p(x(u, v), y(u, v))$，$P(u, v) \equiv q(u, v)|J(u, v)|$ である（**図 1** 参照）。以上の道具を用いて，空間 A と B 間に存在する頑健な不変量について考察する。空間 A の分布 p_1 と p_2，及び，これらの空間 B への写像分布 P_1，P_2 を考える。p_1 と p_2 に対するバタチャリヤ距離は下記式で表される。

$$BD(p_1, p_2) = -\ln \iint \sqrt{p_1(x, y)p_2(x, y)}dxdy$$

*Linear and non-linear transformation invariant representation of information and its use for acoustic modeling of speech. by N. Minematsu, S. Asakawa, and K. Hirose (The University of Tokyo)

これを，空間 B における積分演算へ変換する。

$$
\begin{aligned}
& BD(p_1, p_2) \\
&= -\ln \iint \sqrt{p_1(x,y)p_2(x,y)}dxdy \\
&= -\ln \iint \sqrt{q_1(u,v)q_2(u,v)}|J(u,v)|dudv \\
&= -\ln \iint \sqrt{q_1(u,v)|J|}\sqrt{q_2(u,v)|J|}dudv \\
&= -\ln \iint \sqrt{P_1(u,v)P_2(u,v)}dudv \\
&= BD(P_1, P_2)
\end{aligned}
$$

即ち，空間 A におけるバタチャリヤ距離は，空間 B における対応する二分布間の距離と等しくなる。この不変性は，ヤコビアンによる変数変換が可能であれば，非線形変換を含む，広い変換群に対して成立する。また，カルバックライブラ距離，ヘリンジャ距離などでも成立する一般的性質である。ここで，両空間の写像関数やヤコビアンを求める必要は一切無い。音声認識の話者適応，音声合成の話者変換いずれも，異なる話者間におけるケプストラム空間の写像を定義する操作に他ならない。上記の議論は，その写像に対して，頑健なる不変量が存在することを保証する。

3　多次元音楽としての音声モデリング

相対音感者は，音高変化を伴う音ストリームに対して，その変化のみを感覚し，個々の音を階名として同定する。この能力はヒトのみが有しており，動物は基本的に絶対音感者である。個々の音要素の同定が行なえない幼児は，親の声と自らの声の何が一致した時に「同一」を感覚し，満足するのだろうか？

図 2 に CDEFG と鍵盤を叩いた場合の基本周波数の様子を示す。これを「ドレミファソ」と書き起こす相対音感者は GABCD を聞いても「ドレミファソ」と聞こえる。音高の遷移パターンが同一だからである。即ち，**図 2** の縦軸の数値は相対音感者には無意味である。筆者らが提唱する音声の全体的・構造的表象は，例えば**図 3** に示すように，音声ストリームを分布系列へと変換し，全ての分布間距離を求め，距離行列を算出（幾何学構造を規定）することで求まる。この操作を**図 2** 同様，位相空間で考えたものが，例えば**図 4** である。「あいうえお」という発声を，F1/F2 平面に対して時間軸が延びた三次元空間を用いて示している。音高という一次元物理量に対する変化が音楽であり，音色（この図では F1/F2）という二次元物理量に対する変化が音声である。即ち，多次元音楽として音声を捉えている（音色の表現には，複数次元が必要）。前節の議論は，この遷移パターンを変換不変とすることが可能であることを意味する。なお，線形変換（アフィン変換）は幾何学構造を歪ませる変換として使われる。同様に非線形変換も構造を容易に歪ませる。これらに対して構造不変性が成立するのは，構造が歪まないように空間を歪ませているからである。絶対的な空間・時間は仮定しない。

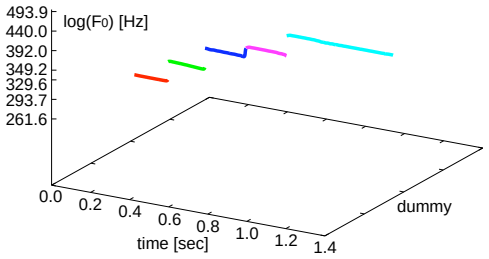

Fig. 2　CDEFG に対する基本周波数の遷移パターン

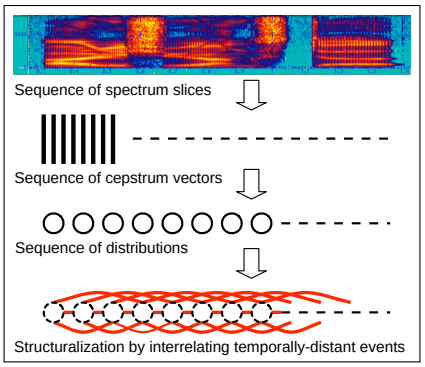

Fig. 3　事象間差のみを抽出して構成される不変構造

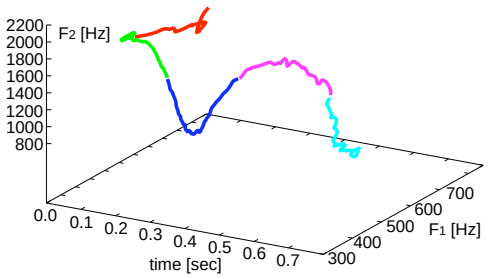

Fig. 4「あいうえお」に対する F1/F2 の遷移パターン

4　種々の検討と考察

相対音感者は時として「全ての長調の曲がハ長調に聞こえる」と主張する。これは彼らが「ド」と聞く音全てが，同一の鍵盤（C の鍵盤）によって生成された，即ち，同一の物理特性を持つとの解釈から来る勘違いである。音声工学は「あ」と聞こえる音全てに物理的同一性を期待して，数万人の音声を集めた訳だが，この勘違いをどう考えるべきなのだろうか？

絶対音感者の中には，C の音で始まらない「カエルの歌」は「カエルの歌」では無いと主張する者もいる。音声の研ぎすまされた絶対音感者は，二話者の「おはよう」の同一性認知が困難となるはずだが，これは，一部の自閉症者に見られる。当然，音声言語は無い。そのような自閉症者が，他者の声をそっくり模倣する例も知られている。当然，そこに意味は無い。

相対音感者の一部は，階名としての音同定が困難である。**図 2** の遷移パターンを追従するだけである。同様に**図 4** のパターンを追従する形で音声を生成すれば，音韻同定が困難となり，文字言語に支障を来すだろう。アインシュタインやグラハムベルで知られる失読症が，似た症状を示す。音声を音韻（文字）列として認知できない（する必要がない）。しかし，音声言語に難は無い。非常に興味深い事実である。

[1] 峯松他, 人工知能学会, SIG-Challenge-0624-6, 35–42, 2006.