「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――

第4章

話し言葉の音声

峯松 信明

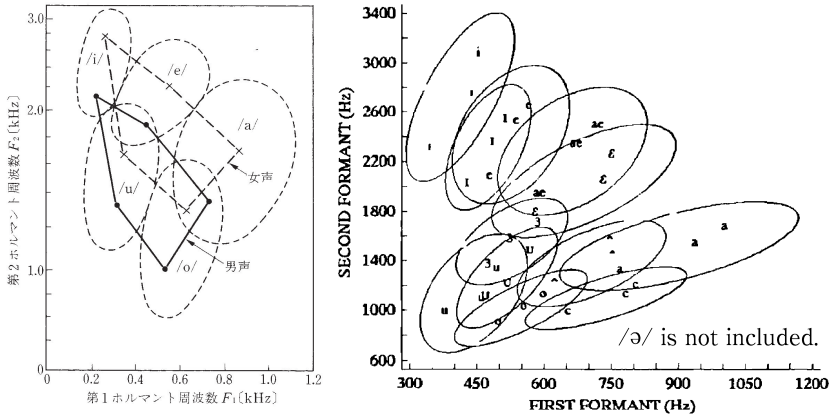
はじめに ～何、この変なタイトル？～

タイトルを見て、多くの読者が首を傾げていることだろう。しかし、十一頁の本記事を読み終えた時に、ほぼ全ての読者に私の意図は通じるもの、と考えている。そう。「「あ」という声を聞いて、それを有限個の音カテゴリの一つとしての母音「あ」であると同定する能力は、音声言語運用の必要条件ではない。」との主張を本稿では展開する（文献1）（文献2）。

そんな馬鹿な、と思われるかもしれない。こんな実験を考えてみよう。身長300㎝の巨人と50㎝の小人に孤立母音を発声してもらう。通常音声学の教科書には、F_1・F_2の母音図が出ている（**図1**参照）。複数の男性／女性のサンプルから、凡そ男性の各母音はこの領域、女性の各母音はこの領域にある、といった図である。フォルマント周波数（共鳴周波数）は声道長に依存するため、身長が50㎝、300㎝という架空の大人を想定した場合、彼らの母音は、通常知られている領域の外に存在する。そのような母音でも、現在の音声分析・再合成技術を使えば非常に高品質な音声として生成できる。さて、聞いたことのない母音音声を孤立提示されて、読者は同定できるだろうか？

文献(5)によれば、これは困難なタスクであることが分かる。しかし、その巨人、小人が無意味モーラ列を単

図１　日本語及び米語の第一、第二フォルマント周波数（文献３）（文献４）



語のように発声した場合、その無意味語の中の母音は凡そ復唱できる様子が示されている（文献６）。つまり、個々の音を孤立提示した場合その同定が難しいのに、意味・統語情報の無い連続モーラ列に対して、その復唱が凡そ可能であり、平仮名として書き起こすことも可能である。意味のあるフレーズが聴覚提示されれば、その復唱／書き起こしは更に容易になるだろう。何故、このようなことが起こるのだろうか？

孤立音同定と音ストリームの復唱・書き起こし ～それって不思議なことなの？～

音声の音響的特徴というのは、性別、年齢、マイク、部屋など様々な要因によって変化する。声を聞いて個体を同定できるのは、音響的に言えば、喉の形状が声の音色を決定し、（デフォルトの）喉形状は人によって異なるからである。言うなれば、母音「あ」の音響的実体は、数十億パターン存在する。マイクの数、部屋の数を考えれば、更に爆発する。このような多様性に富む音声を、我々は頑健に処理できる。音声物理の多様性と音声知覚の不変性。音声科学の世界で古くから議論されている謎である。実は本稿は、この謎を非音声学的に解いた私の

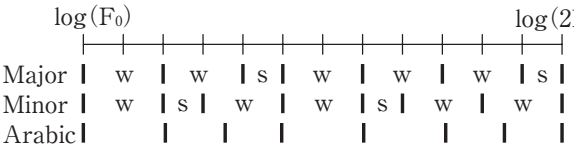
解答の紹介記事である。そして、その解答は「音響音声学は音声言語を議論するための妥当な科学とは必ずしも言えない。」と主張する。

個々の音を有限個のカテゴリの何れかとして同定できないのに、提示された音声を復唱できる。この現象を不思議に思われた方、同様の問いをメロディーに対して考えてほしい。提示されたメロディーをハミングして復唱することは、そのメロディーの個々の音を音名や階名といった音カテゴリ（ドレミ…）として同定することを要求するだろうか？　テレビを賑わすアイドル歌手の何割が、孤立提示された音を正しくドレミに落とせるだろうか？個々の音をシンボルとして同定することなど、歌手デビューの必要条件ではない。

少し話を整理しよう。特にドレミには二つの定義があるので注意が必要である。鍵盤を一つたたいて音を出し、それに対してドレミ同定が正しくできる場合、その人は絶対音感を持つと言い、その人にとってドレミとは、音高の絶対的特性に付けられた名前（音名）である。その一方で、メロディーをドレミで書き起こせるが、カラオケに行って、キー（調）を上げ下げしても、書き起こされたドレミ列が変わらない方々もいる。言語化できる相対音感者である。この場合、ドレミは階名である。絶対音感者の場合、キーを変えれば当然ドレミ列は変わる。音が変わるのだから。そして絶対音感が極端に強い場合、キーを変えると（即ち移調することで）「異なる曲」としての知覚がまず起こるようになる（文献7）。音が変わるからである。相対音感者の場合、このようなことは無い。なお、相対音感者の中にはドレミに落とせない方々もいる（言語化できない相対音感者）。絶対音感、言語化できる相対音感、言語化できない相対音感、読者はいずれの音感をお持ちだろうか？

さて、キーを上げ下げしても書き取るドレミ列が変わらない場合（階名）、どういった音響的キューに基づいてドレミを感じ取っているのだろうか？　音階の個々の音の音程（音高差）を**図2**に示す。長音階の場合八つの音が全音、全音、半音、といった決まった間隔で並んでおり、この音配置はキーを上げ下げしても変わらない。言語化可能な相対音感者は、この音配置の情報に基づいてドレミが聞こえて来る。例えば二つの音が三全音離れていれば、その二音に対して「ファとシ」あるいは「シとファ」という内なる声が聞こえてくる。逆に彼らは、孤立音に対してドレミ同定が出来ない。メロディーの全体像があって初めて、ドレミが聞こえて来る。逆に言えば、全体から入るからこそ、キー（調）に依存しない頑

図2　音高配置のバリエーションとしての長音階、短音階、アラビア音階（w＝全音、s＝半音）



健な（階名としての）音同定が可能となる。

孤立音の同定は出来ない。でも、音ストリームをシンボル列として書き起こせる。

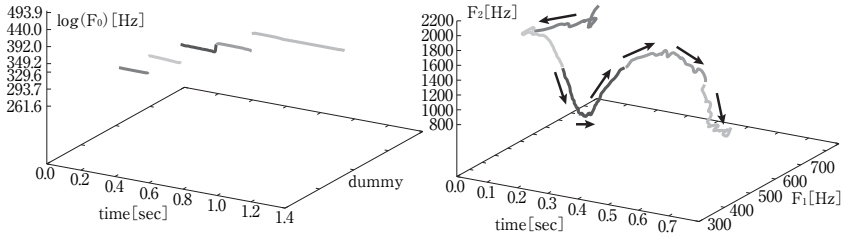
前節でちょっと不思議な現象について述べた。だが、音楽のことを少しかじった者であれば、前節の現象の何が不思議なのか分からなかったはずである。孤立音として提示された場合それが「何であるのか」分からなくても、連続的な音ストリームは復唱できるし、必要であれば、シンボル列として書き起こせる、という現象は「当たり前」なのである。

音声と音楽の類似性 ～音色の体系と音高の体系～

確かにメロディーの場合は「当たり前」だが、メロディーは基本周波数が主要な音響量であるのに対し、音声の場合はそうではない。よって、両者を同一の枠組みで議論するのはおかしい、という批判が聞こえてくる。本節ではこの点について考えてみたい。

音楽（メロディー）を最も簡素化して考えれば、上記したように音高の時間的な動的パターンとなる。一方、音声（ここでは母音に代表される共鳴音を考える）は、音色の時間的な動的パターンである。音色は共鳴と同義であり、それは、喉の形によって決定される。つまり、口の開閉（舌の位置の変化）によって、喉の形状を変化させ、音色（共鳴）を動的に動かしているのが音声である。男女が同じ歌を歌ったとする。男性の声が低いのは（音高の男女差）、男性の声帯が長くて重たいからである。その男女が歌詞を読み上げたとする。男性の声が太いのは（音色の男女差）、男性の喉が長いからである。前者に対して音高の相対音感が当たり前のように議論され、これがあるから、音高の絶対レベルは異なるものの、男女の歌声の音高パターンは同一であると知覚できる。私が本稿で検討したいのは、音色の相対音感である。個々の音の絶対的な音色ではなく、音色の相対的特性や動的パターンを通して音声を認知するからこそ、巨人の「あいうえお」と、小人のそれとが同一の音色パターンである、と知覚できるのではないのか、という議論である。この同一性

図３　F_0の動的変化としてのCDEFGと音色の動的変化としての「あいうえお」

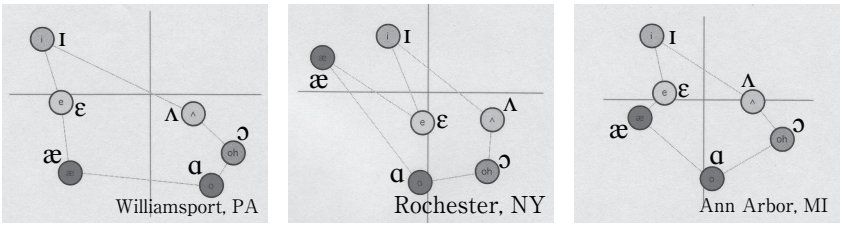


認知は、刺激をシンボル列に変換することを要求しない。言語化できない相対音感者が男女のメロディーを同一であると知覚するように、男女の読み上げ音声を、シンボル列として表象できなくても、両者の同一性を感覚できる機械を作ることは可能なのでは、ということである。そして、その機械は、私の研究室にある。

図3を見て戴きたい。左図はCDEFG（ハ長調のドレミファソ）における音高の動的パターンである。3次元表示しているのは（3次元目はダミー）、隣の「あいうえお」と比較するためである。右図は私の「あいうえお」を音色の動的パターンとして描いている。F_1、F_2、時間の3次元表示をしている。左図でこのパターンを移調しても音高の配置関係は変わらない。だから、言語化できる相対音感者は調に依らず「ドレミファソ」と答える。右図のパターンを女声化するとどうなるだろうか。図1には男声の平均的な5母音配置と女声のそれとを示している。さて、これを2次元の移調として捉えることは不自然な思考だろうか？　音そのものは変わる。だが、音群の配置は変わらず、音群配置の「場所」が移動しているだけだ、と。そして、音をシンボル同定するのではなく、音と音の話者不変の配置情報に基づいて単語音声を同定する機械は作れないだろうか？　その場合、年齢、性別の違いに頑健な処理が可能となるはずである。その機械は、私の研究室にある。

音高における音群配置と、音色における音群配置の議論を更に進めよう。例えば、長音階に対して、その配置をちょっと変えるとそれは短音階となる。また、その他にも色々な音高配置のバリエーションがあり、中世の教会音楽で使われていた。アラビア音階も独特な配置パターンを採択している（図2参照）。西洋音楽をアラビア音階（配置）で弾くと、一見「壊れたピアノでの演奏？」と聞こえるが、中近東出身の留学生は「あ～、懐かしい。」と答える。西洋的には崩れた音高配置こそが彼らの母語的配置なのである。

図4　音色配置のバリエーションとしての米語方言（文献8）



音色の空間における音群配置（母音配置）を考えた場合、それを崩すと何になるだろうか？　音声学を知る者なら、これが西洋の「方言」に対応することはご存知だろう。**図4**には米語の方言における母音群配置を示している。アラビア語の母音に引きずられた母音群で英語を発声すれば、米語母語話者には「壊れた」発声のように聞こえるが、アラビア人にとってみれば「馴染み深い」発声となる。

音高における音群配置と、音色における音群配置の見事なまでの対応が存在しているのである。この事実を既知とした場合「メロディーと音声は、音響量が異なる。よって、両者を同一の枠組みで議論するのはおかしい。」と主張することに、どれほどの価値があるのだろうか？対象とする現象の物理実体（物理量）に縛られることなく、その現象を支配する原理・原則を見極めるのが科学的研究の基本姿勢であるとすれば、音高や音色といった物理量に縛られずに、複数の現象の共通性にこそ目を向けるべきであろう。言語化できる相対音感は個々の音の音高ではなく、音と音の音高差に基づいてメロディー中の個々の音を同定する。冒頭に書いた巨人と小人の実験も、個々の音の音色ではなく、音と音の音色差に基づいて発声中の個々の音を同定している、と考えた方が自然ではないだろうか？

体系（システム）の一員としての音要素 ～音韻論と音声学～

「個々の音は、他の音群から独立して存在しているのではなく、音群は体系、システムとして存在しており、その要素として個々の音があるだけである。」と考える訳だが、これは、音韻論の基本的な考え方である。一方音声学は、他の音がどうであろうと与えられた音の絶対的な物理的特性に目を向ける。私の研究室では、音韻論の価値感に基づき音声を処理する機械を構築した。個々の音は、単独では、それが何であるのか分からない。で

も、単語音声は同定できる。この機械を作る前に、解決しておくべき問題があった。音楽の場合、調不変の音高システムを物理的に定義することは簡単である。基本周波数の対数をとれば、それで調不変の音高システムは議論できる。音声の場合にはそうはいかない。話者不変の音色システムを物理的、数学的に定義することはできるのだろうか？厳密な数式展開は参考文献に任せるとして（文献9）、結論を述べさせてもらう。音色の異なる音群を、話者不変のシステムとして纏め上げる枠組みを数学的に導出している。そして、その上で研究を行なっている。私の研究室にある機械はその枠組みの上で動いている。言うなれば、（構造）音韻論を数学的、物理的に構築している、ということである。対数をとった音高差が調不変量となるように、我々の構築した数学的枠組みでは、音色差が話者不変量となる。そして、観測された音群に対して全ての（二音間）音色差を計測すれば、その音群を話者不変の体系、システムとして定義できる。それが言語であると私は考えている。本雑誌の定期購読者であればピンと来るのでは無いだろうか。「言語は観念的差異と音的差異の体系である。」と述べたのはソシュールである。私はそんな主張を知ることなく、音声を音的差異の体系として纏め上げることで、話者不変の音声物理量を導出した。

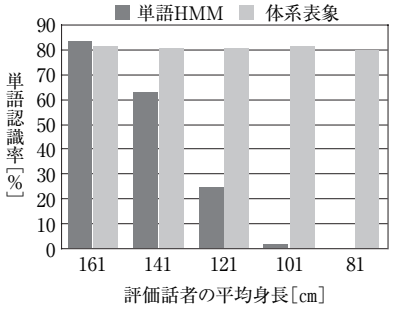
実験的検証　～本当に存在する巨人と小人～

図5に世界一の巨人と世界一の小人（但し大人）の写真を示す。約240㎝と約70㎝である。人が聞く声の約半数が自分の声であることを考えれば、巨人の耳は常に太い声に曝され、小人の耳は常に細い声に曝される。にもかかわらず、会ったその場で彼らは会話を始める（両者はモンゴル人）。音声の物理現象を知る者は、このシーンが不思議でたまらない。両者の声の音響的特性には雲泥の差がある。両者の「あ」には雲泥の差がある。でも両者は笑顔で会話する。相手の声が過去に聞いたことのある声と如何に離れていようとも。

簡単な実験をしてみる。普通の成人男女八名の五母音連結音声（「あえおうい」など。合計一二〇種類ある）を単語として考え、一二〇単語の孤立単語音声認識器を二種類構築した。一方は現在の音声認識技術の中核を成す隠れマルコフモデル（ＨＭＭ）を用いたものである。声の絶対的音響特性を統計的にモデル化し、与えられた音をシンボル同定する技術である。もう一方は私の研究室で開発している、単語内の音群を話者不変の体系として

図５　世界一の巨人と世界一の小人（但し大人）

図６　単語HMM、及び、体系表象による単語音声認識結果



モデル化する技術である。この技術は、孤立音は一切同定できない。評価用データは、認識器構築時に使用した男女八名とは異なる男女八名が五回ずつ発声した一二〇単語である（合計四八〇〇単語発声）。両認識器は入力音声を、一二〇単語の何れかとして同定する。この評価用データを、音声技術を用いることで小人化した場合の結果も含めて**図６**に示す。絶対量に基づく認識器は80cmの小人の声は一切認識できないが、私の研究室で開発している認識器は性能が落ちない。落ちることができない。人間の行なう音声活動の物理的モデルとして、音声学的価値感に基づくモデルと、音韻論的価値感に基づくモデルのどちがが科学的に正しいと読者は考えるだろうか？　ViaVoice という音声認識ソフトウェアをご存知の方も多いと思う。音声は年齢、性別など様々な要因で変化する。音の絶対量に基づいて（音→シンボル変換技術に基づいて）頑健な認識器を作る場合、様々な「あ」という声を集める必要が生じるが、ViaVoice の開発には三五万人の音声を集めたことが発表され、その数字が広告に使われている（文献10）。読者はこれを「すごい技術！」それとも「あきれた技術！」のどちらとして受け止めるのだろうか？　少なくとも私は三五万人の声を聞いて初めて電話越しのお婆ちゃんと会話できるようになった子供を見たことがない。そもそも、全人口が三五万人に満たない言語は多数存在する。

様々な考察 ～音声言語運用の必要条件は何か？～

ヒト以外の霊長類は音高の相対音感が極めて乏しいこ

とが、多くの進化人類学研究の論文で示されている（例えば、文献11）。移調すれば同一性認知が困難となる。彼らは「言語化できない絶対音感者」である。そもそも刺激の絶対量に基づく処理系と相対量に基づく処理系とでは前者の方が実装が容易であり、生物進化においても、音高の絶対量に基づく処理系が古く、相対量に基づく処理は非常に新しい（文献12）。チンパンジーに言語を教える試みは数多く行なわれているが、その殆どが視覚言語である（ジェスチャー、ボタンなど）。声は使わない。チンパンジーに言語教育を試みた研究者の言を借りれば「声はなかなかトークンにならない。」（文献13）。音高は1次元、音色は多次元である。音色の相対処理をヒト以外の霊長類に期待するのは難しいのかもしれない。音色が違えば違う音になるのだろう。

極めて強い（音高の）絶対音感者は、あるメロディーを移調すると、その前後で同一性認知が難しくなる。であれば、母親の「おはよう」と父親の「おはよう」の同一性認知が難しい、極めて強い（音色の）絶対音感者がいても不思議ではない。そのような事例は一部の自閉症者に見られる（例えば、文献14）。当然彼らは音声言語を操ることが極めて困難である。異なる音は異なる音として捉えてしまえば、音声コミュニケーションは破綻する。

実話を一つ紹介しよう。私はメロディーの階名書き起こしと、音声の書き起こしとを並列化させ、色々と思考してきた。ある時、言語化出来ない（シンボル列化できない）相対音感者の音声版を考えた。言語化出来ない相対音感者は「あるメロディーを聞かせます。三番目の音を覚えて下さい。次に別のメロディーを聞かせます。同じ音が出て来たら手を挙げて下さい。」との問いに困惑する。ならば「ある発話を聞かせます。三番目の音を覚えて下さい。次に別の発話を聞かせます。同じ音が出て来たら手を挙げて下さい。」という問いに困惑する人がいても不思議ではない。音ストリームをシンボル列（音韻列）として認知することが困難な方々である。文字言語が極めて難しい方々である。音ストリームの全体像を捉える傾向が強く、相対音感度が高い方々、ということだが、それは日本語やイタリア語などの母音数が少ない言語よりも、英語のように母音数が多い言語に頻繁に見られるはずである。何故なら図1を見れば分かるように、母音数が増えると、母音間の重なりが増えるからである。音声コミュニケーションにおいて絶対量を使い難くなるからである。「音声言語は流暢だし雄弁。頭は良いのかもしれない。でも何故か本が読めない、手紙が書けない。そういう成人が米国や英国には多いはずである。」とい

う予言をした。でも、この予言、人には言えなかった。私の思考が正しければ、彼らは当たり前のように存在するはずなのだが、そんな人が存在することが信じられなかったからである。ある時、勇気を出して（恥をかくことと覚悟で）言語聴覚士に、恐る恐る、聞いてみた。

「音声言語は流暢だし雄弁。頭は良いのかもしれない。でも何故か本が読めない、手紙が書けない。そういう成人が米国や英国に多かったりしませんか？　え～と、教育を受けていないとか、そういう事ではなく、彼らの認知特性として文字言語が何故か難しい……」

「先生、ディスレクシアってご存知なんですか？　特に音韻性のやつ。」

「でぃすれ……何ですかそれ？」

「変だな。先生、今、自分でディスレクシアの説明してたじゃないですか。」

四一年間の人生の中で、あれほど口をあんぐり開けたことは無い。顎が外れるかと思った。これは実話である。私は彼ら（文献15）の存在を、音声の物理学に基づいて予言していた。

音そのもの、即ち、声の絶対的な物理量への着眼を基本とする音響音声学、そして、それを工学として纏め上げ、様々な技術を構築して来た音声工学。これらは科学的にどれだけ正しい活動だったのだろうか？と最近考えることがある。音→（ＩＰＡ）シンボル変換の能力は、音声学者になるためには必須の能力なのかもしれないが、それは音声言語運用の必要条件ではない。音をシンボル化できる能力ではなく、音としては異なる二つの音ストリームの中に、物理的に同一の情報が埋め込まれている（符号化されている）ことを認知する能力の獲得こそ、音声言語を操るための第一歩であると私は考えている。音響音声学は声を議論するには非常に適した科学であるが、音声言語を議論するための妥当な科学とは言えない、と考えている。読者から忌憚の無い意見を頂戴できれば幸いである。

参考文献

1　峯松信明他（二〇〇七）「孤立音を聞いて音韻同定できる能力は音声言語運用に必要か？」（『日本音声学会全国大会予稿集』一三五～一四〇頁）

2　N. Minematsu et al., (2008) "Consideration of infants' vocal imitation through modeling speech as timbre-based melody," in New Frontiers in Artificial Intelligence, LNAI4914, pp.26-39, Springer

3　R. K. Potter et al., (1950) "Toward the specification of speech." J. Acoust. Soc. Am. vol.22, no.6, pp.807-820

4　J. Hillenbrand et al., (1995) "Acoustic characteristics of American English vowels." J. Acoust. Soc. Am. vol.97, no.5,

pp.3099-3111
5　青木美和他（二〇〇四）「スケール変形した日本語５母音の知覚特性」（『日本音響学会秋季講演論文集』２－Ｐ－６、三七三～三七四頁）
6　林芳恵他（二〇〇七）「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」（『日本音響学会春季講演論文集』２－Ｑ－27、四七三～四七四頁）
7　宮崎謙一（二〇〇四）「「絶対音感」はどこまで分かったか？」（『日本音響学会誌』60巻11号、六八二～六八八頁）
8　W. Labov et al. (2001) Atlas of North American English, Walter De Gruyter Inc.
9　峯松信明他（二〇〇七）「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」（『日本音響学会春季講演論文集』１－Ｐ－12、一四七～一四八頁）
10　http://tepia.or.jp/archive/12th/pdf/viavoice.pdf
11　M. D. Hauser et al. (2003) "The evolution of the music faculty: a comparative perspective," Nature Neurosciences, vol.6, pp.663-668
12　D. J. Levitin et al., (2005) "Absolute pitch: perception, coding, and controversies," Trends in Cognitive Sciences, vol.9, no.1, pp.26-33
13　S. Kojima, (2003) A search for the origins of human speech: auditory and vocal functions of the chimpanzee, Trans Pacific Press
14　東田直樹他（二〇〇五）『この地球にすんでいる僕の仲間たちへ』（エスコアール出版社）
15　サリー・シェイウィッツ（二〇〇六）『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～』（Ｐ

ＨＰ研究所）
（みねまつ・のぶあき　東京大学大学院工学系研究科准教授）