1F2-3

多次元音楽としての音声モデリングと音声模倣

Speech modeling as multidimensional music and speech mimicking

峯松 信明*1　　西村 多寿子*2
Nobuaki MINEMATSU　Tazuko NISHIMURA

*1東京大学大学院新領域創成科学研究科
Graduate School of Frontier Sciences, The University of Tokyo

*2東京大学大学院医学系研究科
Graduate School of Medicine, The University of Tokyo

Music is composed of dynamic changes of pitch. Similarly, speech is composed of dynamic changes of timbre. Some of people with relative pitch hear the same sequence of Do, Re, and Mi both in a musical piece and its transposed version. Their perception is key-invariant. This mechanism can also be considered in speech perception where phoneme perception is done within the sound structure in which the target sound is embedded. This paper shows a method of implementing this framework for speech recognition where only the phonic contrasts were extracted for recognizing a spoken word and a single training speaker is basically sufficient for speaker-independent speech recognition. In this framework, speech is viewed as multidimensional music and an isolated sound is not identified at all. Here, it is discussed that this rather strange strategy is not strange at all and some investigations are done as for speech mimcking done by infants.

1. はじめに

音声には，話者・環境・聴取者に起因する，多様な音響歪みが不可避的に混入する。その一方で幼児は，大部分が「母親と父親の音声」という音声資料の提示を通して音声言語を獲得する。その後も，人の聞く音声の半分は自らの音声であることは自明である（スピーチ・チェイン）。即ち，音声言語の使い手は誰でも，音響的に非常に偏った話者性の音声資料の提示を通して，超頑健な音声情報処理能力を身につける訳である。

音響音声学／音声工学では，個々の音韻の音響モデルを，数千・数万という話者の音声を集め，分布としてモデル化することで多様性問題を扱って来た。両者は何が異なるのだろうか？

同一言語内容を異なる二話者が発声したとする。両音声資料の中に音韻「あ」を感覚する部位があったとする。音声学／音声工学のいずれにおいても，その二つの部位に某の同一物理量があることを前提とした議論を繰り返して来た。そしてそのために，多数の話者の「あ」を求める方法論に陥る。この前提は果たして正しいのだろうか？本稿では，音楽を例にとり，この前提が必ずしも妥当ではないことを示し，「あ」という音韻を感覚させる部位に物理的同一性は不必要である事を説明する。これは，この不適切な前提の上に長年に渡って構築された枠組みも不適切であることを主張するものである[1]。

本稿では，この不適切な枠組みに変わる枠組みとして筆者らが提唱している音声表象を解説し，それに基づく音声認識系の構築，更には幼児の音声模倣の謎についても検討を加える。

2. 相対的な音感と音楽・音声

2.1 言語化できる相対音感とその頑健性

図 1 に示す二つの曲をドレミに落とすよう人に依頼した場合，どのような反応が考えられるだろうか。凡そ返答は三通りある。「初めはソーミソドー，次がレーシレソーですね」と答えた場合，その人は絶対音感の持ち主であり，その人にとってドレミは音名である。「両方ともソーミソドーと聞こえてきます」と答えたとすれば，その人は「言語化ができる」相対音感者であり，その人にとってドレミは階名である。この場合，音階の構造（全全半全全全半という音高遷移の枠組み）をメロディーの中に感覚し，例えば長調の場合，主音をドとして知覚し，同様にして，上主音，中音，下属音を，レミファ，として知覚する。即ち，全体の把握があって初めて要素音の同定が可能となり，孤立要素音のみでは一切音の同定は出来ない。また「すみません。ラーラララーとしか歌えません」となった場合，その人は「言語化できない」相対音感者となる。

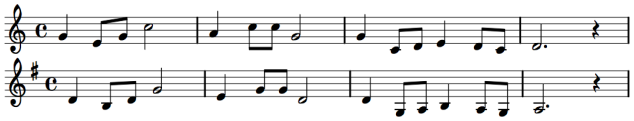

図 1: とあるメロディー（ハ長調）とその移調版（ト長調）

さて，言語化できる相対音感を持っている場合，移調によっていくら「ド」の音高が変わろうとも，彼らは「ド」という「内なる声」を聞く。即ち，メロディー全体の音群構造の把握から入ることで，個々の音の絶対的な物理特性とは無関係に個々の音を音シンボルとして同定する。その結果，歌い手の平均音高に依存せず，超頑健なドレミ書き起こしが可能になる。

さて，同様の枠組みを音声知覚（音韻同定）においても考えることはできないのだろうか？この場合の必要条件は，音群構造の不変性である。階名同定の場合，個々の音は，$\log(F_0)$ 軸において「全全半全全全半」という間隔を置いて配置される。この音群配置が，歌い手の平均音高に依存せず保持されることが頑健な階名同定の必要条件である。タスクとして母音系列同定を考えた場合，音響空間における母音群配置が話者不変になることが必要である。これは満たされるのだろうか？

2.2 音色の動的変化パターンとしての音声

母音の生成は声道（音響管）の共鳴現象である。これは，管楽器における音生成と物理的には等価である。即ち「あいうえお」の違いは，声道形状の変化による共鳴現象の変化である。結局，音色を表現するための最も簡素な物理パラメータはフォルマント周波数となり，ここでは F_1 と F_2 を考える。なお母音同様，複数の管楽器を F_1/F_2 平面上にプロットし，音色配置を示す場合もある。**図 2** に F_0 の動的変化としての CDEFG，及び音色の動的変化としての/aiueo/を示す[2, 3]。前者を移調しても，階名同定が要求する音群配置は不変であり，この動的パターンは上下に移動するだけである。一方/aiueo/の動的パ

連絡先: 峯松信明（東京大学）mine@k.u-tokyo.ac.jp

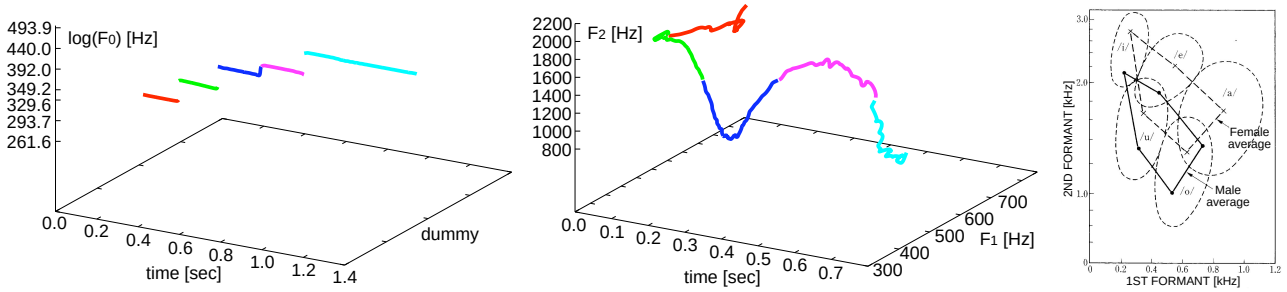

図 2: F_0 の動的変化としての CDEFG と音色の動的変化としての/aiueo/，及び，日本語母音図

ターンであるが，日本語母音図（**図 2**）に示すように，音響音声学では，F_1/F_2 平面で男声の母音構造を移動すると女声の母音構造に重なると言われる。このような単純な写像で変換できれば，母音構造の話者不変性は容易に実現できるが，厳密にはこのような単純な写像で変換できる訳では無い。音声合成の話者変換技術は，話者 A の音響空間と話者 B の音響空間との対応付け（写像）を精密に定義することで実装されるが，音群構造の不変性は，この両空間における不変構造を要求する。

3. 頑健に話者不変な音声の音響的表象

3.1 頑健に話者不変な音声表象の数学的導出[1]

図 3 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。

$$x = x(u, v), \quad y = y(u, v) \tag{1}$$

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。

$$1.0 = \iint p(x, y)dxdy \tag{2}$$

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。

$$\iint f(x, y)dxdy = \iint f(x(u, v), y(u, v))|J(u, v)|dudv \tag{3}$$

$$= \iint g(u, v)|J(u, v)|dudv \tag{4}$$

$g(u, v) \equiv f(x(u, v), y(u, v))$ 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。

$$1.0 = \iint p(x, y)dxdy \tag{5}$$

$$= \iint p(x(u, v), y(u, v))|J(u, v)|dudv \tag{6}$$

$$= \iint q(u, v)|J(u, v)|dudv \tag{7}$$

$$= \iint P(u, v)dudv \qquad p() \text{ in A} \rightarrow P() \text{ in B} \tag{8}$$

$q(u, v) \equiv p(x(u, v), y(u, v))$，$P(u, v) \equiv q(u, v)|J(u, v)|$ である（**図 3** 参照）。以上の道具を用いて，空間 A と B 間に存在する

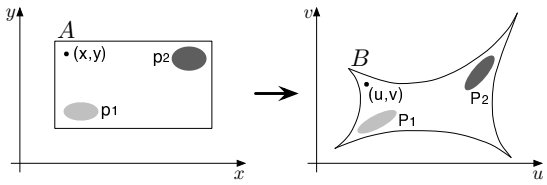

図 3: 一対一対応関係を有する二つの空間 A と B

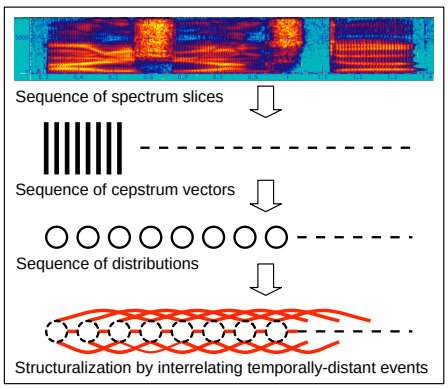

図 4: 事象間差のみを抽出して構成される不変構造

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のように変換される。

$$BD(p_1, p_2) = -\ln \iint \sqrt{p_1(x, y)p_2(x, y)}dxdy \tag{9}$$

$$= -\ln \iint \sqrt{q_1(u, v)|J|}\sqrt{q_2(u, v)|J|}dudv \tag{10}$$

$$= -\ln \iint \sqrt{P_1(u, v)P_2(u, v)}dudv \tag{11}$$

$$= BD(P_1, P_2) \tag{12}$$

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（頑健な不変性）。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。

以上，頑健なる不変項が分布間距離として存在することを数学的に導出した。この不変項に基づいてある発話を不変的に表象した場合，**図 4** に示すように，音声ストリームを分布系列へと変換した後（系列長=N），時間的に離れているものも含め，全ての二分布間距離を求めて $N \times N$ の距離行列として発声を表象することになる。この時，個々の音響事象の絶対的な物理特性は一切捨象する。距離行列は一つの幾何学構造を規定するが，この構造が非言語的特徴に一切不変な音声の物理的

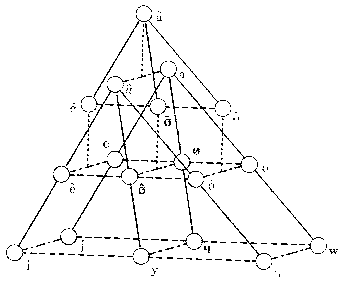

図 5: ヤコブソンによるフランス語の母音・準母音構造

表象である。**図 5** にヤコブソンによるフランス語の母音・準母音構造を示す[4]。構造音韻論では，このような構造が話者に依らず観測されることを主張するが，筆者らが提唱する音声表象は構造音韻論の物理的・数学的解釈に他ならない。

3.2 音高に対する相対処理／音色に対する絶対処理

男女が同一歌詞の歌を歌った時，音高の動的パターンには絶対的な違いがある。男声は低く，女声は高い。これは男性の声帯が長く，重たいがために声帯振動周期が長くなるためである。このような純粋に物理的な要因のために男女間の音高差は生じる。よって，両者の動的パターンの同一性を検討する場合，絶対的な音高知覚では役に立たない。極端な絶対音感者は，移調前後で曲の同一性が感覚できない。

その男女が同一歌詞を読み上げた場合，音色の動的パターンには絶対的な違いがある。男声は太く，女声は軽い。これは男性の声道長が長いがために，共鳴周波数が低くなるためである。このような純粋に物理的な要因のために男女間の音色差は生じる。よって，両者の動的パターンの同一性を検討する場合，絶対的な音色知覚では役に立たない，と記したいところであるが，筆者らの知る限り，全ての音声科学，音声工学の議論は音色に対しては絶対的な処理系を常に構築してきた。

数十年以上に渡る，この不可解な議論はどう解釈すべきか？

言語化できる相対音感者が時にする主張として，「全ての長調の曲がハ長調として聞こえる」という主張がある。彼らは移調不変のドレミ知覚を行なうが，「ド」と聞こえた部位がピアノの C 音の鍵盤を弾いている（即ち物理的に全く同一の音である）と解釈した時に，この勘違いが生まれることになる。数十年に渡る不可解な議論では，男女が読み上げた同一歌詞に対して音韻「あ」を感覚した部位には同一の物理現象が存在するという仮定で議論を繰り返して来た。上記議論を考えるならば，これは勘違いの議論となるのだろうか？例えば，相対音感者の勘違いは，絶対音感者が彼らの相対的知覚特性を解説すれば，彼らは「勘違いであった」と納得する。同様に，音声の（極端な）絶対音感者がいれば，数十年に渡る議論が勘違いか否かを指摘してくれると期待される。音声の極端な絶対音感者は，ある話者の「おはよう」と別話者の「おはよう」との同一性が認知困難となるが，一部の自閉症者がそれに相当する[5]。彼らに解説を期待したいところではあるが，それは甚だ困難である。多くの場合，彼らは音声言語を持たないからである[1]。

なお，人間以外の霊長類は音高に対する相対音感を持っていない。相対音感は音と音の差分に基づく感覚であるため，絶対音感より処理のタスクが高いからである[6]。音高は一次元物理量，音色は多次元物理量であることを考えると，音色の相対音感を持つ霊長類がいたとすれば，それは人間のみであろう。言い換えれば，筆者らの提唱する新しい音声の構造的表象は，人間固有の認知能力に根付いた音声表象と考察できる。

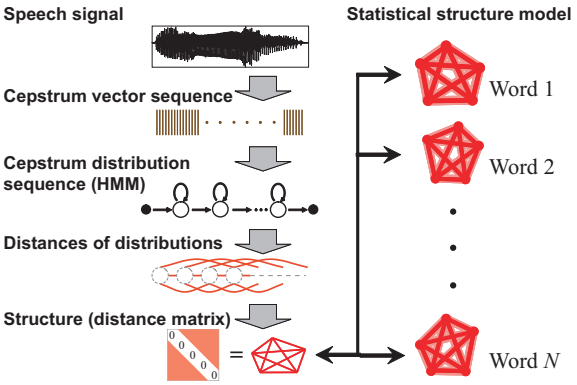

図 6: 音的実体を用いない構造的音声認識

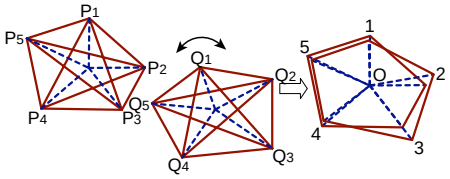

図 7: 回転及び平行移動を通して行なう音響照合

表 1: 音的実体を用いない構造的音声認識結果 [%]

	HMM(4,130)	HMM(260)	提案手法 (8)
単語単位	97.4	82.1	92.6
母音単位	98.8	90.4	96.6

4. 音的実体を全く用いない構造的音声認識

図 4 に示した音声の音的実体を一切捨象した物理表象で，果たして音声認識が可能なのだろうか？既に日本語母音系列を対象として認識実験を行なっている。日本語五母音を入れ替えて構成される連結母音発声（語彙数 120 であるため，PP=120 の孤立単語認識となる）を対象として検討した[7]。

図 6 に音的実体を一切用いない構造的音声認識の枠組みについて示す。入力音声を構造化し，統計的にモデル化された構造的テンプレートと照合する。この時の照合は，基本的に**図 7** に示されるように，片方の構造を回転及び平行移動して両構造を合わせた上での音響照合となる。頑健に話者不変な音声の構造的表象は距離行列，即ち幾何学構造としての表象であるため，この表象に対して行なわれる様々な変換操作（例えば話者変換など）は，回転あるいは平行移動としてこの構造に作用する。例えば，声道長の差異（即ち年齢，性別の差異）は構造の回転として，マイクなどの音響機器の違いは構造の平行移動として解釈される。回転＆平行移動後の音響照合は二つの距離行列のみを用いて計算されるが，これはタンパク質の構造解析などで行なわれている手法と同一の手法となっている。

男女計 8 名に 120 単語を 5 回ずつ発声してもらい，これを用いて 120 単語の統計的構造モデルを作成した。これとは異なる男女 8 名に同様の発声を依頼し，これを評価データとした（合計 4,800 発声）。結果を**表 1** に示す。比較対象として学習話者 260 名, 4,130 名の不特定話者 HMM の結果も示す。現時点では，性能に関しては HMM(4,130) に及んでいないが，フォルマント周波数やスペクトル包絡に代表される音的実体を表象する物理量を一切用いず，音の変化／動きのみを捉えることで，連続発声中の母音の約 97%が，非常に少ない学習話者数で同定できて「しまう」事実は，甚だ驚嘆に値すると考えている。音声の言語情報は，音的実体ではなく，音の変化／動きに対して符号化されていると解釈すべきであろう。

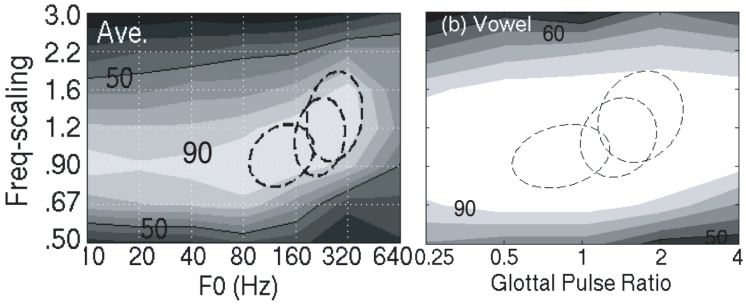

図 8: 孤立母音及び連続音声中の母音同定

5. 母音は階名なのか，音名なのか？

前節までの議論では，音韻認知を階名認知として捉え，音階構造に対応する不変的な音群構造として母音群構造を考えた。しかし，音韻認知と階名認知には否定し難い相違が存在する。階名認知の場合，孤立音を提示してもそれを階名として認知することは不可能である。一方，音韻の場合は容易にそれができる。音韻認知は音声の絶対音感なのだろうか？

図 2 に示す母音図から分かるように，日本語の場合，話者による違いを考えても，母音間の重なりはそれほど大きく無い。しかし，この重なりは容易に増加させることができる。フォルマント周波数は声道長の関数であるため，巨人／小人の声を合成すればよい。通常の領域から外れた孤立母音音声に対する母音同定は可能なのだろうか？もしそれが困難であった場合，音の連続ストリームの中にある母音はどうなのだろうか？孤立母音の場合は不可能であるにも拘らず，ストリーム中であれば容易である場合，これこそ階名認知として考えることができる。

先行研究にその答えを見ることができる[8, 9]。**図 8** 左が孤立母音に対する同定率，右が無意味 4 モーラ単語の中の母音同定率である。縦軸の値 y に対して，$170/y$[cm] が凡そ話者の身長となる。また，右図の横軸の値 x に対して，$160x$[Hz] が基本周波数である。即ち，様々な身長・基本周波数の音声に対する，孤立母音の同定，及び無意味モーラ列中の母音同定の正解率である。図中点線の楕円が 3 つあるが，これは，実在する男性，女性，子供の領域を示す。全ての提示音声は高品質音声分析合成系 STRAIGHT による合成音声である。左図から分かるように，孤立母音提示時（絶対的音認知時）は，実際に人間が存在する領域では 90%を越える結果となる。しかし，その領域を越え始めると同定率は下がり，例えば 65[cm] の小人となると，160[Hz] の音声で同定率は約 20%となる。これはチャンスレベルであり，母音同定は全く不可能の状態になる。

一方，無意味連続モーラ列発声中に母音が置かれると，とたんに同定率が上昇する。65[cm] の小人ですら，約 60%の正解率を呈する。提示単語が有意味語や，高頻度単語であれば，正解率は格段に上昇することが容易に推測される。ガリバーと小人の会話を考えてみる。彼らは連続ストリームとして音声を呈している限りにおいて，互いにコミュニケーションを図ることができるであろう。しかし，個々の音（モーラ）を孤立発声し始めた途端，音声コミュニケーションが破綻する様子が観測されるはずである。言い換えるならば「孤立音として提示された [a] を音韻/a/として同定する能力は音声言語コミュニケーションの必要条件では無い」ということになる。前節でも述べたように，音声の言語情報は個々の音の絶対的物理特性ではなく，音の変化／動きに対して符号化されていると考えるべきである。その意味において，孤立音や連続音声からの切り出し音を音響的に分析し，その絶対的音響特性を議論する作業は，もはや言語的作業ではない。「勘違い」の作業と考えるべきである。

6. 幼児は親の声の何を真似ているのか？

幼児は親の声を模倣しつつ言語を獲得すると言われるが，声そのものを形態模写する幼児はいない。彼らはまだ音韻意識が未熟であるため，聴取した音声ストリームを音韻に区分し，一つ一つの音韻を自らの口で再生することは不可能である。この場合，彼らは親の声の何を真似ているのだろうか？[1] 優秀な九官鳥は聞けば飼い主が分かるようだが[10]，どんなに優秀な幼児であっても，その声を聞いて父親を当てられるお巡りさんはいない。発達心理学は「幼児は個々の音韻を獲得する前に，語全体の音形を獲得する」と主張するが[11]，「語全体の音形」は話者不変の音声表象である必要がある。でなければ，父親・母親の声が出るまで努力する幼児が巷に溢れるはずである。一部の自閉症児が相手の声を形態模倣することが知られている。この場合も，言語が出る以前の状態であるにも拘らず，形態模写を通して音声（彼らにとっては音でしかない）を生成する様子が報告されている[12]。なお，筆者らは「語全体の音形」の物理的定義を発達心理学の各研究者に尋ねて来たが，解答は寄せられていない。筆者らが考える定義は既に述べた通りである。

7. まとめ

「音声物理の多様性と音声認知の容易性」という古典的問題を，音声を多次元音楽として，音韻認知を階名認知として捉えることで解決可能であることを数学的に示し，かつ種々の実験を通してその妥当性を示した。提案する枠組みは音と音の差異のみに着眼し，差異を集めることで構成される不変なる音群構造に着眼する。音楽の場合「全全半全全全半」という間隔で配置される音階構造が重要となるが，この構造を例えば「全半全全半全全」「全半全全全半全」「半全全全半全全」のように音配置を変えることで，ニュアンスの異なる旋律が作られることになる（旋法，モードと呼ばれる）。これと同様の作業を F_1/F_2 平面で考えれば，母音配置を変えることになるが，これが英語圏の方言に相当するのは言うまでも無い。ここまで見事に対応する音高配置構造と音色配置構造であるが，前者に対しては相対音感を当然としつつも，後者に対しては（音声学／音声工学では）絶対音感を前提とした議論のみがなされて来た。これは，孤立音に対して音韻同定が可能であるためであるが，その能力が音声言語コミュニケーションの必要条件では無いことは既述した通りである。絶対的な物理量のみの着眼では，環境変化に頑健に動作するシステムなど構築不可能である。

参考文献

[1] 峯松他, 人工知能学会，SIG-Challenge-0624-6，35–42，2006.
[2] 峯松他，春音講論，1-P-12，147–148，2007.
[3] N. Minematsu, *et al.*, “Speech as multi-dimensional music,” Proc. InterSpeech, 2007 (submitted)
[4] R. Jakobson *et al.*, Notes on the French phonemic pattern, Hunter, N.Y., 1949.
[5] 東田他，この地球にすんでいる僕の仲間たちへ，エスコアール出版社，2005.
[6] D. J. Levitin *et al.*, Trends in Cognitive Sciences, 9, 1, pp.26–33, 2005.
[7] S. Asakawa, *et al.*, “Automatic recognition of connected vowels only using speaker-invariant representation of speech dynamics,” Proc. InterSpeech, 2007 (submitted)
[8] 青木他，秋音講論，373–374，2004.
[9] 林他，春音講論，2-Q-27，474–475，2007.
[10] 宮本，音を作る・音を見る，森北出版，1995.
[11] 早川，月刊言語，35，9，62–67，大修館書店，2006.
[12] 深見，ひろしくんの本 (V)，中川書店，2006.