3E1-06

学習者を記述する音声科学の提案とそれに基づく学習支援

Yet another speech science to describe a student and its application to CALL

峯松 信明
Nobuaki MINEMATSU

東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, University of Tokyo

This paper proposes yet another speech science, which is derived by implementing phonology on physics. Speech events are probabilistically represented as distributions, distance between two events is calculated based on information theory, and all the events are relatively captured as structure. The resulting structure has completely no dimensions to represent static non-linguistic information. A student's pronunciation represented by the new science is purely acoustic and purely linguistic at the same time. This means that the pronunciation can be acoustically matched with a teacher's pronunciation directly and also linguistically matched with the language of English directly. This paper also shows possibility of yet another speech engineering based on the new science and some application examples to CALL (Computer Aided Language Learning) are experimentally investigated.

1. はじめに

音声科学は音声学・音韻論に二分される。音声学は各言語音を，その調音特性・音響特性に基づいて記述することを目的とする。音韻論は各言語を，その音韻群或いは音韻系列に潜む構造・関係に基づいて記述することを目的とする。即ち音声活動に対し，音声学は「音」を記述し，音韻論は「言語」を記述する。本研究では「個人」を単位として音声活動を記述する新しい音声科学を提案する。この音声科学は言語学の一分野として定義される音韻論を物理実装する形で定義され，この科学の上で得られる音声表象は「純粋に音響的であり，純粋に言語的である」という興味深い特性を持つ。学習者の発音をこの科学の上で表象すれば，それは学習者の今を記述する発音カルテとなるばかりでなく，教師発音と「音響的に」比較することが極めて信頼性高く実現できると共に，学習者発音を英語という言語そのものと「言語的に」比較することも可能となる。即ち新しい音声科学の上に新しい音声工学を構築する可能性を有する。

2. 科学無き発音教育

2.1 音声学・音韻論は本当に発音教育に役立つのか？

英語教師を目指す学生は，発音教育のための基礎知識として音声学・音韻論といった科学を（単位上は）習得する。音声学は「音」を記述する科学であり，音韻論は「言語」を記述する科学である。これらの科学は英語という言語の音やその言語に内在する規則を「紹介」するためには非常に優れたツールである。言語を紹介することが教育の目的であるならば，現在の英語教育は何ら問題を抱えていない。非常に優れた紹介をしていると筆者は考えている。紹介だけでは意味がない，とするならば何が必要か？それは，学習者「個人」を記述する科学である。学習者が現在どういう状態にあるのか，その状態にある学習者にはどのような訓練が効果的であるのか，こういった教育を実現するためには，学習者を記述する「文字」が必要である。音声学，音声工学の基礎を育んできた英・米・瑞・蘭といった国々では，外国語の学習が社会問題となっていない。即ち，音声を対象とした物理学は，学習者を記述する文字が社会的に不必要な環境で構築されてきた，という歴史を持つ。

連絡先: 峯松信明，東京大学大学院情報理工学系研究科，mine@gavo.t.u-tokyo.ac.jp

2.2 音響音声学に基づく試みは正しかったのか？

「正しい発音は正しい調音によってのみ可能となる」と考えれば，学習者の記述は調音音声学的な記述を求めることが最も相応しいと考える。しかし，調音運動の測定は高価な医学機器を必要とし，それらを各教室，各家庭に置くことは不可能である。音響調音変換なども工学的には検討されているが，その信頼性・安定性・簡易性は教育使用のレベルからはほど遠い。

音響音声学の登場と共に，音声の音響的側面を簡単に取得することが可能となった。波形・スペクトル表示である。現在考えられている発音教育支援は全てこの「文字」を前提としている。この文字は本当に正しいのだろうか？結論を言えば「雑音混じりの文字」である。波形・スペクトルから得られる情報には，発音の善し悪し以外に，性別，年齢，体格，マイク特性，伝送路特性など種々の情報がある。これらは全て発音学習から見れば雑音である。この雑音の存在が音声システムを不安定にする。音声認識の世界では「沢山データを集めること」でこの雑音を消す，という最もナイーブな方法論が常套手段となっている。不特定話者音響モデルがその代表例である。しかし，話者適応技術が頻繁に必要となる事実は「不特定話者音響モデルは不特定話者を対象にし得ない」ことを意味し，「集めること」による方法論の限界を示唆している。CALL システムも当初は教師，学生に歓迎された。しかし，最近になってその不安定さを批判する報告を目にするようになった [1]。上記事実を知る者からすれば，これはある程度予測された事態である。

3. 「個人」に着眼する音声科学の提案

3.1 音響音声学を「捨てる」

現在の音声工学では，上記した雑音を消す手段として「沢山集めること」「音響モデルの中に含まれる話者性を音声データのそれに合わせること（即ち適応）」「音声データの中に含まれる話者性を 1 に揃えること（即ち正規化）」が広く行なわれている。適応や正規化は，より少量のデータで行なう必要性からその性能には限界がある。また学習者への適応は，不適切な発音に高いスコアを付与し易くなる，という問題も生じる。この雑音問題は，本質的にどこから来るのだろうか？技術の問題であろうか？筆者の考えでは音響音声学による音声表象を「そのまま使う」ことが全ての元凶である。音響音声学に基づく表象を使えば，性別，年齢，体格といった情報が混入することは避

けられない。この「雑音」を表現する次元が残留するまま「集めること」「揃えること」を試みても，当然その次元は残るため限界が生じる。音声認識の現状がその問題点を示している。本来必要なものは「雑音を表現する次元そのものを物理的に消すこと」である。言い換えれば，峯松の英語音声から「峯松であること」を表現する次元を物理的に消し「英語という言語を用いて言語活動を営むに際してどの程度相応しい物理特性を持っているのか」を表現する次元のみを残すこと，である。

3.2 音声学から音韻論へ

そのような物理表象が存在するのだろうか？本論文ではそれが数学的に導出されることを示す。この数学的導出は「音韻論の物理実装」を意図したものである。なぜ音韻論なのか？

音韻論は音声に内在する非言語情報（年齢，性別など）を研究者の頭の中で消し去り，言語音を音素という抽象概念を用いて表現する。そして音素の「系列」や音素の「群」の中に内在する構造・規則を明示することを目的とする。本研究では音素群の中に潜む構造議論に着眼する。ソシュールの構造主義に端を発し，ヤコブソン，ハレといった先駆者により，音素群構造が弁別素性を用いて検討されている。図 1 にロシア語音素に対する構造議論，即ち音素樹型図を示す [2]。この音素分類は，弁別素性を用い，各ノード下の音素サブセットが自然類（言語学的に意味のある群）となるように行なわれる。つまり，言語学的知識を用いたトップダウンクラスタリングである。この場合，用いる知識の差によって異なる分類が行なわれる。音素分類木も音韻論研究の歴史と共に種々の変遷を受けている。

本研究では，この構造議論をボトムアップクラスタリングを用いて考察する。この場合必要なのは知識ではなく，全要素に対する二要素間距離（距離行列）となる。n 個の要素に対して ${}_nC_2$ 個だけ定義される対角線の長さだけに着眼することは，n 点で張られる構造に着眼することに数学的に等しい。n 要素をその構造だけに着眼することは，様々な情報を捨て去る操作として解釈できるが，非言語情報（即ち雑音）を全て「捨て去られる」情報に埋め込むことができれば，物理的に定義されるこの構造が，筆者が求める新しい音声の物理表象となる。

3.3 非言語情報源のモデル化

非言語情報を表現する次元を数学的に消滅させる場合，その情報（源）をモデル化して検討することが望ましい。音声認識で検討される歪み・雑音は大きく三つに分類される。加算性雑音・乗算性歪み・線形変換性歪みである。

加算性雑音の代表例は背景雑音であるが，本研究では加算性雑音は考慮しない。それはこの型の雑音は，雑音源の抹消が物理的に可能であり，不可避な雑音ではないからである。

乗算性歪みは所謂フィルター性の歪みであり，マイク，録音室の特性がその代表例である。GMM による話者モデリング

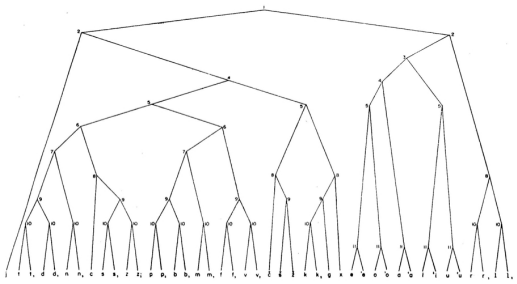

Fig. I-1. Branching diagram representing the morphonemes of Russian. The numbers with which each node is labelled refer to the different features, as follows: 1. vocalic vs. nonvocalic; 2. consonantal vs. nonconsonantal; 3. diffuse vs. nondiffuse; 4. compact vs. noncompact; 5. low tonality vs. high tonality; 6. strident vs. mellow; 7. nasal vs. nonnasal; 8. continuant vs. interrupted; 9. voiced vs. voiceless; 10. sharped vs. plain; 11. accented vs. unaccented. Left branches represent minus values, and right branches, plus values for the particular feature.

図 1: *Halle's tree diagram of Russian phonemes*

を考えれば，話者性の一部も乗算性歪みとなる。この歪みは如何なる発声者，如何なる収録環境を用いても不可避に混入する歪みである（意味論的に零点が存在しない）。音声事象をケプストラムベクトル c で表現すれば，乗算性歪みはベクトル b の加算となり，その結果 $c' = c + b$ となる。

線形変換性歪みの代表例は，話者間の声道長の違いによるスペクトル変形，聴取者間の聴覚特性の違いによるスペクトル変形である。この変形は対数スペクトルに対する周波数軸の非線形伸縮で近似されることが多いが，この非線形伸縮は，ケプストラム領域においては行列 A を掛ける演算となることが数学的に導出される [3]。即ち，$c' = Ac$ である。

結局，音声コミュニケーションにおいて不可避に混入する多種多様な雑音（歪み）は b_i や A_i として記述されることとなり，どのような組み合わせにおいても最終的には $c' = Ac + b$, 即ちアフィン変換として近似できることになる。

3.4 音韻論の物理実装

音韻論は「人，環境に依らず第 3.2 節で示した音韻構造が普遍的に観測される」ことを主張する。即ち音韻論を物理実装するための数学的必要十分条件は以下で示される。

• 音素を音響空間内の点として考えた場合，n 点で張られる構造が如何なるアフィン変換によっても歪まない。

周知のようにアフィン変換は構造を歪ませる変換であり，上記は数学的に不可能である。音韻論は数学的幻影なのか？以下に示すように，情報理論により，音韻論は物理に落ちる。

音素を点で捉えることは物理的に不自然である。ピッチ波形は常に変動しているように，また，たとえ同一話者であったとしても，同じ音響事象を二度繰り返すことができないように，音素を点で表現することはできない。分布である。また，${}_nC_2$ 個の対角線の長さ（距離行列）が等しい二つの構造は合同であることを考慮すると，必要十分条件は以下に書き換えられる。

• 二分布間距離が，如何なる（共通の）アフィン変換によっても一切変化しない。

この性質を満たす距離尺度としてバタチャリヤ距離がある。分布 $d_x(c)$，$d_y(c)$ に対して，以下の式で定義される。

$$BD(d_x(c), d_y(c)) = -\ln \int_{-\infty}^{\infty} \sqrt{d_x(c) d_y(c)} dc \qquad (1)$$

即ちバタチャリヤ距離は，二つの分布を独立事象と見なした時に，二事象が同時に生起する確率（同時確率）に対する自己情報量として定義される。そしてこの距離は如何なるアフィン変換によっても変化しない。結局行列 A を掛ける演算は構造の回転となり，ベクトル b を足す演算は構造のシフトとして幾何学的に解釈される。人間の成長による音声の音響変化が声道長の延びによる変化だけであると仮定すると，子供から大人にかけての音声の音響変化は音声構造の回転として捉えられることになる。約 15 年ほどかけてゆっくり音声構造は回転している。

以下，本音声表象で学習者個人を記述することによって何が可能になるのか，という点に絞り現在までに得られている結果を概観する。個々の結果の詳細は参考文献を参照して戴きたい [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]。

4. 10 億人の学習者の記述に向けて

4.1 1 学習者の記述

ERJ (English Read by Japanese) データベース [15] 中の各話者に対して，特定話者音響モデルを構築し，各話者の音素

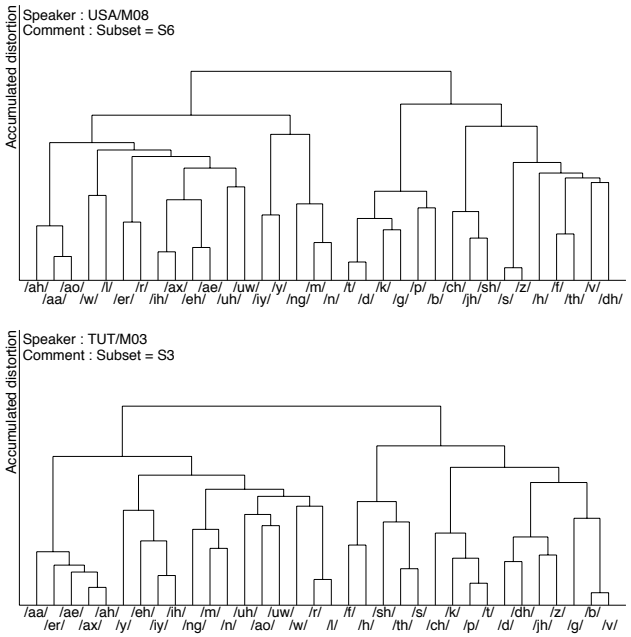

図 2: 米語話者（上）と日本人（下）に対する音素樹型図

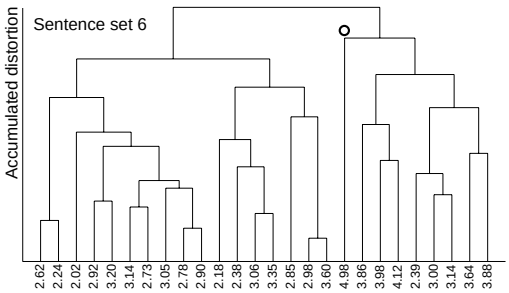

図 3: 文セット 6 の話者に対する分類

樹型図を作成した。図 2 に米語話者と日本人による音素樹型図の例を示す。日本人樹型図には，母音縮退や子音置換，更には schwa の不適切な発声の様子がそのまま表現されている。英語音声学によれば，母音は弱勢となると schwa に近づくと言われるが，これは母音構造のサイズが小さくなることと同値である。既に強勢・弱勢母音を用いた分析により，構造のサイズが調音努力におよそ対応することを実験的に確認している。なお，樹型図は距離行列を視覚化する一手法に過ぎず，多次元尺度法など，より良いインターフェイスを模索する必要がある。

4.2 2 学習者間の距離尺度

2 学習者を比較する場合，距離行列の各要素を比較することも可能である（局所的比較）。ここでは，2 者の差異をスカラー量で表現することを考える（大局的比較）。距離行列をベクトルと見なして計算される 2 者間のユークリッド距離が，近似的に，A と b に対する適応処理を施した後の，対応する音素間距離の平均値に比例することを実験的に導いている。

4.3 全学習者の分類

2 学習者間距離がスカラー量で表現されると，任意の学生母集団に対し，その学生群をボトムアップクラスタリングすることが可能となる。例えば英語学習者は地球上に約 10 億人存在すると言われるが，彼らを個人の単位で記述し，全学習者を分類することは技術的には十分に可能である。実際には，距離行列の様子は読み上げさせる文セットや読み上げ方にも依存して

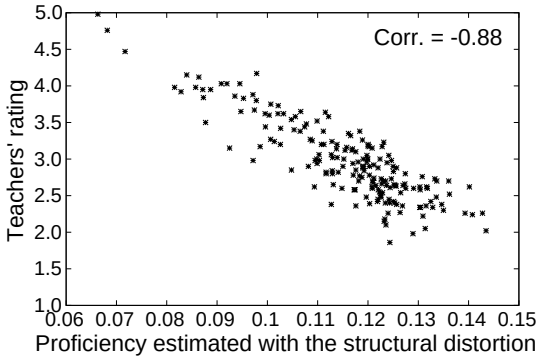

図 4: 音韻構造歪みに基づく自動発音評定

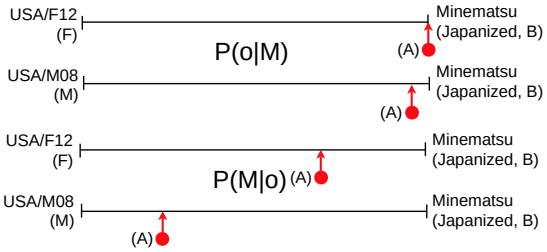

図 5: $P(o|M)$ 及び $P(M|o)$ に基づく自動評定結果

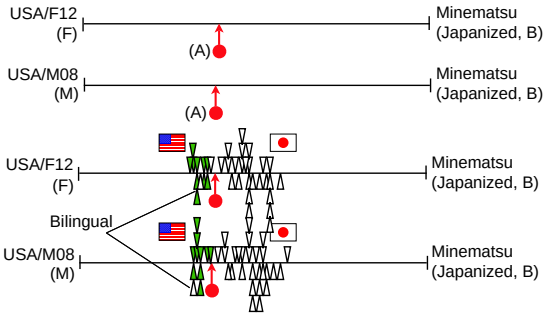

図 6: 音韻構造歪みに基づく自動評定結果

くるため，細かく収録条件を制御することが必要である。ERJ における文セット 6 に対して行なった学習者の分類の様子を図 3 に示す。リーフノードの数値は各学生の発音スコア（5 点満点，米国人教師による評定）である。約 30 人の日本人を 3 つに分類した時に，一人の学生で 1 クラスタが形成されているが，この話者は 202 名中の唯一のバイリンガル話者である。

5. 音韻構造歪みに基づく発音評定

5.1 教師・学生間における音韻構造歪みの定量化

2 者間の比較ができるということは，学習者と教師とを，その音韻構造の歪みのみに基づいて比較できることを意味する。上述した 2 話者間距離を用いて ERJ の日本人学生を自動評定した結果が図 4 である。非常に良好な関係が得られている。

5.2 発音評定における究極の安定性を求めて

学習者の構造化によって，雑音（非言語情報）はどのくらいそぎ落ちているのだろうか？筆者は英語劇の舞台役者としての経験を持ち，母語話者の役を演じる（発音する）ための筋肉武装から英語発音に取り組んだ一人である。しかし純粋な日本人である以上，日本語訛りの英語を発声することも容易である。そこで，以下に示す興味深い実験を行なった。

筆者による英語音声（通常の英語 A，故意に訛った英語 B），及び米語話者二名による英語音声（男性 M，女性 F），合計

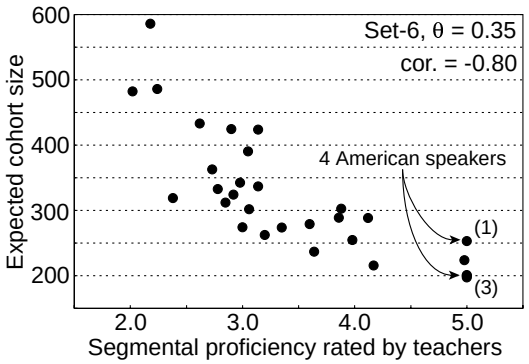

図 7: 語彙密度に基づく発音の自動評定

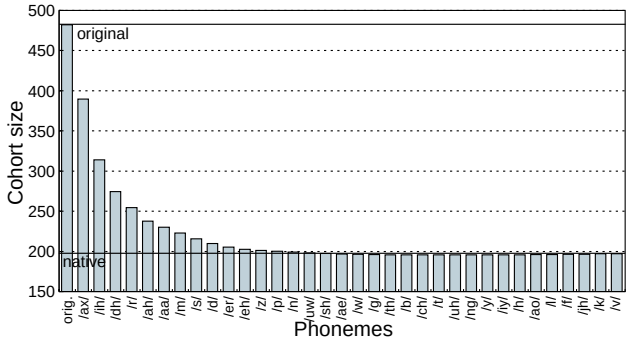

図 8: 関係の部分置換に基づく語彙密度の減少

4 種類の英語音声を用意した。次に B, M, F から音響モデル（HMM）を作成した。A をこれら三種類の発音と比較した場合，A は B よりも M, F に音響的に近いと判定されるのだろうか？英語教育学的に A は B よりも F に近いと判定されるべきであるとするならば，この条件が音声工学的には，最も困難な条件である。音響的類似性を判定するための尺度としては，1) 音声認識で広く使われる音響スコア $P(o|M)$ を使用，2) 入力話者と音響モデルとの相性を正規化する目的で広く使われる音響スコア $P(M|o)$ を使用，3) 本研究で提案する構造歪みに基づく音響スコアを使用，の 3 種類を用いた。

結果を図 5，図 6 に示す。$P(o|M)$ の場合，A は限りなく B に近いと判定される。全ての音響的特徴を考慮して判定していることを考えれば，これは当然である。$P(M|o)$ 使用時は，M と F とで評定が異なっている。この不安定性が CALL システムが批判される理由である。構造歪みに着眼したスコアでは，安定した判定が行なわれている。図 6 には，同一文セット（セット 6）を読み上げた米語話者，他の日本人の位置も示しているが，バイリンガル話者以外の全ての話者が，筆者（A）以上に筆者（B）に「音響的に」近いと判定されている。これは，スペクトルの比較を原理とする方法論では不可能である。

6. 言語構造歪みに基づく発音評定

6.1 英語特有の語彙構造との相性に基づく発音評定

日本人英語では音素の混同が頻発する。元来音素数が少ない日本語の音だけを用いて英語の発音を試みれば，音素混同が生じるのは当然である。音素混同は異なる単語の音響的類似性を高め，音響空間における語彙密度の増加を促す。語彙密度が増加すれば，それは不明瞭な英語となる。提案する学習者表象（距離行列）及び英語語彙だけを用いて語彙密度を計算することが可能である。発音教育の目標を native-sounding な発音ではなく intelligible な発音とする声が高まってきているが，intelligible な発音を物理的に定義することが困難であることも事実である。「語彙密度が低い発音がより intelligible である」と考えればその定義は可能となる。図 7 に推定された語彙密度と教師による発音スコアとの相関の様子を示す。本手法は，学習者と教師との間の音響的な比較を一切行なわず，学習者の発音が対象とする言語にとってどの程度「相性が良いのか」という観点から評定した結果であると解釈できる。

6.2 矯正すべき音素順序の推定

提案する学習者表象（距離行列）は，非言語情報を一切含まないため，学習者・教師間で「関係の部分置換」が意味を持つ。即ち，行列の一部を置換することで語彙密度の大幅な低減が可能である場合，その関係の是正が，学習者にとってまず初めに矯正すべき対象であると解釈できる。図 8 に学習者 RYU/F06 に対して得られた音素の矯正順序を示す。日本人が頻繁に混同する音素が優先度高く抽出されている様子が分かる。

7. まとめ

本研究ではまず，確率論，相対論，そして情報論に基づく音韻論の物理実装を通して定義される新しい音声科学を提案した。そこでは「音」や「言語」ではなく「個人」を単位とした音声活動の記述が可能となる。そして，その科学の上に構築される工学が，発音学習に対して非常に効果的に寄与できる可能性を具体的な事例を通して示した。参考文献欄に，既発表の論文リスト（の一部）を掲載している。興味のある読者は参照して戴きたい。語学以外の応用についても検討を開始している。

参考文献

[1] A. Neri *et al.*, “Automatic speech recognition for second language learning: how and why it actually works,” Proc. ICPhS, pp.1157–1160 (2003)

[2] M. Halle, “The sound patterns of Russian,” The Hague: Mouton (1959)

[3] M. Pitz *et al.*, “Vocal tract normalization as linear transformation of MFCC,” Proc. EUROSPEECH, pp.1445–1448 (2003)

[4] N. Minematsu, “Yet another acoustic representation of speech sounds,” Proc. ICASSP (2004)

[5] N. Minematsu, “Mathematical evidence of the acoustic universal structure in speech,” Proc. ICSLP (2004, submitted)

[6] N. Minematsu, “Yet another speech science to describe the pronunciations of individual students,” Proc. ICSLP (2004, submitted)

[7] N. Minematsu, “Pronunciation assessment based upon the phonological distortions observed in language learners’ utterances,” Proc. ICSLP (2004, submitted)

[8] N. Minematsu, “Pronunciation assessment based upon the compatibility between a learner’s pronunciation structure and the target language’s lexical structure,” Proc. ICSLP (2004, submitted)

[9] N. Minematsu *et al.*, “Speech communication based upon the acoustic universal structure in speech,” Proc. ICSLP (2004, submitted)

[10] 峯松信明，“音声に内在する音響的普遍構造とそれに基づく語学学習者モデリング”，電子情報通信学会音声研究会，SP2003-179，pp.25-30 (2004)

[11] 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”，電子情報通信学会音声研究会，SP2003-180，pp.31-36 (2004)

[12] 峯松信明，“音響的普遍構造と言語的普遍構造の整合性に基づく発音明瞭度の評定”，電子情報通信学会音声研究会，SP2003-181，pp.37-42 (2004)

[13] 峯松信明他，“音声に内在する音響的普遍構造とそれに基づく音声コミュニケーション”，第三回「話し言葉の科学と工学」ワークショップ論文集，pp.143-150 (2004)

[14] 峯松信明他，“音韻論の物理実装に基づく新しい音声の音響的表象”，電子情報通信学会音声研究会 (2004-6)

[15] N. Minematsu *et al.*, “Development of English speech database read by Japanese to support CALL research,” Proc. Int. Conf. Acoustics (ICA’2004), pp.557-560 (2004)