社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

音声の構造的表象とその距離尺度

峯松　信明†　　志甫　　淳††　　村上　隆夫†††　　丸山　和孝†††　　広瀬　啓吉†††

† 東京大学大学院新領域創成科学研究科 〒277-8562 千葉県柏市柏の葉 5-1-5
†† 東京大学大学院数理科学研究科 〒204-0013 東京都目黒区駒場 3-8-1
††† 東京大学大学院情報理工学系研究科 〒113-0033 東京都文京区本郷 7-3-1
E-mail: {mine,murakami,maruyama,hirose}@gavo.t.u-tokyo.ac.jp, shiho@ms.u-tokyo.ac.jp

あらまし　音声に不可避的に混入する静的な非言語的特徴を表現する次元を有しない，音声の構造的表象が提案されている（音響的普遍構造）。音声事象を全て分布として記述し，全ての二分布間距離を正規化相互相関として求め，事象群全体を一つの構造として捉える。得られた構造はアフィン変換でモデル化される静的な非言語的特徴によって歪むことがない。これは言語学的には構造音韻論の物理実装，認知心理学的には音声ゲシュタルトとして解釈できる物理表象である。本稿では，異なる 2 つの発声が各々構造的に表象された場合の距離尺度，即ち構造間距離尺度の導出を行なう。まずユークリッド空間に存在する 2 つの N 点構造間距離を導出し，次にその近似解について検討する。
キーワード　音響的普遍構造，非言語的特徴，構造音韻論，ゲシュタルト，構造間距離尺度

Structural representation of speech and its distance measure

N. MINEMATSU†, A. SHIHO††, T. MURAKAMI†††, K. MARUYAMA†††, and K. HIROSE†††

† Graduate School of Frontier Sciences, The University of Tokyo,
5-1-5, Kashiwanoha, Kashiwa, Chiba, 277-8562 Japan
†† Graduate School of Mathematical Sciences, The University of Tokyo,
3-8-1, Komaba, Meguro, Tokyo, 204-0013 Japan
††† Graduate School of Information Science and Technology, The University of Tokyo
7-3-1, Hongo, Bunkyo, Tokyo, 113-0033 Japan
E-mail: {mine,murakami,maruyama,hirose}@gavo.t.u-tokyo.ac.jp, shiho@ms.u-tokyo.ac.jp

Abstract　A novel and structural representation of speech is recently proposed, where the dimensions of inevitable non-linguistic features are diminished. This representation is called the acoustic universal structure. Every speech event is described as distribution and distance between any two events is calculated as normalized cross correlation. Then, a global structure is composed of all the distances. The structure is invariant with any of a single affine transformation, which is the simplest model of the non-linguistic features. This structural representation can be viewed linguistically as physical implementation of structural phonology and psychologically as speech Gestalt. This paper introduces a distance measure between two structures. First, the measure is investigated between two structures in an euclidean space. Next, the measure is obtained approximately between two structures in an noneuclidean space.
Key words　the acoustic universal structure, non-linguistic features, structural phonology, Gestalt, distance measure between two structures

1. はじめに

音声は常に歪んでいる。性別，年齢，体格は話者によって異なる。収録系，伝送系，再生系の特性は音響機器によって異なる。更に，聴取者の聴覚特性も人一人違う。話すことは歪みであり，また，聞くことは歪みである。にも拘らず，音声は最も容易なコミュニケーション・メディアである。何故だろうか？『音声知覚の容易性と音声物理の多様性』[1] 音声研究者を長年悩ませてきた古典的な問題である。音声科学は音素に対する音響不変量の不在から「調音運動」にその答えを求め [2]，音声工学は「膨大な音声データと統計モデリング」にその答えを求めている [3]。しかし現時点でこれらの方法論は解を見い出すに至っていない。何れの手法も「音素を物理的に定義する」ことを目的とした方法論であるが，言語学は必ずしも，音素

に対して物理的対象物を求めてはいない。例えば [4] には，A phoneme is one element in the sound system of a language having a characteristic set of interrelations with each of the other elements in that system. The phonemes cannot be defined acoustically and they are a set of abstractions. と示されている。音素は他の音素との関係によって初めて意味を持つ実体であると考えれば，「個々の音素を特定する不変量」を求める必要性は失われる。また，音声コミュニケーションの目的は意味の伝達である。語や形態素を単位として不変量が存在すれば，音素を単位として不変量が存在する必要も失われる。

この古典的問題に対して，1) 上記の音素定義に基づき，2) 凡そ語を単位として音響的不変量が存在し，3) その不変量は音声事象群の全関係を抽出し，語を一つの全体的な構造として表象することによって求まることが，数学的に示されている [5]。本稿の目的は 2 つの構造の距離尺度を導出することである。

2. 不可避的な非言語的特徴の工学的モデル化

音声に混入する歪み・雑音は加算性雑音，乗算性歪み，線形変換性歪みに分類される。ここで加算性雑音は，その消去が可能であるため不可避ではない。よって本稿では考慮しない。

乗算性歪みはマイクや伝送特性など，フィルタリングとしての歪みであり，話者性が GMM でモデル化されることを考慮すると，話者性の一部も乗算性歪みとなる。これらはケプストラムベクトル c に対するベクトル b の加算となる（$c'=c+b$）。

声道長差異（個体のサイズ差，年齢差）によるフォルマントシフトは，スペクトルの周波数ウォーピングとして捉えられる。また聴取者毎に異なる聴覚特性も，周波数ウォーピングとなる。連続かつ単調な周波数ウォーピングはケプストラム次元では行列 A を掛ける演算となる（線形変換性歪み）[6] （$c'=Ac$）。

結局，不可避的な歪みによる統合歪みも一次変換 $c'=Ac+b$ となり（即ちアフィン変換），本稿ではこれを，音響的歪みの最も簡素な工学的モデルとして採用する。

3. 音響的普遍構造

空間内に存在する n 点に対して，${}_nC_2$ 個全ての二点間距離を求めると，その n 点で張られる構造は一意に規定される。即ち事象群に対して，全ての二事象間距離を求めると，その事象群を構造的に表象することになる。ケプストラム空間内の n 点に対して構造を考えた場合，その構造は非言語的特徴によって不可避的に歪む。何故なら，非言語的特徴はアフィン変換としてモデル化されるからである。この不可避的に歪む構造は，空間を歪ませることによって，不変な構造として定義可能となる。

構造不変の定理：意味のある記述が分布（ガウス混合分布）としてのみ可能な物理現象を考える。分布群に対して，全ての二分布間距離を求める（距離行列）。二分布間距離として，バタチャリヤ距離，カルバック・ライブラ距離，ヘリンガー距離などを用いた場合，各分布に対して単一の任意一次変換を施しても，二分布間距離は不変である。即ち距離行列は不変であり，その結果，構造も一切不変となる。

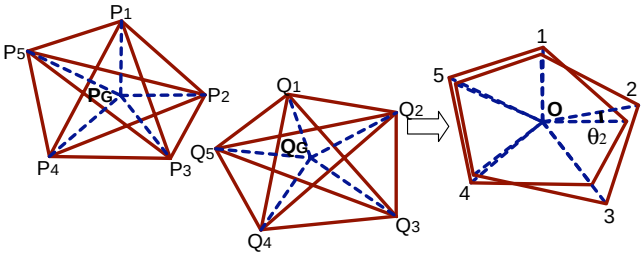

図 1　回転とシフトによる構造の移動と照合
Fig. 1　Structural matching with shifts and rotations

音声事象を全て分布として描き，分布距離を例えばバタチャリヤ距離（正規化相互相関と等価）として求め，事象群全体を一つの構造として表象すると，その構造は単一の任意一次変換に対して不変であり，凡そ非言語的特徴に依存しない普遍的な構造となる（音響的普遍構造）。なお，本定理が定義する歪んだ空間の詳細な定義については [5] を参照して戴きたい。

4. 構造間距離尺度の導出

4.1 ユークリッド空間における構造間距離

一次変換 $Ac+b$ によって構造は歪まない。これは A を掛ける演算は構造の回転（鏡像も含む），b を足す演算は構造のシフトに相当することを意味する。上記したように，構造を不変とするためには空間を歪ませる必要があるが，ここではまずユークリッド空間における 2 つの構造と，回転のみを許す直交行列を考え（**図 1** 参照），距離尺度を解析的に導出する。

N 次元ユークリッド空間（原点は O）内の 2 つの M 点の組を考える。これらを $\{P_1,P_2,...,P_M\}$，$\{Q_1,Q_2,...,Q_M\}$ とする。距離行列 P，Q を $P=(\overline{P_iP_j})_{1\leqq i,j\leqq M}$，$Q=(\overline{Q_iQ_j})_{1\leqq i,j\leqq M}$ とおく。直交行列 A とベクトル $\vec{b}$ に対して，$O\vec{Q}'_i = AO\vec{Q}_i+\vec{b}$ を定め，構造間差異 $D(A,\vec{b})$ を次のようにおく。

$$D(A,\vec{b}) = \sum_{i=1}^{M} \overline{P_iQ'_i}^2$$

A と $\vec{b}$ を動かした時の $D(A,\vec{b})$ の最小値を，P，Q の成分のみを用いて表すことを考える。

Step 1　$\{P_1,P_2,...,P_M\}$，$\{Q_1,Q_2,...,Q_M\}$ の座標を一組ずつ求める。まず $P_1=O$ とする。次に $\overline{P_1P_2}$ が正しい値になるように P_2 をとる。次に $\overline{P_1P_3}$，$\overline{P_2P_3}$ が正しい値になるように P_3 をとる。これを繰り返して $P_1,P_2,...,P_M$ の座標を一組求めることができる。$Q_1,Q_2,...,Q_M$ についても同様である。

Step 2　平行移動により，$\{P_1,P_2,...,P_M\}$，$\{Q_1,Q_2,...,Q_M\}$ の重心が原点 O に重なるようにする。平行移動後の 2 つの構造に対して，$N\times M$ 行列 S，T を $S=(O\vec{P}_1,O\vec{P}_2,...,O\vec{P}_M)$，$T=(O\vec{Q}_1,O\vec{Q}_2,...,O\vec{Q}_M)$ と定義する。

Step 3　N 次正方行列 S^tTT^tS の固有値を，重複も含めて $\alpha_1,\alpha_2,...,\alpha_N$ とする。この時，求める構造間差異の最小値は，

$$\sum_{i=1}^{M}\left(\overline{OP_i}^2+\overline{OQ_i}^2\right) - 2\sum_{i=1}^{N}\sqrt{\alpha_i}$$

となる。以下，上記の証明を行なう。まず $D(A,\vec{b})$ を展開する。

$$
\begin{aligned}
D(A,\vec{b}) &= \sum_{i=1}^{M} |AO\vec{Q}_i - O\vec{P}_i - \vec{b}|^2 \\
&= \sum_{i=1}^{M} \left(|AO\vec{Q}_i|^2 + |O\vec{P}_i|^2 + |\vec{b}|^2 \right) \\
&\quad -2\sum_{i=1}^{M} AO\vec{Q}_i \cdot O\vec{P}_i - 2\sum_{i=1}^{M} \left(AO\vec{Q}_i - O\vec{P}_i \right) \cdot \vec{b} \\
&= \sum_{i=1}^{M} \left(|O\vec{Q}_i|^2 + |O\vec{P}_i|^2 \right) - 2\sum_{i=1}^{M} AO\vec{Q}_i \cdot O\vec{P}_i \\
&\quad +M|\vec{b}|^2 \qquad \left(\sum_{i=1}^{M} O\vec{P}_i = \sum_{i=1}^{M} O\vec{Q}_i = \vec{0} \right)
\end{aligned}
$$

よって $D(A,\vec{b})$ が最小値になるとすれば，それは $\vec{b}=\vec{0}$ の時である。結局 2 つの構造を重心が重なるように重ね合わせ，一方を回転させる形で対応する二点間距離（$\overline{P_iQ_i'}^2$）の総和が最小となればよい。これは $E(A)=\sum_{i=1}^{M} AO\vec{Q}_i \cdot O\vec{P}_i$ とおき，A を動かした時の $E(A)$ の最大値を考察すればよい。

$A=(a_{\alpha\beta})_{1\leqq\alpha,\beta\leqq N}$，$O\vec{P}_i=(p_{\alpha i})_{1\leqq\alpha\leqq N}$，$O\vec{Q}_i=(q_{\alpha i})_{1\leqq\alpha\leqq N}$ と成分で書き，$c_{\alpha\beta}=\sum_{i=1}^{M} p_{\alpha i}q_{\beta i}$ とおくと，$E(A)=\sum_{\alpha,\beta} c_{\alpha\beta}a_{\alpha\beta}$ となる。A が直交行列であるとは

$$\sum_{\beta} a_{\alpha\beta}^2 - 1 = 0 \qquad (1 \leqq \alpha \leqq N) \tag{1}$$

$$\sum_{\beta} a_{\alpha\beta}a_{\alpha'\beta} = 0 \qquad (1 \leqq \alpha < \alpha' \leqq N) \tag{2}$$

と書ける。条件 **(1)**，**(2)** の下で $E(A)$ の最大値を考える。**(1)**，**(2)** の左辺を $f_\alpha(A)$，$f_{\alpha\alpha'}(A)$ とおく。Lagrange の未定係数法より，$E(A)$ が極値をとるような $A=(a_{\alpha\beta})$ において，ある $\lambda_\alpha(1 \leqq \alpha \leqq N)$ 及び $\lambda_{\alpha\alpha'}(1 \leqq \alpha < \alpha' \leqq N)$ において，

$$\frac{\partial}{\partial a_{\alpha\beta}} \left\{ E(A) - \lambda_\alpha f_\alpha(A) - \sum_{\alpha'\neq\alpha} \lambda_{\alpha\alpha'} f_{\alpha\alpha'}(A) \right\} = 0 \tag{3}$$

を満たすものが存在する。但し，$1 \leqq \alpha,\beta \leqq N$ であり，$1 \leqq \alpha' < \alpha \leqq N$ の場合には，$\lambda_{\alpha\alpha'} = \lambda_{\alpha'\alpha}$，$f_{\alpha\alpha'}(A) = f_{\alpha'\alpha}(A)$ とおいた。**(3)** を計算して，

$$c_{\alpha\beta} - 2\lambda_\alpha a_{\alpha\beta} - \sum_{\alpha'\neq\alpha} \lambda_{\alpha\alpha'} a_{\alpha'\beta} = 0 \quad (1 \leqq \alpha,\beta \leqq N) \tag{4}$$

を得る。ここで $N \times N$ 行列 Λ を，

$$\Lambda\text{の第 } (\alpha,\alpha') \text{ 成分} = \begin{cases} 2\lambda_\alpha & (\alpha = \alpha') \\ \lambda_{\alpha\alpha'} & (\alpha \neq \alpha') \end{cases}$$

と定義すると Λ は対称行列であり，また **(4)** より

$$\Lambda A = (c_{\alpha\beta})_{1\leqq\alpha,\beta\leqq N} = S^tT$$

を得る。すると

$$\Lambda^2 = \Lambda^t\Lambda = S^tTT^tS$$

である。一方 $E(A)$ は，

$$E(A) = \sum_{\alpha,\beta} c_{\alpha\beta}a_{\alpha\beta} = \mathrm{Tr}(S^tT^tA) = \mathrm{Tr}(S^tTA^{-1}) = \mathrm{Tr}(\Lambda)$$

と計算される。S^tTT^tS の固有値を重複をこめて $\alpha_1,...,\alpha_N$ とし，Λ の固有値を重複をこめて $\beta_1,...,\beta_N$ とすれば，$\beta_1^2,...,\beta_N^2$ は $\alpha_1,...,\alpha_N$ の並べ替えであり，また $E(A)=\sum_{i=1}^{N}\beta_i$ である。このような $E(A)$ の中で最大なのは $\sum_{i=1}^{N}\sqrt{\alpha_i}$ である。よって求める $E(A)$ の最大値は $\sum_{i=1}^{N}\sqrt{\alpha_i}$ となり，これより構造間差異の最小値は以下のように導出される。

$$\sum_{i=1}^{M} \left(\overline{OP_i}^2 + \overline{OQ_i}^2 \right) - 2\sum_{i=1}^{N} \sqrt{\alpha_i} \tag{5}$$

4.2 非ユークリッド空間における構造間距離

構造不変の定理が定義する音響的普遍構造を用いた構造間距離を考える場合，前節で解析的に導出した距離尺度は直接的には使えない。それは，空間が非ユークリッド空間（多様体）であり，三角不等式が必ずしも成立しないため，例えば **Step 1** の操作が行なえない。そこで，何らかの近似的な手法を用いて，物理的に意味のある距離尺度を定義する必要がある。ここでは，実験的に得られる近似式を元にその距離尺度を考える [7]。

日本人 202 人によって発声された英語音声約 60 文（／人）より [8]，話者毎に英語母音の monophone（5 状態 3 分布）を作成した。音素間距離を対応する状態間のバタチャリヤ距離の平均値で定義した。このようにして，各話者毎に種々の日本人訛りによる歪みが混入した（日本人訛りは言語的な特徴であり，構造化によって消失しない）英語母音構造が規定される。この 202 種類の母音構造を用いて距離尺度を検討する。

まず，一つの構造内に観測される特徴について考える。今，N 次元ユークリッド空間内に M 個の点があるとする（$\{P_i\}$, $1 \leqq i \leqq M$）。重心を P_G とした場合，次式が真となる。

$$\frac{1}{M}\sum_{i<j}^{M} \overline{P_iP_j}^2 = \sum_{i}^{M} \overline{P_iP_G}^2 \tag{6}$$

上記の性質はバタチャリヤ距離では近似的にも成立しない。しかしバタチャリヤ距離の平方根を分布間距離尺度として使用すると，例えば上記の 202 種類の母音構造を用いた場合，**式 (6)** の両辺は**図 2** に示す関係を呈する（但し計算の都合上，**式 (6)** の両辺を M で割り，更に平方根をとったものをプロットして

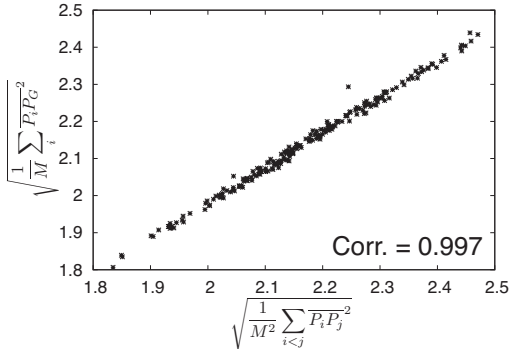

図 2　バタチャリヤ距離の平方根が持つ特性

Fig. 2　Properties of square root of Bhattacharyya distance

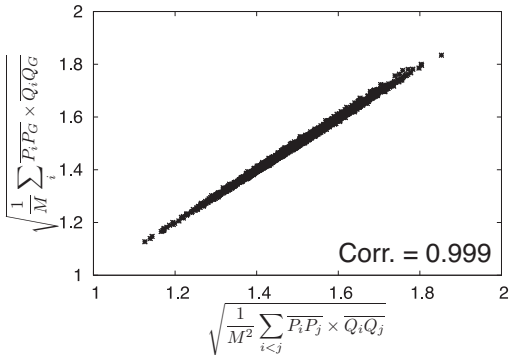

図 3　二つ構造間においてバタチャリヤ距離の平方根が持つ特性
Fig. 3　Properties of square root of Bhattacharyya distance found in two structures

いる。**図 3** も同様である）。バタチャリヤ距離の平方根を使用すれば，**式 (6)** が近似的に満たされることが分かる。

二つの構造 $\{P_i\}$，$\{Q_i\}$ を考える。但し P_{i_0} と Q_{i_0} は同一音素である（構造を構成する点は構造間で対応がとれる）。202 種類の母音構造に対して，以下の式が近似的に成立する。

$$\frac{1}{M}\sum_{i<j}^{M}\overline{P_iP_j}\times\overline{Q_iQ_j}\approx\sum_{i}^{M}\overline{P_iP_G}\times\overline{Q_iQ_G} \tag{7}$$

実際に 202 人の日本人に対して，同一文サブセットを読み上げた任意の 2 人に対して上式を求めたものが**図 3** である。

式 (6)，**式 (7)** より，以下の関係式が導かれる。

$$\begin{aligned}&\frac{1}{M}\sum_{i<j}^{M}(\overline{P_iP_j}-\overline{Q_iQ_j})^2 &(8)\\ =\ &\frac{1}{M}\sum_{i<j}^{M}\overline{P_iP_j}^2-2\overline{P_iP_j}\times\overline{Q_iQ_j}+\overline{Q_iQ_j}^2\\ \approx\ &\sum_{i}^{M}\overline{P_iP_G}^2-2\overline{P_iP_G}\times\overline{Q_iQ_G}+\overline{Q_iQ_G}^2\\ =\ &\sum_{i}^{M}(\overline{P_iP_G}-\overline{Q_iQ_G})^2 &(9)\end{aligned}$$

式 (8) は音素間距離行列をベクトルと見なし，行列間距離をベクトル間のユークリッド距離として求めたものに相当する。**式 (9)** の物理的解釈を考える。**式 (9)** は二つの構造の絶対座標系の中での位置，方向とは無関係に算出される量である。今，二つの構造の重心が重なるように両者をシフトさせる（その時の重心を O とする。**図 1** 参照）。その後，対応する点と点が重なるよう，いずれかの構造を重心周りに回転させ，$\sum|\theta_i|$ が最小となる方向を決定する。なお，θ_i は**図 1** にあるように $\angle P_iOQ_i$ である。ここで $|\theta_i|$ が十分に小さい場合は，

$$\left|\overline{P_iO}-\overline{Q_iO}\right|\approx\overline{P_iQ_i}\quad\text{（但し，シフト\&回転後）}$$

であり，回転&シフトによって $\sum_i|\theta_i|$ が十分に小さくなれば

$$\frac{1}{M}\sum_{i<j}^{M}(\overline{P_iP_j}-\overline{Q_iQ_j})^2\approx\sum_{i}^{M}\overline{P_iQ_i}^2 \tag{10}$$

が成立する。構造のシフトが乗算性歪みを，構造の回転が線形変換性歪みを表現することを考えると，**式 (10)** は上記両歪みに関する完全な適応をかけた後の音素間距離の平均値（の近似値）が，MLLR [9] や SAT [10] が行なう最尤推定を行なうことなく，また，音声の物理的実体を直接的に用いずに，構造と構造との比較のみによって導かれることを意味する。

なお，**式 (9)** において $P_G=Q_G=O$ とすると，

$$\sum_{i=1}^{M}\left(\overline{OP_i}^2+\overline{OQ_i}^2\right)-2\sum_{i=1}^{M}\overline{OP_i}\times\overline{OQ_i}$$

となり，式の形からも**式 (10)** との近似性が窺われる。

5. ま と め

構造的に表象された音声事象群に対して，2 つの構造間距離を検討した。ユークリッド空間にある 2 つの構造については構造間距離が解析的に求まるが，非ユークリッド空間においては三角不等式が満たされないなどの理由により，ユークリッド空間における解析解はそのまま使用することはできない。そこで実験的に得られた分布間距離の特性に基づいて，近似的に構造間距離尺度を導出した。結果として，バタチャリヤ距離の平方根で距離尺度を定め，距離行列を一つのベクトルと見なして算出されるユークリッド距離が，構造のシフトと回転後の対応する二点（分布）間距離の総和の最小値を近似できることが示された。言い換えれば，距離行列のユークリッド距離は比較的明確な物理的対象を持つことが示された。ケプストラム係数とは異なり，LPC 係数などはそのユークリッド距離が明確な物理的対象を持たない。LPC 係数よりもケプストラム係数の方が種々の音声処理においてより高い性能・精度を示す理由の一つがここにあると考えれば，この距離行列は「距離尺度化の容易さ」という意味において有益な特徴量である。

文　献

[1] K. Johnson and J. W. Mullennix, Talker Variability in Speech Processing, Academic Press (1997)
[2] 柏野牧夫，“音声知覚の運動理論をめぐって”，音講論，1-2-10, pp.243–246 (2004)
[3] 徳田恵一，“隠れマルコフモデルによる音声認識と音声合成”，情報処理，vol.45, no.10, pp.1005–1011 (2004)
[4] H. A. Gleason, An introduction to descriptive linguistics, New York: Holt, Rinehart & Winston (1961)
[5] 峯松信明他，“構造不変の定理とそれに基づく音声ゲシュタルトの導出”，信学技法 SP2005 (2005-5，発表予定)
[6] M. Pitz *et al.*, “Vocal tract normalization as linear transformation of MFCC,” Proc. Eurospeech, pp.1445–1448 (2003)
[7] 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)
[8] 峯松信明他，“英語 CALL 構築を目的とした日本人及び米国人による読み上げ英語音声データベースの構築”，日本教育工学会論文誌，vol.27，no.3，pp.259–272 (2004)
[9] C. J. Leggetter *et al.*, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” Computer Speech and Language, vol.9, pp.171–185 (1995)
[10] T. Anastasakos, *et al.*, “A compact model for speaker-adaptive training,” Proc. ICSLP’96, vol.2, pp.1137–1140 (1996)