Vol.2013-MUS-99 No.35
2013/5/12

音声信号における特徴量分離と情報分離

峯松 信明[1,a)]

概要：波形として観測される音声信号には「誰が」「何を」「どのように」話しているのか，など様々な情報が同時に存在している。人間は通常，この音声信号からいとも簡単に特定の情報を抽出し，適切な応答を返すことができる。同様の処理を計算機実装する場合，情報分離（あるいは所望の情報のみに着眼する）機能を，1) その情報に直接対応する特徴量を探求する，2) 着眼しない情報に対応する特徴成分を正規化する，3) 教師有りの識別的変換を一般の特徴量に適用する，4) 着眼しない特徴成分に適合するようモデルパラメータを調節する，5) 確率定義に従って着目しない情報を隠す，などの方策が可能である。本稿では「人間らしい汎化能力の高い，柔軟なシステム」構築を目的とした場合，特徴量レベルで情報分離を考えることの重要性を述べる。また，世界英語データベースを用いた発音分類を例にとって，その有効性を述べる。

キーワード：情報分離，言語・非言語情報，構造的表象，変換不変性，f-divergence，世界英語分類

Feature separation and information separation in speech signals

Minematsu Nobuaki[1,a)]

Abstract: In speech signals observed as waveforms, there exist various kinds of information simultaneously, such as “what” is said by “whom” in “what way”. Humans can generally focus on a specific kind of information exclusively and respond adequately. When one tries to realize this function of information separation or exclusive payment of attention on a machine, he can take a strategy such as 1) finding good features corresponding to a specific kind of information, 2) normalizing irrelevant components in used features, 3) supervised and discriminative transformation of general features, 4) adaptation of model parameters that corresponds to irrelevant information, and 5) hiding irrelevant information in used features using the definition of probability. In this paper, if one wants to realize human-like machines with good generalization ability, the author claims that feature-level information separation is important. Further, some interesting examples of feature-based information separation are shown in the task of world English pronunciation clustering.

Keywords: Information separation, linguistic and non-linguistic information, structural representation, transform-invariance, f-divergence, world English clustering

1. はじめに

音声の生成過程を，二つの過程（音源の生成と声道による共鳴）に分離するソース・フィルタモデルが音声認識，合成の分野で広く用いられている。しかし，声道の共鳴特性は語彙（言語的情報）によっても，話者（非言語的情報）によっても変形する。音声認識システムは，音声から言語的情報を抽出することを目的とするが，音響特徴量 o としては，声道の共鳴特性（スペクトル包絡）が使われることが多い。そのため，話者独立なシステムを構築する場合，1) 特徴量 o をラベル情報を使って識別性の高い特徴量へ変換したり，2) 非言語的情報に相当する特徴成分を正規化したり，3) 音響モデル $P(o|w)$ を多数話者から統計モデルとして構築して話者 s（確率変数）を隠したり，4) 入力話者の声質に対して逐一モデル適応（修正）を施すなどが一般的である。また，最近流行の多層の人工ニューラルネットを用いた DNN による音響モデルでは，話者の違いに対する頑健性・不変性が実験的に報告されている [1]。

[1] 東京大学大学院
7-3-1, Hongo, Bunkyo-ku, Tokyo 113–8656, Japan
[a)] mine@gavo.t.u-tokyo.ac.jp

音声信号は様々な要因によって変形を被るが，人間はそれを頑健に処理することができる（極めて高い汎化能力）。例えば [2] には，幼児の言語発達において，当初は，聴取した音声の非言語的情報と言語情報が分離できず，ある話者の音声にだけに適切に反応することがあるが，やがて，誰の声であっても言語メッセージの同一性を認識するようになる様子（汎化能力の発達）が記述されている。その一方，発達障害の一つである自閉症などでは，汎化能力が健常者ほど発達せず，母親以外の音声に対応することが難しい小学生児童の様子が報告されている [3]。成人になっても，カラオケや外国語の発音練習の時に，声帯模写以外の真似が分らない自閉症者の例もある [4]。

本稿では，これらの事例を踏まえ，高い汎化能力を有する頑健かつ柔軟な，人間らしいシステム開発を目的とした場合，特徴量レベルで情報を分離する必要があることを主張し，発音分類を例にとってその有効性を述べる。

2. 特徴量分離に基づく情報分離

音声は一次元信号（波形，数値列）として観測されるが，その中に，言語・パラ言語・非言語的な様々な情報が符号化されている。多様な情報を適切に反映しつつ数値列を導出するのが音声合成であり，その数値列から多様な情報を的確に抽出するのが音声認識・理解である。

これらの技術を構築する場合，人間の聴覚は音声信号の位相成分には鈍感であるとの知見から，位相情報を切り離し，振幅スペクトルに着眼する場合が多い。更に，音韻情報は音高情報と独立であるとの事実より，ソース・フィルタモデルに基づいて調波構造を切り離し，包絡特性のみに着眼することが多い。Fig. 1 はこの模式図であり，情報・特徴量の分離は二段階までしか行わないのが一般的である。

話者独立単語音声認識，テキスト独立話者認識を考える。包絡特性 o は，単語 w（言語情報），話者 s（非言語情報），何れにも依存する。統計的音響モデルを考えた場合，単語認識の場合は $P(o|w)$ を，話者認識の場合は $P(o|s)$ を推定することになる。統計モデルでは，認識対象とは独立な要因を期待値（周辺化）操作で隠すことが多い。

$$\begin{aligned} P(o|w) &= \sum_s P(o,s|w) = \sum_s P(o|w,s)P(s|w) \\ &\approx \sum_s P(o|w,s)P(s) \\ P(o|s) &= \sum_w P(o,w|s) = \sum_w P(o|w,s)P(w|s) \\ &\approx \sum_w P(o|w,s)P(w) \end{aligned}$$

言語情報と非言語情報は，そもそも独立した情報である。にも拘わらず，それらを運ぶ音響的対象物（特徴量）として，対応した特徴量（o_w や o_s）を考えずに，共通項 $P(o|s,w)$ に対する期待値操作で各々の音響モデルを導出する。各種

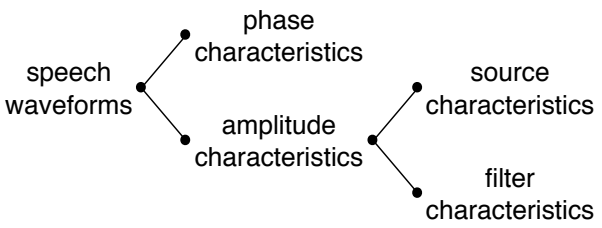

図 1 情報分離に基づく特徴量抽出
Fig. 1 Feature extraction based on information separation

情報が同居する様子を同時確率として捉え，特定の情報に着眼する場合は，それ以外の情報を周辺化（全積分）によって隠す戦略はベイジアンアプローチの常套手段である。このように話者独立（不特定話者）モデルと呼ばれる統計モデルは，話者を隠したモデルであり，振幅スペクトルの位相独立性，包絡特性のピッチ独立性とは意味が異なる[*1]。

一方数式上は，調波構造や位相も統計的な独立性を用いて隠すことはできる。調波を h，位相を p とすれば，

$$P(o|w) \approx \sum_{s,h,p} P(o|w,s,h,p)P(s)P(h)P(p)$$

は，単語 w に対する「話者・調波・位相に独立・非依存」な音声波形 o の統計的音響モデルとなる。更に周辺化を行わず，その話者・調波・位相に対する波形モデル $P(o|w,s_0,h_0,p_0)$ を推定すれば，話者・調波・位相適応したモデルとなる。

筆者は，波形ベースの統計モデルや，適応モデルを用いた研究例を知らないが，これは，精度の高低以前に，抽出すべき言語的情報に対して凡そ独立な要因である，調波や位相を物理的（音響的）に分離して特徴量を定義する常套手段が存在する（Fig. 1）からであると考える。

ある観測量が複数の情報を伝達する場合に，特定の情報のみを抽出することを考える。この時，第 1 節に示したように，技術的には種々の対策を講じることが可能である。適切な特徴量定義が知られていれば特徴量で分離し，それが困難な場合は，隠したり（周辺化），合わせたり（適応化）すればよい。そして最終的には，精度が高くなるよう，各種手法を組み合わせればよい。しかし，以下の節で主張するように，技術構築の目的を「定型発達を遂げた人間が行うような」汎化能力の高い柔軟な情報処理の構築を目的とした場合，上記の選択は慎重に行うべきであると考える。

3. 言語獲得過程に見られる敏感さ・鈍感さ

人間の音声知覚における位相への鈍感さ，音声から音韻を把握する場合に見られる音高への非依存性（鈍感さ）について第 2 節で言及した。さて，幼児の言語獲得は他個体の発声を真似る「音声模倣・学習」を基本とするが [5]，この時，音声のどの側面に敏感に，あるいは鈍感に音声を模倣・学習するのだろうか？例えば話者情報に敏感に模倣す

[*1] 前者は統計的独立性，後者は物理的独立性と言ったところか。

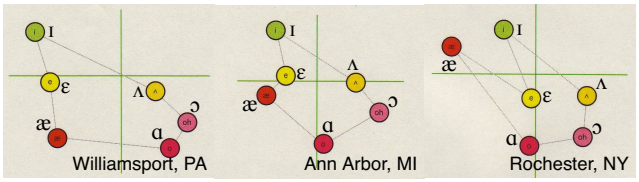

図 2 米語方言における母音分布 [8]
Fig. 2 Vowel distribution of regionally accented AE [8]

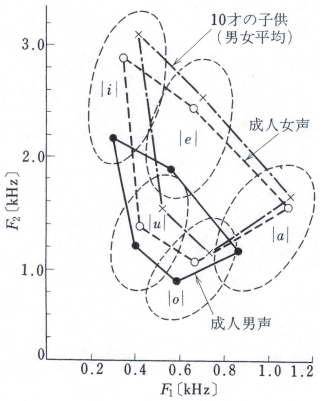

図 3 日本語語母音の分布 [9]
Fig. 3 Distribution of the five Japanese vowel [9]

るとすれば，彼らの模倣行為は声帯模写的になるはずであるが，健常者の場合はそうならない。模倣とは言え，音響的には非常にずれた模倣を常に行なう。

近年の言語獲得研究によれば [6], [7]，幼児が敏感な音声情報・特徴として，音の群が作る分布的特性（distributional properties）を指摘している。これは，bimodal なスペクトル分布から生成されるような音刺激環境にあれば，各分布から得られる二音を異なる音として知覚するようになるが，unimodal（分離が曖昧）な場合は，上記と物理的には同一の二音を区別しなくなる，という実験結果に基づく。分布的特性という言葉から容易に想起される，言語獲得において幼児・子供が敏感に模倣・学習する音声情報として方言性を挙げることができる。日本語の場合，方言の違いはアクセントやイントネーションなどの韻律的特徴に表出する傾向にあるが，広く一般には（韻律的特徴も含め），個々の母音の音質が変化することを指すことが多い [8]。

Fig. 2 に米語方言における 6 種類の母音の分布状況を示す [8]。一方 Fig. 3 には日本語における 5 母音の分布状況（但し成人男女，子供別）を示す [9]。いずれも，F_1，F_2 を軸とした母音分布であるが，米語方言は声道長の正規化後の図である。これらから類推されるように，母音の分布状況は（同一方言内）では話者によっておよそ不変であるが，方言間では大きく変わりうることが分る。言語獲得は，より厳密に考えれば「必ず」方言獲得であり，方言の差異に敏感であることは自明である。話者の違いも方言の違い同様，音色の違いである。しかし通常，言語獲得は，Fig. 3 に示される話者性による音色の違いには鈍感に，Fig. 2 に

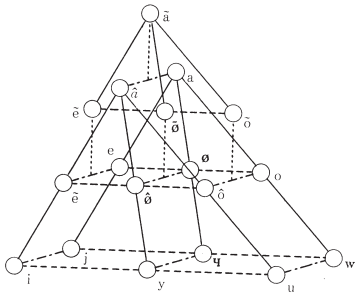

図 4 構造主義的音韻論における仏語母音体系 [11]
Fig. 4 The French vowel system in structural phonology [11]

示される方言性による音色の違いには敏感に行なわれる。なお，自閉症児はしばしば方言を話さないことが報告されているが，非常に興味深い事実である [10]。

この敏感さ・鈍感さ（依存性・非依存性）を技術的に実装する場合，ラベル（教師）を用いて特徴量変換したり，沢山集めて隠したり，頻繁に合わせたり，といった戦略よりも，[6], [7] での検討同様，敏感な情報に対応した特徴量を検討することが，人間を理解し「人間らしい」処理系を構築するには必要であると考える。例えば幼児の言語獲得は，（音声認識器の学習とは異なり）教師・ラベルの無い状態で開始される。また，幼児が聴取する音声の大部分は両親や育児者の声であり，多数話者音声資料による期待値操作，全積分というのは考え難いからである。

さて，上記で示した敏感さは「音の体系に対する敏感さ」と言えるが，古典的な音韻論である構造的音韻論 [11] においても類似した主張が見られる。「偶発的な要因による音の変形ではなく，これらの変形の中に見られる音声の共通項（denominator）」への着眼を促しており，Fig. 4 に示す母音体系（仏語母音と準母音）を提案している。

一方音声信号を，そこに内在する（と思われる）共通項にまで還元せずに，言語・非言語情報が混在する包絡特性 o を基本的な特徴量とした場合，例えば音声合成システムは，声帯模写システムと呼ぶべき機能を提供していることに気付かされる。Blizzard Challenge では，学習話者の個人性が正確に再現されているか否かも評価対象である [12]。

4. 写像不変量に基づく音声の構造的表象

扱うメディアが音声であれ，非音声であれ，センサーを通して計算機入力される場合に様々な要因による歪み，バイアスが容易に混入する。その結果，特徴量が変形する訳だが，これは写像として捉えることができる。即ち，前節で検討した音声の体系を写像不変量で記述できれば，それを，頑健な体系的特徴量として利用することができる。

連続かつ可逆な任意の写像 f に対して（$\hat{\boldsymbol{x}} = f(\boldsymbol{x})$），写像不変性を持つ特徴量を考える。このような問題設定に対して，[13] では，二分布間の距離尺度の一つである f-divergence が写像不変であること（十分性），また，可逆

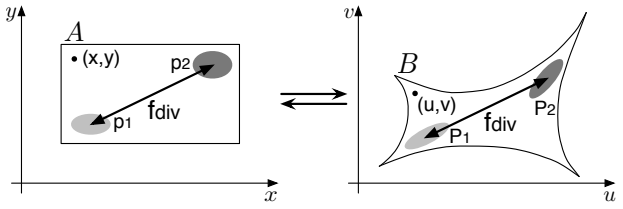

図 5 f-divergence の写像不変性
Fig. 5 Transform-invariance of f-divergence

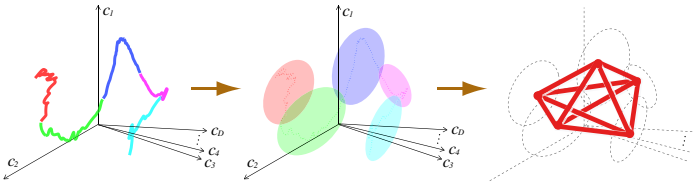

図 6 一発声に対する構造化
Fig. 6 Utterance to structure conversion

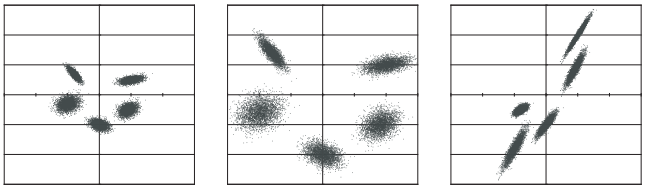

Fig. 6 Feature modification by linear transform
図 7 線形変換による音声特徴の変形

かつ任意の写像に対する不変量は f-divergence になること（必要性）を証明している（Fig. 5）。

$$f_{div}(p_1, p_2) = \oint p_2(\boldsymbol{x}) g\left(\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\right) d\boldsymbol{x}$$

この f-div. のみを用いて発声を表象することを考える。特徴量空間にて発声（軌跡）を分布系列に変換し，全ての二事象間の f-div. を計測して得られる距離行列を取得すれば，それは写像不変表象となる（Fig. 6）。一般に距離行列は一つの幾何学的形態を規定するため，この不変表象を，音声の構造的表象，あるいは音声構造と呼んでいる [14], [15]。

写像による軌跡の変形が，凡そ空間内の平行移動である場合（ケプストラム空間で言えば，定ベクトル $\vec{b}$ の足し算，即ち，定常的な乗算性歪み）は，その軌跡の速度成分（Δ 特徴量）は不変量となる。しかし，例えばケプストラム空間において声道長を伸縮させる変換は回転性の高い行列となり [16]，この場合は，軌跡の方向性が声道長に大きく依存することになる。結局のところ，f-div. で計測された事象間距離（スカラー量）のみが不変量となる。

なお，方言間の違いも写像として捉えることは可能である。言い換えれば，話者性の違いを表現する写像群には不変であり，方言間の違いを表現する写像群には非不変となる（都合のよい）不変性が必要になる。例えば各事象を単一ガウス分布で近似すれば，不変性は線形変換に限定され（Fig. 7 参照），更に特徴量を次元分割して複数の特徴ストリームを定義し，各ストリーム毎の構造表象を考えれば，帯行列を使った線形変換に対してのみ不変性が成立する。このように，不変性を制限して利用することができる。

音高系列（メロディー）を調不変に表象するには，音高差を特徴量とすればよい（音高の相対音感）。この場合，音高は一次元であり回転できず，音高差の方向性も不変性を持つ。しかし音色は多次元であり容易に回転するため，速度成分ではなく，Fig. 6 に示す特徴量定義が必要となる。音高の相対音感を次元拡張しているという意味で，音声の構造表象は音色の相対音感と考えることができる。音高の相対音感は各話者固有の声帯の重さ・長さに起因する音高バイアスを取り除き，音色の相対音感は各話者固有の声道形状に起因する音色バイアスを取り除いているだけである。

5. 構造的表象に基づく世界英語発音分類

これまで音声の構造的表象は，外国語発音分析 [17], [18] や発音誤り検出 [19]，音声認識 [20]，音声合成 [21] に応用されてきたが，近年，世界英語発音に対する話者を単位とした自動分類にも応用されており [22]，その結果を紹介する。

5.1 世界英語（World Englishes）と英語発音学習

英語を母語あるいは外国語として話す話者は世界に約 20 億人いる。多くは外国語として英語を学んでおり，その発音には母語が大きく影響する（訛り）。英語を公用語としている国／地域も同様であり，更には，英語や米語を英国訛り，米国訛りとして捉えることもできる。英語は国際コミュニケーションにおける唯一の共通語であり，各話者は自身の言語背景に起因する話者固有の「訛り」を有している。即ち英語発音の訛りとは「顔のようなもの」として捉えられる。皆違う「顔」を持ち，正しい顔／誤った顔など存在しないように，皆違う「訛り」を持ち，正しい訛り／誤った訛りなど存在しない，と世界英語では考える [23], [24]。

従来英語発音教育では，母語話者の発音をモデルとすることが多かったが，近年では世界英語の考え方が浸透しており，コミュニケーションに支障がなければ，訛りは自らの identity であると考える教師も多い。このような英語感に立てば，学習者に提供すべき情報は「学習者の発音が母語話者とどう違うのか」ではなく，「学習者の発音が他者とどう違うのか，世界中の英語発音の中でどのように位置づけられるのか」に関する情報であると言える。

訛りは国や地方を単位として議論されることが多いが，厳密な最小単位は個人であり，個人を単位とした（約 20 億人の）世界英語発音分類（地図）の自動構築の可能性を検討している。英語地図が構築できれば，自身の発音を客観的に捉えられ，更にインターネット上の英語音声コンテンツを地図とリンクすれば，世界英語ブラウザが構築できる。

"Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station."

[pʰ|iːz kʰɑːl stɛlʌ æsk hɚ ɾə bɹĩŋ ðiːz θĩŋz wɪθ hɛɹ fɹɐ̃m ðə stoːɹ sɪks spũːnz əɣ fɹɛʃ snoʊ pʰiːɻ faːɣ θɪkˈ slæːbz ə blụː ʧiːɻ ɛ̃n mɛɪbi ə snækˈ foɹ hɚ bɹʌðə bɑb wi ɑlso niːd ə smɑːl pʰ|æstɪk snɛɪkˈ æ̃n ə bɪgˈ tʰɔɪ fɪog fõɪ ðə kʰiɪdʒ ʃi kʰĩn skụːp ðiːɻ θĩŋz̃ ĩntʰụ θɪi ɹɛ̧d bæːgz æ̃n wi wɪl miːt hɛɹ wɛ̃ntsdi æt ðə tɹeĩn steɪʃĩn]

図 8 読み上げ用パラグラフと IPA 書き起こし例
Fig. 8 The elicitation paragraph and its IPA transcription

5.2 Speech Accent Archive

Speech Accent Archive (SAA) [25] は，世界中の英語話者（学習者，母語話者）に特定のパラグラフを朗読させ，その朗読音声とその IPA 書き起こし（補助記号を使った詳細な書き起こし）を提供しているコーパスである。パラグラフと書き起こし例を Fig. 8 に示す。現在世界中より収集した 1,700 名以上のデータを提供しており，これを用いて発音地図構築の技術的検討を行った。

ある集合の自動分類は，任意の二要素間の距離，即ち全要素に対する距離行列を計算することで可能となる。世界英語発音分類の場合，任意の二話者間の発音間距離を計測すれば良い。SAA は特定のパラグラフを読ませており，また，IPA 書き起こしも提供しているので，この目的には都合の良いコーパスである。二話者の IPA 書き起こし間の距離を計測すれば，距離行列は得られる。しかし，IPA 書き起こしは手間のかかる作業であり，特定のパラグラフを読ませた N 人の音声資料セットのみから，$N \times N$ の発音間距離行列を推定する技術が必要である。本研究では，IPA-based な二話者間距離を参照距離とし，当該二話者の音声データのみから，この参照距離を予測（回帰）する技術的枠組みを検討した。SAA は世界中からボランティアベースで音声を集めており，その録音環境は様々である（電話音声もある）。このような音声群から，発音だけに関する二話者間距離を推定する。ここに構造的表象を応用する。

なお，SAA は発声中に言い直しがあっても訂正せず，そのまま IPA 書き起こしを行っている。本研究ではこれらの発声は手動で除外した。また，背景雑音レベルが非常に高いサンプル，語順を間違えて読み上げているサンプルも除外した。以下の検討は最終的に得られた 381 人の音声，${}_{381}C_2$=72,390 通りの発音間距離の推定を検討している。

5.3 IPA 書き起こしに基づく参照距離の算出

SAA に含まれる IPA 書き起こしは，補助記号も使われており，音声記号数は 153 種類であった。IPA 書き起こし間の距離は DTW により求めるが，この場合，153×153 の音声記号間距離行列が必要となる。熟練の音声学者 1 名に全音声記号を音声化してもらい（3 回／記号），これを使って HMM を構成した。なお分散は共有させている。次に，二つの HMM 間距離を状態間のバタチャリヤ距離[*2]の平均で定義し，音声記号距離行列を得た。これを用いて，任意の話者対（任意の IPA 書き起こし対）間の，IPA-based な発音間距離を DTW により求め，参照距離とした。

5.4 ベースラインシステム

構造的表象に基づく手法との比較のために，下記の参照距離予測を事前に行った。参照距離は，1)IPA を用いた手動書き起こし，2)DTW による自動距離計算により得られている。1) のプロセスを自動化できれば，参照距離の自動計算は可能である。入力音声を IPA 記号列へと変換するシステムは存在せず，ここでは，米語音素の音響モデル（HMM）を用いた連続音素認識器を代用した。これにより，全ての音声は米語音素系列に置き換わる。また 2) の DTW も，米語音素 HMM より計算した 43×43 の音素距離行列を用い，得られた音素系列間距離を発音間距離とした。

まず，音素認識率 100%の場合を想定して音素系列間距離を求めた。この場合，IPA 記号列を簡単な変換表により米語音素列へと変換した。IPA 記号は 153 種類もあり，これを 43 種類の音素へと変換すれば，様々な言語情報・発音情報が消失する。得られた発音間距離と第 5.3 節で定義した参照発音間距離との相関は 0.86 となった。

次に実際の音素認識器を利用した。WSJ より構築した monophone HMM を用い，入力音声以外の 380 発声から構成されたネットワーク文法[*3]を用いた。本実験で用いた音声は殆どが非母語話者の音声であり，また収録環境も様々であり，得られた音素正解率は 46.1%であった。各発声に対して得られた音素系列と DTW を用いて，発音間距離を推定した。参照発音間距離との相関は僅か 0.04 であった。

5.5 構造表象と SVR による参照発音間距離予測

構造表象を使う場合，SAA パラグラフ読み上げ音声を分布系列，即ち HMM 化する必要がある。ここでは，パラグラフを 9 つに分割し（文や句），各々に対して音素数 ×3 だけの状態を有する HMM を構成した。まず，利用した音声資料全体から構築される不特定話者 HMM を 9 区間に対して各々構築した。これを UBM（Universal Background Model）として利用する。この UBM を当該発声で MLLR 適応して（クラス数 32），各話者の各発声を HMM 化した。

各話者に対して 9 個の HMM が構築され，各 HMM 毎に，3 状態を音素（相当）の単位と仮定して 3 状態単位でバタチャリヤ距離の平均値 $d(i,j)$ を求めた。

[*2] f-div. の一種である。
[*3] 各単語の発音バリエーションを 380 音声から取得し，単語単位で構成したネットワーク文法。

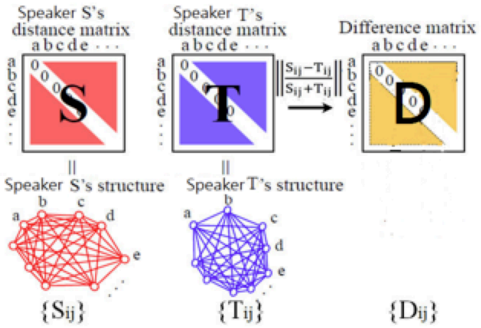

図 9 差行列 $\{D_{ij}\}$ の計算

Fig. 9 Calculation of differential matrix $\{D_{ij}\}$

$$d(i,j) = \sqrt{\frac{BD_1(i,j) + BD_2(i,j) + BD_3(i,j)}{3}}$$

i，j は音素インデックスであり，$BD_n(i,j)$ は音素 i，j 間の第 n 状態でのバタチャリヤ距離である。各話者に対して 9 種類の（音素単位の）距離行列が得られる。なお距離行列の上三角部分の要素数は全部で 2,804 であった。これがある話者の発音様態を表現する特徴ベクトル次元数である。

二話者 S，T の距離行列 $\{S_{ij}\}$，$\{T_{ij}\}$ に対して，その差を表現する差行列 $\{D_{ij}\}$ を以下のように求めた（Fig. 9）。

$$D_{ij}(S,T) = \left| \frac{S_{ij} - T_{ij}}{S_{ij} + T_{ij}} \right|$$

二話者間の発音間差異も 2,804 次元の特徴量となる。

得られた発音間差異特徴量を使って，参照距離に対する回帰モデルを学習する。ここでは，識別的な回帰である SVR を用いた。実装系としては LIBSVM の ϵ-SVR を使用した。なお，カーネル関数は RBF である。

72,390 通りの話者対を二分し，2-fold な交差検定を行なった。参照発音間距離と，構造表象及び SVR によって予測された発音間距離との相関は 0.77 となった。100%の音素認識装置には及ばないが，利用可能な実際の音素認識器を用いた実装（相関 0.04）よりも遥かに高い相関値を示した。本研究で使用した特徴量定義では，各話者の各音がどのような音色（スペクトル特性）を持っているのかは全く使っていない。音群の中の関係性だけを取り上げ，それを特徴量としている点は非常に興味深いと考えている。

本稿では特徴量の分離に基づく情報の分離に主眼を置いているが，構造的表象は特徴量次元数が容易に増加するため，これまでの応用例においても何らかの識別モデルに基づく次元削減・最適化が必要であったことを言及しておく。

6. おわりに

人間が示す，非常に高い音響的汎化能力を有する音声処理系の実現に向け，特徴量レベルで情報分離を実現する手法の一つとして構造的アプローチを紹介した。世界英語発音分類を例にとり，その有効性について述べた。

参考文献

[1] A. Mohamed *et al.*, “Understanding how deep belief networks perform acoustic modelling,” *Proc. ICASSP*, 4273–4276 (2012)
[2] R. S. Newman, “The level of detail in infants’ word learning,” *Current Directions in Psychological Science*, 17, 3, 229–232 (2008)
[3] 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 (2005)
[4] 綾屋他, 発達障害当事者研究, 医学書院 (2008，但し, 著者らの対談を含む)
[5] P.K. Kuhl, *Nature Reviews Neuroscience*, 5, 831–843, 2004.
[6] J. Maye *et al.*, “Infant sensitivity to distributional information can affect phonetic discrimination,” *Cognition*, 82, B101–B111 (2002)
[7] J. F. Werker *et al.*, “Infant-directed speech supports phonetic category learning in English and Japanese,” *Cognition*, 103, 147–162 (2007)
[8] W. Labov *et al.*, *Atlas of North American English*, Mouton and Gruyter (2005)
[9] 中川他，音声・聴覚と神経回路網モデル, オーム社 (1990)
[10] 橋本，“広汎性発達障害（自閉症スペクトラム）”，母子保健情報，63，1–5 (2011)
[11] ローマン・ヤコブソン他，言語音形論，岩波書店 (1986)
[12] S. King *et al.*, “The Blizzard Challenge 2009”, *Proc. Blizzard Challenge 2009 Workshop*, 2009.
[13] Y. Qiao *et al.*, “A study on invariance of f-divergence and its application to speech recognition,” *IEEE Transactions on Signal Processing*, 58, 7, 3884–3890 (2010)
[14] N. Minematsu *et al.*, “Speech structure and its application to robust speech processing,” *Journal of New Generation Computing*, 28, 3, 299–319 (2010)
[15] 峯松他, “音声に含まれる言語的情報を非言語的情報から音響的に分離して抽出する手法の提案　～人間らしい音声情報処理の実現に向けた一検討～”, 電子情報通信学会論文誌, J94-D, 1, 12–26 (2011)
[16] D. Saito *et al.*, “Rotational properties of vocal tract length difference in cepstral space,” *Journal of Research Institute of Signal Processing*, 15, 5, 363–374 (2011)
[17] 鈴木他，“音声の構造的表象と多段階の重回帰を用いた外国語発音評価”，情報処理学会論文誌，52，5，1899–1909 (2011)
[18] 峯松他，“音声の構造的表象に基づく学習者分類の検証と発音矯正度推定の高精度化”，情報処理学会論文誌，52，12，3671–3681 (2011)
[19] T. Zhao *et al.*, “Automatic Chinese pronunciation error detection using SVM with structural features,” *Proc. Spoken Language Technology*, 473–476 (2012)
[20] M. Suzuki *et al.*, “Discriminative reranking for LVCSR leveraging invariant structure,” *Proc. INTERSPEECH* (2012)
[21] D. Saito *et al.*, “Structure to speech – speech generation based on infant-like vocal imitation –,” *Proc. INTERSPEECH*, 1837–1840 (2008)
[22] H.-P. Shen *et al.*, “Speaker-based pronunciation clustering of World Englishes based on pronunciation structure analysis,” *IEICE Technical Report*, SP2012-116, 7–12 (2013)
[23] D. Crystal, *English as a global language*, Cambridge University Press, New York (1995)
[24] J. Jenkins, *The phonology of English as an international language*, Oxford University Press (2000)
[25] Speech Accent Archive, http://accent.gmu.edu