社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－ＳＬＰ－69（31）
2007／12／20

始終端特徴パラメータの線形結合による調音結合の分析

西 宏之† グエン・ヴァン・ドン†

† 崇城大学情報学部
熊本市池田４丁目22の１
E-mail: †nishi@cis.sojo-u.ac.jp

あらまし 調音結合は，人間の音声生成機構の構造的特徴とその運動能力に起因し，人間同士の音声言語コミュニケーションにおいて，音声の明瞭性および自然性を与える重要な現象であると考えられる．しかしながら，音声認識の立場から見ると，調音結合は音声の特徴パラメータが連続的に変化することを意味し，音素や音節などの言語的情報と特徴パラメータとの写像関係が不明確となることから，認識処理の困難さの原因の一つとなっている．従来の音声認識手法では，上記調音結合の問題を解決するために，前後の音素環境を考慮した音素HMMを用いるなど，大量のデータに基づく統計・確率的な手法が主流であった．これらのシステムでは，学習データが十分に提供できるアプリケーションでは高い認識率を確保できるという特長がある反面，十分な量の学習データが準備できない場合や，外来語などで従来にない音素の連続を含む単語などが出現した場合に，対応が困難となるという問題がある．また，調音結合を，ある音素から別の音素への遷移過程と見た場合に，どのような軌跡により音響空間内を移動しているのかという知見が陽に得られないという点も問題である．本報告は，２重母音を対象とし，第１母音から第２母音への遷移を詳細に分析するとともに，その知見をもとに，調音結合を，始点と終点の特長パラメータの線形結合で表現することを試みたものである．さらに，第１母音から第２母音に至る途中に，おどり場としての特長パラメータの存在を仮定するモデル(おどり場モデル)に基づく調音結合モデルを提案する．

キーワード 音声認識，音節HMM，特定話者，話者適応，トレーニング

Analysis of co-articulation using linear combination of characteristic parameters in the first and the last frames

Hiroyuki NISHI† and Nguyen VAN DON†

† SOJO University Conputers and Infomations Science Department
4-22-1, Ikeda, Kumamoto city
E-mail: †nishi@cis.sojo-u.ac.jp

Abstract The co-articulation is originated in a structural feature of human body and the moving ability of man's speech generation mechanism and is thought to be an important phenomenon to give distinctness and the naturalness of speech in oral communications. However, the co-articulation is one of the causes of the difficulty of speech recognition because feature parameters change continuously, and the relationship between language information and the feature parameters becomes indefinite. This report describes the analysis result of the transition from the 1st vowel to the 2nd one in detail for the diphthong, and using the analysis results, proposes Landing Model that assumes the feature parameters go through the landing on the way from the 1st vowel to the 2nd vowel.

Key words speech recognition, syllable HMM, speaker dependent, speaker adjustment, training

1. はじめに

現在，多くの音声認識システムでは，音響モデルとして，前後の音素環境を考慮した不特定話者音素HMM(以下，トライフォンモデルと呼ぶ)が用いられている(例えば[1][2])．トライフォンモデルは，HMMの出力確率，遷移確率ともに前後の音素環境を考慮して学習されるため，平均的には正解候補に対して高い尤度を出力できるものの，１つの状態で複数のフレームの出力確率を計算するため，状態の中心部以外では，出力確率が小さくなるという問題点がある．

そこで，例えば，子音-母音間の誤差を小さくするために，音節など，より大きな単位のモデル [3] [4] [5] を用いることで誤差を小さくする試みがなされている．

報告者らも，音節を単位とする単語音声認識において，HMM の分散を最適化する手法，無音部分を削除することで，尤度の低下を小さくする手法について報告した [6] [7].

しかし，これらの手法は基本的に 1 音素あたり 3 状態程度の HMM で表現するという枠組みは変わらず，本来連続的に変化する特徴パラメータを，高精度に表現できる枠組みとなっていないことが指摘されている．そこで，離散的な状態を，連続的に表現する手法として，HMM のトラジェクトリを HMM に組み込む手法等が試みられている [8].

従来の研究手法は，学習データが十分に提供できるアプリケーションでは，一定以上の音声認識率を確保できる利点があったが，十分な量の学習データが準備できない場合や，外来語・新語などで従来にない音素の連続を含む単語などが出現した場合に，対応が困難となるという問題がある．また，調音結合を，ある音素から別の音素への遷移過程と見た場合に，どのような軌跡により音響空間内を移動しているのかという音声科学的立場の知見が陽に得られないという点にも不満がある．

すでに，孤立発話した音節から，モルフィングによって調音結合を表現する方法について報告したが，第 1 音素から第 2 音素に直線的に遷移することを前提としたため，実際の 2 重母音の特徴パラメータ軌跡を高精度に表現できるまでには至っていない [9] [10].

本報告では，2 重母音を対象とし，はじめに第 1 母音から第 2 母音への特長パラメータの遷移を詳細に分析する．その知見をもとに，調音結合を，始点と終点の特長パラメータの線形結合で表現することを試みる．さらに，第 1 母音から第 2 母音に至る途中に，おどり場としての特長パラメータを仮定する，おどり場モデルに基づく調音結合モデルを提案する．

2. 調音結合モデルの仮説と分析手法

2 重母音の調音結合は，図 1 に示すように，始点となる第 1 母音から，第 2 母音への連続的な変化によって実現すると考えられる．特徴パラメータの音響空間内では図 2 のように表現される．

したがって，その調音結合モデルは，連続的な変化の過程を高精度に表現できるものでなければならない．

調音結合モデルとして下記の仮説を立てる．

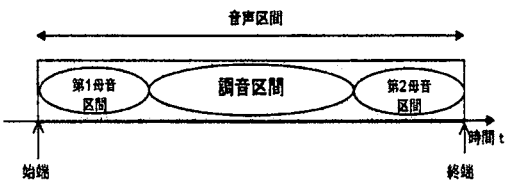

図 1　2 重母音の調音結合の考え方

[定義]

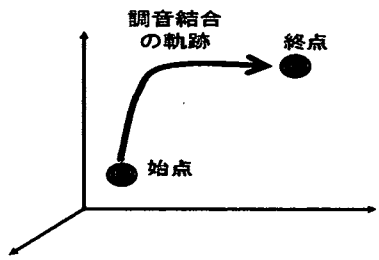

図 2　軌　　跡

p_1：　第 1 母音の代表点の特徴パラメータ

p_2：　第 2 母音の代表点の特徴パラメータ

i：　調音区間始点を開始点 (0) とするフレーム番号

$p(i)$：　調音区間の第 i フレームの特徴パラメータベクトル

$\alpha(i)$:　第 i フレームの特徴パラメータの第 1 母音の成分係数

$\beta(i)$:　第 i フレームの特徴パラメータの第 2 母音の成分係数

N:　調音区間のフレーム長 (フレーム数)

[仮説]

調音区間の第 i フレームの特徴パラメータ ($p(i)$) は次式で与えられるものとする．

$$p(i) = \alpha(i)p_1 + \beta(i)p_2 \qquad (1)$$

$$p(0) = p_1$$

$$p(N-1) = p_2$$

式 (1) は，調音区間内の音響特徴パラメータは，第 1 母音と第 2 母音の線形結合で表現できるという仮説を表す．

$\alpha(i)$ は，第 i フレームにおける特徴パラメータ中の第 1 母音成分の割合を表し，$\beta(i)$ は，第 2 母音の成分の割合を表す．$\alpha(i)$，$\beta(i)$ ともに i の関数である．

始点 ($i = 0$) では，$\alpha(0) = 1$，$\beta(i) = 0$ であり，i が大きくなるにつれて，$\alpha(i)$ は小さくなり，$\beta(i)$ は大きくなる．終点に達すると，$\alpha(N) = 0$，$\beta(N) = 1$ となる．

この仮説を検証するための分析手法を以下に述べる．

はじめに，2 重母音音声から，調音の開始点フレームを表す始点フレーム，調音の終了を表す終点フレームを決定し，始点フレームより前および，終点フレームより後ろのフレームは分析対象外とする．

次に，式 (1) によって計算された調音結合特徴パラメータと，実際のフレームの特徴パラメータとの距離値 () が最小となるように，$\alpha(i)$，$\beta(i)$ を決定する．

すなわち，2 つの特徴パラメータ x と y との距離を，

$$D(x,y) = \sum_{i=0}^{N-1}(x_i - y_i)^2$$

(x_i は，特徴パラメータ x の第 i 次の値)

とすると，$\alpha(i)$，$\beta(i)$ は次式のように表される．

$$(\alpha(i),\beta(i)) = arg \min_{\alpha(i),\beta(i)} D(p(i), \alpha(i)p_1 + \beta(i)p_2) \qquad (2)$$

得られた $\alpha(i)$, $\beta(i)$ の, $i=0,1,2,...,N-1$ における値の変化が, 調音結合の挙動を示す. ただし, ここで, 式 (1) によって計算された特徴パラメータと, 実際のフレームの特徴パラメータとの距離値の大きさを吟味し, 上記仮説が, 成り立つか否かの検証を行う. 成り立たない場合は, 仮説の補正を行うこととする.

3. 実　　験

3.1 実験条件

2. に述べた仮設の検証を行うための実験条件を表 1 に示す. 日本人成人男性 6 名の, 20 種類の 2 重母音を各 1 回収録した.

表 1 実験条件

項目	実験条件
分析条件	16kHz サンプリング,16bit フレーム長 : 30ms フレームシフト : 10ms 窓関数 : ハミング窓
特徴パラメータ	パワーおよびケプストラム 16 次
音声区間検出しきい値	無音区間パワー値に 10dB を加えた値
収録音声	2 重母音 20 種類 各 1 回 あい, あう, あえ, あお いあ, いう, いえ, いお うあ, うい, うえ, うお えあ, えい, えう, えお おあ, おい, おう, おえ
発声者	日本語を母国語 とする成人男性 6 名 HYA,KIM,YAM,YOK,YON
第 1 母音の 特長パラメータ	音声区間始端から 5〜7 フレーム目の平均値
第 2 母音の 特長パラメータ	音声区間終端から前 5〜7 フレーム目の平均値
分析に用いる 計算値	①式 (2) により計算される α (i), β (i) ②上記 α (i), β (i) における $D(p(i),\alpha(i)p_1+\beta(i)p_2)$ ③ $D(p(i),p_1)$ ④ $D(p(i),p_2)$

収録した音声から, 無音区間パワー値に 10dB を加えた値をしきい値として, 音声区間を切り出し, ハミング窓をかけた後, ケプストラムを求めた.

切り出された音声区間の最初から 5〜7 フレームの平均値を第 1 母音の特徴パラメータとした. 同様に, 終端フレームから前方の 5〜7 フレームの平均値を第 2 母音の特徴パラメータとした.

仮説を吟味するため, 下記①〜④の値を計算した.

①線形結合パラメータ : 式 (2) により計算される $\alpha(i)$, $\beta(i)$, すなわち, 式 (1) の左辺と右辺の値の差が最小となる場合の $\alpha(i)$, $\beta(i)$

②入力フレームの特徴パラメータと計算値との距離 : ①の $\alpha(i)$, $\beta(i)$ における $\alpha(i)p_1+\beta(i)p_2$ と, 入力フレームの特徴パラメータ $p(i)$ との距離値 $D(p(i),\alpha(i)p_1+\beta(i)p_2)$

③ $p(i)$ と第 1 母音特徴パラメータとの距離　入力フレームの特徴パラメータ $p(i)$ と第 1 母音特徴パラメータ p_1 との距離 $D(p(i),p_1)$

④ $p(i)$ と第 2 母音特徴パラメータとの距離　入力フレームの特徴パラメータ $p(i)$ と第 2 母音特徴パラメータ p_2 との距離 $D(p(i),p_2)$

①, ②は, 仮説による特徴パラメータの軌跡が実際の調音結合とどの程度合っているかを確かめるためのものである. ③, ④は全音声区間の中から, 調音区間を切り出すために求めた. 具体的には始点からの距離が一定値を越えた時点で調音が開始したと判断し, 終点からの距離が一定値以内になった時点で調音が終了したと判断する.

3.2 実験結果

表 1 に示した条件で音声の収録, ケプストラム分析を行い, さらに①〜④の値を計算した. 図 3 は話者 HYA の 2 重母音「あい」の表 1 における① ($\alpha(i)$, $\beta(i)$) の計算結果, 図 4 は, ② $D(p(i),\alpha(i)p_1+\beta(i)p_2)$, ③ $D(p(i),p_1)$, ④ $D(p(i),p_2)$ である. 図 3 より, $\alpha(i)$ は, 第 1 母音の代表として設定した, 第

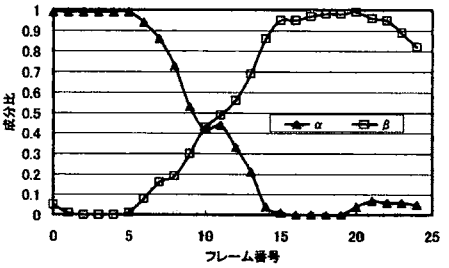

図 3　計算結果の例 1〜話者 HYA の「あい」の α, β の変化

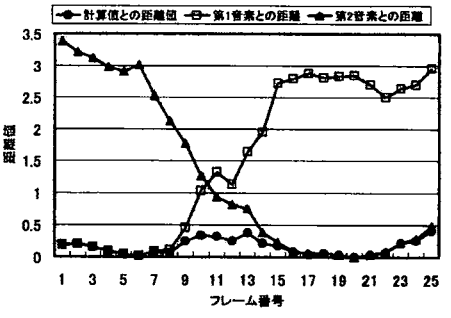

図 4　計算結果の例 2〜話者 HYA の「あい」の第 1 母音, 第 2 母音, 計算値と入力との距離値

5〜7 フレームで, 1 に近い値を示し, 以後徐々に減少して第 2 母音である 20 フレーム目付近で 0 に近い値に達していることが読み取れる. $\beta(i)$ は逆に, 第 5 フレームにほぼ 0 の値から出発し, 徐々に増加して, 20 フレーム付近で 0.9 を超える値に到達している.

一方, 図 4 からは, 図中第 1 母音と入力フレームとの距離値から 9 フレーム付近で第 1 母音から第 2 母音への調音が開始さ

れ，第2母音と入力フレームとの距離値から17フレーム付近で調音が終了していることが読み取れる．また，計算値と入力フレームとの距離値とから，調音の中間点付近では計算値と入力フレームとの距離値が若干大きくなり，第1音素と第2音素の線形結合により調音結合が表されるという仮説は，概ね正しいものの，不完全な領域が存在することがわかった．

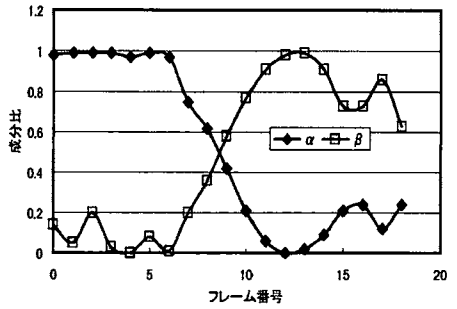

図5　計算結果の例3〜話者YAMの「うえ」のα，βの変化

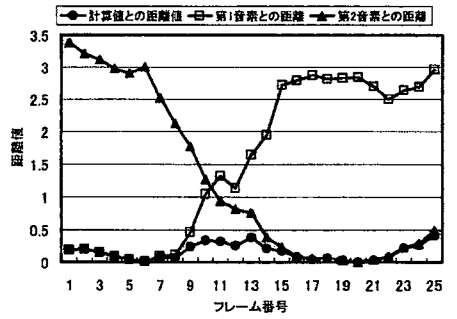

図6　計算結果の例4〜話者YAMの「うえ」の第1母音，第2母音，計算値と入力との距離値

この傾向は，図5および図6等，他の2重母音，他の発声者についても同様に見受けられた．

そこで，次節では，収集および計算により得られたデータを詳細に分析し，仮説を補正することを試みる．

4. 分析と考察

4.1 分析のためのデータの考察

4.1.1 調音区間

本報告では，調音区間内のパラメータの変動を分析するために，調音の開始時点，終了時点を次のように定義する．

調音の開始点　第1母音と入力との距離が0.1を超えた時点

調音の終了点　第2母音と入力との距離が0.1を下回った時点

調音区間長　調音の開始点と終了点の時間差

第1母音および第2母音の特長パラメータは，表1に示したように，始端および終端から5〜7フレーム目の平均値で与えられる．各入力フレームの特長パラメータと第1母音，第2母音の特長パラメータとのユークリッド距離を計算して，上記調音の開始点，終了点を求める．

2重母音の種類により，調音区間長がどの程度異なるかを詳細に分析し，調音結合をモデル化するための基礎データとして考察に用いることとする．

4.1.2 第1母音−第2母音間距離 $D(p_1,p_2)$

第1母音と第2母音との特長パラメータのユークリッド距離 ($D(p_1,p_2)$) である．この値は第1，第2母音の種類に依存する．本データは第1母音から第2母音に調音される際に，その移動量を示すものであることから，本データが調音結合に何らかの影響を与える可能性がある．調音結合に関する知見をうることを目的に，本データと他のパラメータとの相関を分析することとする．

4.1.3 入力と計算値との距離の平均値 ϵ_{ave}

3.2の最後に述べたように，計算値と入力フレームとの距離値とから，調音の中間点付近では計算値と入力フレームとの距離値が若干大きくなり，第1音素と第2音素の線形結合により調音結合が表されるという仮説には不完全な領域が存在することがわかった．

そこで，計算値と入力フレームとの距離値の調音区間内の平均値を測定し，この値と他のパラメータとの関係を調査する..

4.2 分析結果

4.に述べた分析の結果を表2に示す．

表2の各値は，4.の項目ごとに，6名の話者のデータを平均したものである．

表2　分析結果

2重母音の種類	音声長の平均 (ms)	調音区間長の平均 (ms)	第1-第2母音間距離の平均	ϵ_{ave} の平均値
あい	248	98	2.615	0.293
あう	240	82	0.814	0.185
あえ	252	107	1.587	0.211
あお	233	102	1.231	0.166
いあ	248	112	2.506	0.243
いう	275	135	1.471	0.204
いえ	260	107	0.970	0.146
いお	247	108	2.576	0.367
うあ	253	107	1.140	0.198
うい	260	120	2.515	0.303
うえ	272	117	1.114	0.169
うお	267	132	1.022	0.316
えあ	280	137	1.632	0.188
えい	268	105	0.869	0.153
えう	292	143	0.859	0.18
えお	258	115	1.932	0.252
おあ	262	122	0.671	0.23
おい	238	115	2.204	0.443
おう	280	125	0.774	0.248
おえ	283	133	1.713	0.289
平均	261	116	1.511	0.239

[調音区間長]

音声長は，「あう」「あお」「おい」が，他の種類の2重母音に比べて，短さに関して有意であったが，それ以外は特段の大きな

差は見られなかった．

調音区間長は，「あう」が他の2重母音に比べて，短さに関して有意であった．一方，「うお」および「えう」が，他より長いことに関して有意であった．また，音声長と調音区間長の相関性を図7に示す．相関係数は0.8であり，音声長と調音区間長

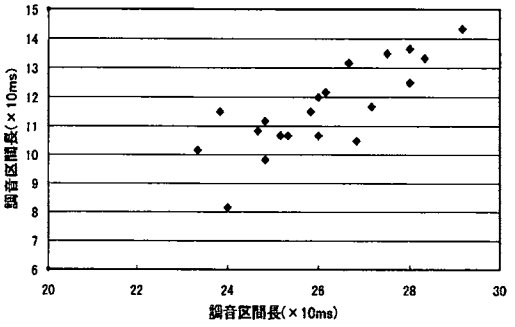

図7　音声長－調音区間長

には強い相関性が認められる．発話内容が2重母音であることから，音声長は発話速度に反比例する．すなわち，調音区間長は発話速度に反比例して伸縮することがわかる．このことから，調音結合をモデル化する際，発話速度を反映したパラメータの動きを表現できる構造となっていることが必要であることが確認された．

[第1母音－第2母音間距離 $D(p_1,p_2)$]

第1母音－第2母音間距離の計算結果を図8に示す．

有意差検定の結果，値の大きい方として，「あい」「いあ」「いお」「うい」「おい」「おえ」が，小さい方として，「あう」「うお」「えい」「えう」「おあ」「おう」が，危険率5％で有意であることが分かった．

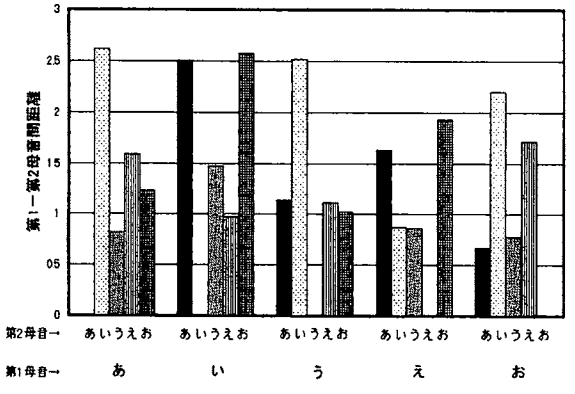

図8　第1－第2母音間距離 $D(p_1,p_2)$

第1－第2母音間距離は調音開始から終了までに移動すべき特徴パラメータ空間内の直線移動距離を表すことから，この値は移動に要する時間に影響があると考えられる．そこで，第1－第2母音間距離と調音区間長との相関を計算した．その結果を図9に示す．しかし，相関係数は-0.1であり，両者の相関

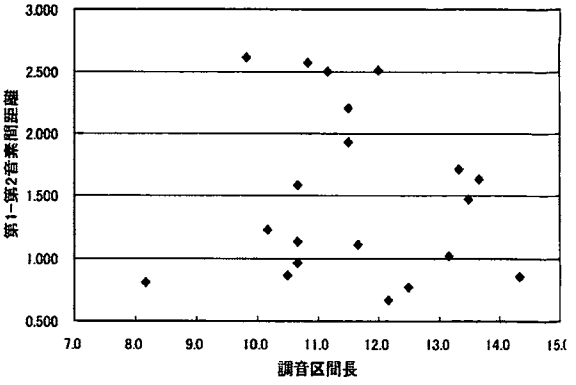

図9　調音区間長と第1-第2母音間距離の相関

性は認められない．これは，発話速度や調音に要する時間は，発話のリズムに大きく支配され，第1母音と第2母音の距離に関係しないことを示している．

[調音区間の入力フレームと計算値との距離の平均値 ϵ_{ave}]

図10に，調音区間の入力フレームと計算値との距離の平均値を示す．入力フレームの特徴パラメータと，対応する計算結果との距離を，調音区間にわたって平均したものである．値の大

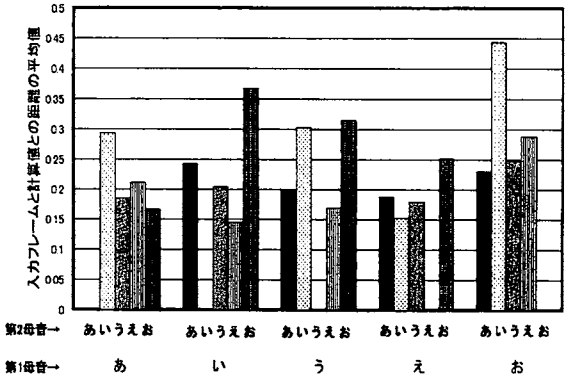

図10　調音区間の入力フレームと計算値との距離の平均値 ϵ_{ave}

きい方として，「あい」「いお」「うい」「うお」「おい」が，小さい方として，「あお」「いえ」「うえ」「えい」「えう」が，危険率5％で有意であることが分かった．

図10を，図8と比較すると，部分的な違いを除くと，大小関係が類似していることがわかる．両者の相関を図示すると，図11のようになる．相関係数は0.634であり，弱いながらも相関性が認められる．

この関係から，第1母音と第2母音特徴パラメータの線形結合により2重母音の調音結合が表されるという仮説の，不完全な領域を説明することが可能である．すなわち，第1母音と第2母音間の距離が比較的小さい場合は，両母音間の直線距離と実際の特徴パラメータの軌跡のずれが小さいことから，仮説が相対的によく成り立つ．一方，第1母音と第2母音間の距離

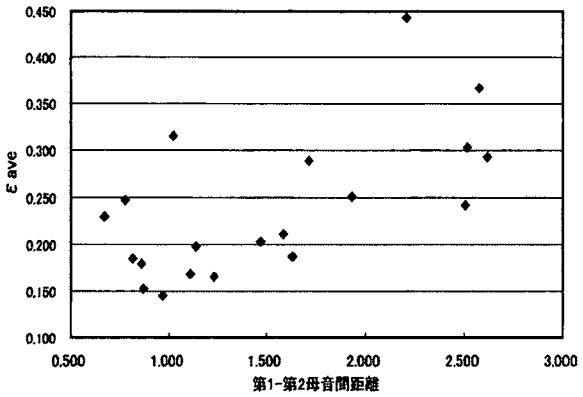

図 11 ϵ_{ave} と $D(p_1, p_2)$ の相関

が比較的大きい場合は，両母音間の直線距離と実際の特徴パラメータの軌跡のずれが大きくなり，仮説の不完全性が強調されることになると考えられる．

5. おどり場モデルの提案

第 1-第 2 母音間の直線軌跡と実際の特徴パラメータの軌跡のずれを吸収するために，次に述べるおどり場モデルを考察する．おどり場モデルの概念図を図 12 に示す．前記の仮説では，第 1 母音と第 2 母音の線形結合であったが，このおどり場モデルでは，両者の間におどり場となる第 3 の特徴パラメータが存在すると仮定する．第 3 のパラメータは第 1 母音，第 2 母音，$\alpha(i)$, $\beta(i)$ に依存し，以下の方法で求められるパラメータである．

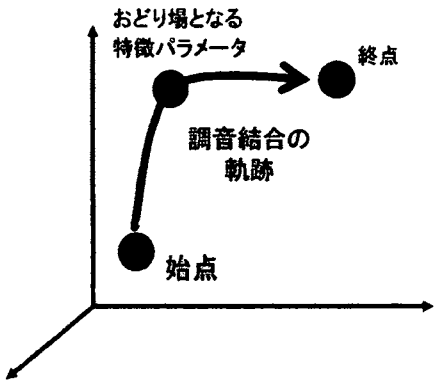

図 12 おどり場モデル概念図

おどり場パラメータを p_l とすると，

$$p_l = ave(p(i) - \alpha(i)p_1 - \beta(i)p_2) \tag{3}$$

で，与えられる．すなわち，$\alpha(i)p1 + \beta(i)p2$ で予測された軌跡と実際の $p(i)$ との誤差を調音区間内で平均している．

この p_l は第 1，第 2 の両母音に依存するので，以後 $p_l(i,j)$ と記す．

i,j は第 1，第 2 の母音番号である．$p_l(i,j)$ は，対応する母音 i,j ごとに保存される．

この結果，新たな仮説として，おどり場モデルによる下記の式が導かれる．

$$p(i) = \alpha(i)p_1 + \beta(i)p_2 + \gamma(i)p_l \tag{4}$$

$$\alpha(i),\beta(i),\gamma(i) = arg \min_{\alpha(i),\beta(i),\gamma(i)} D(p(i), \alpha(i)p_1 + \beta(i)p_2) + \gamma(i)p_l \tag{5}$$

式 (4)(5) により得られた $\alpha(i), \beta(i), \gamma(i)$ を用いて，第 1 母音，第 2 母音から，その調音区間におけるパラメータの軌跡を予測し，例えば，DP 標準パターンを作成することは容易である．

6. ま と め

2 重母音を対象とし，第 1 母音から第 2 母音への特長パラメータの遷移を詳細に分析した．その知見をもとに，調音結合を，始点と終点の特長パラメータの線形結合で表現することを試みた．さらに，第 1 母音から第 2 母音に至る途中に，おどり場としての特長パラメータを仮定し，さらに精度の良いおどり場モデルに基づく調音結合モデルを提案した．今後は，提案したおどり場モデルを用いた計算値と，実際の入力パラメータとの距離値を評価し，単母音のみから 2 重母音の音声認識用標準パターンを作成する手法を確立し，実際の認識実験に発展させる．

文 献

[1] 鹿野他編著:"IT Text 音声認識システム", Ohmsha 社, pp.133〜pp.134, 2001 年
[2] 板橋秀一編著:"音声工学", 森北出版, pp.219, 2005 年
[3] 中川聖一, 花井建衆, 山本一公, 峯松信明:"HMM に基づく音声認識のための音節モデルと triphone モデルの比較", 信学論, D-II, Vol. J83-D-II, No.6 pp.1412-1421, 2000 年 6 月
[4] 山本一公, 池田太郎, 松本弘:"コンパクトで高精度な音節モデルの検討", 音学講論, 1-9-22, 2002 年秋季研究発表会, pp.43, 2002 年 9 月
[5] 緒方淳, 有木康雄:"日本語話し言葉音声認識のための音節に基づく音響モデリング", 信学論, D-II, Vol. J86-D-II, No.11 pp.1523-1530, 2003 年 11 月
[6] 西宏之, 江藤貫峰:"音節 HMM 特定話者音声認識における音声区間検出方法の影響", 平成 18 年度電気関連学会九州支部連合大会, 06-1P-03, pp.151, 2006 年 9 月
[7] 西宏之, 江藤貫峰:"特定話者音節 HMM の標準偏差補正と無音削除処理による認識率改善", 情報処理学会研究報告, 2006-SLP-64(20), pp.113-118, 2006 年 12 月
[8] 南康浩, エリックマクダーモット, 中村篤:"カルマンフィルターによる音声認識のための特徴量トラジェクトリ生成法", 2005 年度音響学会春季研究発表会, 1-5-17, pp.33, 2005 年 3 月
[9] 西宏之, グエン・ヴァン・ドン:"単母音パラメータのモルフィングによる二重母音標準パターンの作成", 平成 19 年度電気関連学会九州支部連合大会, 05-1A-15, pp.45, 2007 年 9 月
[10] 西宏之, グエン・ヴァン・ドン:"単母音パラメータのモルフィングによる二重母音標準パターンの合成", 平成 19 年度音響学会秋季研究発表会, 3-3-18, pp.181, 2007 年 9 月