二言語に渡る構造的表象に基づく音声・言語変換の実験的検討*

◎見原 隆介，齋藤 大輔，峯松 信明，広瀬 啓吉 (東大)

1 はじめに

近年のコンピュータを用いた音声合成分野においては，テキスト (文字情報) から音声を合成する Text-to-Speech（TTS）方式が主流になっている．この方式では音韻列と音響事象の対応関係が統計的に学習されるが，ある程度大規模なデータベースの構築とラベリングが必要である．音声合成技術を幅広くアプリケーション応用していくには，話者性と言語性とを独立させて処理できる枠組みが望ましい．例えば，ある特定の話者・言語の音声から任意の話者・言語の音声を合成できる枠組みが実現すれば，外国語学習や機械翻訳への応用が期待されるほか，海外映画の吹替音声を役者本人の声質で作成するといった用途も可能となる．そうした枠組みを実現する手法として音声変換技術が応用されており，日英バイリンガル話者と日本語話者の音声から日本語での話者変換関数を作成し，それを英語音声に適用するという言語性の異なる音声を対象とした話者変換が真下らによって提唱された [1].

任意の話者性，任意の言語性の音声合成を行う上では，言語性・話者性の情報と音響的特徴との対応関係を明示的に分析・モデル化する必要がある．そうすることで音声の音響的特徴から「その話者であること」に対応する部分を削除する事も可能になる．近年，峯松らによって構造的表象が提案されているが，これは音響事象間の相対関係のみを表象しており，話者性を削除した音声の音響的表象となっている [2]．またこの構造的表象に基づき，ある話者の日本語母音群が構成する構造表象に対して別話者の幾つかの母音サンプルを適用することで，別話者の音声を合成する手法が齋藤によって提唱されている [3].

本研究では，文献 [3] の合成手法の対象を日本語・英語の二ヶ国語に拡張し，日英二ヶ国語に渡る構造を抽出し，これに対して別話者の一方の言語音声をサンプルとして与え，もう一方の言語音声を合成することを試みる．

2 音声の構造的表象

2.1 非言語的情報による音響事象の変形

音声の音響的特徴は，非言語的情報による変形を不可避的に受けるが，その変形は乗算性変形と線形変換性変形に大別できる．乗算性変形はスペクトルに対する乗算で表現される変形であり，ケプストラム空間においては加算演算 $c' = c + b$ として表現される．この種の変形の例としてマイクロフォンの音響特性が挙げられる．

一方，線形変換性変形はケプストラム空間において行列 A による線形変換 $c' = Ac$ として表現されるものである．スペクトルにおいて話者の声道長差異や聴覚特性差異を表現する周波数ウォーピングは，ケプストラム空間においては近似的に線形変換によって表現できることが示されている [4]．よって，話者ごとの声道形状特性や聴覚特性の差異は線形変換性変形として表現できる．

音声は必ずある話者によって発声され，必ず収録機器を通して収録されるため，これらの変形は音声には不可避であると言える．以上により，これら非言語的情報による音響的特徴の変形は $c' = Ac + b$ というアフィン変換によって近似的にモデル化できる．

2.2 音声の構造的表象

ユークリッド空間において N 角形の形状は ${}_NC_2$ 本の頂点間距離を定めることで一意に決定できる．即ち音響事象群においても，全事象間の距離を求めることで (距離行列) 事象群全体を構造的に表象できる．しかし，音響事象をケプストラム空間の点として捉え，点間距離行列で構造を構成した場合，話者性の違いによってその構造が必ず歪められる．それは，非言語的情報による変形がアフィン変換でモデル化されるからである．

ここで，点間距離の代わりに分布間距離を用いることを考える．分布間距離の一つである Bhattacharyya 距離（以下，BD）は，任意の二つの確率密度分布 $p_1(x)$ と $p_2(x)$ について，下式 (1) によって求まる．

$$BD(p_1, p_2) = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(x)p_2(x)}\, dx \qquad (1)$$

この時，二つの分布に対して共通のアフィン変換を施しても，二分布間の BD は変換前後で不変である．この不変性は非線形変換においても成立する [2]．即ちケプストラム空間における音響事象群の関係を分布間距離で表すことで，非言語性変形におよそ不変な構造的表象を求めることができる．

---

*Cross-lingual voice conversion based on structural representation of speech. by MIHARA Ryusuke, SAITO Daisuke, MINEMATSU Nobuaki, HIROSE Keikichi (The University of Tokyo)

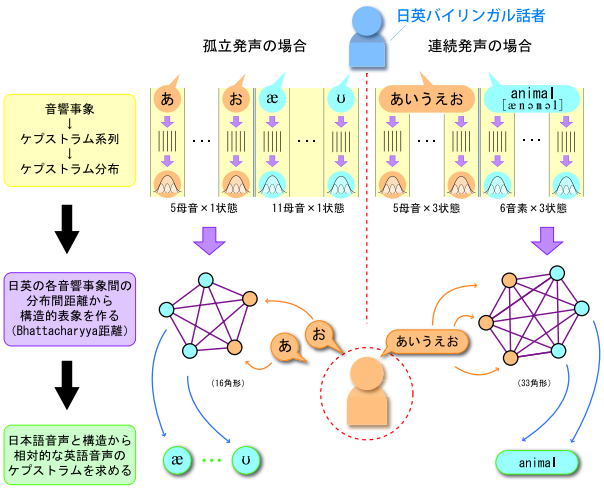


Fig. 1　音声の構造化および音声変換の枠組み

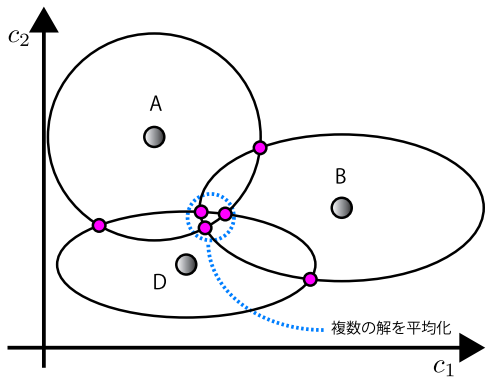


Fig. 2　解析手法に基づく解の導出 (2 次元の場合)

2.3　一発声の構造化

一発声から構造的表象を抽出する流れを Fig. 1 上段に示す．音声波形から短時間スペクトル系列を求め，ケプストラム系列に変換する．ケプストラム系列もまた時系列信号であるため，適当な時間長で区切られたケプストラム系列は，同一の音響事象の分布からのサンプル列として見なすことができる（各分布に対応する時間長は分布によって異なる）．これらのケプストラム分布群全ての分布間距離を求めることで一発声から構造的表象を得る (Fig. 1 中段)．

3　構造的表象に基づく音声合成

3.1　解析的手法によるベクトル解の探索

構造的表象は音声から話者の声道長などの情報を削除した表象となっている．これに対して，声道の情報を構造に戻すことで音声を生成する枠組みが齋藤らによって提案された．ここでは，声道形状の話者性に対応するパラメータを求める操作をケプストラム空間における解探索問題として定式化し，さらに探索の高速化が検討されてきた [5]．構造的表象は音響事象群の距離行列によって構成されるため，それだけでは各事象を音響空間に定位することはできない．しかし，既に発声された一部の音響事象を与えれば，ケプストラム空間におけるその他の音響事象の存在領域を相対的に推定できる．

二つの音響事象がガウス分布が $\mathcal{N}(\mu_1,\Sigma_1)$, $\mathcal{N}(\mu_2,\Sigma_2)$ となる場合，BD の計算式 (1) は次の式 (2) ようになる．

$$BD=\frac{1}{8}\mu_{12}^T\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}\mu_{12}+\frac{1}{2}\ln\frac{|(\Sigma_1+\Sigma_2)/2|}{|\Sigma_1|^{\frac{1}{2}}\,|\Sigma_2|^{\frac{1}{2}}} \quad (2)$$

($\mu_{12}=\mu_1-\mu_2$,$|\Sigma|$ は Σ の行列式を表す．)

式 (2) のうち BD，μ_2，Σ_1，Σ_2 を既知であると仮定し，これを μ_1 が満たすべき方程式であるとすると，解 μ_1 は多次元空間における楕円体を描く．ここで 2 次元の場合を考え，初期条件として二つの音響事象 $\mathcal{N}(\mu_a,\Sigma_a)$，$\mathcal{N}(\mu_b,\Sigma_b)$ から，音響事象 p の平均 $\mu_p=(\mu_p^x,\mu_p^y)$ を求めることを考える．すると，式変形により μ_p に対する以下の連立方程式 (3) が定まる．この時，解 μ_p は二つの楕円体の交点として求まる．

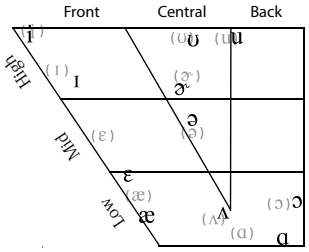
$$\begin{cases} BD_a-\epsilon_a=\displaystyle\sum_{s\in\{x,y\}}\frac{1}{4({V_p}^s+{V_a}^s)}(\mu_p^s-\mu_a^s)^2 \\ BD_b-\epsilon_b=\displaystyle\sum_{s\in\{x,y\}}\frac{1}{4({V_p}^s+{V_b}^s)}(\mu_p^s-\mu_b^s)^2 \end{cases} \quad (3)$$

但し，音響事象の分散共分散行列は対角としその成分を V^x，V^y と表記している．p の分散項も既知としている．また簡潔な表記のため式 (2) の右辺第二項を ϵ，2 次元の x，y 成分を右肩の添字で表している．

この時，二つの楕円の交点は一般には二つ，長軸および短軸の配置により最大で 4 つ求まる．そのため音響事象を一意に求めるにはさらに方程式が必要となる．一般に n 次元空間において n 個の超楕円体だけでは交点をただ一つに定めることはできない．よって n 次元において一つの音響事象の定位には $n+1$ 個以上の音響事象が必要となる．2 次元の場合の例を Fig. 2 に示す．図中の楕円の中心は，それぞれ初期条件として与えた音響事象である．得られた楕円の交点が最も縮退している箇所がターゲットとなる音響事象の存在領域として考えられ，その領域における交点群の平均ケプストラムを解とする [3].

3.2　二言語に渡る構造からの音声合成

バイリンガル話者から得た日本語音声と英語音声を網羅する構造的表象に対して，初期条件として任意の日本人の日本語音声を与え，3.1 の手法を用いて英語音声のケプストラムベクトルを探索する．そして得られたケプストラムに，予め抽出したピッチ，時間長，パワーを適用し，話者性を与える形で音声を合

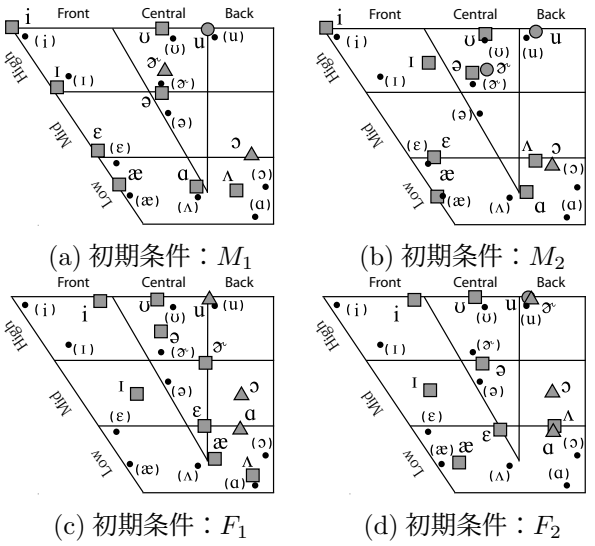


黒色灰色：データクローズドの評価結果

Fig. 3　母音図における話者 M_1 の英語 11 母音

Table 1　共鳴音のみから成る英単語 20 語

airline	align	aluminium	animal
annually	anyway	anywhere	aura
euromoney	learn	millenium	millionaire
oil	onion	railway	rumania
unknown	woman	yummy	rumor



(a) 初期条件：M_1　(b) 初期条件：M_2

(c) 初期条件：F_1　(d) 初期条件：F_2

※●：M_1 の収録音声，□：円唇なし，△：弱い円唇，○：強い円唇

Fig. 4　M_1 の構造に基づく合成音声の評価

成する．本研究では孤立発声母音音声と連続発声音声の 2 種類の合成を試みた．連続発声においては調音結合を考慮して音素数の 3 倍の状態数でケプストラム系列を分割し，構造抽出を行った (Fig. 1).

4　合成実験

4.1　実験条件

バイリンガル話者から得た日本語音声と英語音声から網羅的な構造的表象を抽出し，日本語音声を初期条件に英語音声のケプストラムを推定した．以降，構造的表象と初期条件とを与える話者をそれぞれ構造提供者，初期条件提供者と区別する．構造的表象を抽出するために，収録音声に対して STRAIGHT[6] を用いたスペクトル分析を行い，40 次のケプストラムを得た．また，ケプストラムと同時にパワー，継続長，ピッチも STRAIGHT の分析を元に抽出し，構造解析より推定されたベクトル解 (ケプストラム) に適用して音声を合成した.

構造に基づく音声認識で研究された次元分割を応用した多次元空間における解探索が検討されているが [7, 3]，今回は孤立発声母音音声と連続発声音声のいずれもケプストラムを 1 次元ずつ計算・推定しており，式 (2) を用いて一つの初期条件から解候補を算出した.

4.2　孤立発声母音音声の合成

孤立発声母音の構造的表象の導出には，バイリンガル話者によって孤立発声された日本語 5 母音 (a，i，u，e，o) および英語 11 母音 (Fig. 3) の孤立発声母音を一発声ずつ用いる．今回は 4 名（男性話者 M_1，M_2，女性 F_1，F_2）のバイリンガル話者についてそれぞれ日英母音の網羅的構造表象を導出した．ここでは一発声全体 (即ち孤立母音) を単一の音響事象とみなし，構造的表象は 16 種類の音響事象群の距離行列として定められる (16 角形)．今回，M_1 を構造提供者として使用し，初期条件提供者 4 名 (M_1 を含む) とを組み合わせてケプストラムの推定および音声合成を行った.

4.3　連続発声音声の合成

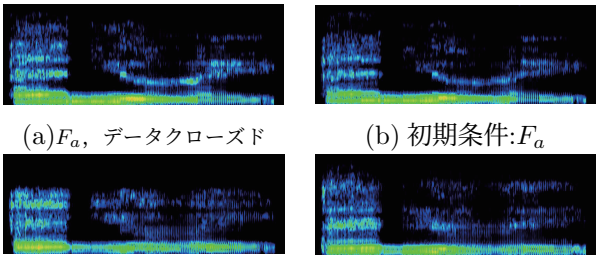
孤立母音に続き，連続母音列の音声についても合成実験を行った．構造的表象を抽出する為，日本語 5 母音の連続発声音声/aiueo/および共鳴音のみから構成される英単語 20 語 (Table. 1) を選定し，1 名の女性バイリンガル話者 (F_a) の音声を収録した.

これらの音声に対して，日本単語/aiueo/と英単語 (音素数 n) をそれぞれ 15 個，$3n$ 個のケプストラム分布に分割し，$15 + 3n$ 角形の構造的表象を抽出した．次に，上記と異なる日本語音声の 15 個のケプストラム分布を初期条件として，英単語音声の $3n$ 個の平均ケプストラムベクトルを推定した．今回，構造提供者 F_a のほかに日本語話者男女 1 名ずつ (M_a,F_b) による/aiueo/の音声を用い，3 人の初期条件提供者について合成を行った.

5　評価実験

5.1　孤立母音音声の評価

孤立発声母音の合成音声を音声学者による母音図へのプロットによって評価した．本実験では，男性話者 M_1 を構造提供者として使用した．まず，初期条件提供者として M_1 を用い，かつ構造抽出に用いた



(a)F_a，データクローズド　(b) 初期条件:F_a

(c) 初期条件:F_b　(d) 初期条件:M_a

Fig. 5　合成音声 “annually” の評価

日本語母音音声を使って英語母音のケプストラムを推定した (データクローズド)．データクローズドの場合は理論的に真値が得られる．Fig．3 に M_1 の構造推定に使った英語母音を黒で，データクローズドの日本語母音から推定された英語母音を灰色で示す．両者のずれは，主観評価におけるプロットの揺らぎ，誤差の指標と考えられる．

一方，Fig．4 は構造抽出時とは異なる日本語 5 母音の音声を初期条件とした合成（データオープン）の評価結果であり，(a) は M_1 の話者クローズド・データオープン，(b)〜(d) は M_2，F_1，F_2 の日本語音声を初期条件とした話者オープンの結果を示す．話者クローズド・データオープンの結果 (a) が収録音声と近い分布をとっているのに対し，他者の音声を用いた (b)，(c)，(d) は収録音声と比較してずれが大きくなっている，また異性話者の音声を用いた (c)，(d) の方がその傾向がやや強く見られる．

5.2　連続発声音声の評価

連続発声音声の合成音声をスペクトログラムを用いて比較した (Fig．5)．(a) はデータクローズド，(b)〜(d) は F_a，F_b，M_a の日本語音声を初期条件としたデータオープンの合成音声のスペクトログラムである．元音声とほぼ一致する (a) と比較し，(b)〜(d) のスペクトログラムも目的の英単語音声の概形はとれており，言語情報は十分に再現できていると言える．しかし，ターゲットとする話者性については十分に生成できていなかった．

6　考察

今回の実験においては，ケプストラム空間を 1 次元ごとに分割し，各部分空間で構造解析，目的の音声のケプストラムを推定した．部分空間への分割は，BD の強すぎる不変性を緩和する目的で行っており [7]，これを音声合成にも応用し，適切な構造の制約条件の下での音声合成手法が検討されている [8]．今回の 1 次元ごとの分割は話者不変性を極端に抑制したケースである．孤立発声母音の評価結果として，同一話者内での推定計算は非常に精度が高く，続いて同性話者，異性話者の順に精度が落ちていた．連続発声音声についても，目的の英単語音声の言語情報は十分に再現にできたが，非言語的情報の再現は十分でなかった．また，孤立発声母音と同様，初期条件提供者の話者性が推定の精度に影響を与えている事を確認した．更に連続発声音声の場合，音素数 n の単語を $3n$ 個の分布に分割し構造化しているが，2 話者の/aiueo/を 15 状態に分割した時に，m 番目の分布がどの音素に当てはまるのかという点で発声の間にずれが生じる可能性があり，その場合には適切な時間合わせを行う必要も生じる．

7　おわりに

非言語的情報による音声の変形におよそ不変な音響的物理表象である構造的表象に基づく音声合成の応用として，構造抽出の対象とする音声の言語を増やし，初期条件とは異なる言語の音声を合成することを試みた．今回の孤立発声母音および連続発声音声の合成によって，初期条件によって精度に差があるが，任意の話者の日本語音声を初期条件に英単語音声を合成できることを確認した．今後は，部分空間の次元数を増やし，合成音声に対してより強く話者性を反映する推定方式を検討する予定である．

謝辞　孤立発声母音の合成音声の評価実験に際し，中央大学経済学部の牧野武彦先生に多大なご協力をいただきました．この場を借りて厚く御礼申し上げます．

参考文献

[1] 真下美紀子 他, 秋季音講論, 1-P-17, pp. 389-390, 2001.
[2] 峯松信明 他, 春季音講論, 1-P-12, pp. 147-149, 2007.
[3] 齋藤大輔 他, 信学技報, SP2008-40, vol. 108, No. 116, pp. 115-120, 2008.
[4] M. Pitz and H. Ney, IEEE Trans. Speech and Audio Processing, vol. 13, pp. 930-944, 2005.
[5] 齋藤大輔 他, 秋季音講論, 1-P-2, pp. 399-402, 2007.
[6] 河原英紀, 音響学会誌, vol. 63, No. 8, pp. 442-449, 2007.
[7] S. Asakawa *et al*., ICASSP’08, pp. 4097-4100, 2008.
[8] D. Saito *et al*., “Optimal event search using a structural cost function - improvement of structure to speech conversion -”, INTERSPEECH’09 (to appear).