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.

# 音声の構造的表象に基づく異言語間・異話者間の音声変換手法

見原　隆介†　　齋藤　大輔††　　峯松　信明†　　広瀬　啓吉†

† 東京大学大学院情報理工学系研究科 〒 113-8656 東京都文京区本郷 7-3-1
† 東京大学大学院工学系研究科 〒 113-0033 東京都文京区本郷 7-3-1
E-mail: {mihara,dsk_saito,mine,hirose}@gavo.t.u-tokyo.ac.jp

**あらまし**　音声は話者の声道形状の特性や音響機器の特性などの非言語的特徴によって変形するが，この非言語性の変形に対して凡そ不変な音声の構造的表象が提案されている．これは音声の物理的実体を捨象し，その音響空間内での相対的な動きのみを捉えた物理表象である．また，この構造的表象に基づく音声合成の枠組みが提案されている．この枠組みでは音声を発話内容（語形）と発話者の身体性とに分離して捉え，語形に対して発話者の身体性を付与する（戻す）形で音声を生成しており，幼児の音声模倣をモデル化したものといえる．本研究では，この音声合成系の対象を複数の言語に拡張し，任意の話者性と任意の言語性を独立に処理できる音声合成系を検討する．即ち，語学教師によって発声された日英二ヶ国語の音声を網羅する構造的表象に対して，日本語母語話者の身体性を日本語を通して付与することで，未修得言語の語形を音声として実体化することを検討する．この合成系による合成音声の評価を行うとともに，異なる条件下で合成した音声同士の比較を行い，合成手法の改善についても考察する．
**キーワード**　音声の構造的表象, 話者不変, 音声模倣, 言語変換，話者変換

# Cross-speaker and cross-language voice conversion based on structural representation of speech

Ryusuke MIHARA†, Daisuke SAITO††, Nobuaki MINEMATSU†, and Keikichi HIROSE†

† Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
† Graduate School of Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
E-mail: {mihara,dsk_saito,mine,hirose}@gavo.t.u-tokyo.ac.jp

**Abstract**　Speech acoustics easily vary due to non-linguistic factors such as speaker differences and microphone differences. The authors already proposed a structural representation of speech, where these variations can be effectively removed. This representation discards absolute and static properties of speech events and captures only their relative and dynamic features. Recently, a new framework of speech synthesis based on this structural representation has been proposed. Here, an utterance is characterized by two separate attributes, speaker-independent speech form and speaker-dependent embodiment features found in that utterance. On this framework, a new utterance is generated by realizing a given speech form acoustically and this realization is enabled by providing speaker-dependent embodiment features. This generation process can be viewed as implementation of infants' vocal imitation on a machine. In this report, based on this structural representation, cross-speaker and cross-language voice conversion is implemented. A speaker's two utterances of Japanese and English are modeled as a single speech form. By providing another speaker's Japanese utterance of the same content, the English utterance of that speaker will be generated using the speech form. In this report, the performance of the proposed method is evaluated and its problems are also made clear.
**Key words**　structural representation of speech, speaker-invariance, vocal imitation, language conversion, speaker conversion

## 1. はじめに

人間がコミュニケーションを円滑に行う上で，音声は多くの情報の伝達を担っており，音韻，話者性，感情などによって多様に変化する．音声の持つ情報は言語的情報，非言語的情報，パラ言語的情報に大別できるが，通常，音声を用いたアプリケーションでは，それらのいずれか一つに着目し，その他の特徴をある程度の規模の学習データによって平均化することで統計的にモデルを構築している．

ここで，音声の伝達する情報として，話者性と言語性を取り上げる．例えば，統計的手法によって任意の話者性に対応できる音声合成系を実現するには，言語性を合わせた上で十分な規模のデータベースを構築する必要がある．しかし，音声を幅広くアプリケーション応用してゆくには，話者性と言語性を独立に処理できる枠組みが望ましい．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・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，別話者の残りの母音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，この合成手法の対象を日本語・英語の二ヶ国語に拡張し，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音声変換手法を提案する．

本稿では，音声の時間構造，およびケプストラム分布の分散共分散行列に焦点を絞り，それぞれが本研究の合成系に与える影響について考察する．

## 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．

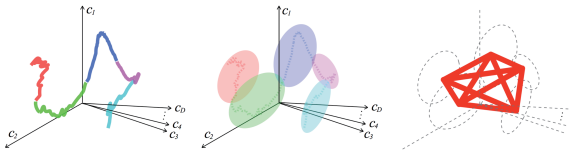


図 1　音声の動きを捉えた音声表象
Fig. 1　Speech representation by capturing only speech dynamics.

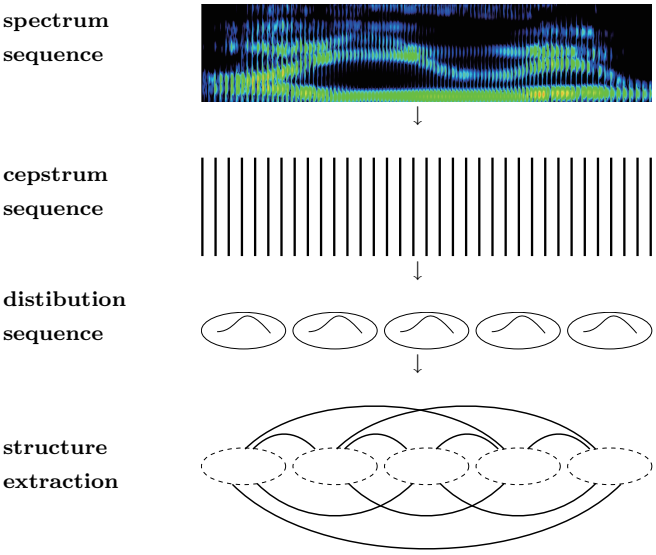



図 2　音声からの構造的表象の抽出
Fig. 2　Structure extraction from one utterance.

### 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，全事象間の距離行列を求めることで，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である Bhattacharyya 距離 (以下，$BD$) を用いることを考える．任意の二つの確率密度分布 $p_1(x)$ と $p_2(x)$ の $BD$ は下式 (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，

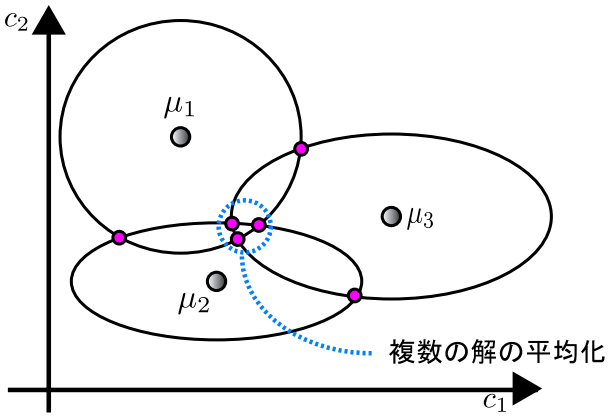



図 3　解析手法に基づく解の導出 (2 次元の場合)
Fig. 3　Solution of the searching probrem.

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と同時に，時間的に連続する音声のダイナミクスを距離行列として捉えることができる (図 1)．

### 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図 2 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同一区間内のケプストラム系列は，同一の音響事象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区間によって異なる)．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．実際には，一発声から隠れマルコフモデルを推定し，分布間距離行列を得る．

## 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 (身体性) 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，さらにそ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，ターゲット音の音響事象の存在領域をケプストラム空間において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．

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

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

式 (2) のうち $BD$，$\mu_2$，$\Sigma_1$，$\Sigma_2$ を既知であると仮定し，これ

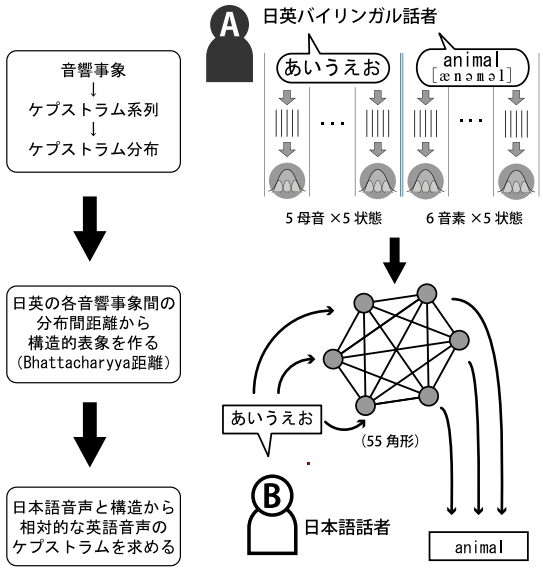



図 4　音声の構造化に基づく音声変換
Fig. 4　Voice conversion based on structuralization of speech.

を $\mu_1$ が満たすべき方程式であると解釈すると，解 $\mu_1$ 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 $\mu_p$ に対する以下の連立方程式が定まる．この時，解 $\mu_p$ は二つの楕円体の交点として求まる．

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

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 $\epsilon$，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 次元の場合の例を図 3 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

本研究では 3.1 の枠組みに則り，複数の言語に渡る音声変換を検討している．複数の発声に渡る構造的表象を構築するためには，各発声の非言語的特徴が一致している必要がある．そのためには，同一の話者，同一の収録環境で収録され，更に両方の言語について十分に正しい音韻性を伴った発声である必要があるため，バイリンガル話者によって発声された二ヶ国語の音声セットを用いることが望ましい．本研究ではバイリンガル音声の代わりに語学教師によって発声された日英二ヶ国語の音声

を収録し，いずれも正しい音韻性を伴った音声であるものとして用いる．

本研究では日本語と英語を対象の言語とし，日本語 5 母音の連続発声音声/aiueo/を初期条件に，英語音声を合成する．本研究では，発声が声道の特性に依存する音響事象のみ対象とし，英語音声は共鳴音のみから成る “I owe you one.” を選んだ．

続いて，調音結合を考慮して音素数の 5 倍の状態数で各音声のケプストラム系列を分割し，日本語音声と英語音声の二つに渡る構造を抽出し (図 4)，3.1 の手法により，日本語を初期条件に英語音声を構成する音響事象を定位させる．以降，構造的表象や初期条件を与える話者をそれぞれ構造提供者，初期条件提供者と記述する．そして得られたケプストラムベクトルは，目的の英語音声の各音響状態に対応する平均ベクトルである．この平均ベクトル群に予め抽出したピッチ，状態継続長，パワーを適用して音声を合成する．

### 3.3 特徴量空間分割

構造的表象を用いた音声認識において，構造の “過剰な不変性” のために，異なる単語を同一とみなす問題が指摘されていた．これは構造的表象がケプストラム空間で絶対座標を持たない幾何構造であるために自由度が高く，場合によっては全く別の単語と幾何構造が一致する危険があるためである．この問題に対し，朝川らは特徴量空間を分割することで話者の違いにのみ適切に不変性を成立させる方法を提案した [6]．今後，分割された部分空間の次元数をブロックサイズと表記する．また，齋藤らはこれを音声合成に応用し，適切な構造の制約条件の下での音声合成手法を提案している [7]．構造的表象に基づく音声合成において，ブロックサイズを 1 とした場合，話者不変性は極端に抑制され，初期条件の話者性はほとんど反映されない．ブロックサイズを 2 に設定し，ケプストラムの次元ごとの関係を制約条件として与え，話者不変性の適用を行う．

### 3.4 反 復 推 定

本研究では，ケプストラムの推定計算において，連立方程式 (3) から得られた解候補群の密集度から最適解の存在領域を推定している．この解の推定計算の精度を向上させるために，同様の推定計算を反復して行っている．図 3 のように各初期条件を中心とする楕円の交わりを考えた時，楕円同士の交差の状況によっては，ケプストラム空間において適切でない領域を解の存在領域と判定してしまう恐れがある．解の誤判定を回避するために，推定計算によって求まった平均ケプストラムベクトルを初期条件とし，再び同様の推定計算を行う [8]．この反復推定を重ねることで，誤って推定された解を適切な解へ修正してゆくことができる．

### 3.5 周波数ウォーピング

本研究を行うにあたり，構造提供者の音声に周波数ウォーピングをかける事によって，模擬的に異なる話者性の音声を作成した．話者の声道長の変化は，音声のスペクトル表現における周波数ウォーピングとして考えることができる．周波数ウォーピングにおける変換前後の正規化角周波数を $\omega$，$\hat{\omega}(0 \leqq \omega,\ \hat{\omega} \leqq \pi)$ とする．このとき，$z = e^{j\omega}$，$\hat{z} = e^{j\hat{\omega}}$ として，周波数ウォーピングとして式 (4) の 1 次全域通過関数を考える．

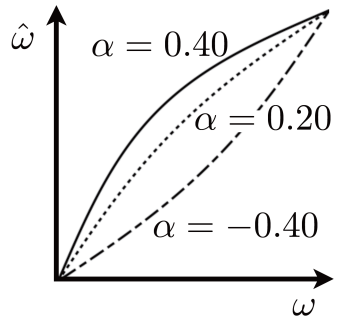



図 5　周波数ウォーピング関数 ($-0.4 < \alpha < 0.4$ の場合)
Fig. 5　Frequency warping function ($-0.4 < \alpha < 0.4$).

$$\hat{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \tag{4}$$

このとき $\alpha$ は $|\alpha| < 1$ の実数であり，$\alpha < 0$ の場合，周波数軸が低域に変換され声道長は長くなる．一方，$\alpha > 0$ の場合，周波数軸は広域に変換され，声道長が短くなる (図 5)．以降 $\alpha$ をウォーピングパラメータと呼ぶ．江森らは上記の周波数ウォーピングを元に，そのケプストラム空間における記述を導出している [9]．この時，ケプストラムベクトルに対して，周波数ウォーピングを施す線形変換は式 (5) で表現される．

$$c' = Ac,\quad A = \begin{pmatrix} 1 & \alpha & \alpha & \alpha \\ 0 & \alpha - \alpha^2 & 2\alpha - 2\alpha^3 & \dots \\ 0 & -\alpha + \alpha^3 & 1 - 4\alpha^2 + 3\alpha 4 & \dots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \tag{5}$$

また，ケプストラムベクトル $c$ の変換が行列 $A$ による線形変換の式 (5) で表される時，ケプストラムの平均 $\mu$ と分散 $\Sigma$ は式 (6) のような変換を受ける．

$$\begin{cases} \mu' = A\mu \\ \Sigma' = A^T \Sigma A \end{cases} \tag{6}$$

なお，式 (3) において，分散共分散行列を対角行列として計算しているのは，ケプストラム系列を分布系列化する際に，次元ごとに個別に平均と分散を計算しているためである．分散共分散行列が対角である場合，ブロックサイズが 2 以上の時であっても，平均ケプストラムの各成分は互いに計算上の影響を及ぼさない．一方，既に求まっている対角分散共分散行列に式 (6) のウォーピングをかけた場合，変換後の分散共分散行列 ($\Sigma'$) が非対角化される．

音声の構造的表象の枠組みにおいては，話者性の差異を変換行列 $A$ で近似的に表現している．任意の話者性に対応する音声合成系を考える上で，音声合成系のパラメータとして全角の分散共分散行列を取り扱うべきであるかを検討する必要がある．今回は，初期条件として与える音声にウォーピングにより話者性の変更を施し，それに伴う初期条件の分散共分散行列の全角化について，合成系へ与える影響を確認する．

## 4. 評 価 実 験

日本人 11 名（男性 8 名，女性 3 名）を対象に聴取実験を行った．

表 1　聴取実験の評価基準
Table 1　Judgement criteria for the listening experiments.

(a) 実験 1：合成音声の音韻性の評価基準

| | |
|---|---|
| 5. | ターゲット音声と遜色なく，発話内容もよくわかる |
| 4. | 劣化はわずかに認められるが，発話内容は十分にわかる |
| 3. | 劣化が気になるが，発話内容はわかる |
| 2. | 劣化が気になり，発話内容もわかりにくい |
| 1. | 劣化がひどく，発話内容がわからない |

(b) 実験 2：分散共分散行列の変換による変化の評価基準

※ A…平均・分散共分散行列ともに変換したもの，　B…平均のみ変換したもの

| | |
|---|---|
| 5. | A の方がわずかに自然性が高い |
| 4. | A の方がわずかに自然性が高い |
| 3. | 自然性の差は見られない |
| 2. | B の方がわずかに自然性が高い |
| 1. | B の方が自然性が高い |

### 4.1 合 成 条 件

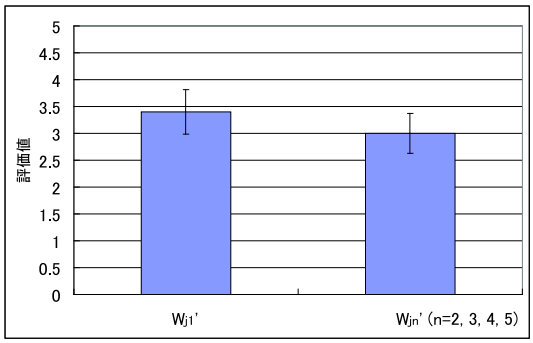
語学教師から得た日本語音声と英語音声から網羅的な構造的表象を抽出し，日本語母音の 25 個のケプストラム分布を初期条件として，3.1 の枠組みで英語音声の各音響事象の平均ケプストラムベクトルを推定した．今回の実験ではブロックサイズを 2，反復学習回数を 10 回と定めた．

構造提供者とする 1 名の女性話者 $F$ から，日本語 5 母音の連続発声音声/aiueo/および共鳴音のみからなる英文 “I owe you one.” の音声を各々，5 回ずつ収録した．これらはほぼ同一の発話スタイルで発声されたものである．以降，日本語の各発声を $W_{j1} \sim W_{j5}$，英語の各発声を $W_{e1} \sim W_{e5}$ と表記し区別する．英文音声の構造化は，その音素数が 9 であることから 45 状態で行った．

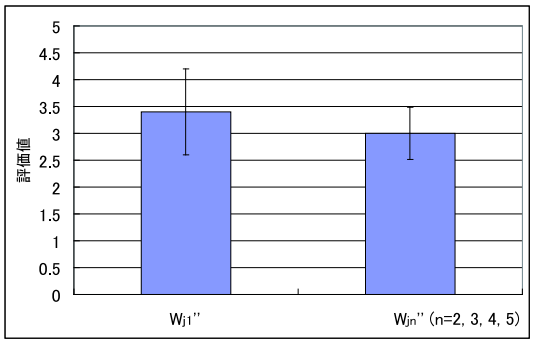
まず，$W_{j1}$ と $W_{e1}$ から，$W_{j1}$ と $W_{e1}$ を構成する音響事象を網羅する構造を抽出した．事象数は $25+45=70$ である．この構造を $[W_{j1}+W_{e1}]$ と記述する．構造的表象を抽出するために，収録音声に対して STRAIGHT [10] を用いたスペクトル分析を行い，40 次のケプストラムを得た．また，ケプストラム 0 次項 (パワー)，状態継続長，ピッチも STRAIGHT の分析を元に抽出した．構造的制約条件 $[W_{j1}+W_{e1}]$ に対して与える初期条件 (身体性) は，$W_{j1} \sim W_{j5}$ から抽出したケプストラム分布に対して式 (6) のウォーピングをかけたものを用いる．$\alpha = 0.14$ を用いた分布群を $W'_{j1} \sim W'_{j5}$，$\alpha = -0.14$ を用いた分布群を $W''_{j1} \sim W''_{j5}$ で表す．

$[W_{j1}+W_{e1}] \oplus W_{j1}$ を，構造 $[W_{j1}+W_{e1}]$ に対して構造推定時に用いた事象群 $W_{j1}$ を初期条件とした音声合成とする．この時，$[W_{j1}+W_{e1}] \oplus W'_{j1}$ は時間構造が同一の他話者の発声を初期条件として用いて合成することを意味し，$[W_{j1}+W_{e1}] \oplus W'_{jn}(n=2\ldots5)$ は時間構造の自然なゆらぎが混入した他話者の発声を用いて合成することを意味する．

なお合成時には，構造の抽出に用いた $W_{e1}$ から抽出したパワー，状態継続長をケプストラム系列に適用した．また $F_0$ についても同様に，$W_{e1}$ から得た $F_0$ 系列を適宜変更して適用した．




(a) “I owe you one.”($\alpha = 0.14$) の場合




(b) “I owe you one.”($\alpha = -0.14$) の場合

図 6　実験 1:合成音声の音韻性の評価結果
Fig. 6　Results of experiment 1.

### 4.2 実験 1：合成音声の音韻性評価

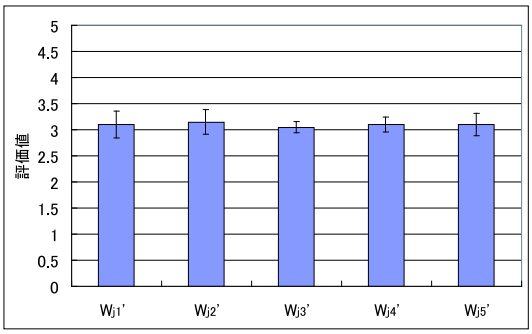
まず，ウォーピング音声を初期条件とした合成音声の音韻性を DMOS(Degradation Mean Opinion Score) によって評価した．被験者にはまず合成のターゲットとなる参照音声を提示し，続いて評価対象の音声を聞かせることで，参照音声に対して音韻性にどの程度の劣化が見られるかを表 1(a) の 5 段階で評価させた．ここで，合成音声 $[W_{j1}+W_{e1}] \oplus W'_{jn}$ に対する参照音声は $[W'_{jn}+W'_{e1}] \oplus W'_{jn}$ によって作成した．これは構造を介した分析再合成を行うことに相当する．

### 4.3 実験 2：分散共分散行列の影響の評価

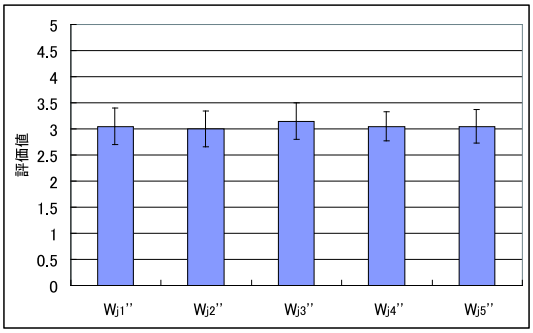
$[W_{jn}+W_{en}]$ と $[W'_{jn}+W'_{en}]$ は構造の不変性により一致する．しかし，$W_{jn}$，$W_{en}$ の音響事象の分散共分散行列を対角行列で推定した場合，式 (6) より $W'_{jn}$，$W'_{en}$ のそれは非対角行列となる．合成時の初期条件として与える事象の分散共分散行列として，非対角行列の対角成分のみを用いる場合と，全成分を用いる場合とが可能であり，この両者を比較する．即ち $[W'_{jn}+W'_{e1}] \oplus W'_{jn}$ を行う時に用いる初期条件 ($W'_{jn}$ の音響事象) として，1)$W_{jn}$ の事象群のケプストラム平均，分散共分散行列を共に式 (6) で修正する場合と，2) ケプストラム平均のみを修正する場合を比較する．後者の場合，分散共分散行列は対角のままである．評価基準は表 1(b) のように実験 1 と異なる 5 段階を設定した．

## 5. 結　　果

実験 1 の結果として評価値の平均と 95% 信頼区間を図 6 に示す．上図 (a) が $\alpha = 0.14$，下図 (b) が $\alpha = -0.14$ の場合で




(a) “I owe you one.”($\alpha = 0.14$) の場合




(b) “I owe you one.”($\alpha = -0.14$) の場合

図 7　実験 2:分散共分散行列の影響の評価結果
Fig. 7　Results of experiment 2.

ある．両図において，初期条件の時間構造に「ずれ」がない場合 ($W'_{j1}$，$W''_{j1}$)，他と比べて評価が高い．

また，$W'_{j1}$ の場合も含めて評価値が 3 前後に集中しており，音韻性を損なってはいないものの，音質については良好な結果が得られていない．これまでにも，構造提供者と初期条件提供者の話者性の差の度合いによって，合成音声に特有の音質が付与される傾向があることを確認している [11]．この問題の解決が今後の課題の一つとなる．

続いて，実験 2 の結果を図 7 に表す．全角の分散共分散行列を用いた場合に若干の聞こえの違いは生じたものの，自然性比較の評価値の平均は 3 前後に集中しており，全体を通して目立った音質の変化は見られなかった．

## 6. ま と め

日英二言語に渡る構造的表象に基づく音声変換の枠組みにおいて，初期条件として与える音声の時間構造と，ケプストラムの分散共分散行列に焦点を当て，それぞれが合成音声に与える影響を確認した．時間構造については，構造とのずれが比較的わずかな場合でも合成音声の音質に影響を与えることから，初期条件とする音声へ適切な時間合わせを行うことが重要であることを確認した．一方，初期条件として与えるケプストラムの分散共分散行列の全角化については，本研究の枠組みにおける合成音声への目立った影響は確認できなかった．ケプストラムの平均ベクトルは声道形状の特性に強く影響するが，分散については個人の身体性に依存して大きく変わるものではなく，話者性の再現のために分散共分散行列に処理を行う必要性は低いものと思われる．今後の課題としては，上記した時間合わせの問題，及び，本手法特有の音質劣化に対して，解決策を講じていきたい．

## 文　献

[1] 真下美紀子他, “混合正規分布モデルに基づく声質変換法の日英言語間への適用”, 秋季音講論, 1-P-17, pp.389-390, 2001.

[2] 峯松信明他, “線形・非線形変換不変の構造的表象とそれに基づく音声の音響モデリングに関する理論的考察”, 春季音講論, 1-P-12, pp.147-149, 2007.

[3] 齋藤大輔他, “構造的表象からの音声合成とそれに基づく音声模倣に関する研究”, 信学技報, SP2008-40, Vol.108, No.116, pp.115-120, 2008.

[4] M. Pitz and H. Ney, “Vocal tract normalization equals linear transformation in cepstral space”, IEEE Trans. Speech and Audio Processing, Vol.13, No.5, pp.930-944, 2005.

[5] 齋藤大輔他, “音声の構造的表象を入力した音声合成に対する基礎的検討”, 秋季音講論, 1-P-2, pp.399-402, 2007.

[6] S. Asakawa et al., “Multi-stream parameterization for structural speech recognition”, ICASSP’08, pp.4097-4100, 2008.

[7] D. Saito et al., “Optimal event search using a structural cost function - improvement of structure to speech conversion -”, INTERSPEECH’09, pp.2047-2050, 2009.

[8] D. Saito et al., “Improvement of Structure to Speech Conversion Using Iterative Optimization”, Proc. Speech and Computer (SPECOM), pp.174-179, 2009.

[9] 江森正, 篠田浩一, “音声認識のための高速最ゆう推定を用いた声道長正規化”, 電子情報通信学会論文誌 D-II, Vol.J83-D-II, No.11, pp.2108-2117, 2000.

[10] 河原英紀, 音響学会誌, “Vocoder のもう一つの可能性を探る –音声分析変換合成システム STRAIGHT の背景と展開– ”, Vol.63, No.8, pp.442-449,2007.

[11] 見原隆介他, “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”, 秋季音講論, 3-P-19, pp.403-406, 2009.