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.

構造評価関数を用いた構造的表象からの音声合成系の高精度化

齋藤　大輔†　　喬　　宇††　　峯松　信明††　　広瀬　啓吉††

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

あらまし　音声は年齢，性別，声道長や音響機器などの非言語的特徴によって変形し，多様性に富んでいる．筆者らはこれらの非言語的な音響変形におよそ不変な音声の構造的・抽象的表象を提案してきた．この表象は音声の動きのみに着眼した物理表象である．先行研究において，音声の構造的表象に基づく音声合成の枠組みを提案し，その基礎的検討を行ってきた．提案する枠組みでは音声発話を発話内容（語形）と発話者の身体性に分離して捉え，生成に際しては話者不変の語形に発話者の身体性を付与する事で合成音声を得る．これは，幼児の音声模倣に対応する音声合成のモデルといえる．本稿では提案する枠組みと幼児の音声模倣の対応について考察し，加えて構造評価関数とそれに基づく音響事象の推定法（音響空間における定位法）を導入する事で，従来手法における幾何学的アプローチと比べて，技術的な改善を試みた．連続音声を対象とした音声合成実験を行い，主観評価実験の結果から，提案手法において高次の特徴量分割手法を導入した場合における品質の向上を確認した．

キーワード　構造的表象, 話者不変性, 音声模倣, 言語獲得, 構造評価関数

Improvement of structure to speech conversion based on a structural cost function

Daisuke SAITO†, Yu QIAO††, Nobuaki MINEMATSU††, and Keikichi HIROSE††

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

Abstract　Speech acoustics vary due to differences in age, gender, vocal tract length, microphone, and so on. The authors recently proposed a structural and abstract representation of speech, where these variations were effectively removed. This representation captures only dynamics of speech. In our previous study, using this abstract representation, a new framework of speech synthesis was proposed and some fundamental investigations were carried out. In this new framework, an utterance is modeled using two separate attributes; one corresponding to what is known as speech Gestalt, which is a speaker-invariant speech form, and the other to the embodiment seen in vocal tubes, which characterizes speaker differences. Acoustic signals are generated by using the Gestalt as constraint conditions and the vocal tube embodiment as initial conditions. In other words, the Gestalt can be acoustically realized only when the speaker's embodiment is provided. This new framework can be regarded as an implementation of infants' vocal imitation. In this study, by following the initial investigations, we improve accuracy and efficiency in acoustic realization of the Gestalt based on a structural cost function. Experiments of generating continuous utterances of Japanese vowels show the validity of the proposed method.

Key words　structural representation, speaker invariance, vocal imitation, language acquisition, a structural cost function

1. はじめに

近年の音声合成システムは，テキスト列を入力，音響信号を出力とする Text-to-Speech 変換 (TTS) が主流となっている．TTS では音韻列を音声の表象として考え，その上で漢字仮名混じり文と音韻列との対応関係，および音韻と音響信号との対応関係を統計的手法により学習する．このとき構築されるモデルは多くの場合，異音の音響モデル (triphone)，すなわち声そのもののモデルである．構築されるモデルには話者性が不可避的に内在する事になる．

一方，幼児の言語獲得のプロセスは音声模倣 (vocal imitation) が基盤となるが，上記のように声そのものを模倣して言語を獲得する訳ではない．幼児が両親の音声の音響的実体そのものを模倣する事は声道形状の差異から不可能である．父親の「おはよう」と母親の「おはよう」を真似しても同じ本人の「おはよう」となるように，幼児は何らかの抽象化を通して音声模倣を行っていると考えられる．ここで [おはよう] という音響信号を /おはよう/ という話者不変の音韻列に変換し，それを自らの口で音声化しているとの議論も可能であるが，発達心理学はこれを否定する [1]．そもそも幼児は音韻的意識が希薄であり，語から個々の音韻を抽出する能力が完成するのは小学校入学前後といわれている [2]．すなわち音声の抽象化として音韻列を考えるの不適切であると考えられる．

幼児の音声模倣を説明しうる音声合成を考えた場合，幼児が音声模倣に際して参照している話者不変の抽象的表象を物理的，音響的に求める必要がある．発達心理学は「幼児は単語全体の語形・音形（語ゲシュタルト [3], [4]）を獲得し，その後，個々の分節音を獲得する」と主張する [5]． 筆者らはこれまでこの「語ゲシュタルト」の音響的定義と考えられる，話者に不変な音声の構造的かつ抽象的表象を提案してきた [6]．これは音声の音響的実体そのものは直接用いず，実体間の関係性のみをモデル化することで，非言語性変形に対して不変性を有する音声表象である．筆者らは既にこの話者不変の音声表象を用いた音声認識システムについて検討を行ってきた [7]〜[9]．加えてこの表象に基づく音声合成についても，初期検討を行ってきた [10], [11]．

構造的表象からの音声合成の定式化として，構造表象を制約条件とし，身体性に対応するいくつかの初期条件を与える事で，音響空間の探索問題を考えてきた [11]．しかしこれまで時系列として並ぶ個々の音響事象を独立に探索していたため，全体として構造の制約条件を満たしていなかったと考えられる．また推定された音響事象間の関係も考慮されていなかった．加えて，解析的に探索問題を解く計算量が膨大であるという問題点もあった．そこで本稿では，構造の制約条件を適切に表現する構造評価関数とそれに基づく音響事象の導出を提案する．さらに反復的な推定法を導入する事で従来の実装よりも高精度な合成を試みた．

2. 音声の構造的表象

2.1 非言語的特徴による音響的実体の変形

音声の音響的実体は非言語的特徴による変形を不可避に伴う

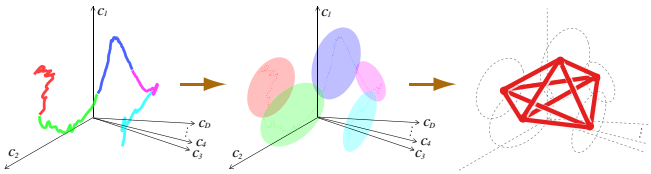


図 1 音声の動きのみを捉えた音声表象

Fig. 1 Speech representation by capturing only speech dynamics.

が，これらは大きく乗算性変形と線形変換性変形に分けられる．

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る．ケプストラム空間では，この種の変形は加算演算 $c' = c + b$ として表現される．マイクロフォンの音響特性がその典型例である．また話者の声道形状差異も一部近似的に乗算性変形であると考えられる．音声は必ず発話者を伴い，音響機器によって収録されるため，これらの変形は不可避である．

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 $c' = Ac$ で表現される変形である．スペクトル表現においては，話者の声道長差異や聴取者の聴覚特性差異は周波数ウォーピングとして考えられる．周波数ウォーピングはケプストラム空間において線形変換で記述されることが示されている [12]．すなわち声道長差異や聴覚特性差異は近似的に線形変換性変形として扱うことができる．

以上をまとめると，非言語的特徴による音声の音響的実体の変形に対する最も簡素なモデルは，ケプストラム空間におけるアフィン変換 $c' = Ac + b$ で得られる．これらの A, b が話者や収録環境によって多様に変化し，音声の音響的実体が様々に変形する事になる．

2.2 音声の構造的表象

話者の違いに不変な特徴を導入するには，上述の非言語的特徴による変形によって，影響を受けない特徴が必要となってくる．声質変換の研究においては，異なる話者間の変形を表現する空間写像を推定する事で，音声の話者性を変化させている．このような写像に対する不変性をもつ特徴を用いて発話を表現する事で，話者不変の音声表象を得る事ができる．ここで，式 (1) で表される確率分布間の距離を考える．

$$f_{div}(p_i, p_j) = \oint p_j(x) g\left(\frac{p_i(x)}{p_j(x)}\right) dx \qquad (1)$$

式 (1) は f ダイバージェンスと呼ばれ，二つの分布に可逆かつ連続な共通の変換が施された場合でもその値は変化しない(注1) [13]．話者変換は空間写像と考えられるため，式 (1) を用いることで，話者不変な発話の表現を導出できる．

今，音響空間において一発声が N 個の音響事象で表されているとする．空間において N 角形の形状は ${}_NC_2$ 個の全ての頂点間距離を規定する事で一意に定めることができる．すなわち事象群に対して，全ての事象間距離を求めることでその事象群を構造的に表象することになる．この事象間距離として f ダイバージェンスを用いる．

(注1)：このとき，写像に線形性は要求されない．非線形写像であっても同様の性質が成立する．

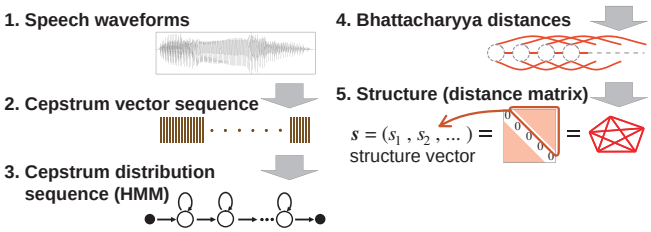


図 2　音声からの構造的表象の抽出

Fig. 2　Structure extraction from one utterance.

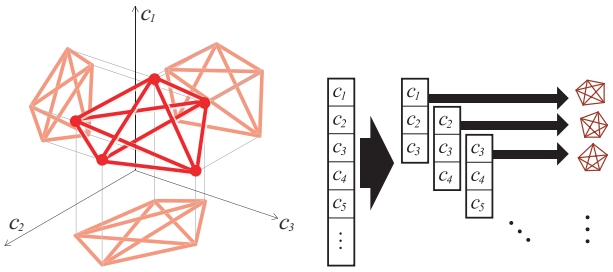


図 3　特徴量空間分割によるマルチストリーム化

Fig. 3　Multiple stream structuralization by parameter division.

すなわちケプストラム空間において音響事象を分布として捉え，音響事象群を f ダイバージェンスによる「**分布間距離**」のみによって定義することで，変換不変，すなわち非言語性変形におよそ不変な構造を求める事ができる．この表象による音声表現は図 1 のように，音響空間における音声の動きを距離関係のみで記述しているといえる．

2.3　一発声の構造化

一発声を構造的表象で記述する場合を考える．図 2 に一発声の音声からの構造的表象の抽出の流れを示す．音声の時系列信号は，まず短時間スペクトル系列からケプストラム系列へと変換される．得られたケプストラム系列もまた時系列信号であるが，これを適当な時間区間において音響事象の分布としてとらえ，その分布の時系列へと変換する（このとき各分布に対応する時間長は分布によって異なる）．これら系列中の各分布に対して全ての組み合わせの分布間距離を求めることで一発声が構造化される．

2.4　特徴量空間分割

構造の不変性は変換の線形・非線形を問わず成立する，構造的表象に基づく音声処理はこの不変性に依って話者性を効果的に取り除く枠組みであるが，一方で不変性によって単語間の差異を表すような変換に対しても不変となり，単語識別力の低下を招く可能性がある．そのため不変性を抑制する必要がある．ここでは，線形変換性歪みを表す行列 A について，その帯行列性に着目する [8]．ケプストラムベクトルに対して，連続する w 個の要素から構成される w 次元ベクトルを構成し，これを低次から要素をずらす事で複数の特徴量ストリームを構成する．そして，構造不変性が各部分空間においても成立すると仮定し，各ストリームから構造を抽出する．これにより過剰な不変性を抑制する事で単語識別力を向上させることができる [8]．以降 w をブロックサイズ（BS）と呼ぶ．図 3 は特徴量空間分割の一例を示している．図 3 の左図のように射影された部分空間での構造も制約条件として含めることで，構造の回転に制約を与えることになる．

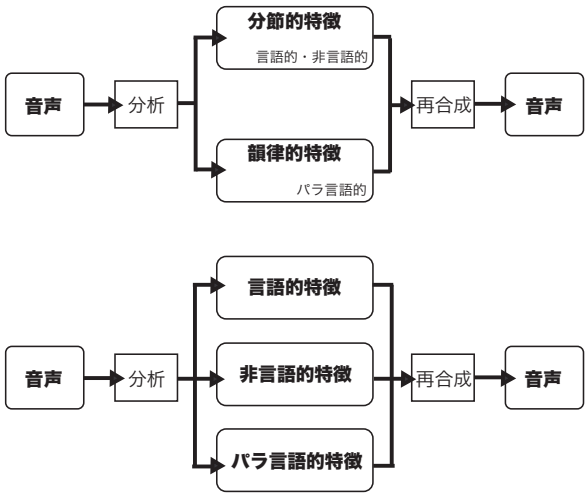


図 4　従来の分析再合成系の枠組み（上）と提案する枠組み (下)

Fig. 4　The conventional framework for analysis-resynthesis and the proposed one with three separate kinds of features.

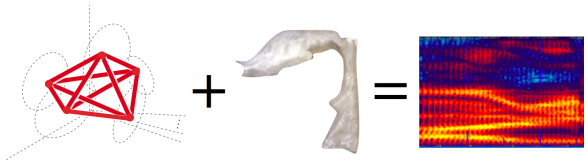


図 5　構造と身体特性によって音響信号が得られる枠組み

Fig. 5　Structure + vocal tube = speech sounds.

3.　構造的表象からの音声合成

3.1　非言語的要因をも分離する分析再合成系

本稿で提案する音声合成の枠組みは，生成対象の語形に対して，発声者の身体性（声道形状特性）を与える・戻すことで初めて音が生まれるという合成系である [14]．分析再合成系でこの考えを示すと図 4 のようになる．従来の分析再合成系では音声を分節的特徴（主にスペクトル包絡に対応し，言語情報・非言語情報を伝搬）と韻律的特徴（主にピッチ，パワー，継続長に対応し，パラ言語情報を伝搬）に分解する．一方提案する枠組みはこれをさらに細分化し，言語的特徴，非言語的特徴，パラ言語的特徴に分ける枠組みである．この時，言語的特徴とパラ言語的特徴を与えても音は生成されない．生成する話者は非言語的特徴の担い手であるからである．この担い手の音響特性（具体的には声道形状），更には伝送媒体のチャネル特性が与えられて初めて，聞き手が聴取できる音響信号が生まれる．このことは幼児が両親の発話全体の語形を獲得し，自らの発声器官を使って言葉を発する過程をモデル化したものといえる．これを模式的に示したものが図 5 である．

3.2　ケプストラム空間の解探索

声道形状のパラメータとして調音器官の制御パラメータが考えられる．しかし調音パラメータは複雑であり，ケプストラム

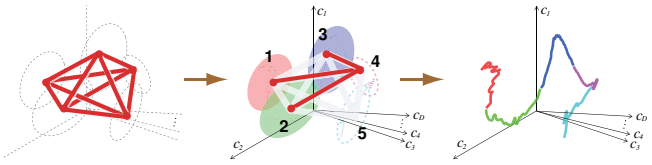


図 6　解探索による構造的表象を制約とする音声合成の枠組み
Fig. 6　Search for the next target under structural constraints.

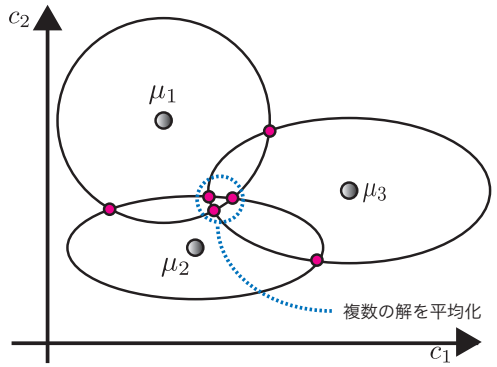


図 7　幾何学的アプローチによる解の導出（2 次元の場合）
Fig. 7　Solution of the searching problem.

空間との対応関係も明確でない [15]．これまで筆者らはケプストラム空間の解探索問題として定式化を行ってきた [10], [11]．即ち構造的表象による制約条件に対して，既に生成された音響事象を初期条件として与えることで，次の時刻の音響事象をケプストラム空間から探索することで求める．

この枠組みを図 6 に示す．図 6 では，3 個の音響事象を初期条件として与え，構造的表象による制約条件から残りの音響事象を探索する．これは図 1 の逆過程と考える事が出来る．また音声模倣の枠組みでいえば，構造的表象による制約条件が両親の発話から得られ，初期条件が幼児から得られることになる．

3.3　幾何学的アプローチによる解の導出

本節では，音響空間の探索問題に対するこれまでのアプローチについて述べる [11]．二つの音響事象がガウス分布 $\mathcal{P}_1 = \mathcal{N}(\mu_1, \Sigma_1), \mathcal{P}_2 = \mathcal{N}(\mu_2, \Sigma_2)$ の場合，f ダイバージェンスの一種である，バタチャリヤ距離（BD）は以下のように定式化される．

$$BD(\mathcal{P}_1, \mathcal{P}_2) = \frac{1}{8}\mu_{12}^T V_{12}^{-1}\mu_{12} + \frac{1}{2}\ln\frac{|V_{12}|}{|\Sigma_1|^{\frac{1}{2}}|\Sigma_2|^{\frac{1}{2}}} \qquad (2)$$

ただし $\mu_{12} = \mu_1 - \mu_2, V_{12} = \frac{\Sigma_1+\Sigma_2}{2}$ である．今，D 次元のケプストラム空間を考える．式 (2) において，μ_2 と Σ_1 および Σ_2 が既知であると仮定する．このとき式 (2) を満たす μ_1 をケプストラム空間の適切な位置に配置する．このとき，μ_1 の描く軌跡は D 次元空間における楕円体を描く．同様に別の音響事象 $\mathcal{P}_i$ との距離によって得られる制約条件 $BD(\mathcal{P}_1, \mathcal{P}_i)$ によって μ_1 は異なる楕円体を描く．よって初期条件として幾つかの音響事象が与えられた上で，構造的表象を制約条件として音響事象 $\mathcal{P}_1$ を空間内に定位する解探索問題は，幾何学的にはこれらの楕円体の交点を解とすることで解く事ができる．すなわち D 次元ベクトルに対する連立方程式を解く事によって導出できる．ただし式 (2) は二次形式のため，その連立方程式は一般に複数の解を持つ．よってこの曖昧性を解消するため，複数の連立方程式から得られた解を統合する事によって，求める音響事象を導出する．図 7 は $D = 2$ における，解の導出の様子を示している．図中の 3 つの楕円中心 (μ_1〜μ_3) は，初期条件として与えた音響事象となる．得られた楕円の交点として最も縮退している箇所について平均する事で，求める音響事象を得る事ができる．

3.4　構造評価関数

前節では，先行研究における幾何学的アプローチによる解の導出について述べた．しかし上述の定式化には 2 つの問題がある．一つは連立方程式の解法についてである．今，D 次元空間において m 個の初期条件，およびそこから得られる m 個の構造制約条件から音響事象を推定する場合を考える．$D > m$ の場合，得られる連立方程式は不良設定問題となり解を得る事ができない．一方 $D < m$ の場合，得られる連立方程式群は ${}_mC_D$ 個となり，特に D が高次元の場合に計算量が非常に大きくなってしまう．加えて，複数の解候補から解を統合する際，最適解が保証されない．

二つ目の問題として，個々の音響事象を独立に推定していることが挙げられる．すなわちこれまでの実装においては，複数の音響事象を推定する場合，推定対象と初期条件の制約のみが俠侶され，推定された音響事象間の関係性が考慮されていない．これにより真に構造の制約条件を満たす解が推定されていないと考えられる．また推定事象が隣接する際に連続性が考慮されておらず，自然性が劣化する可能性もある．

これらの問題に対処するため，本稿では構造評価関数とそれに基づく音響事象推定を提案する．さらに推定事象間の関係性を考慮するための反復解法による高精度化を導入する．今，全ての分散共分散行列が既知と仮定する．この時，平均ベクトル μ を m 個の初期条件と関連する構造制約から導出することを考える．この際，以下の式で表される評価関数を導入する．

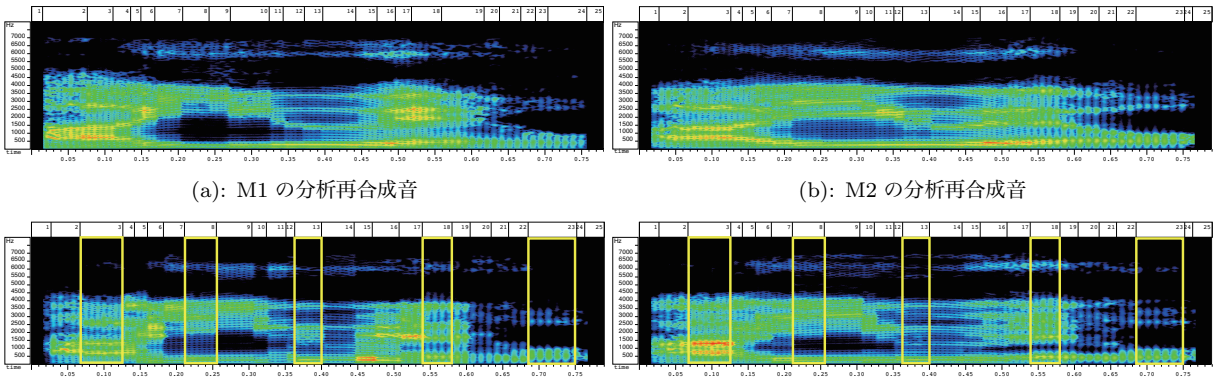
$$J(\boldsymbol{\mu}) = \sum_{i=1}^{m}\{bd(\boldsymbol{\mu}, \boldsymbol{c_i}) - BD_i\}^2 \qquad (3)$$

ここで BD_i は音響事象 i と推定対象の音響事象との間の構造制約条件となる距離を表す．一方 $bd(\boldsymbol{\mu}, \boldsymbol{c_i})$ は音響事象 i と現在推定されている μ によって計算される距離となる．上記の評価関数は目標となる構造との差を定量的に表しており，本稿では構造評価関数と呼ぶ．これを最小化する μ を得るためには以下の更新式を用いて，$\Delta\mu$ が十分小さくなるまで更新する．

$$\left(\frac{\partial^2 J}{\partial \mu^2}\right)\Delta\mu = \left.\frac{\partial J}{\partial \mu}\right|_{\mu_n} \qquad (4)$$

$$\mu_{n+1} = \mu_n - \Delta\mu \qquad (5)$$

一方，不連続性に対する対応として，反復解法による推定を導入する．基本的な考えは推定された音響事象を初期条件として再推定を行うというものである．今 n 個の推定対象と m 個の初期条件があると考える．最初，n 個の推定対象を独立に推定する．次に，これら n 個の音響事象から一つを選び，残りの



(a): M1 の分析再合成音　　(b): M2 の分析再合成音

(c): 先行研究 [11] の合成結果（構造制約:M1, 初期条件:M2）　　(d): 提案手法による合成結果

図 8　合成音声のスペクトログラム; (a)(b): M1, M2 の分析再合成音; (c)(d): 先行研究および提案手法による合成結果

Fig. 8　Spectrograms of resynthesized speech (a)(b) and synthesized speech (c)(d).

$n-1$ を初期条件として再推定を行う．これを n 個全てに対して行う．最後は $n+m$ 個の音響事象を等価に扱い，同様に再推定を繰り返す．

4. 合成実験

4.1 実験方法

提案手法の有効性について調べるため，日本語 5 母音の連続発声を用いて実験を行った．これは各母音が一度ずつ出現するもので，/aiueo/，/ieuao/ など 5!=120 単語が存在する．成人男女各 3 名（それぞれ話者 M1，M2，M3，F1，F2，F3 とする）についてこれらの発声を収録した．これらの発声について STRAIGHT [16] に基づくスペクトル分析を行い，このスペクトルから 40 次のケプストラムを得た．同時に発声のピッチ，パワー，継続長も STRAIGHT の分析を基に得た．

話者 M1 および F1 の発声について，図 2 の流れに基づいてこれらのパラメータから構造を抽出した．ケプストラム系列から分布系列への変換に際しては，一発声に対して MAP 推定に基づく HMM を用いる [8]．この時，一発声を 25 個の分布系列へと変換し，${}_{25}C_2 = 300$ 個の距離情報から不変構造を抽出した．さらに構造的表象に基づく音声認識で提案されている特徴量分割も導入した [8]．この際，部分空間の次元数（ブロックサイズ）を 1 から 5 まで変化させ，各々の部分空間で構造を抽出した．この構造が模倣対象となる単語の語形となり．探索問題は個々の部分空間で個別に計算される．

一方，話者 M2，M3 および F2，F3 についても図 2 の流れで分析を行った．ただし構造表象の抽出は行わず，それぞれの発声を 25 個の分布系列へと変換する．M1 および F1 から得られた不変構造を制約条件，M2，M3，F2，F3 のそれぞれについて 5 つの分布（3，8，13，18，23 番目）を初期条件として用いて，残りの音響事象分布を，提案する枠組みで推定した．反復解法の反復回数は 2 回とした．その結果得られた 25 個の分布と，ピッチ，パワー，継続長の情報を用いて STRAIGHT の枠組みで音声を再合成した．音声模倣の枠組みで言えば，本実験における話者 M1 および F1 は両親，残りの話者は幼児として考えられ，本実験はこれらの話者で音声模倣を実装していることに相当する．

4.2 実験結果

実験によって得られた合成音声の一例を図 8 に示す．図中 (a) は構造抽出に用いた話者 M1 の発声の分析再合成音，(b) は初期条件を提供した話者 M2 の発声の分析再合成音である．(c) は従来の幾何学的アプローチによって M1 の構造および M2 の初期条件から合成した音声である [11]．単語は/aiueo/ であり，ブロックサイズは 1 とした．一方 (d) は，提案手法による合成音である．ブロックサイズは 4 とした．(c)(d) のうち枠囲いされた部分が初期条件として与えた分布に対応する．これらを比較すると構造的表象に基づく合成音声 (c)(d) が (b) に近いスペクトルを導出できている．実際に聴取してみるとこれらの発話内容は /aiueo/ と容易に知覚する事ができる．さらに (c) と (d) を比較すると，(c) にはスペクトルの不連続が見られるが，(d) は連続的なスペクトル変化をしていることがわかる．

なお従来では，ブロックサイズが大きくなると現実的な時間での合成が困難であったのに対し [11]，提案手法はブロックサイズが 4 の場合でも問題なく合成する事が可能となっている．

5. 主観評価実験

5.1 実験方法

合成実験で得られた音声について，主観評価実験を行った．提案手法において特に特徴量分割における部分空間の次元数に着眼して，合成音声の自然性を評価した．特徴量分割によって音声認識における単語識別力が向上する一方で，話者不変性が抑制される問題が報告されている [17]．本研究における構造表象に基づく音声合成においてこの問題が合成音声の品質に与える影響を確認することを目的とする．

聴取実験は 11 人の被験者が参加した．それぞれの被験者は提案手法および従来手法で合成された/aiueo/ の発声を評価した．評価はブロックサイズの異なる 2 つのサンプルを用いた一対比較で行った．2 つのサンプルの構造制約条件話者（M1, F1）および初期条件話者（M2, M3, F2, F3）は共通とした（計 8 通り）．ブロックサイズとしては従来手法の $BS = 1, 2$ および提案手法の $BS = 1 \ldots 5$ の 7 種類（${}_7C_2 = 21$ ペア）について評

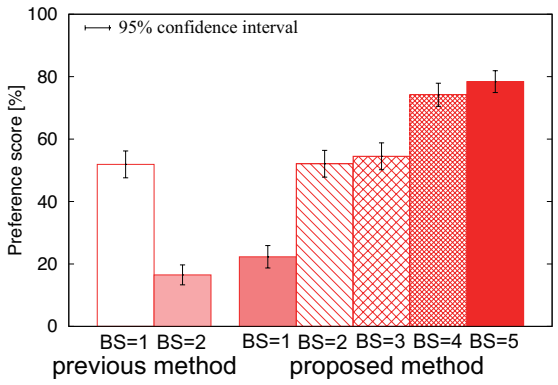


図 9　主観評価実験の結果

Fig. 9　Results of subjective evaluation.

価した．最終的に各被験者は $8 \times 21 = 168$ ペアを比較した．

5.2 実験結果

上記聴取実験の結果を図 9 に示す．図 9 は各手法のプリファレンススコアを示したものである．従来手法ではブロックサイズの増加に伴って合成音の品質が劣化していることがわかる．一方で提案手法ではブロックサイズが大きくなるに従って，大幅な品質の向上が見られた．特に $BS=4$ および $BS=5$ の場合は従来手法を大きく上回っていることがわかる．この結果は，ブロックサイズが大きい場合に話者不変性がよく保持され，構造評価関数による適切な制約条件のもとで，最適な音響事象が探索されたためと考えられる．一方，従来手法において，ブロックサイズの増加により品質が劣化するのは，連立方程式による求解が高次になればなるほど困難になるためと考えられる．

6. おわりに

本稿では，非言語的特徴による音声の変形に対しておよそ不変な物理表象である音声の構造的表象とそれに基づく音声合成の枠組みについて述べた．提案する枠組みでは，音声発話をその発話内容と発話者の身体特性に分離して捉える．音響信号の生成の際には，発話内容に対して，明示的に発話者の身体特性を与える事で音声合成を実現する．提案手法は，幼児の音声模倣のモデルとして捉える事ができる．先行研究において，音響空間の探索問題として提案する枠組みを定式化し，幾何学的アプローチによる解法を導入した．一方本稿では，先行研究の定式化の問題点への対応として，構造評価関数とそれに基づく音響事象の推定手法を提案した．さらに反復解法による推定を導入する事でより高精度な音声生成を可能とした．主観評価実験によって従来手法では見られなかった高次の特徴量分割による品質の向上を確認した．

今後は，幼児の言語獲得に重要な役割を果たしている韻律的特徴について [3]，提案する枠組みの中で取り扱っていく．さらに提案する枠組みにおける子音を含めた音声合成や本手法に適した話者性のモデリングについて検討を行っていく予定である．

7. 謝　辞

本研究は，東京大学グローバル COE「セキュアライフ・エレクトロニクス」の若手研究ファンドの支援を受けて行われた(注2)．

文　献

[1] 内田伸子編, “発達心理学キーワード”, 有斐閣双書, 2006.
[2] 天野清: “子どものかな文字の習得過程”, 秋山書店, 1986.
[3] 早川勝廣: “言語獲得と育児語”, 月刊言語, vol.35, no.9, pp. 62–67,2006.
[4] N. S. トルベツコイ, “音韻論の原理”, 岩波書店, 1958.
[5] 加藤正子: “特集にあたって”, コミュニケーション障害学, vol. 20, no. 2, pp. 84–85, 2003.
[6] N. Minematsu *et al.*: “Theorem of the invariant structure and its derivation of speech Gestalt,” SRIV2006, pp. 47–52, 2006.
[7] Y. Qiao *et al.*:“Random discriminant structure analysis for continous Japanese vowel recognition,” ASRU2007, pp. 576–581, 2007.
[8] S. Asakawa *et al.*: “Multi-stream parameterization for structural speech recognition,” ICASSP’08, pp. 4097–4100, 2008.
[9] N. Minematsu *et al.*: “Implementation of robust speech recognition by simulating infants’ speech perception based on the invariant sound shape embedded in utterances,” Proc. Speech and Computer (SPECOM), pp. 35–40, 2009.
[10] 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検討”, 信学技報, SP2008-40, pp. 115–120, 2008.
[11] D. Saito *et al.*: “Structure to speech conversion –speech generation based on infant-like vocal imitation,” Proc. INTERSPEECH, pp. 1837–1840, 2008.
[12] M. Pitz and H. Ney: “Vocal tract normalization equals linear transformation in cepstral space,” IEEE Trans. Speech and Audio Processing, vol. 13, pp. 930–944, 2005.
[13] Y. Qiao *et al.*, “f-divergence is a generalized invariant measure between distributions,” Proc. INTERSPEECH, pp. 1349–1352, 2008.
[14] 峯松信明他: “孤立音 [あ] を聞いて/あ/と同定する能力は音声言語に必要か？”, 信学技報, SP2007-30, pp. 37–42, 2007.
[15] 錦戸信和他: “GMM を用いた通常発話状態と特異発話状態の弁別”, 日本音響学会春季講演論文集, 1-Q-28, pp. 319-320, 2007.
[16] H. Kawahara *et al.*:“Restructuring speech representations using a pitch-adaptive time-requency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication, vol. 27, pp. 187–207, 1999.
[17] 松浦良他: “構造表象を用いた音声認識におけるパラメータ共有とその効果”, 信学技報, SP2008-52, pp. 55–60, 2008.

(注2)：http://www.ee.t.u-tokyo.ac.jp/gcoe