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

齋藤　大輔† 　朝川　智†† 　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
†† 東京大学大学院新領域創成科学研究科 〒277–8561 千葉県柏市柏の葉 5–1–5
††† 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,asakawa,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．筆者らはこれまでに，これらの非言語性の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，生成に際しては語形に発話者の身体性を付与する事で音声合成を実現する．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，加えて解析的手法を導入する事で，初期検討における音響空間の全探索と比べて，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，少ない初期条件によって合成対象の話者性を持った音声が得られることを確認した．

キーワード　構造的表象, 話者不変, 音声模倣, 言語獲得, 探索問題

Proposal of structure-to-speech conversion and its application to implementation of infants' vocal imitation

Daisuke SAITO†, Satoshi ASAKAWA††, 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 Frontier Sciences, The University of Tokyo
5–1–5 Kashiwano-ha, Kashiwa-shi, Chiba 277–8561, 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,asakawa,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 by two separate attributes; one corresponding to what is known as speech Gestalt, which is speaker-invariant, 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 considered. 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 by using an analytical method. Experiments of generating continuous utterances of Japanese vowels show the validity of the proposed method.

Key words　structural representation, speaker invariant, vocal imitation, language acquisition, searching problem

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) と呼ばれるが，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，各音韻を獲得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]．すなわち前述した音声の抽象化と音韻による音声表象は独立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], [8]．加えてこの表象に基づく音声合成についても，音響空間の探索問題としての定式化を通してその初期検討を行ってきた [9], [10]．

本稿では，幼児の音声模倣との対応関係に言及しながら，構造的表象に基づく音声合成について述べる．この音声合成の枠組みは，ある音声発話をその発話内容と発話者の身体性の二つに分離して捉えている．一方音響信号の生成に際しては話者不変の発話内容に発話者の身体性を付与する事で音声合成を実現する．提案する音声合成システムは幼児の音声模倣のモデルとして解釈可能である．加えて本研究では，解析的手法を導入することで，初期検討における音響空間の全探索と比べて精度面，速度面での向上を試みた．

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．

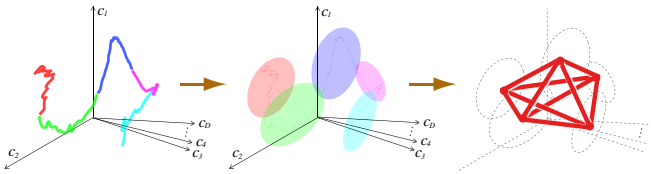


図 1 音声の動きのみを捉えた音声表象
Fig. 1 Speech representation by capturing only speech dynamics.

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 $c' = A\,c$ 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 [11]．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 $c' = A\,c + b$ 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

ユークリッド空間において 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．しかしケプストラム空間において N 点の「**点間距離**」によって構造を規定した場合，その構造は非言語的特徴によって不可避に歪む．なぜなら，非言語的特徴はケプストラム空間におけるアフィン変換としてモデル化され，アフィン変換は特殊な場合を除けば，構造を歪ませる変換である為である．しかしこの不可避に歪む構造は空間自体を歪ませる事で不変構造として定義することができる．

「**分布間距離**」の一つである Bhattacharyya 距離 (以下 BD と記述) を考えた場合，任意の二つの分布の確率密度関数を $p_1(x), p_2(x)$ として以下で表される．

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

二つの分布に対して共通のアフィン変換 $Ac + b$ を施した場合，BD は変換前後で不変となる．なおこの不変性は非線形変換においても成立する [12]．

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「**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．

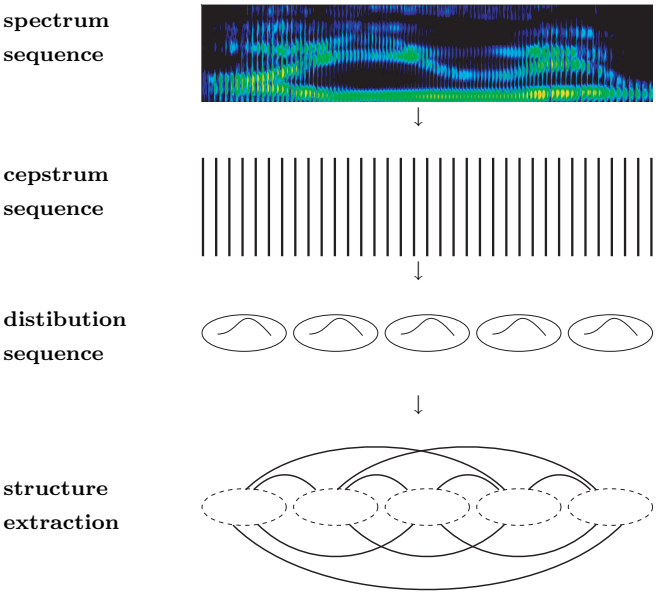


図 2　音声からの構造的表象の抽出
Fig. 2　Structure extraction from one utterance.

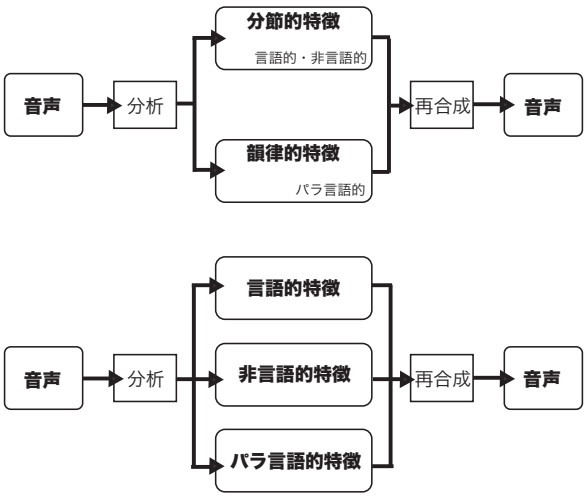


図 3　従来の分析再合成系の枠組み（上）と提案する枠組み (下)
Fig. 3　The conventional framework for analysis-resynthesis and the proposed one with three separate kinds of features.

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．

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 [13]．分析再合成系でこの考えを示すと図 3 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 4 である．

図 4　構造と身体特性によって音響信号が得られる枠組み
Fig. 4　Structure + vocal tube = speech sounds.

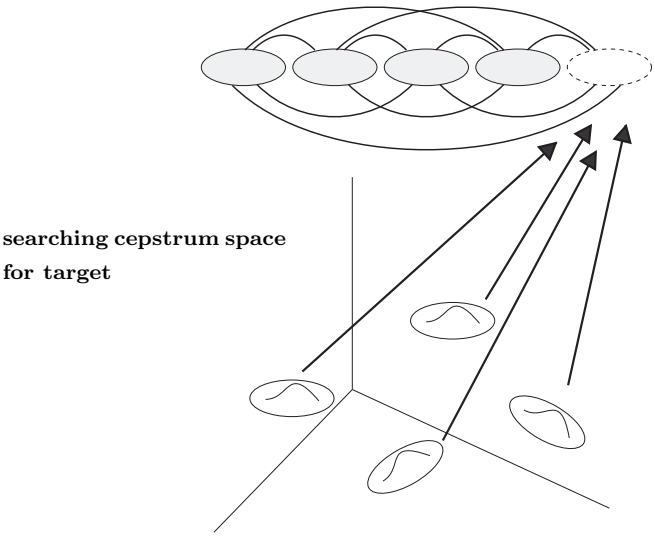


図 5　解探索による構造的表象を制約とする音声合成の枠組み
Fig. 5　Search for the next target under structural constraints.

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，ケプストラム空間との対応関係も明確でない [14]．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 [9], [10]．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 5 に示す．この際，構造に基づく音声認識で検討されている次元分割の手法 [8] や対象話者の音響空間をガ

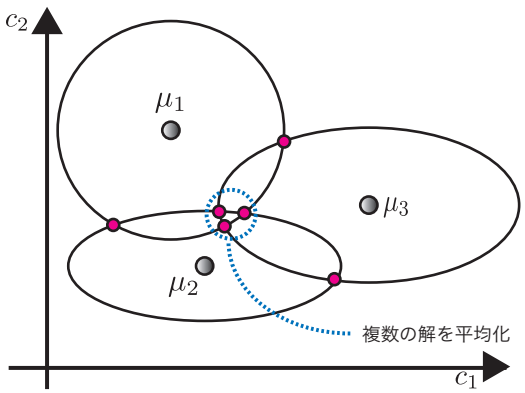


図 6　解析的手法に基づく解の導出（2 次元の場合）
Fig. 6　Solution of the searching problem.

ウス分布で近似し，その分散項で探索空間を規定する事で余剰な解探索を制限していた．しかし各次元の解像度と計算量がトレードオフの関係となり十分な品質が得られない問題があった．

3.3　解析的手法に基づく解候補の導出

本研究では解析的手法に基づき解候補を予め導出することで構造的表象を満たす音響事象を高速に推定し，かつ品質を向上させる手法について検討する．

二つの音響事象がガウス分布 $\mathcal{N}(\mu_1,\Sigma_1),\mathcal{N}(\mu_2,\Sigma_2)$ の場合，式 (1) は以下のようになる．

$$BD(p_1,p_2)=\frac{1}{8}\mu_{12}^{T}V_{12}^{-1}\mu_{12}+\frac{1}{2}\ln\frac{|V_{12}|}{|\Sigma_1|^{\frac{1}{2}}|\Sigma_2|^{\frac{1}{2}}} \qquad (2)$$

ただし $\mu_{12}=\mu_1-\mu_2, V_{12}=\frac{\Sigma_1+\Sigma_2}{2}$ である．今，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 の描く軌跡は多次元空間における楕円体を描く．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 p_1 を空間内に定位する解探索問題は，幾何学的には以下のような手続きで考える事ができる．

（1）　制約条件となる構造的表象に基づいて，式 (2) のような方程式群を導出する．

（2）　初期条件となる音響事象を得られた方程式群に代入する．

（3）　得られた式によって，求める音響事象がとりうる軌跡を描く．

（4）　得られた楕円の交点を求める事で，構造的表象を制約とする解探索問題を解く事ができる．

今 2 次元の場合に，初期条件の音響事象 $A=\mathcal{N}(\mu_a,\Sigma_a), B=\mathcal{N}(\mu_b,\Sigma_a)$ から音響事象 p の平均 $\mu=(\mu_p^x,\mu_p^y)$ を求めたいとする．ただ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 と表記する．p の分散項も既知とする．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す．式変形により μ に対する以下の連立方程式が定まる．

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

これは 2 次元において楕円の交点を求めることに相当する．この時二つの楕円の交点は一般には 2 つ，長軸および短軸の配置により最大で 4 つ求まる．そのため一つの音響事象を求める為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 次元における一つの音響事象の定位には少なくとも $n+1$ 個の音響事象が必要となる．2 次元の場合における提案手法の枠組みを図 6 に示す．図中の楕円の中心は，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．

4.　合 成 実 験

4.1　実 験 方 法

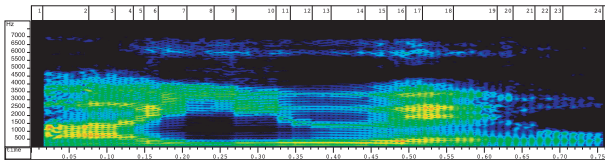
提案手法の有効性について調べるため，日本語 5 母音の連続発声を用いて実験を行った．これは各母音が一度ずつ出現するもので，/aiueo/，/ieuao/ など 5!=120 単語が存在する．成人男女各 3 名（それぞれ話者 M1，M2，M3，F1，F2，F3 とする）についてこれらの発声を収録した．これらの発声について STRAIGHT [15] に基づくスペクトル分析を行い，このスペクトルから 40 次のケプストラムを得た．同時に発声のピッチ，パワー，継続長も STRAIGHT の分析を基に得た．

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次元分割手法も導入した [8]．この時，部分空間の次元数（以下ブロックサイズとよぶ）は 1 および 2 とし，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．

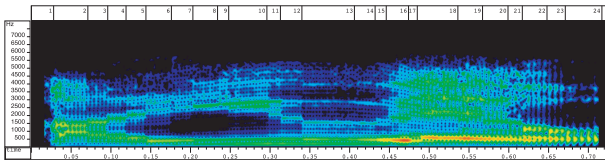
一方，話者 M2，M3 および F2，F3 についても図 2 の流れで分析を行った．ただし構造表象の抽出は行わず，それぞれの発声を 25 個の分布系列へと変換する．M1 および F1 から得られた不変構造を制約条件，M2，M3，F2，F3 のそれぞれについて 5 つの分布（3，8，13，18，23 番目）を初期条件として用いて，残りの音響事象分布を，提案する枠組みで推定した．その結果，既知情報として初期条件の分布，ピッチ，パワー，継続長，推定された情報として探索解である 20 個の分布が得られる．これらの情報を用いて STRAIGHT の枠組みで音声を再合成した．音声模倣の枠組みで言えば，本実験における話者 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

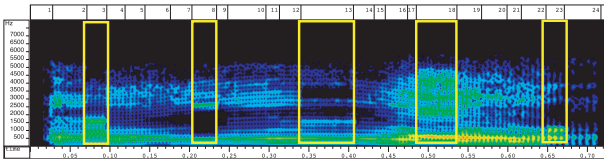
実験によって得られた合成音声の一例を図 7 に示す．図中 (a) は構造抽出に用いた話者 M1 の発声の分析再合成音，(b) は初期条件を提供した話者 F2 の発声の分析再合成音である．(c) は提案手法によって M1 の構造および F2 の初期条件から合成した音声である．単語は/aiueo/ であり，ブロックサイズは 1 とした．(c) のうち枠囲いされた部分が初期条件として与えた分布に対応する．これらを比較すると提案手法による音声が ほぼ



(a): M1 の分析再合成音



(b): F2 の分析再合成音



(c): 提案手法による合成音

図 7　合成音声の一例

Fig. 7　One example of synthesized speech.

(b) のスペクトルを再現していることがわかる．実際に聴取してみると (c) の音声の発話内容は /aiueo/ と容易に知覚する事ができる．さらに話者性についても F2 の話者性を正しく再現できていることがわかる．これらの結果から構造表象に話者性を与えるという提案する枠組みによって音声を生成可能である事が示された．

なお従来，全探索によって音響事象を推定する場合，初期条件が 5 つでは膨大な空間探索を必要とし，現実的な時間では連続発声の合成は困難であった [9]．一方，本手法においては同様の条件によって合成が可能となっており，速度面において大幅な向上が図られたといえる(注1)．

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．提案する枠組みで評価すべき項目は以下の二つであると考えられる．

（1） 本手法で得られた音声について，合成すべき単語（母音列）が正しく生成されたかどうか．

（2） 本手法で得られた音声について，その話者性が目標とすべき話者のものとなっているかどうか．

（1）について評価するため，合成実験で得られた音声について，ヘッドホン聴取による書き取りテストによって了解度評価を行った．この際，各母音が一度ずつ出現する 5 母音連続発声である旨はあらかじめ伝えた．使用したサンプルはブロックサイズ 1 および 2 の場合について，各話者の組み合わせ（計 8 通り）につき 60 単語ずつを選定した．総数は $2 \times 60 \times 8 = 960$ サンプルとなる．

一方，（2）について評価するために，ABX 法により合成音声の話者性を評価した．この際，正解となる話者の分析再合成音および比較用の同一性の別話者（構造抽出話者およびもう一方の初期条件提供話者のいずれか）の分析再合成音をランダムな順序で提示する．その後，提案手法による合成音声を提示し，先のサンプルのどちらに近いかを評価してもらった．用いたサンプルはブロックサイズおよび話者の組み合わせの異なる 32 サンプルである．この際，最初に参照させる 2 つの音声の単語として /aiueo/ を，提案手法による音声サンプルについてはランダムに異なる単語を提示した．（1），（2）それぞれの聴取実験の条件を表 1 に示す．

(注1)：孤立母音による予備実験で，約 20000 倍の高速化を実現している．

表 1　聴取実験の実験条件

Table 1　Experimental conditions for the listening tests.

(a): 了解度評価（1）の為の聴取実験条件

被験者	日本人成人男性 4 名
聴取実験サンプル	提案手法による合成音 960 サンプル
実験法	ヘッドホン聴取による書き取りテスト

(b): 話者性評価（2）の為の聴取実験条件

被験者	日本人成人男性 4 名
聴取実験サンプル	提案手法による合成音 32 サンプル
	対応する合成対象の分析再合成音 32 サンプル
	同一性の別話者の分析再合成音 32 サンプル
実験法	上記 32 組のサンプルによる ABX 法

5.2 実験結果

上記聴取実験の結果を表 2 に示す．表 2(a) は書き取りテストの結果を示したものである．ここで正解とは 4 人の被験者のうち 3 人以上の被験者の回答と，模倣対象の単語が一致した場合を表している．表中の M→F は，構造提供話者が男性，初期条件提供話者が女性であることを表している．表 2 によれば，構造提供話者および初期条件話者が共に男性の場合に 84 %の正答率，全体的にはブロックサイズが 1 の時に，平均 70 %の正答率となった．一方構造抽出話者と初期条件提供話者が異なる性別の場合は正答率が低くなる傾向があった．またブロックサイズが 2 の場合は，ブロックサイズ 1 の場合に比べて正答率が低くなったが，性別が異なる場合はその影響は少なかった．

一方表 2(b) は，話者性評価実験の結果を示したものである．表 2(b) は，被験者 4 名中 3 人以上が初期条件提供話者に近いと評価した数を示している．表 2(b) によれば，ブロックサイズ 1 の場合の方がブロックサイズの 2 の場合に比べて，特に異性別間での音声模倣に際して，正しく話者性を再現していることがわかる．

表 2　聴取実験の結果．表中，M→F は構造抽出話者が男性（Male），初期条件提供話者が女性（Female）であることを表す．その他も同様．減少率はブロックサイズの増加に伴う相対的な正答率の低下を表す．

Table 2　Results of the listening tests.

(a):書き取りテストの結果

	総単語数	正答率（ブロックサイズ 1）	正答率（ブロックサイズ 2）	減少率
全体	480	0.70	0.63	10.0 %
M→M	120	0.84	0.76	9.5 %
M→F	120	0.59	0.57	**3.4** %
F→M	120	0.60	0.56	**6.7** %
F→F	120	0.78	0.65	16.7 %

(b):話者性評価の結果

	サイズ 1	サイズ 2
全体	11/16	7/16
M→M	3/4	3/4
M→F	2/4	1/4
F→M	3/4	1/4
F→F	3/4	2/4

6. 考　　察

合成実験，および主観評価実験の結果について考察する．まず全体的な正答率は，先行研究による孤立母音の聴取実験の場合に 63 %（既知母音数 4 の場合）であったのに比べて向上が見られた．特に発話全体に占める初期条件の割合が先行研究よりも少ないことから，発話の全体的特徴を考慮した音声合成が有効に機能していると考えられる．また同一性別間での音声模倣における正答率と異性別間における正答率とで，異性別間での正答率が低い傾向がみられた．これは性別の違いによる発話スタイルの違いが構造の違いを与えていること [16]，およびブロックサイズ 1 の場合には話者性の違いに対する構造不変性が制約されていることが考えられる．実際ブロックサイズを 2 とした場合に，同一性別間と異性別間の正答率の差が小さくなっていることから，話者性不変性を保つ最適なブロックサイズの存在が示唆される．一方，ブロックサイズを大きくした場合に全体の正答率は低下していた．これはブロックサイズが大きい場合，安定した方程式の解を得るためにより多くの初期条件を必要とすることが考えられる．すなわち，ブロックサイズが 2 の場合，身体性を表現するのにより多くの初期条件で構造を定位する必要がある．これは話者性評価実験で，ブロックサイズが大きい場合に曖昧性が大きくなっていることからも示唆される事項である．

なお主観評価実験において，二つの音韻を逆に取り違える音韻交代的な現象が多く見られた．これは同一単語でも被験者によって起こる場合，起こらない場合があった．これは人間が聴取の際に全体的特徴を少なからず利用していることを示唆しており，音韻による評価が必ずしも適切ではないことを意味する．

7. 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，初期検討として音響空間の全探索によって提案する枠組みの基礎的検討を行ってきたが，速度面，精度面に課題があった．一方今回，解析的手法を導入する事で音響空間探索の高速化，高精度化を図る手法を提案した．その結果，提案する枠組みの中で，従来では困難だった少数の初期条件からの連続発声の音声合成を実現した．実際に合成した音声について主観評価実験を行い，提案手法によって妥当な音声合成が実現可能である事を示した．

今後は，繰り返し演算によって構造制約を満たすように音響事象を最適化する手法やより大きなブロックサイズによる合成実験を行う必要がある．また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も [3]，提案する枠組みの中で取り扱っていく予定である．

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] 齋藤大輔他, “音声の構造的表象を入力とした音声合成に対する基礎的検討”, 日本音響学会秋季講演論文集, 1-P-2, 2007.
[10] 齋藤大輔他: “構造的表象からの音声生成に関する基礎的検討”, 信学技報, SP2007-80, pp.55–60, 2007.
[11] 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.
[12] 峯松信明他:“線形・非線形変換不変の構造的情報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”, 日本音響学会春季講演論文集, 1-P-12, pp. 147–149, 2007.
[13] 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.
[14] 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.
[15] 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.
[16] N. Minematsu *et al.*:“Para-linguistic information represented as distortion of the acoustic universal structure in speech,” Proc. of ICASSP’2006, vol.1, pp.261–264, 2006.