# 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*

◎齋藤大輔, 朝川智, 峯松信明, 広瀬啓吉 (東大)

## 1 はじめに

近年の音声合成システムは，与えられたテキスト列を音響信号として出力する 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) ，すなわち音そのもののモデルである．

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，上記のように音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事象を物理的，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」の音響的定義となる，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，非言語性歪み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]．

本稿では，この構造的表象に基づく音声合成システムについてその枠組みを提案する．提案する枠組みは，音の実体モデルを持ちテキストを入力とする従来の音声合成システムとは大きく異なる．発話全体の語形を考え，それに対して身体特性，収録機器の伝送特性を与える事で初めて，聞き手が聴取する音響信号が生成される [9]．本枠組みは音声模倣のモデルとして解釈可能である．以下本稿ではその基礎的検討としてケプストラム空間の解探索に基づく音声合成を行い，提案する枠組みの妥当性を考察する．

## 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歪み

音声の音響的実体は非言語的特徴によって不可避的に歪むが，これらは大きく乗算性歪みと線形変換性歪みに分けられる．

乗算性歪みは，スペクトルに対する乗算で表現される歪みである．ケプストラム空間では，この種の歪みは加算演算 $\mathbf{c}' = \mathbf{c} + \mathbf{b}$ 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歪み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，これらの歪みは不可避である．

線形変換性歪みはケプストラム空間において行列 $\mathbf{A}$ による線形変換 $\mathbf{c}' = \mathbf{A}\,\mathbf{c}$ で表現される歪みである．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 [10]．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歪みとして扱うことができる．

以上をまとめると，音声の音響的実体に不可避的に混入する非言語的特徴は，ケプストラム空間においてアフィン変換 $\mathbf{c}' = \mathbf{A}\,\mathbf{c} + \mathbf{b}$ で表現される．これらの $\mathbf{A}, \mathbf{b}$ が話者や収録環境によって多様に変化し，音声の音響的実体に様々な歪みが混入する事になる．

### 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

---

*A fundamental study of structure-to-speech conversion.
by SAITO Daisuke, ASAKAWA Satoshi, MINEMATSU Nobuaki, and HIROSE Keikichi (The University of Tokyo)

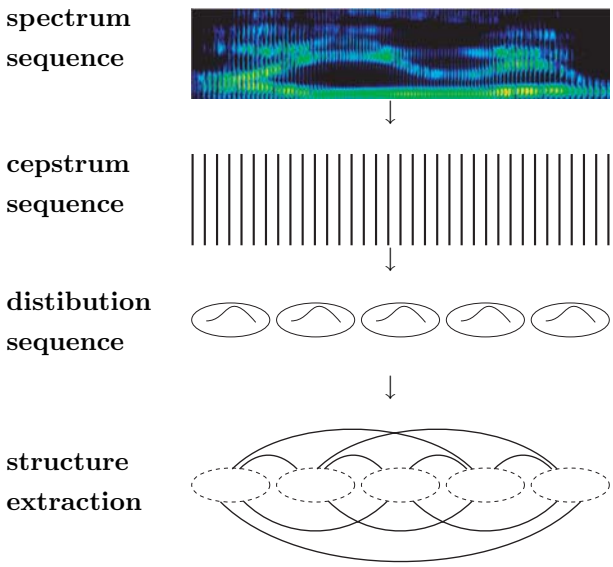



Fig. 1　音声からの構造的表象の抽出

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．

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

二つの分布に対して共通のアフィン変換 $\mathbf{Ac}+\mathbf{b}$ 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 [11].

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，すなわち非言語性歪みにおよそ不変な構造を求める事ができる．

### 2.3　一発声の構造化

一発声を一つの構造的表象で記述する場合を考える．Fig. 1 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．

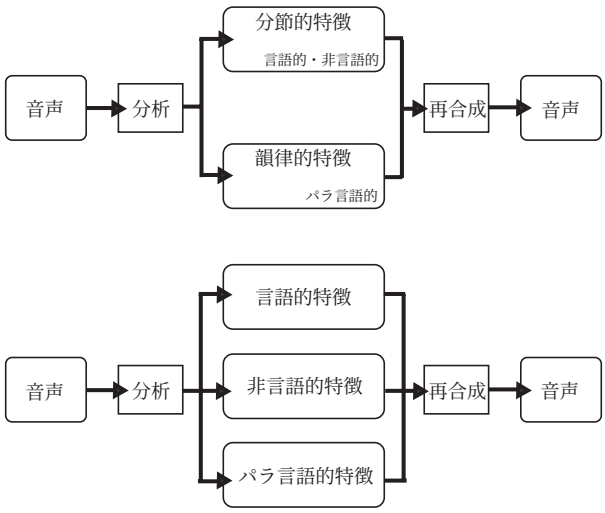



Fig. 2　従来の分析再合成系の枠組み（上）と提案する枠組み (下)

## 3　構造的表象に基づく音声合成

### 3.1　非言語的要因をも分離する分析再合成系

従来の音声合成では，学習話者による数百〜数千文の音声試料を学習データとして異音と音との対応を学習する．そして入力としてテキストを与えた場合に対応する音列として音ストリームを出力する．この場合得られるのは学習話者の声である．

幼児の音声模倣では両親の音声を模倣しても，自らの声で言葉を返す．第 1 節で述べたように，幼児の音声模倣では両親の発声全体の語形を真似，自らの声で返していると考えられている．

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）を与えることで初めて音が生まれるという合成系である [9]. 分析再合成系でこの考えを示すと Fig. 2 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（主には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．

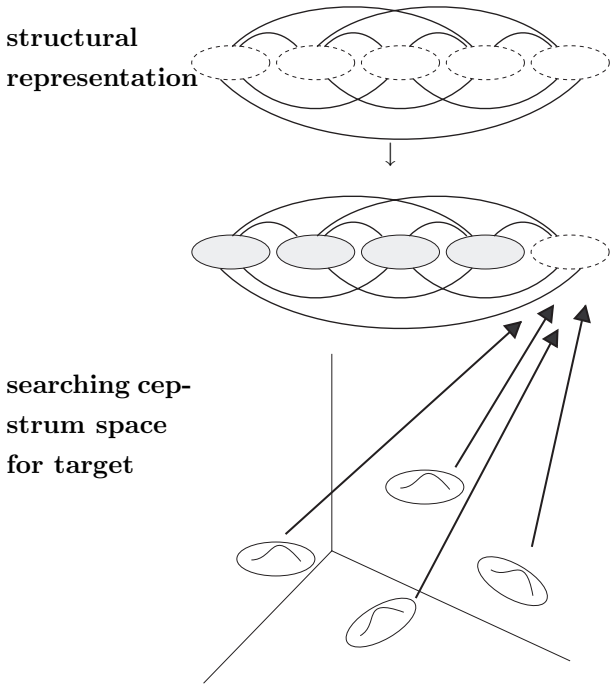



Fig. 3　解探索による構造的表象を制約とする音声合成の枠組み

### 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はない [12]．そのため，今回は提案する音声合成の枠組みの基礎的検討として，ケプストラム空間における制約条件を満たす解の探索問題として定式化する．

今，構造的表象の途中までが声として出力された場面を考える（状態 $s_{t-n}, \cdots, s_{t-1}$ までが出力済み）．このとき次の出力は状態 $s_t$ を声にする操作で得られるが，これを，$s_{t-n}, \cdots, s_{t-1}$ の音的実体 $o_{t-n}, \cdots, o_{t-1}$ からの距離制約を使ってケプストラム空間を探索して求める．提案する解探索による音声合成の枠組みを Fig. 3 に示す．

### 3.3　探索空間の制限

音響特徴量としてのケプストラムベクトルは 12〜25 次元程度の多次元ベクトルである．そのため適切な解探索のためには，探索対象となる音響空間に適切な制限を加える必要がある．今回，発声全体の平均と分散を用いて探索空間を制限することを考える．

単一のガウス分布においては平均を $\mu$，標準偏差を $\sigma$ とした時，$[\mu - 3\sigma, \mu + 3\sigma]$ の区間の値をとる確率は約 97%となるため，統計学ではこの区間を全空間として扱うことが多い．よって本研究では各々の次元について探索範囲をこの $\pm 3\sigma$ 区間に限定する．

Table 1　音響分析条件

| | |
|---|---|
| sampling | 16 bit / 16 kHz |
| window | length 25 ms / priod 5 ms |
| parameter | Mel cepstrum (1 to 10) [$\alpha = 0.42$] |

## 4　合成実験

### 4.1　実験方法

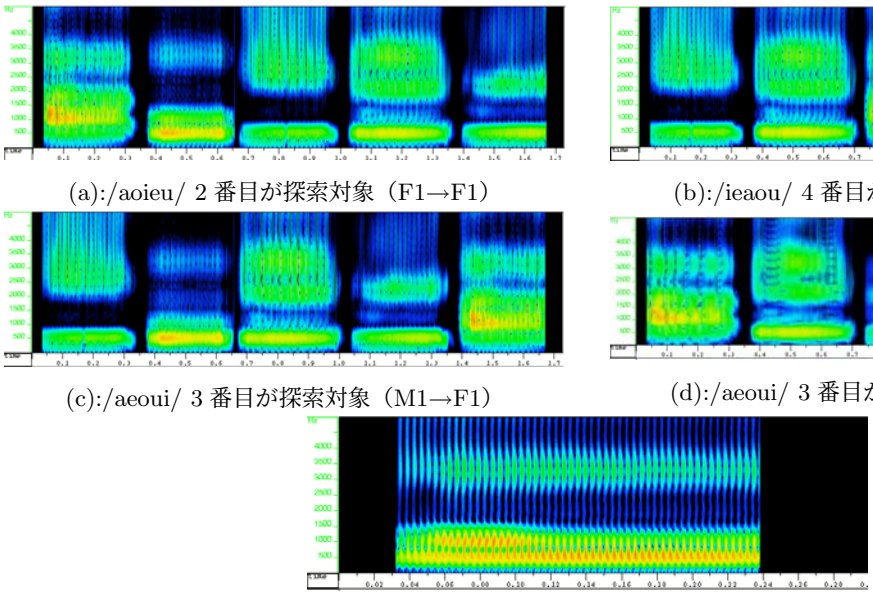
提案する枠組みによって音声合成が可能であることを確認するため，日本語 5 母音の孤立発声（/a/, /i/, /u/, /e/, /o/）を用いて実験を行った．成人男女各 1 名（それぞれ話者 M1，F1 とする）の日本語 5 母音の孤立発声を収録した．これらの発声について Table 1 に示す音響分析条件でケプストラム分析を行った．同時にそれぞれの発声のピッチ，パワー，継続長についても分析した．これらのパラメータを用いて以下に示す手順で構造抽出と解探索を行った．

1. 各母音発声を平均ベクトルと対角共分散行列で表される多次元ガウス分布として，最尤推定を行い各パラメータをモデル化する．
2. 分布間距離を求めて構造を抽出する．
3. 5 母音のうち，4 つを既知，残りの 1 つを探索対象とする．本実験では分散項は既知とし，平均ベクトルを探索範囲で変化させる．なお各次元について $\pm 3\sigma$ の範囲で 6 点に量子化し，これをその次元の探索点とする．
4. 探索対象以外の 4 母音との Bhattacharyya 距離を求め，所望の分布間距離との差が事前設定した評価誤差以内のときにそれを解とする．
5. 得られたケプストラムの解と事前に抽出したピッチ，パワー，継続長の情報から音声を合成する．

この際 2 名の音声について，(a):同一話者内における分析再合成の場合，(b):構造抽出の話者と合成対象の話者が異なる場合について実験を行った．

### 4.2　実験結果

話者 F1 を合成の対象話者とした場合の結果について示す．Table 2 は探索対象の母音について評価誤差と解の個数の関係を示したものである．候補総数は $6^{10} \simeq 6 \times 10^7$ であり，構造的表象を制約条件としていくつかの解を導出できていることがわかる．一方合成音声のスペクトルを Fig. 4 に示す．(a) および (b) は構造として話者 F1 自身のものを用いた分析再合成，(c) および (d) は構造抽出話者を話者 M1 とした再合成の結果である．探索対象はすべて/o/ であり，参考のため (e) に分析時のパラメータを用いて再合成した/o/のスペクトルを示している．(a)(c) およ



(a):/aoieu/ 2 番目が探索対象（F1→F1）

(b):/ieaou/ 4 番目が探索対象（F1→F1）

(c):/aeoui/ 3 番目が探索対象（M1→F1）

(d):/aeoui/ 3 番目が探索対象（M1→F1）

(e):/o/の分析再合成音（参考）

Fig. 4　合成実験の結果．探索対象はすべて/o/.

Table 2　各探索母音における評価誤差と解個数の関係

(a):同一話者の分析再合成

| vowel | 2.0 | 1.0 | 0.8 | 0.6 | 0.4 | 0.2 |
|---|---|---|---|---|---|---|
| /a/ | 3749 | 251 | 114 | 34 | 7 | 0 |
| /i/ | 931 | 63 | 29 | 11 | 3 | 0 |
| /u/ | 3277 | 258 | 111 | 35 | 7 | 0 |
| /e/ | 288 | 20 | 7 | 3 | 0 | 0 |
| /o/ | 871 | 36 | 13 | 7 | 1 | 0 |

(b):異なる話者の構造を用いた場合

| vowel | 2.0 | 1.0 | 0.8 | 0.6 | 0.4 | 0.2 |
|---|---|---|---|---|---|---|
| /a/ | 121 | 0 | 0 | 0 | 0 | 0 |
| /i/ | 1613 | 94 | 38 | 12 | 2 | 0 |
| /u/ | 1088 | 51 | 17 | 5 | 0 | 0 |
| /e/ | 133 | 9 | 3 | 1 | 0 | 0 |
| /o/ | 995 | 74 | 36 | 18 | 2 | 0 |

び (b)(d) は後者がより構造制約を厳密に満たす解である．予備的な聴取実験の結果，複数の解の中で (b), (d) は異なる音韻として知覚されたが，(a)，(c) では音韻性を保っていることが確認された．これは 1 次元あたりの解像度が 6 点と少ない為と考えられる．また話者性については全て F1 のものと知覚された．

## 5　おわりに

本稿では，非言語性歪みにおよそ不変な特徴量である音声の構造的表象に基づく音声合成の枠組みを提案した．音声の構造的表象は話者性を含まない物理表象であるため，音声合成に際しては明示的に身体性を与える必要がある．提案する枠組みにおける音声合成を検証するため，構造的表象を制約条件とし，既に音化された事象をもとに次の音響事象を推定する形で問題を定式化した．実際にケプストラム空間の解探索を行い，提案手法によって一定の音韻性を保った音声を合成できることを示した．

提案手法は，幼児の音声模倣のモデルとして捉えることができる．合成実験において，異なる話者の構造を用いた音声合成においても，構造抽出した話者の話者性の影響を受けず合成音が生成されており，提案する枠組みの妥当性を示唆している．

今後は本枠組みの連続発声への適用や，声道形状特性を直接的に入力パラメータとする手法について検討を行う予定である．既に [8] では連続母音系列を対象とした認識系で優れた結果を出しており，また構造推定時にも “過剰な不変性” 問題を解消する為に次元分割などの手法を提案している．これらの手法は余剰な解空間を排除する事が可能であり，合成の枠組みでも検討する予定である．

## 参考文献


[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, “子どものかな文字の習得過程”, 秋山書店, 1986.
[3] 早川, 月刊言語,35,9,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, コミュニケーション障害学, 20, 2, pp.84–85, 2003.
[6] N. Minematsu et al., Proc. SRIV’2006, pp.47–52, 2006.
[7] 朝川他, 信学技法, SP2006-105, pp.119–124, 2006.
[8] 朝川他, 音講論（秋）, 3-Q-10, 2007（発表予定）.
[9] 峯松他, 信学技法, SP2007-30, pp.37–42, 2007.
[10] M. Pitz, H. Ney, IEEE Trans. Speech and Audio Processing, Vol.13, pp.930–944, 2005.
[11] 峯松他, 音講論（春）, 1-P-12, 2007.
[12] 錦戸, 党, 音講論（春）, 1-Q-28, 2007.