

構造的表象からの音声合成における音響空間探索の高速化*

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]。加えてこの表象に基づく音声合成についても、「発話の語形に身体特性を与える・戻すことで音響信号を得る」という枠組みを提案し、音響空間の探索問題としての定式化を通してその検討を行ってきた [9, 10]。

従来の探索問題の定式化においては、上記の構造表象を制約条件とし、制約を満たす音響事象を音響空間を線形探索することで得ていた。しかし解探索の解像度と演算時間のトレードオフから十分な品質の音声を高速に求めることは困難であった。そこで本稿では提案する枠組みにおける音声合成において、解析的手法を取り入れることで音響空間探索の高速化を試みる。

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免的に歪むが、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 $c' = c + b$ 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 A による線形変換 $c' = Ac$ 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 [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免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 $c' = Ac + b$ で表現される。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

* Efficient search for acoustic events to accelerate structure-to-speech conversion.

by SAITO Daisuke, ASAKAWA Satoshi, MINEMATSU Nobuaki, and HIROSE Keikichi (The University of Tokyo)

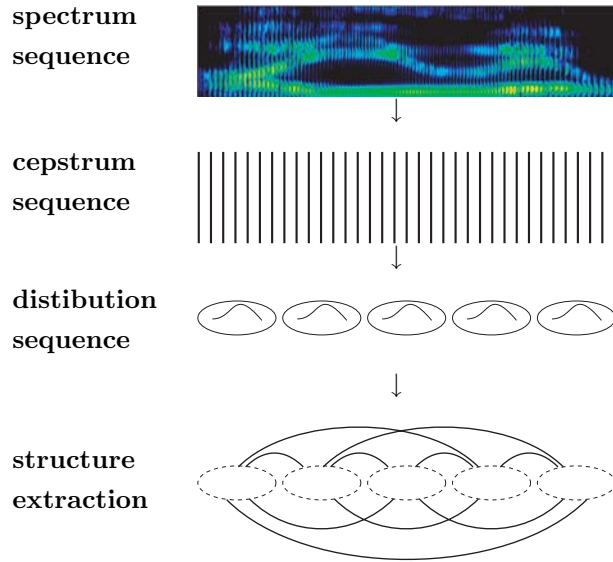


Fig. 1 音声からの構造的表象の抽出

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．

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

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

一発声の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．

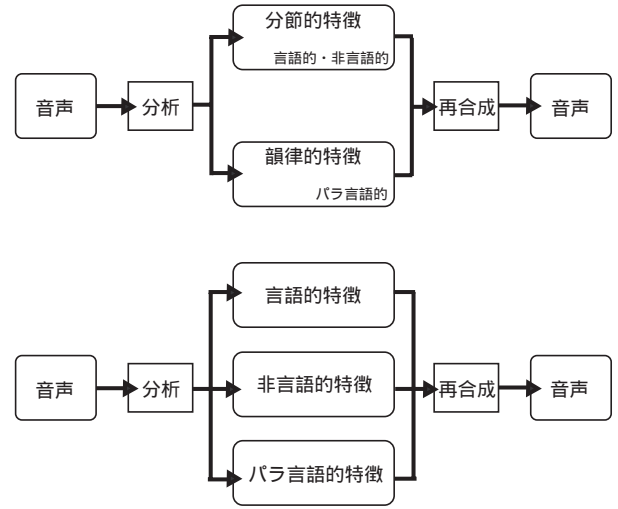


Fig. 2 従来の分析再合成系の枠組み（上）と提案する枠組み（下）

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 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．

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 Fig. 3 に示す．この際，構造に基づ

searching cep-
strum space
for target

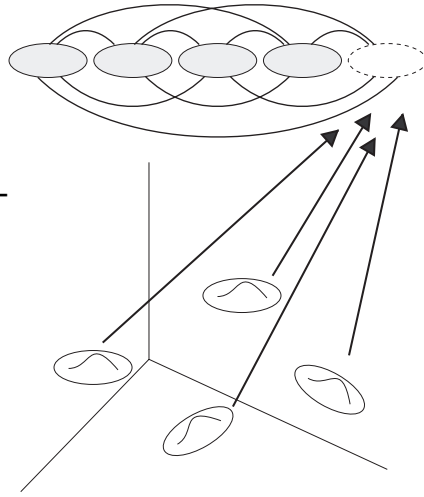


Fig. 3 解探索による構造的表象を制約とする音声合成の枠組み

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 解析手法に基づく解候補の導出

本研究では解析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。

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

ただし $\mu_{12} = \mu_1 - \mu_2, V_{12} = \frac{\Sigma_1 + \Sigma_2}{2}$ である。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、上記の μ_1 を変数とする連立方程式の解を求める事に相当する。今 2 次元の場合に、初期条件の音響事象 $A = \mathcal{N}(a, V_a), B = \mathcal{N}(b, V_b)$ から音響事象 p の平均 $\mu = (c_x, c_y)$ を求めたいとする。ただし p の分散共分散は対角で既知としその成分を V_x, V_y とする。簡潔な表記のため式 (2) の右辺第二項を ϵ と書くと、式変形により μ に対する以下の連立方程式が定まる。添字は 2 次元の x, y 成分に対応する。

$$\begin{cases} BD_a - \epsilon_a = \sum_{d \in \{x, y\}} \frac{1}{4(V_d + V_{ad})} (c_d - a_d)^2 \\ BD_b - \epsilon_b = \sum_{d \in \{x, y\}} \frac{1}{4(V_d + V_{bd})} (c_d - b_d)^2 \end{cases} \quad (3)$$

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

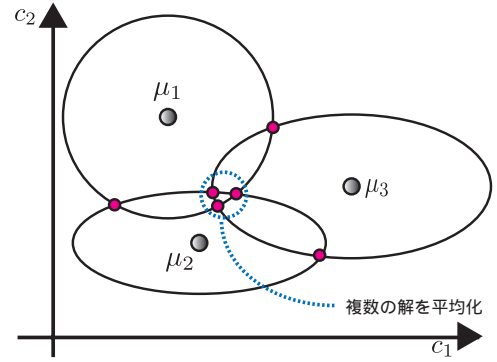


Fig. 4 解析手法に基づく解の導出 (2 次元の場合)

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。

提案する解析手法によるアプローチについて述べる。まずブロックサイズ m において、 n 個の音響事象を初期条件とする場合、探索対象の音響事象に対する連立方程式は $n C_m$ 個得られる。これらから得られた複数の解について、最も縮退している近傍について平均し、これを求める音響事象とする。2 次元の場合における、提案手法の枠組みを Fig. 4 に示す。

4 合成実験

4.1 実験方法

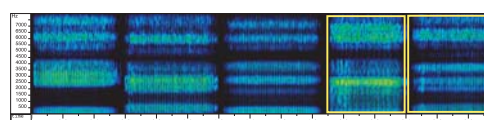
提案手法の有効性について調べるため、日本語 5 母音の孤立発声 ($/a/, /i/, /u/, /e/, /o/$) を用いて実験を行った。成人男女各 2 名 (それぞれ話者 M1, M2, F1, F2 とする) の日本語 5 母音の発声を収録した。これらの発声について STRAIGHT [15] に基づくスペクトル分析を行い、このスペクトルから 40 次のケプストラムを得た。同時に発声のピッチ、パワー、継続長も STRAIGHT の分析を基に得た。

これらのパラメータから Fig. 1 の流れで構造を抽出した。さらに既知母音の数および次元分割のブロックサイズを変化させ、従来の線形探索に基づく手法および提案手法で音響空間探索と音声の合成を行った。構造を提供する話者は上記の 4 名、初期条件を提供する話者を上記のうち男女 1 名ずつ (M1, F1) とした。幼児の音声模倣の枠組みでいえば、構造提供話者は母親に、初期条件提供話者は幼児に対応する。

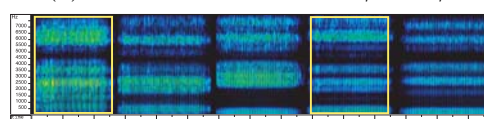
また音質について評価するため、生成したサンプルと初期条件とした既知母音によって構成される 5 モーラ単語を作成し成人男性 5 名による書き取りテストを行った。単語知識の影響を除くため、あらかじめ被験者に 5 モーラの出現の重複を許す旨を伝えた。

Table 1 聴取実験の結果．括弧内はブロックサイズ，表の値は全モーラ一致の正答率を表す．

	総単語数	線形探索 (1)	線形探索 (2)	提案手法 (1)	提案手法 (2)
既知母音数 1	40	0.05	—	—	—
既知母音数 2	80	0.20	—	0.14	—
既知母音数 3	80	0.45	0.40	0.44	0.28
既知母音数 4	40	0.83	0.73	0.65	0.73



(a): 従来手法による合成音 /ieuao/



(b): 提案手法による合成音 /aeiou/

Fig. 5 合成音声の一例

4.2 実験結果

実験によって得られた合成音声の一例を Fig. 5 に示す．Fig. 5 は構造提供話者として話者 F2，初期条件提供話者として話者 M1 の音声を用いている．探索および解析によって母音 /a/, /o/ を推定した．Fig. 5(a) は従来の線形探索による結果，Fig. 5(b) は提案手法による結果を示している．ともに同様のスペクトルが得られている事がわかる．

一方計算時間について比較する．筆者らの実験環境において，ブロックサイズを 1 とした場合，従来手法では既知母音数と話者の組み合わせを変化させた 240 セットを作成するのに，3 時間程度の時間を要していた．一方提案する手法では同様の条件において，30 秒程度で終了する．およそ 20000 倍の高速化を実現したことになる．

続いて聴取実験の結果を Table 1 に示す．括弧内の数字は特徴量分割におけるブロックサイズを表している．各欄の数字は 5 名の話者のうち 3 名以上が 5 モーラ全て一致させた場合を正答として正答率を求めている．数値の存在しない欄については探索時間が膨大または方程式の個数が不足しているため実験を行っていない．Table 1 から，提案手法による合成音の了解度は従来の線形探索の結果を若干下回る結果となった．従来手法ではブロックサイズが大きい場合，探索時間とのトレードオフにより各次元を低い解像度で探索しなければならない．一方提案手法では，解像度，探索時間の両面で従来手法のような制約を受けないため，このことが有効に作用していることがわかる．提案手法の了解度の結果が従来手法より

若干低い点については，得られた解候補をもとに構造による制約を満たすまで繰り返し演算を行う等により改善する事が可能であると考えられる．

5 おわりに

本稿では，構造的表象からの音声合成における音響空間探索について，解析的手法を導入することでその高速化を行う手法を提案した．従来の線形探索に基づく定式化に比べて，およそ 20000 倍の高速化を実現し，合成実験および聴取実験によって，従来手法では検証が困難だったブロックサイズの増加した場合による音声合成を実現する事ができた．

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．また提案手法では，初期条件として与えた音響事象数が計算時間にほとんど影響しないため，連続音声における構造的表象の基づく音声合成について，より少ない音響事象から一発声全体を合成できる可能性がある．この点についても検討していく予定である．

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] Y. Qiao et al., Proc. ASRU2007, pp.576-581, 2007.
- [8] S. Asakawa et al., ICASSP2008, 2008 (to appear).
- [9] 齋藤他，音講論（秋），1-P-2, 2007.
- [10] 齋藤他，信学技報，SP2007-80, pp.55-60, 2007.
- [11] M. Pitz, H. Ney, IEEE Trans. Speech and Audio Processing, Vol.13, pp.930-944, 2005.
- [12] 峯松他，音講論（春），1-P-12, 2007.
- [13] 峯松他，信学技報，SP2007-30, pp.37-42, 2007.
- [14] 錦戸，党，音講論（春），1-Q-28, 2007.
- [15] H. Kawahara et al., Speech Communication,27, 3-4, pp.187-207, 1999.