

反復解法に基づく構造的表象からの音声合成の 高精度化に関する検討*

◎齋藤大輔, 喬宇, 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話者非依存の表象と考え、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), 即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]。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、音声合成システムについても検討を行っている [7, 8, 9, 10]。

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 [9]。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。そこで本稿では、従来法 [10] で推定された音響事象を初期値として、再度探索を繰り返す事で、より構造表象の制約条件に合致する音声を生成する手法を試みる。

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 [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

* Improvement of structure-to-speech conversion by iterative calculation.

by SAITO Daisuke, QIAO Yu, MINEMATSU Nobuaki, and HIROSE Keikichi (The University of Tokyo)

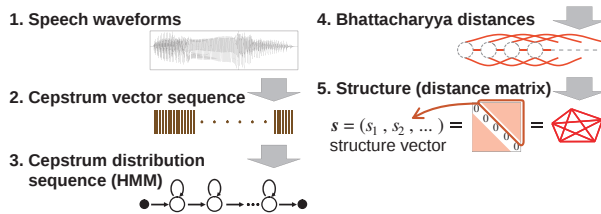


Fig. 1 音声からの構造的表象の抽出

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 $Ac+b$ 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 [13]。

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。

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

我々が提案している構造的表象からの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]。従来の分析再合成系では音声を分節の特徴 (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) と韻律の特徴 (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

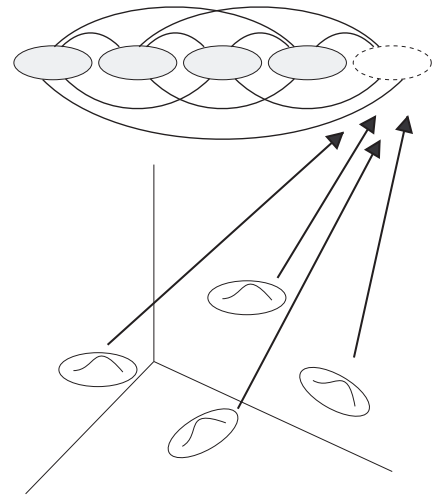


Fig. 2 解探索による構造的表象を制約とする音声合成の枠組み

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 解探索としての定式化

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 [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 [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. 2 に示す。

上記の探索問題は複数の楕円が与えられた場合の交点導出問題として考えることができる [10]。今、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

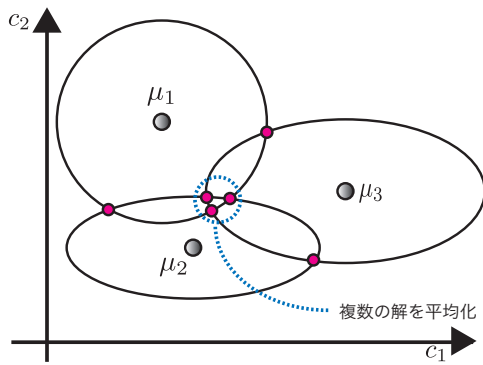


Fig. 3 解析手法に基づく解の導出 (2次元の場合)

条件の音響事象 $A = \mathcal{N}(\mathbf{c}_a, \Sigma_a)$, $B = \mathcal{N}(\mathbf{c}_b, \Sigma_b)$ から推定対象の音響事象 $p = \mathcal{N}(\mathbf{x}, \Sigma)$ の平均 $\mathbf{x}$ を求めたいとする。ただし全ての共分散行列 ($\Sigma, \Sigma_a, \Sigma_b$) を既知とする。簡潔な表記のため式 (2) の右辺第二項を ϵ , 音響事象 A, B について第一項の V^{-1} と定数項をあわせた係数行列を $\mathbf{A}, \mathbf{B}$ と書くと、式変形により $\mathbf{x}$ に対する以下の連立方程式が定まる。

$$\begin{cases} BD_a - \epsilon_a = (\mathbf{x} - \mathbf{c}_a)^t \mathbf{A} (\mathbf{x} - \mathbf{c}_a) \\ BD_b - \epsilon_b = (\mathbf{x} - \mathbf{c}_b)^t \mathbf{B} (\mathbf{x} - \mathbf{c}_b) \end{cases} \quad (3)$$

ただし右肩の $()^t$ はベクトルの転置を表す。上記の式は正値の2次形式であり、連立方程式を解く事は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。

式 (3) に基づく交点の導出について述べる。まず d 次元の空間 (以下ブロックサイズ d とよぶ¹) 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 nC_d 組得られる。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. 3 に示す。

3.3 反復解法による高精度化

上記の手法は、推定対象となる音響事象を各々、独立に推定している。今、発声全体が N 個の音響事象で表されているとし、推定対象を m 個、初期条件を n 個 (すなわち $m+n=N$) とすると、全ての制約条件 $N C_2 (= m C_2 + n C_2 + mn)$ のうち mn 個だけしか考慮されていない。この際推定対象間の距離関係 ($m C_2$)

¹ここでは構造における音声認識で提案されている次元分割手法 [8] における部分空間の次元数をブロックサイズと呼んでいる。

が考慮されていないため構造の制約条件を真に満たす推定となっていないといえる。また推定事象間の連続性が満たされず、自然性が劣化する問題があった [11]。そこで一回の探索で得られた解を新たな初期条件として、逐次的に推定を繰り返す事で構造制約条件を満たす音響事象群を推定することを考える。今回提案する逐次推定の手順を以下に示す。ブロックサイズを d とする。

1. n 個の初期条件をもとに m 個の音響事象を独立に推定する。個々の音響事象には $n C_d$ 組の連立方程式が用いられる [10]。
2. 推定された m 個の音響事象のうち一つに着目する。その他の $m-1$ 個との距離関係をもとに得られた $m-1 C_d$ 組の連立方程式を用いて解を推定する。これを m 個全てについて行う。
3. 初期条件 n 個および推定された音響事象 m 個を全て併せた $N (= m+n)$ 個について、一つに着目しその他の $N-1$ 個との距離関係を 2. と同様にして推定していく。
4. 3 を適宜繰り返す。

提案手法により全ての音響事象間の関係性を考慮した、真に構造制約を満たす音響事象の推定が可能になると考えられる。

4 合成実験

4.1 実験手法

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 (/aiueo/) を用いて実験を行った。成人男女各3名 (それぞれ話者 M1, M2, M3, F1, F2, F3 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 [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, F1 の発声について、25 個の分布系列に変換後、ブロックサイズを2として、構造を抽出した。初期条件を提供する話者 M2, M3, F2, F3 については、25 個の分布系列に変換し、このうち5つ (3, 8, 13, 18, 23 番目) を初期条件とした。これらを初期条件として提案する反復による推定を10回まで繰り返した。

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. 4 に示す。Fig. 4 は構造提供話者として話者 M1, 初期条件提供話者として話者 F2 の音声を用いている。Fig. 4

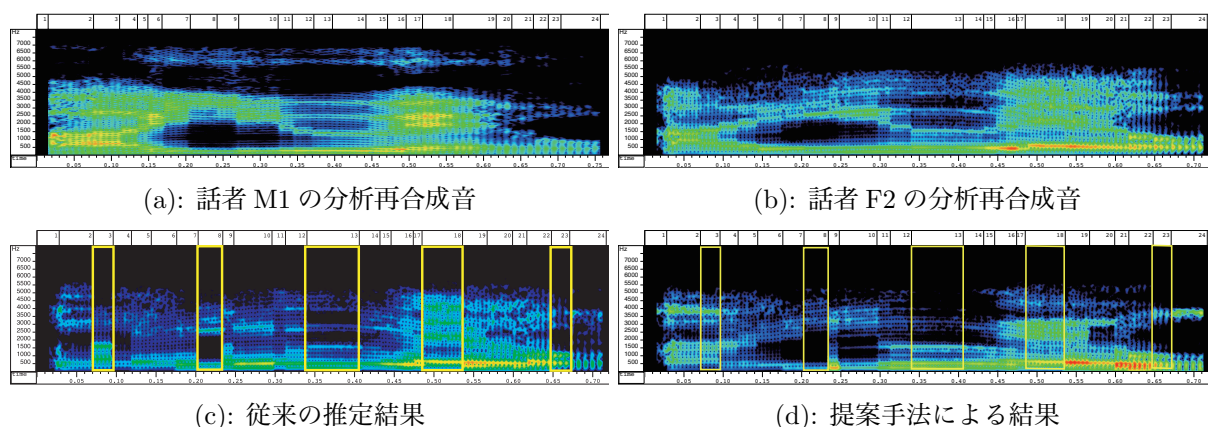


Fig. 4 合成音声の一例; (a)(b): 構造提供話者 M1, 初期条件話者 F2 の分析再合成音; (c): 反復を行う前の結果; (d): 提案手法による結果

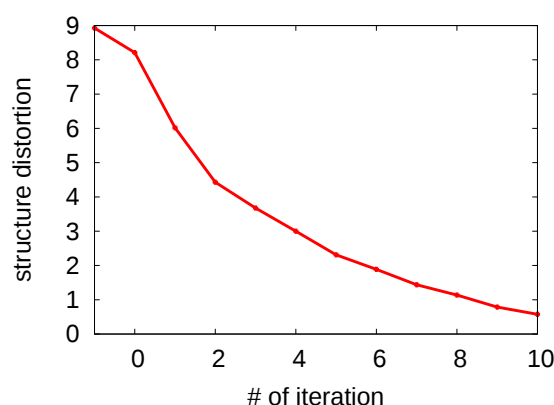


Fig. 5 反復回数と構造歪みの変化の関係

(a)(b) はそれぞれ M1, F2 の分析再合成音である。Fig. 4(c) および (d) はそれぞれ反復推定を行う前のもの [10], および提案手法により 10 回の反復推定を行ったものである。枠で囲われた部分は初期条件として与えた音響事象である。Fig. 4 から、反復推定により時間方向のスペクトルの連続性が向上しているとともに、(c) と比較して、(d) のスペクトルがより (b) に近くなっていることがわかる。また今回生成した音声サンプルについて予備的な聴取実験を行った結果、いずれのサンプルについても提案手法による品質の向上が確認できた。

また反復による構造制約条件の充足について、推定された音響事象が構成する構造と制約条件の構造との差異を Fig. 5 に示す。Fig. 5 は構造提供話者を M1, 初期条件提供話者を F2 とした場合の、推定の反復回数と構造差異の関係を表している。Fig. 5 から、提案手法によって構造の制約条件をより充足する音響事象が推定されていることがわかる。

5 おわりに

本稿では、構造的表象からの音声合成の高精度化のために、反復解法による推定法を提案した。従来の手法では個々の音響事象が独立に推定されており、制約となる構造を真に充足する音声を得る事が出来ていなかった。提案手法では推定された音響事象間の関係性まで考慮する事で、構造制約条件を満たし、かつ音響事象の時間的連続性を確保した音声を推定する事が出来る。実際に合成された音声は従来手法よりも聴感上の品質向上が確認できた。

今後の課題として、大規模な聴取実験により提案手法の有効性を確認する事が上げられる。また構造制約条件を数学的に定式化する事でより高速かつ高精度な推定手法を実現する予定である。加えてピッチや継続長を同様の枠組みに組み込むことで、人間の音声生成と対応のとれた自然な音声合成の実現を検討していく。

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, pp.4097-4100, 2008.
- [9] 齋藤他, 信学技報, SP2007-80, pp.55-60, 2007.
- [10] D. Saito et al., INTERSPEECH, pp.1837-1840, 2008.
- [11] 齋藤他, 信学技報, SP2008-40, pp.115-120, 2008.
- [12] M. Pitz, H. Ney, IEEE Trans. Speech and Audio Processing, Vol.13, pp.930-944, 2005.
- [13] 峯松他, 音講論 (春), 1-P-12, 2007.
- [14] 峯松他, 信学技報, SP2007-30, pp.37-42, 2007.
- [15] 錦戸, 党, 音講論 (春), 1-Q-28, 2007.
- [16] H. Kawahara et al., Speech Communication, 27, 3-4, pp.187-207, 1999.