3F3-5

音声の不変表象に基づく語ゲシュタルトの物理的解釈とそれに基づく幼児の音声模倣の実装

Physical interpretation of word Gestalt based on invariant representation of speech and its application to implement of infants' vocal imitation

齋藤 大輔[*1] Daisuke SAITO　朝川 智[*2] Satoshi ASAKAWA　峯松 信明[*1] Nobuaki MINEMATSU　西村 多寿子[*3] Tazuko NISHIMURA　広瀬 啓吉[*4] Keikichi HIROSE

[*1]東京大学大学院工学系研究科 Graduate School of Engineering, The University of Tokyo

[*2]東京大学大学院新領域創成科学研究科 Graduate School of Frontier Sciences, The University of Tokyo

[*3]東京大学大学院医学系研究科 Graduate School of Medicine, The University of Tokyo

[*4]東京大学大学院情報理工学系研究科 Graduate School of Information Science and Technology, The University of Tokyo

In this paper we propose a new framework of speech generation by imitating "infants' vocal imiation". Most of the speech synthesizers take a phoneme sequence as input and generate speech by converting each of the phonemes. However infants usually acquire speech generation ability without text or phoneme. As developmental psychology states, from the utterances of their parents, they acquire the holistic sound pattern of words, called word Gestalt, and reproduce it with their vocal tubes. In our previous studies, the word Gestalt was defined physically and a method of extracting it from an utterance was proposed. In this paper, a method of converting the word Gestalt back to speech is proposed and evaluated. Unlike a reading machine, our proposal simulates infants' vocal imiation.

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．このような枠組みを人間の音声言語活動と対比すれば，これは大量の読み上げ音声とその書き起こし文との対応を学習し，与えられた文章を目標話者の声で読み上げていることに相当する．

一方幼児の音声言語獲得過程を考えた場合，上記の枠組みとは本質的に異なっている．まず幼児が聞く音声はその大部分が「母親と父親の音声」である．幼児はこれらの音声を模倣することで音声言語を獲得するが，結果発声する音声はまさに「幼児の声」そのものである．幼児が両親の音声の音響的実体そのものを模倣することは声道形状の違いから不可能である．このとき幼児は両親の音声に対して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．

幼児は両親の声の何を真似ているのか．例えば両親の「おはよう」という音声を模倣する場合，[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．発達心理学はさらに「幼児は単語全体の語形・音形を獲得し，その後，個々の分節音を獲得する」と主張する [3]．この単語全体の語形・音形は語ゲシュタルトと呼ばれる [4, 5]．これまでに筆者らはこの「語ゲシュタルト」の物理的解釈となる，音声の不変表象を提案してきた [6]．また筆者らはこの表象を用いた音声認識システムについても検討を行ってきた [7]．本稿では音声の不変表象について説明し，さらにこの表象に基づき，幼児の音声模倣の実装としての新しい音声合成方式について述べる．

連絡先: 齋藤大輔（東京大学大学院工学系研究科）
dsk_saito@gavo.t.u-tokyo.ac.jp

2. 音声の不変表象と語ゲシュタルト

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話者や環境の違いと言った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．音声工学でスペクトル情報を効率的に表現するために用いられるケプストラム特徴量空間（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 [8]．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 $c' = Ac + b$ 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 音声の構造的不変表象

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

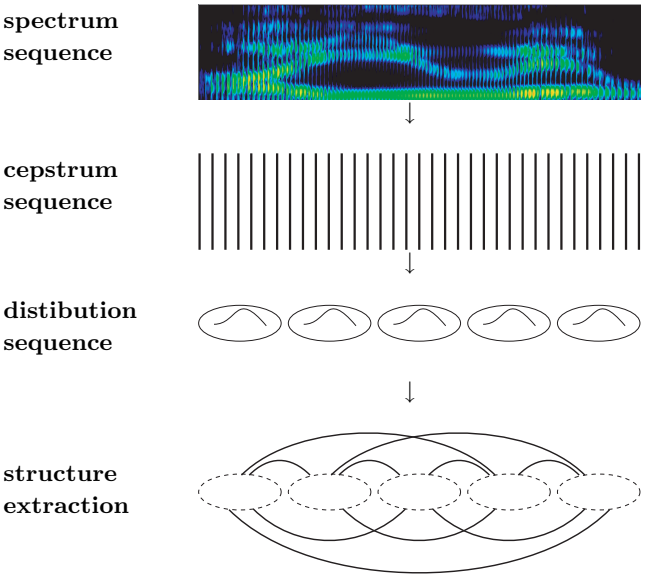


図 1: 音声の不変表象の抽出

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 [9]．

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，時間的に離れた事象まで含めて，これらの音響事象群の「**分布間距離**」関係を抽出することによって非言語性歪みにおよそ不変な音声の構造的不変表象を得る事ができる．このとき，個々の音響事象の絶対的な物理特性は一切捨象する．

2.3 一発声の語ゲシュタルト

一発声を一つの不変表象で記述する場合を考える．図 1 に一発声の音声からの不変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．このときこの音声の構造的不変表象は時間的に離れた事象も含めて一発声全体をモデル化しており，単語全体の語形・音形を捉えていると考える事ができる．さらに個々の音響事象の絶対的な物理特性を捨象しているため，これは話者不変の表象である．すなわち筆者らの提案する音声の不変表象によって語ゲシュタルトを物理的に解釈する事が可能となる．筆者らはこの音声の不変表象を音声認識に応用し，子供の声でも頑健に認識できることを確認している．

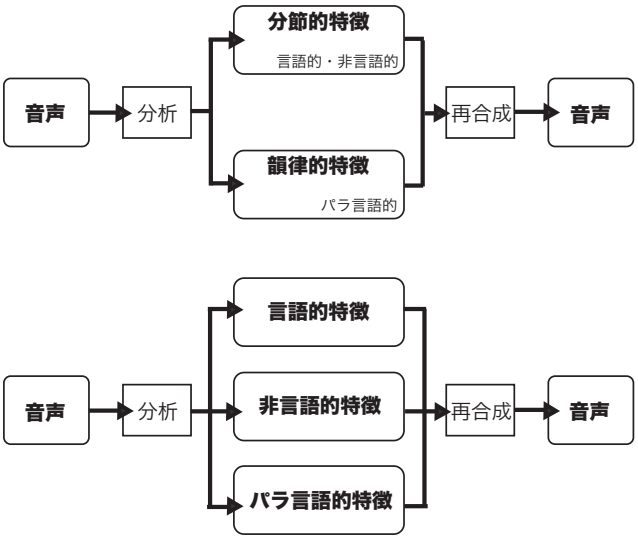


図 2: 従来の分析再合成系の枠組み（上）と提案する枠組み (下)

3. 幼児音声模倣の実装としての音声合成

3.1 非言語的要因をも分離する分析再合成系

本稿では音声の不変表象に基づいて幼児の音声模倣を解釈し，その実装としての新しい音声合成システムについて提案する．幼児の音声模倣では前述の通り，両親の発声全体の語形を真似，これを自らの声で返していると考えられる．これは語形の獲得に際しては，親の声から身体性を取り除き，発話時には発話したい語形に対して幼児の身体性，すなわち声道形状特性を与えることで音声が生成されていると解釈することができる．

この幼児の音声模倣に基づいて音声合成の枠組みを考える．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 [10]．音声の分析再合成系でこの考えを示すと図 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．音声の不変表象はこのうちの言語的特徴およびパラ言語的特徴を担う事になる．すなわちこの時，言語的特徴とパラ言語的特徴のみ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．すなわち提案する音声合成の枠組みは幼児の音声模倣の実装として捉えることができる．

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

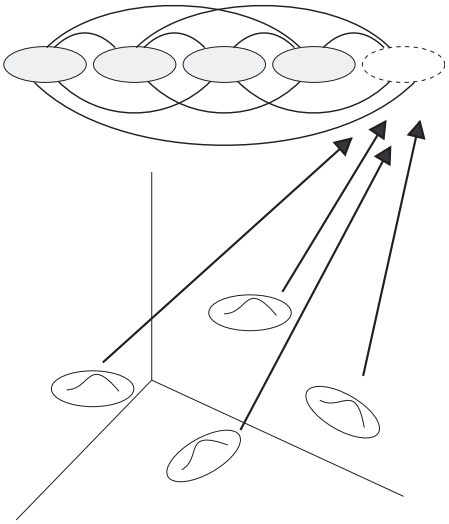


図 3: 解探索による不変表象を制約とする音声合成の枠組み

ム空間との対応関係も明確でない [11]．そこで提案する音声合成を実現するにあたり，ケプストラム空間の解探索問題としての定式化を行う．すなわちケプストラム空間において，不変表象の構造の制約条件に対して，既に生成された音響事象の絶対的な物理特性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．これは空間における不変構造を初期条件の絶対量によって音響空間に定位することに相当する．幼児の音声模倣で言えば発話の語形と声道形状特性（音響空間の座標系）をもとに音声を生成していることにあたる．

この枠組みを図 3 に示す．図の上部には発話対象となる語の不変構造が与えられており，既にいくつかの音響事象が生成されている．このとき不変構造を制約条件，音響事象を初期条件として次の音響事象を音響空間から探索する．

3.3 解析的手法に基づく解の導出

本節では上記のように定式化した探索問題を実際に解く事について言及する．

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．このとき一方の音響事象を固定したとき，式 (1) を満たすもう一方の音響事象の軌跡は多次元空間における楕円体を描く．すなわち初期条件として幾つかの音響事象が与えられた上で，不変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，楕円体の交点を求める問題として解釈する事ができる．これは上記の 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 $A=\mathcal{N}(a,V_a), B=\mathcal{N}(b,V_a)$ 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．

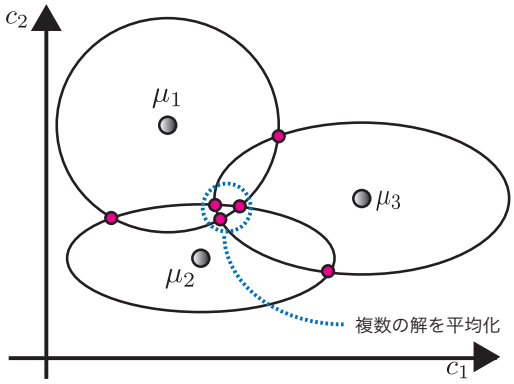


図 4: 解析手法に基づく解の導出（2 次元の場合）

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

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．

提案する解析的手法によるアプローチについて述べる．音響空間の次元数 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 ${}_nC_m$ 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図 4 に示す..

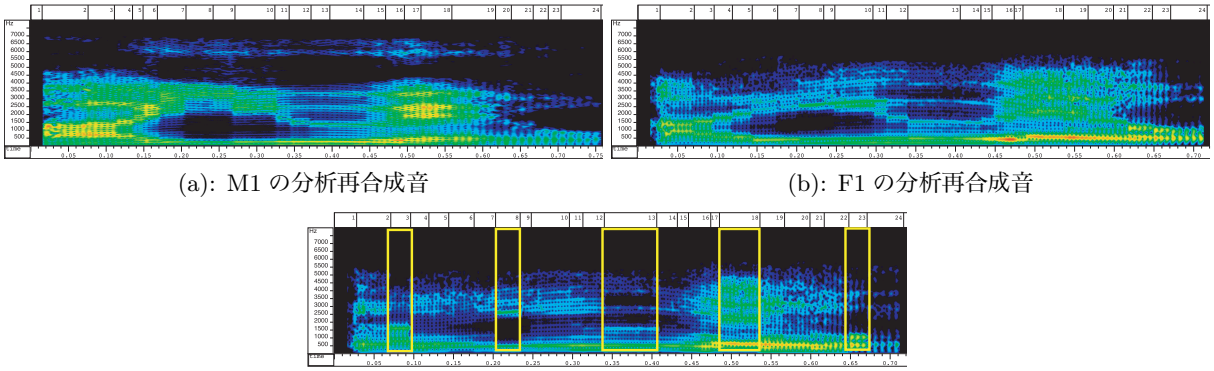
4. 合成実験

4.1 実験方法

提案する音声合成システムによって実際に音声が生成できることを示すため，日本語 5 母音の連続発声 /aiueo/ を用いて実験を行った．成人男性 2 名（それぞれ M1，M2）および成人女性 1 名（F1）の発声を収録した．これらの発声について STRAIGHT[12] 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話者 M1 の発声について図 1 の流れで不変構造を抽出した．ケプストラム系列から分布系列への変換には，一発声に対して MAP 推定に基づく HMM を用いる [7]．このとき一発声を 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表象を抽出した．この不変表象を模倣する単語の語形とした．

一方話者 M2 および F1 についても図 1 の流れで分析を行った．ただし不変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 から得られた不変表象を制約条件，M2，F1 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．最終的に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 個の分布を用いて STRAIGHT の枠組みで音声を再合成し



(a): M1 の分析再合成音　　(b): F1 の分析再合成音

(c): 提案手法による合成音

図 5: 実験結果

た．この実験は話者 M1 を親，話者 M2，F1 を幼児として音声模倣を実現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 5 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 F1 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 F1 の初期条件から合成した音声である．(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 (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 F1 の話者性を正しく再現できていることがわかる．これらの結果から不変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．

5. おわりに

本稿では幼児の音声模倣について考察し，その実装としての音声合成システムについて提案した．音声模倣において，幼児は両親の発話全体の語ゲシュタルトを獲得し，その後個々の音韻を獲得する．本稿ではこの語ゲシュタルトの物理的解釈となる音声の不変表象について説明し，これをもとに音声模倣の実装としての音声合成システムを構築した．さらに提案手法の妥当性を探索問題としての定式化と実験を通して示した．提案するシステムを従来の音声合成の枠組みと比較したとき，どちらがより適切に人間の音声言語活動を説明しているかは明確であろう．今後は幼児の言語獲得に重要な役割を果たしていると考えられる韻律的特徴についても，提案する枠組みのなかで取り扱っていく予定である．

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] 天野清: “子どものかな文字の習得過程”, 秋山書店, 1986.

[3] 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.

[4] 早川勝廣: “言語獲得と育児語”, 月刊言語, vol.35, no.9, pp.62–67,2006.

[5] N. S. トルベツコイ, “音韻論の原理”, 岩波書店, 1958.

[6] N. Minematsu *et al.*: “Theorem of the invariant structure and its derivation of speech Gestalt,” SRIV2006, pp.47–52, 2006.

[7] S. Asakawa *et al.*: “Multi-stream parameterization for structural speech recognition,” ICASSP2008, pp.4097–4100, 2008.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.