# 合成音声における個人性表現のための音声特徴量の抽出

杉浦 千加志*, 藤嶋 梓*, 坂田 聡**, 上田 裕市* （熊大院自*, 熊大工**）

## 1 はじめに

音声合成システムはマンマシンインターフェイスとして有効な手段であり、その多くは規則合成方式による TTS（Text-To-Speech）が主である。最近の合成方式の特徴は大量の実音声波形素片からデータベースを作成しているので音質は良いが、データベース作成に比較的多くの時間と労力を要してしまうので、例えばある話者の声を合成させたいといった要求に対し、対応が困難であるという短所がある。本研究はこの問題を解決すべく、少数の実音声試料を分析,利用することでその話者の個人性を有する音声合成手法の確立を目的としている。ここでの個人性とは声の質を表しており、イントネーションの個人差による質の違いにつては現段階では考察の対象としない。

一般に音声は定常的発声による母音と主に過渡的発声による子音との 2 つに大別できる。本研究では子音の多くは母音発声時の定常的声道の形状を基本とし、その形状を変化させることなどにより合成可能であると仮定した。また子音よりも母音の方が音声信号中に占める割合が高いことから、先ず個人性特徴量を母音音声の質に求めることとした。母音は声帯で生成された準周期的な音源信号が声道において共振され口唇から放射されることで発声される。フォルマント合成はこの一連の生成過程をモデル化したものであり、そのモデルパラメータが母音の個人性を特徴づけるものであるので、本研究では音声の合成方式としてソースフィルタモデルに基づくターミナルアナログ合成方式を採用した。

本稿はフォルマント合成によって任意の話者の母音音声をその個人性特徴を損なうことなく合成することを中間目標とし、それを実現するため実音声からの音声特徴量の抽出方法とその結果について報告する。

## 2 原理

### 2.1 合成方式と個人性特徴量について

本研究では音源特性と放射特性を声帯音源微分波として LF モデル[1]によってモデル化し、これを駆動部とするソースフィルタ方式によって音声を合成している。声道特性には単共振フィルタを 6 つカスケード接続したもの（以下声道フィルタ）を用いた（図 1）。単共振フィルタの連結数 6 は本研究での音声のサンプリング周波数 12kHz（帯域 6kHz）において観測され得る共振の最大数としている。

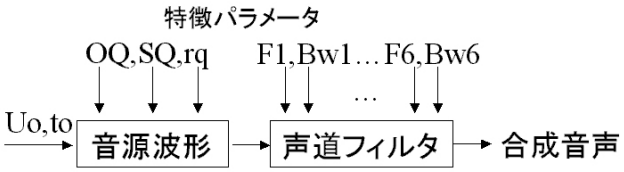

図 1：合成方式の簡易図

音源モデルの特徴を表す波形パラメータは声帯開放時間率 OQ,声帯開放速度率 SQ,負

Speechfeatureextractionforexpressionofspeaker ’sindividuallyinsynthetic speech.

By Chikashi Sugiura, Azusa Fujishima, Yuichi Ueda, Tadashi Sakata.
（KumamotoUniversity ）

の立ち上がり部分の割合を表す rq の 3 つであり、それ以外のパラメータに振幅 Uo とピッチ周期 to がある。声道モデルの特徴を表すパラメータは第 1 から第 6 までのフォルマント周波数ならびに帯域幅の計 12 個のパラメータである。このモデルによって音声を合成するには、まず音源モデルにおいて特徴パラメータ OQ,SQ,rq を固定値で与え、振幅やピッチを制御することによって音源微分波形を生成する。その音源微分波形を個人毎のある母音の特徴パラメータを与えた声道フィルタに通すことにより音声を合成する。

2.2 個人性特徴量の抽出方法

個人性特徴量の抽出は声道モデリング部と音源モデリング部の２つから成る。予備実験として声道パラメータに逆フィルタ制御法[2]によって推定したフォルマント値を使い、音源波形における個人差がそれ程大きいものではないという推測のもと音源パラメータに一定値を使い音声を合成した。しかしこの方法ではフォルマント（特に帯域幅）の推定誤差が大きい場合、合成音声と実音声とのスペクトル誤差が比較的大きく、結果としてそれが合成音声と実音声との聴感上の差異や波形の違いにつながっていた（以後従来手法)。そこでこの聴感上の差異を低減すべく、合成音声スペクトルと実音声スペクトルとの整合性を考慮した抽出処理を提案した。図２にそのブロック図を示す。

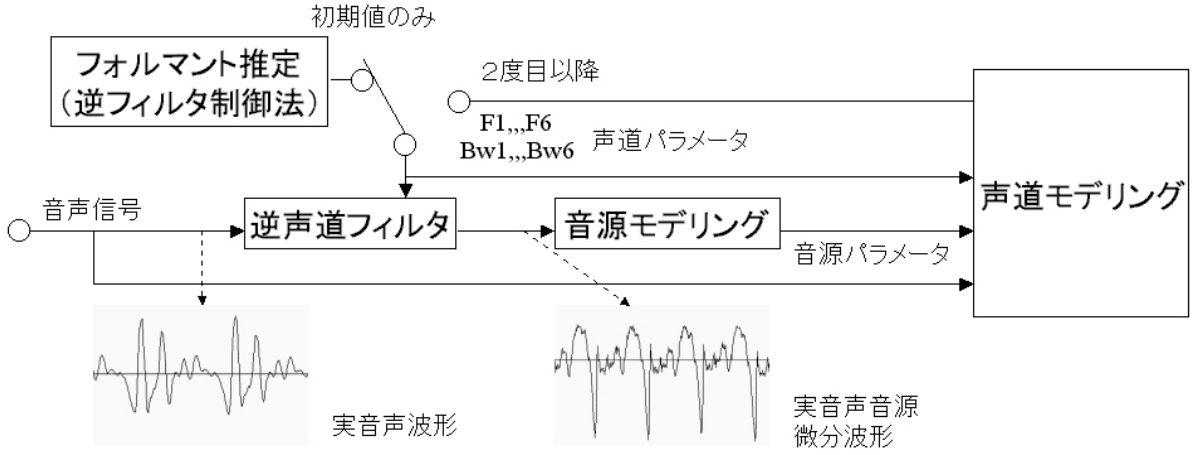


図２：提案する特徴量抽出処理のブロック図

この提案手法の改良点は、声道モデリング部では前述した合成音声と実音声のスペクトルの整合性を高めるために GA(GeneticAlgorithm )[3]によって声道パラメータ空間（12 次元）での最適パラメータ探索を行う点、音源モデリング部では音源波形における個人性を考慮するため音源パラメータ抽出を行う点、さらにこれら 2 つの処理結果は互いに依存関係にあるため（後述）両モデリングを数回繰り返しその中で実音声と合成音声のスペクトル距離が最小であるパラメータの組を最終的な結果とするという点である。

この手法ではまず音源モデリングを行う。音源モデリング部は逆声道フィルタに実音声を通して実音源微分波形を生成し、それと LF モデル音源微分波形との二乗誤差を最小にするパラメータセットを探索するという方法で音源パラメータを抽出する。逆声道フィルタの係数であるフォルマント値の初期値には逆フィルタ制御法によって得られた値を用い、2 回目以降は声道モデリングの結果を用いる。次に声道モデリング部では抽

出された音源パラメータを使って LF モデル音源波形を用意し、実音声と合成音声のスペクトルを誤差評価の対象とする GA によって声道パラメータを抽出する。以下この処理を数回繰り返し、最良のパラメータセットを最終的な抽出結果とする。

### 2.3 GA と A-b-S 法を使った声道モデリング

声道モデリングのための GA では個体数 500、個体ビット数は（各パラメータ 8 ビット）96 ビット、世代数 200、選択はルーレット選択とエリート保存選択により行い交叉は交叉確率 0.4 で 2 点交叉、突然変異確率を 0.06 とした。パラメータ毎の値の範囲については、フォルマント周波数はある程度所望の値付近である必要があるので音源微分波生成時の値を中心とする前後 50Hz とし、帯域幅は 10Hz から 20000Hz の範囲での対数周波数値（1.0 - 4.3）とした。初期個体群はランダムに生成した。適合度は処理内部で音声を合成し、それを分析して得られたスペクトルと同じ分析によって得られた実音声スペクトルとの二乗誤差の逆数とした。これは分析誤差による抽出精度の低下を防ぐ為である。ここでの分析とはスペクトル分析のことを指すが、同時にピッチによる高調波成分の除去（スペクトル平滑化）も行っている（図３）。誤差計算は実音声と合成音声とのスペクトル距離を、スペクトルピーク付近の重みを重くする二乗誤差によって算出した。抽出する分析フレームは 256 点(21.33[ms])とした。

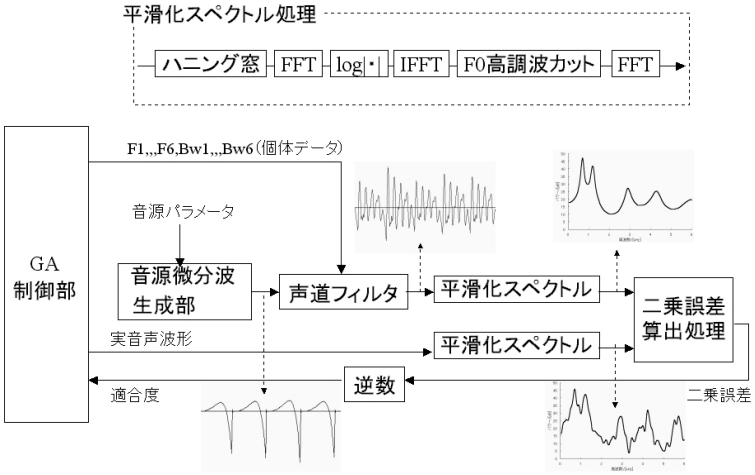


図３：GA における適合度算出ブロック図

## 3 評価実験

### 3.1 パラメータ抽出例

抽出実験の音声試料にはある成人男性の単母音音声５母音、時間長約１秒を用意し、その中のあるフレームデータを用いた。音声資料はサンプリング周波数 48kHz、防音室内で録音したものをカットオフ周波数 6kHz の FIR ローパスフィルタに通したあと 12kHz にダウンサンプリングして作成した。表１に従来手法と提案手法それぞれによって特徴パラメータを抽出した値を使って合成した音声と実音声とのスペクトル距離を示す。表より提案手法の方が実音声とのスペクトル誤差が近いことがわかる。

表１：誤差評価値例（単位[dB]）

| | 従来手法 | 提案手法 |
|---|---|---|
| /a/ | 3.40 | 2.45 |
| /i/ | 5.23 | 2.04 |
| /u/ | 3.84 | 3.17 |
| /e/ | 2.66 | 2.08 |
| /o/ | 2.94 | 1.59 |

3.2 聴取実験

従来手法と提案手法それぞれの手法によって抽出された特徴パラメータを使って合成された音声と実音声との聴取比較実験を行った。音源微分波形の振幅は一定値、ピッチパタンは実音声を分析して得られたものを用いた。これはピッチの違いによる聴感上の差異を無くすためである。実験環境は防音室で呈示方法はヘッドホンによって行った。被験者は成人男性 7 名成人女性 3 名の計 10 名、実験方法は[実音声 - 合成音声 1 - 実音声 - 合成音声 2]の順に聴いてもらい、評価方法を 1 の方が実音声に近い,2 の方が実音声に近い,どちらでもないの 3 つから選択してもうこととした。合成音声 1 と 2 は従来手法による合成音声と提案手法による合成音声をランダムに振り分けた。実験結果を図 4 に示す。この結果から提案手法が良い結果を示すものもあれば、従来手法とほぼ変わらないものもあることがわかる。この原因のひとつに提案手法による第 1 フォルマントの帯域幅が従来手法と比較して大きな値として抽出される傾向にあることが考えられる。誤差評価はスペクトルピーク付近に重みを付しているため、ある程度人間の聴覚特性を考慮しており、また表 1 からスペクトル誤差は従来と比較し小さくなっている。しかし帯域幅が大きな値として抽出されるという結果から、この誤差評価における重み関数の重みの度合い、または重みの付け方が結果として聴感上の差異につながってしまったと考察される。今後この問題を解決するための方法として重み関数の改良や、誤差評価式そのものを改良することが考えられる。

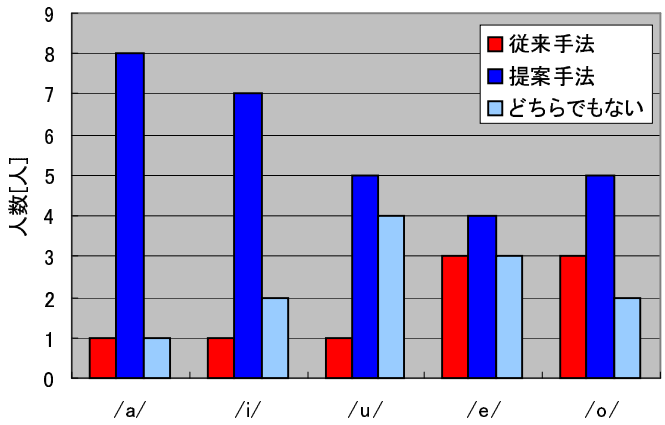

図 4 ：聴取実験結果

4 まとめ

個人性を有する音声合成手法の確立を目的とする研究において、それを実現する為の音声特徴量抽出手法について報告した。声道パラメータの抽出方法は発声器官のモデルに基づいたターミナルアナログ合成方式において GA を使ったものであった。結果は良好なものもあるが従来手法と比較し完全に優れているとは言えず、まだ改良の余地がある。今後は誤差評価方法の改良により精度の高い特徴パラメータの抽出を行う予定である。

参考文献

[1] G.Fant and Q.Lin, "Frequency domain interpretation and derivation of glottal flow parameters", STL -QPSR, 2-3 pp.1 -21(1988)

[2] Akira Watanabe, "Formant Estimation Method Usin g Inverse -Filter", IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING Vol.9 pp.314-326(2001)

[3] 坂和正敏,田中雅博 共著, “ソフトコンピューティングシリーズ 1 遺伝的アルゴリズム”, 日本ファジィ学会