電子情報 67

フォルマントの高精度推定に基づく高品質かつ柔軟な音声合成

西澤信行

内 容 梗 概

本論文では、AR-HMMモデリングに基づく高精度な母音音声の分析手法を提案し、これを用いた高品質かつ柔軟な音声合成の実現に関する検討を行う。今日、比較的容易に高い品質が得られることから、テキスト音声変換システム等では波形接続方式による音声合成が広く用いられている。しかし、朗読調でない音声、例えば対話音声や感情音声の合成には、韻律的特徴の自由な制御が不可欠であり、さらには声質の制御も要求される。これを波形編集方式で実現するためには、非常に大量の波形データの蓄積が必要であり、将来的に蓄積の大きさというハード面の問題が無視できるとしても、音声収録の負担は依然として大きな問題である。

一方、大規模な蓄積を必ずしも必要としない音声合成方式として、音声生成過程を音源と調音フィルタに分解して考えるソース・フィルタモデルに基づく手法が知られている。ソース・フィルタモデルに基づく音声合成において、両者の特性を明確に捉えることができれば、それぞれを独立に制御することによる柔軟な音声合成の実現が期待される。従来、ソース・フィルタモデルに基づく多くの音声分析合成システムでは、線形予測分析等の手法により、自然音声波形を白色化するフィルタパラメータを求め、合成時には、その逆特性を調音フィルタの特性として与え、一方、音源には、簡単のためにインパルス列と白色雑音源を組み合わせたものが用いられることが多かった。

しかしこのようなモデルで、音源と調音フィルタを制御することは実際には容易ではない。なぜならば、モデル上の音源は実際の音声生成過程における音源の特徴の一部しか表現しておらず、調音フィルタにも生成過程における音源由来の成分が含まれているため、結果的に、音源と調音フィルタを独立に制御することが出来ないためである。このことを無視して独立に制御した場合、例えば音源の基本周波数を大きく変化させた際に極端な品質低下が起きるといったような問題が生じる。従って、生成機構における音源と声道伝達特性という形で自然音声の特徴を分離することができれば、もちろん両者は相互に関連しており、完全に独立には制御できないにせよ、従来よりも独立に制御した際の品質の低下を抑える

ことができると期待される。

そのような音源・フィルタ分離を行う音声合成手法として、声帯音源波形形状を数式により表現したモデルである声帯音源波形モデルを駆動音源として用いる、フォルマント合成 (ターミナルアナログ方式による音声合成) は代表的な手法である。特にフォルマントは音声の周波数領域における特徴を記述する上で比較的優れた特徴量であり、これをパラメータとするフォルマント合成は、パラメータを広い範囲で操作した場合においても合成音声品質の低下が小さいため、柔軟な音声合成に適していると考えられる。

フォルマント合成により、特に母音型音声については、声帯音源波形モデルの利用と、ARX(Auto-regressive with exgenious input) モデルにより合成回路のモデル化を行い自然音声から合成器のパラメータを推定する手法により、比較的高い品質の合成音声が得ることができている。しかし、フォルマント合成には幾つかの問題がある。その1つは分析に基づく子音波形の生成が困難であること、もう1つは母音についてもフォルマントの制御方法が明らかではないことである。本論文ではこれらの問題について論じる。

まず、子音波形生成の困難さについてであるが、子音については生成過程との対応性を求めると、比較的複雑な構成の合成回路が必要となり、自然音声波形からそのパラメータを精度良く推定することが困難であった。そこで、本論文においてはまず、合成システム開発を容易にすることを目的として、分析が困難な子音については自然波形を直接利用する手法について検討を行った。そして知覚実験の結果から、そのような2つの音声合成手法を組み合わせたことによる、極端な品質低下が生じないこと、また、音声合成における柔軟性は母音合成において重要であり、波形利用により子音合成の柔軟性が失われた場合においても、合成システム全体として柔軟性を有していることを確認した。この結果に基づき、本論文において以降では主に母音型音声をその対象とする。

一方、フォルマントの制御方法に関する問題に対しては、将来的にはパラメータ制御に統計的なモデルを利用することが考えられる。近年、音声認識で用いられる HMM(隠れマルコフモデル) を音声合成に用いる、という手法が注目されている。この場合、音声認識で広く用いられている音響特徴量であるメルケプストラムを音声合成にも用いる手法が一般的であるが、パラメータとしてフォルマントの周波数・帯域幅を用いることも可能である。メルケプストラムと比較し、フォルマントは特に母音音声の音響的特長を良く表現する特徴であり、より少ない音声サンプルで、より柔軟な音声の合成が期待される。しかし、メルケプストラム

の推定とは異なり、フォルマント合成のパラメータ推定に必要な、声帯音源波形モデルとARXモデルに基づく音源・フィルタ特性の同時推定問題は非線形問題であり、安定した分析結果を得ることが容易ではない。そのため、音声合成に必要な高精度なモデルを作ることが困難である。

そこで本論文では、声帯音源波形モデルを用いず、より自由度が高く取り扱いが容易なループ状のHMMをAR過程の分析誤差のモデルとして用いる手法であるAR-HMMモデリング基づく、より安定した音源・フィルタ特性の分離手法を提案する。AR-HMMモデリングは佐宗らにより提案された手法であり、線形予測分析において定常的な白色雑音と仮定されている残差波形の統計的性質を、HMMで表現することで、より精密にモデル化を行う手法である。この手法により、周波数軸上において調波成分として現れる、音声波形の周期性の影響がより精密に表され、音声のスペクトル包絡特性推定の際に、線形予測分析において問題となる音源の基本周波数の影響を受けにくい音声分析が実現される。AR-HMMモデリングで用いられるループ状のHMMは周期性を有する音源波形の表現に適したものであるが、音源波形の特徴を表すための制約としては不十分であるため、本論文では、AR過程により表現される極配置を制限することにより、周期性の分離だけでなく、より生成機構との対応性に優れた音源・フィルタ特性の分離を行う手法を導入する。この際の制約条件としては、声道伝達特性が共振特性のみの積で表現されるという仮定を採用した。AR-HMM推定自体が逐次近似によるパラメータ推定を必要とするが、提案手法においては、AR-HMM分析は反復的に行われ、モデルにおけるAR部に実極が現れなくなるまで、AR次数を減らしていき、一方でその分の特徴はHMMにおいて表現されるように分析は誘導される。この手法は、線形予測分析の結果得られる複素共役な極、すなわち共振特性と、声道伝達特性における共振特性との間に対応関係がある、との前提に基づくもので、本手法により、音源特性の影響が含まれない、よりソース・フィルタモデルによる音声合成に適したフォルマントの特徴が推定される。

そして提案手法の妥当性を評価するための実験を行った。まず自然発話中に含まれる母音音声に対し、線形予測分析、AR-HMM分析、提案手法でそれぞれ分析を行い、各母音ごとに、推定された音源特性・フィルタ特性の32次ケプストラム空間におけるパラメータの広がりについて調べた。その結果、提案手法により、分布の小さいパラメータが得られることが確認された。さらに各ケプストラム次数毎にその分散を調べたところ、他手法と比較し、1次のケプストラム係数の分散が特に小さくなっていることが判った。1次のケプストラム係数はスペクトラム傾斜

成分に大きく関係するパラメータであり、音声のスペクトル傾斜は主に音源特性に由来するものであることから、より適切に音源特性を取り扱うことができる、と考えられる。また、フォルマント合成による母音音声に対し、線形予測分析と提案手法で分析を行い、それぞれ比較した。結果、提案手法は逐次近似推定でありながら、線形予測分析と比較し、より安定した分析結果を返すことが判った。このことより、提案手法は、大量の音声分析に有効な手法であることが明らかとなった。ところで音声合成においては、最終的な評価は合成音品質が基準となる。このため、客観評価だけでなく主観評価が重視される。提案手法による分析の妥当性を評価するため、推定フォルマントに対する逆フィルタ波形により駆動されるフォルマント合成器を構築し、それを用いた分析再合成音に対する評価を行った。この際、TD-PSOLA 法により音声波形自体にピッチ変換を施したものと、提案手法により分離された推定音源波形に対しピッチ変換を行い、この波形でフォルマント合成器を駆動したものを比較し、ピッチ変換率の点で分析合成手法が優れていることを確認した。ピッチ変換に対して有効な同様の手法は他にも存在するが、それらの手法の多くがノンパラメトリックなスペクトル包絡表現となっているのに対し、本論文における分析合成系ではパラメトリックな表現となっており、ある程度の合成音品質を保ったまま、合成音声のスペクトル包絡を自由に制御することが可能であることが示された。

目 次

図 目 次

表 目 次

第１章

序論

1.1 本研究の背景

音声は人間が用いる相互の情報伝達手段のうち、最も基本的なものの一つである。道具や特殊な能力は不要であるという特徴をもち、また、他の手段との併用も可能であるため、その応用範囲は広い。そして、これを人間と機械との情報伝達手段として用いようとする試みも、以前から行われている。

これまで、音声合成技術は一般に言語情報の伝達をその主たる目的としており、朗読音声の合成を対象としたものが一般的であったが、近年では、対話音声や感情音声のように、言語情報以外の情報を含む合成音声のニーズが高まりつつある。音声合成技術もまた、この要求を支えるべく、より柔軟な音声の合成へ向けた研究が期待されている。

これまで様々な音声合成の方式が検討されてきたが、計算機の利用可能な記憶容量の大幅な増加に伴い、波形レベルでの大量の音声の蓄積を用いた波形接続方式による音声合成が、比較的容易に実現できるようになった。

波形接続方式は、あまり複雑な信号処理を行わないことにより高い品質の得られる方式であり、今日最も主流の音声合成手法であると言える。しかし、合成音声の声質の制御等、より柔軟な音声の合成を行うためには、あらかじめ蓄積しておく波形素片の数は更に莫大なものとなり、その作成に要するコスト等の点からも、その実現が非常に困難である、という問題がある。

これに対し、波形以外のより効率的な音響的特長を用いて音声を扱う手法が考えられる。そのような手法として、信号処理的な音声分析に基づきその特徴を抽出し、合成時にはその特徴量を用いて合成する手法である合成分析方式がある。分析合成方式では特徴パラメータを操作することで声質の変換等も可能であるが、特にパラメトリックな特徴量に基づく分析合成方式ではパラメータの非線形な制御ができることから、例えば、一部の音素の蓄積から、合成に必要な全てのパターンを生成することも可能であると考えられる。

しかし、特にパラメトリックな特徴を利用する分析合成方式の品質は、現時点では波形接続方式や、ノンパラメトリックな特徴に基づく分析合成方式と比較し一概に低く、さらなる品質の向上が望まれている。さらに、パラメータの非線形な制御のためには音韻・声質との間との関係がより明確な特徴量を用いて音声を扱うことが望ましいと考えられる。このためには、現在分析合成方式でよく用いられている音源をインパルス列や白色雑音と仮定し、調音フィルタに全極型のフィルタやLMA(対数振幅近似)フィルタ[4]を用いるといった様な、比較的単純なモデ

ルに基づくものではなく、声帯音源波形モデルと極零型フィルタを組み合わせる等の、より直接的に音声の生成機構と対応するモデルを用いることが必要になる。

ところが、これらのモデルは一般に非線形なものとなり、そのパラメータの自動推定は容易ではない。推定においては、音声の特徴に関する知見から、厳密な制約条件を定めること等が有効であると考えられるが、しかし、現在、自動分析に必要なだけの知見を持っているとは言い難い状況にある。

実際、例えばフォルマント音声合成では、自然音声の分析結果に基づく合成器制御を行うことができず、制御方法を合成音の品質を基準として経験的に決めざるを得ない場合が少なくない。すなわち、実際には分析合成方式として成立していない場合が生じる、ということである。そしてこの結果として合成音品質は一般に低下する。

確かに、合成音声の品質が高いものであるならば、自然音声の特徴を反映していなくても良い、という考え方も可能である。しかし、その場合、音声の品質を向上させるためのパラメータは、主観評価に基づき、試行錯誤的に決定することになり、その試みが仮にうまくいくとしても、それに至るまでのコストは非常に大きなものとなることが予想される。

一方、現在 sinusoidal modeling[5] や STRAIGHT[6][7][8] といったノンパラメトリックな分析合成手法により、かなり高品質な合成音声が得られている。しかし、これらの方式はパラメトリックな分析合成方式と比較し、音声変形における柔軟性では大きく劣る。

今後、特に音声合成における柔軟性が重視されると予測されるため、現在有望であると考えられる音声合成研究のアプローチはパラメトリックな分析合成手法の高品質化であると考えられる。しかし、これまでに挙げたような問題を解決する必要があり、そこで本研究においては、合成方式のモデルの再検討から始め、柔軟かつ高品質な音声合成方式の構築を目指す。

1.2 本研究の目的

柔軟かつ高品質な音声合成を実現するためには、既存の音声合成手法それぞれについて、長所・短所を詳細に検討し、新たな音声合成の枠組みを構築する必要がある。

本研究においては、まずパラメトリックな手法のメリットが、今日では直接的

な情報圧縮ではなく、音声変形の可能性であることを重視する。この観点では、パラメトリック表現のために複雑なモデルを構築した場合に、パラメータ間の関係が複雑なものとなり、結果的に実用合成システムにおいてパラメータの制御ができないのであれば、そのような特徴をパラメトリックに表現する意味はないことになる。

最終的に本研究において、パラメトリック表現が困難かつ適当でない特徴について積極的にノンパラメトリック表現を導入する。これによりパラメトリックな音響的特徴に基づく分析合成手法のロバスト性を上げシステム構築に要するコストを下げることにより、高い柔軟性と高い精度の両方を備えた合成システムの構築を可能とすることを目指す。

特に柔軟性は重要である。想定する具体的な状況として、合成可能な声質 (あるいは話者性) を追加する際に必要な波形素片数を減らすことが挙げられる。現在の自然音声波形素片接続の枠組みでは、例えば合成可能な話者性を追加しようとすると、特に音声波形収録のコストについて、新規に合成システムを開発するのと同様のコストを要する。今後も、多様な表現のために、収録音声波形素片数は増えていく傾向があり、収録のコストが占める割合も増加しつつある。具体的にはこれを下げることが目的の1つであるといえる。

このために本研究においては具体的に次の2つに着目する。1つがパラメトリック表現が適当ではない音素の検討であり、もう1つが、パラメトリック表現が適当な音素において、パラメトリック表現が適当ではない音響的特長の検討である。

1.3 本論文の構成

本論文の構成は次のようになっている。

第2章では音声合成における波形生成手法の研究動向について、その背景としての音声の基本的性質に関する知見や信号処理技術等も含め述べる。第3章では子音波形に波形接続を用いる音声合成方式の検討を行う第4章ではAR-HMMモデリングに基づく母音音声分析手法について論じる。第5章では第4章で論じた音声分析の自動化を実現するための手法を提案し検討を行う。第6章では高品質かつ柔軟な音声合成を実現するための分析合成系の検討・検証を行う。そして、最後に第7章にて結論を述べる。

第 2 章

音声合成における波形生成手法の研究

動向

2.1 はじめに

本章では音声合成における波形生成手法に関し、その動向について、関連する音声学的知見、合成システム全体から見た際の位置付け、信号処理技法、合成回路制御手法など関連する領域も含め概観する。

2.2 音声の基本的性質

大規模蓄積を用いずにより柔軟な音声合成を行うためには、音声の性質を利用していくことが必要である。本節では音声の基本的な性質について、音響的な性質と生成機構の特徴の両面から述べる。

2.2.1 音声の音響的性質

音声のスペクトルを観察すると、一般に数個の共振特性が見られる。これは声道の共振特性に対応したもので、フォルマントと呼ばれる。また各フォルマントは周波数の低いほうから第1フォルマント、第2フォルマント、…と呼ばれる。

また音声のうち、声帯振動を伴うものを有声音、伴わないものを無声音と言うが、有声音の場合には声帯振動の周期性によるハーモニックな成分がスペクトルに見られる。声の高さは、一般にこの声帯振動の振動周波数により決まり、この振動周波数は F_0 とも呼ばれる。

一般にフォルマントが音素を特徴づけており、特に母音の場合は、第3フォルマントまでで、その音響的特徴を表現することができる。しかし、実際には同一音素でも話者により大幅に変動することが知られている。

また、会話音声のように連続して発声される場合は、声道の形が急に変化できないために、フォルマントは前後の音素の影響を受ける。この現象は一般に調音結合と呼ばれる。

さらに、後述するように、鼻音のように鼻腔への分岐が生じたり、破裂音や摩擦音のように、音源位置が声道の中間にあるような音声では、声道伝達特性に零点の特性が見られる。しかし、特に子音は定常的な部分を持たないことが多く、母音との調音結合の影響を受けて、その音響的性質は大きく変動する。

ちなみに、音声は一般に非定常なものであるが、特に有声音の場合は、20ないし40ms程度の区間で見た時は、ほぼ定常と見なすことができる。

2.2.2 音声の生成機構

人間の音声生成は一般に、音源の発生、調音、放射の3段階より成る。人間の発声器官は図2.1に示すように、肺(lung)、気管(trachea)、喉頭(larynx)、咽頭(pharynx)、鼻腔(nasal cavity)、口腔(oral cavity)等から成っており、これらは全体として一つの連続した管を構成している。喉頭より上の部分は声道(vocal tract)と呼ばれ、顎、舌、口唇などを動かすことにより、種々の形に変化する。鼻腔は軟口蓋を持ち上げることにより、咽頭や口の奥から遮断される。

腹筋が横隔膜を押し上げることにより、肺から押し出された空気は、気管を通った後、咽頭の声門(glottis)を通る。通常の呼吸の時は、声門すなわち左右の声帯(vocal cords)の間隔が大きく開いているが、声を出そうとするとき、声帯が接近しこの間を空気が通り抜けようとするため、空気流と声帯の相互作用により、声帯が周期的に開いたり閉じたりして、ほぼ規則的な空気の断続が生ずる。これが音声の音源であり、これを声帯音源(glottal source)という。声帯の緊張が大きく、かつ肺からの空気圧が高いとき、声帯の開閉周期、すなわち振動周期が短くなり、音程が高くなり、逆の場合には低くなる。これが声の高さであり、この振動周波数を一般に基本周波数(fundamental frequency)、あるいはF_0という。このF_0の時間的な変化により、アクセントや、イントネーションが付加される。

以上は母音型音声の生成過程であるが、他にも二種類の音声の生成過程がある。一つが、舌や口唇によって声道のある部分に狭い場所を作り、そこを空気流が通り抜ける時に乱流を生じさせ雑音的な音を生成するものであり、これは摩擦音と呼ばれる。もう一つの方法は、舌や口唇で声道を遮断し空気流を一時的に止め、圧力が十分高まったところでこれを急に開放しインパルス的な音を生成するもので、これは破裂音と呼ばれる。これらの音の生成は、声帯の振動の有無とは独立に行われる。

この他、軟口蓋が下がり、口腔のいずれかの位置で空気流を遮断することで、鼻腔にも空気流が供給され、声道に分岐が生じた形になることがある。このときの音声を鼻音という。また、母音発声時に軟口蓋が下がり鼻腔が声道の一部を形成することもある。この時の音声を鼻音化母音という。

図2.2は母音の、図2.3は子音のスペクトルの例である。

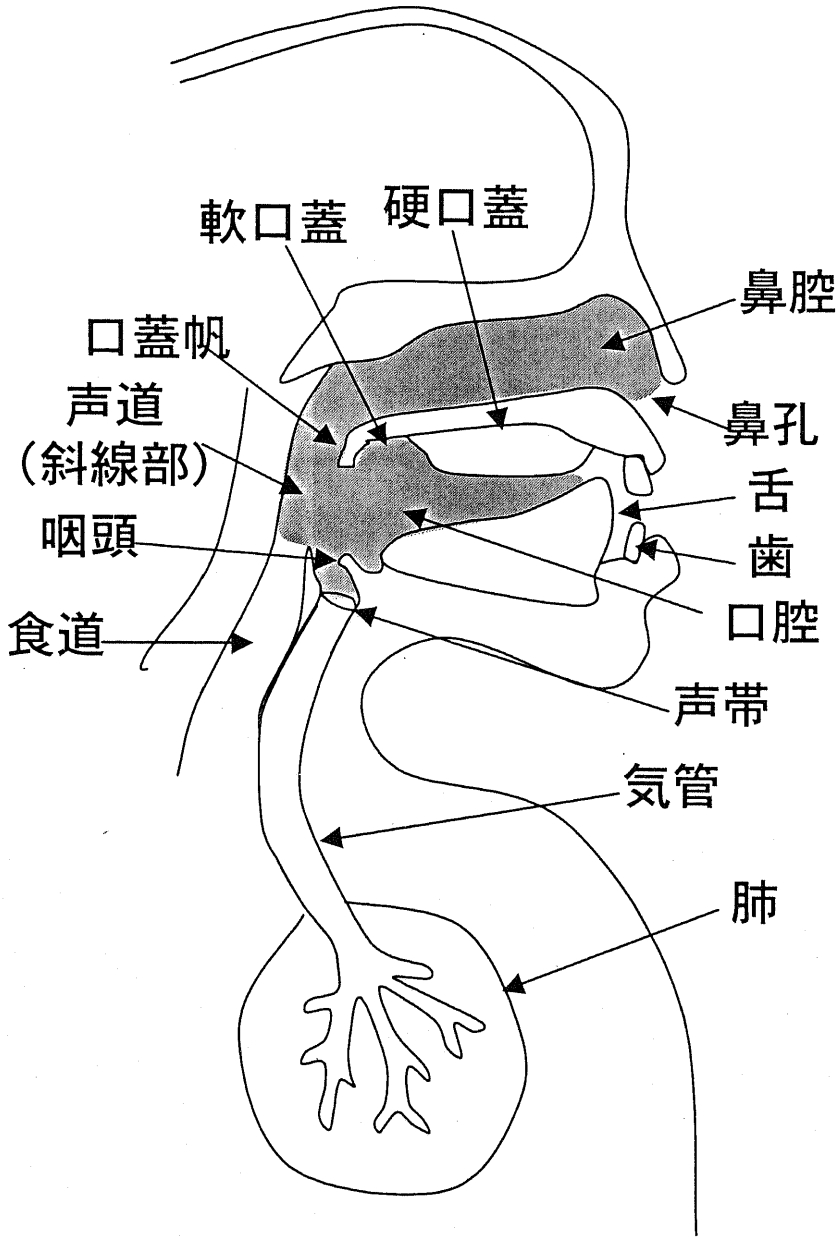


図 2.1: 人間の発声器官 [2]

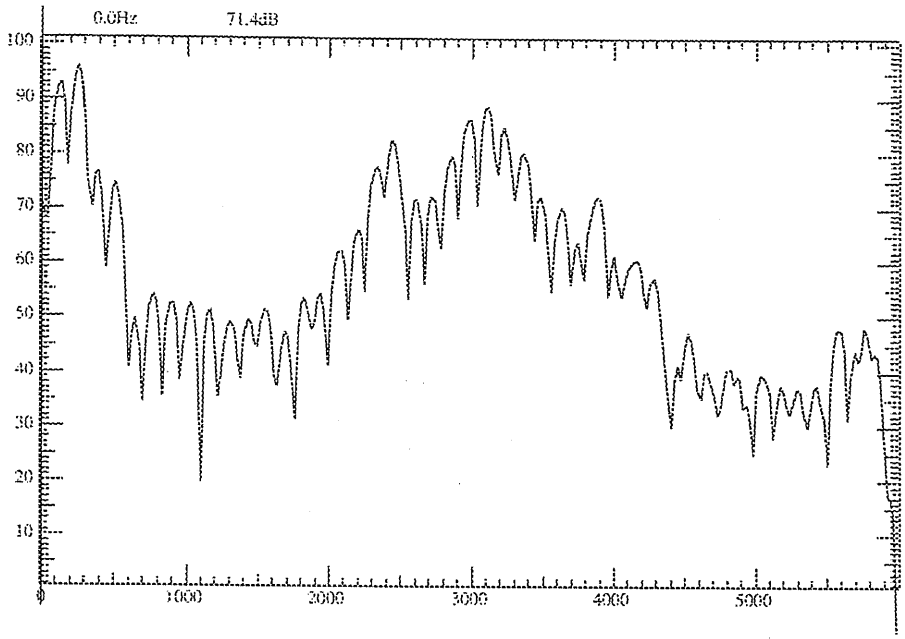


図 2.2: 母音の例 (男声 /isu/)

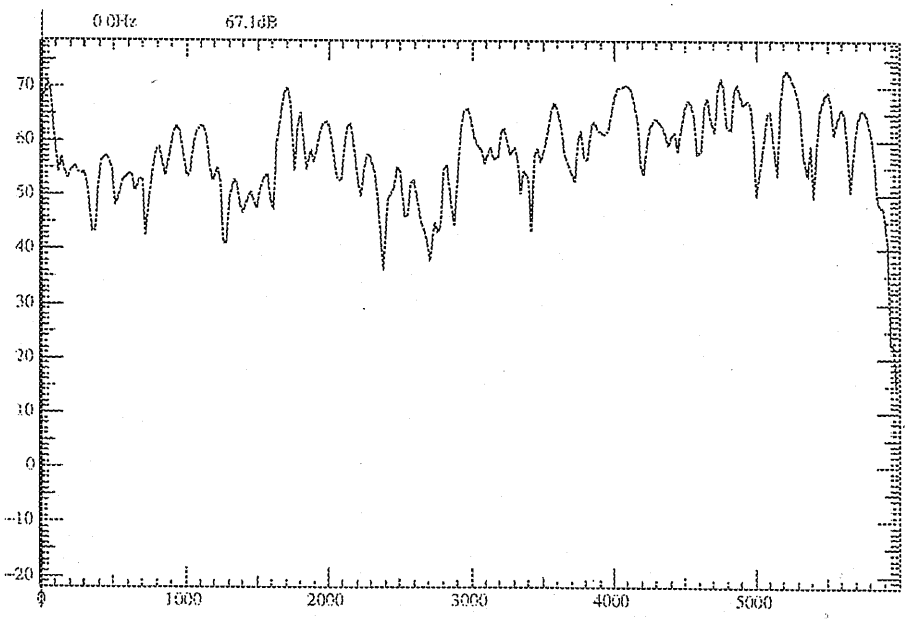


図 2.3: 子音の例 (男声 /isu/)

2.2.3 線形分離等価回路モデル

一般に音声のモデル化の基礎となるのは、線形分離等価回路モデルである。これは音声の生成過程を音源 $G(s)$ と声道伝達特性 $T(s)$ と放射特性 $R(s)$ に分離し、それらの縦続接続で音声の生成を表現するのものである。

母音・子音・鼻音は、それぞれ以下のようにモデル化されることが多い。

母音 声道は分岐を含まない。音源・声道・放射の縦続接続である。

子音 声道の中間に音源がある。声道の声帯側の端は固定端反射であるとする。

鼻音 声道の途中にそれぞれ口腔と鼻腔への分岐がある。

これらをそれぞれ図 2.4 に示す。

2.2.4 声道伝達特性

音源生成・声道による調音・放射の 3 過程のうち、言語音にもっとも影響が大きいと考えられるのは声道による調音である。

一般に母音型の声道の伝達関数 (体積流伝達比) は以下の様に与えられる。

$$T(s) = \frac{K}{\prod_{i=1}^{m}(1-\frac{s}{s_i})(1-\frac{s}{s_i^*})} \tag{2.1}$$

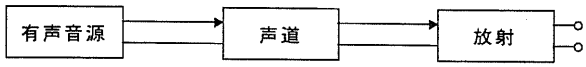
$$\begin{cases} s_i = \sigma_i + j\omega_i \\ s_i^* = \sigma_i - j\omega_i \end{cases} \tag{2.2}$$

これに対し、子音や鼻音化母音の場合は、音響管の分岐に伴い、声道伝達特性において極に加え零点の特性が現れることが知られている。

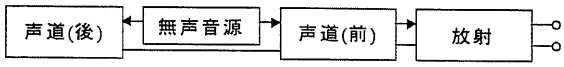
2.2.5 放射特性

放射については、無限大平面バッフルに取り付けたピストン音源からの放射で近似し、放射インピーダンスを $L-r$ 直列回路で近似することが行われる。このとき、単位自由振動面あたりの正規化インピーダンスは、振動面の半径を a、角周波数 ω と音速 c の比 $k=\omega/c$ に対して、$ka \ll 1$ のとき、

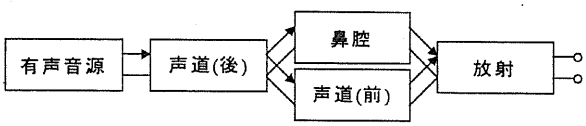
$$Z_r = \frac{(ka)^2}{2} + j\frac{8ka}{3\pi} \tag{2.3}$$



(a) 母音モデル



(b) 子音モデル



(c) 鼻音モデル

図 2.4: 音声生成のモデル

と近似できる。この第1項は音波の放射によるエネルギー損失を表し、第2項は声道が開口面と等しい断面積をもって$8a/3\pi$だけ延長されていることと、等価であることを示している。この放射インピーダンスにより、全ての共振周波数は一定の割合で低下し、帯域幅は増加する。

放射特性は、実際には6dB/octの微分回路で近似することが多い。

2.3 音声合成システムの構成

今日、音声合成処理を用いたシステムの代表的なものは、テキスト音声合成(text-to-speech, TTS)システムである。一般にテキストから音声を合成するためには、言語処理、音韻処理、音響処理の各段階の処理が必要になる。ただし具体的な処理は、合成システムによる違いがあり、また近年の統計的手法に基づく音声合成では、用いる特徴(コンテキスト)の種類を増やす傾向にあり、処理の区分を明確に出来ない場合も多いが、基本的には以下のようにまとめられる。

言語処理 形態素解析によりテキストを単語単位に分解し、品詞情報を得るとともに、統語・意味解析等により、統語情報等の言語情報を抽出する。

音韻処理 発音辞書等により、テキストを実際の発音の表記に変換する。また言語情報等から韻律的特徴を合成する。

音響処理 音韻処理の結果に基づき、予め蓄積しておいたデータを選択・接続し、音声波形を生成する。

本章において以下では、上記のうち特に音響処理に注目し、話を進める。

2.4 波形生成手法

任意のテキストから音声を合成するためには、少なくとも対象とする言語で用いられる全ての音素の合成が出来なければならない。そのような用途に用いることが可能で、信号処理により波形生成が可能な音声合成方式として、

- 波形接続方式
- 分析合成方式

- フォルマント合成方式
- 声道アナログ方式

が知られており、以下ではそれぞれについて説明する。

2.4.1 波形接続方式

自然音声波形から合成単位の波形を切り出したものを蓄積しておき、合成時に蓄積波形を波形上で接続する方法である。接続に伴う歪みが生じるが、蓄積波形それ自身の品質の低下はなく、高い品質の音声が容易に得られるが、実際に高品質の音声を合成するためには、合成したい音声と蓄積音声素片との間の差異を小さくしなければならず、そのために、大量の蓄積が必要となる。

従来、非常に大量の波形素片の蓄積が必要となるために、音声の基本周波数を制御することが困難で、その用途が限られていたが、PSOLA(pitch synchronous overlap and add)[9][10] 法による音声の基本周波数の変換が、変換率はそれほど大きくできないものの、高い品質で行われるようになり、高品質の音声の合成が、実用的な規模の蓄積で可能になった。

現在主流の音声合成方式である。

2.4.2 分析合成方式

基本的に波形をそのまま蓄積する波形編集方式とは異なり、分析合成方式は、何らかの方法で音声波形を分析した結果を蓄積し、合成時にはその分析結果から逆に波形を合成する方法である。

音声の特徴を利用した分析を利用することで、効率的な蓄積が可能であり、また分析結果の特徴量を操作することで音声の変形も可能であるが、一般に分析・合成時の誤差等から品質の低下が生じる、という欠点もある。

分析合成方式における分析は、線形予測分析やケプストラム分析により、音声のスペクトル包絡成分を抽出し、白色化された残差成分については、そのビットレートを削減する、というものが一般的であるが、さらに残差成分については、有声音の場合は残差成分のピッチ成分のみに注目しこれをインパルス列で、無声音については残差成分を白色雑音と仮定し、これを切り替えながら音声の合成を行う、ということも行われる。(図 2.5)

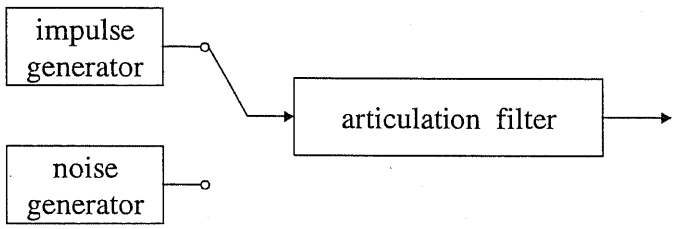


図 2.5: インパルス駆動による分析合成方式による合成器

合成時にこのインパルス列の基本周期を制御することで、合成音声の基本周波数を大きく制御することが可能である。

しかし、この方法は残差成分の誤差分を全て捨てているために、合成音声の品質の低下が生じる。そこで、より高い品質を得るために、一周期の声帯振動を複数のインパルス駆動で近似する、マルチインパルス駆動も検討されている。

近年では、HMM に基づく統計的処理等により、パラメータを制御することも検討されている。

2.4.3 フォルマント合成方式

声道の伝達特性を模擬することで、音声を合成する方法である [11][12]。声道の伝達特性はフォルマントとしても観測される複数個の極や、零点の特性を掛け合わせたものとして表現されるが、これに対応する合成回路を構成し、これを声帯音源波形モデルや摩擦音源に相当する白色雑音や破裂音源に相当するインパルスで合成回路を励振し、音声を合成する。特に母音の合成においては、声道伝達特性を共振回路の縦続接続で模擬することが一般的である。共振回路として

$$\begin{aligned} H(z) &= \frac{A}{(1 - Bz^{-1} - Cz^{-2})} \\ B &= 2e^{-\sigma_n T} \cos \omega_n T \\ C &= -e^{-2\sigma_n T} \\ A &= 1 - B - C \end{aligned} \tag{2.4}$$

の特性を持つ回路等が用いられるが、この共振回路は音声の音響的特徴としてみられるフォルマントに対応したものであり、フォルマント合成とも呼ばれる。合成回路と音声の特徴との対応は、その他の分析合成方式よりもより直接的である。

声道の鼻腔への分岐が生じた場合や、破裂音や摩擦音のように音源位置が声道の途中にあるような場合に、声道伝達特性に零点の特性が見られるが、このような場合について、共振回路の並列接続で近似的に表現する合成器 (並列型合成器) と、反共振回路の直列接続で直接的に表現する合成器 (直列型合成器) がある。図 2.6、図 2.7 はそれぞれ並列型合成器、直列型合成器の例であるが、特にこの図における共振器・反共振器は決まったものでなく、適当な個数は合成すべき音声により異なる。

実際には、ターミナルアナログ合成器の回路に相当するモデルを置き、自然音声波形からそのモデルパラメータを推定することも行われ [18][19]、この場合は、

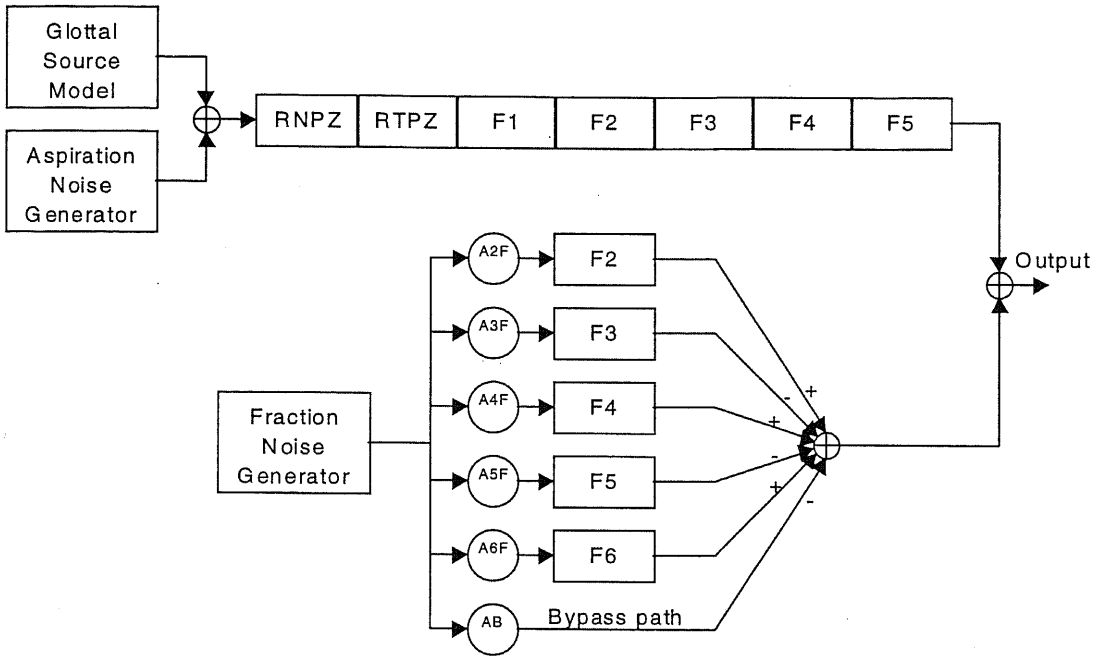


図 2.6: フォルマント合成方式による音声合成器 (直列/並列型)[12]

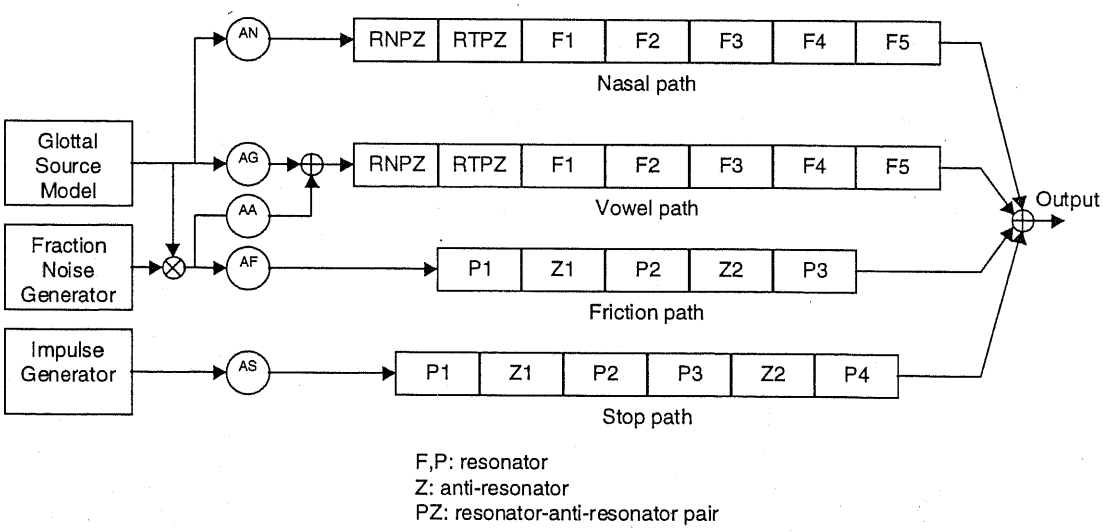


図 2.7: フォルマント合成方式による音声合成器 (直列型)[13]

先述の分析合成方式に分類することも可能である。しかし、分析合成方式でしばしば用いられる音声波形のスペクトル包絡成分とピッチ成分の分離が、必ずしも音声生成機構との直接的な対応を持たないのに対し、フォルマント合成のモデルに基づく分析結果は、音声生成機構との対応性に優れ、その推定パラメータを音響学的知見に基づき操作することも比較的容易である。しかし、パラメータの値が合成音品質に大きく影響することから高品質な音声合成と柔軟な制御を両立させることは一般に困難である。

2.4.4 声道アナログ方式

声道の物理的なモデルを作成し、音波の伝搬特性までさかのぼって合成する方式である。従来の声道断面積のレベルで表現されたモデルにとどまらず、音声器官の構造的なモデル化により、ターミナルアナログ方式よりも、より生成機構と直接的な対応を持つパラメータでの表現が可能で、規則合成には有効であると考えられる。

しかし、一般に音声器官の形状測定が必要であることから、このモデルの作成および規則抽出は困難で、現状では、高い品質の合成音声を得ることが難しい。本論文では、声道アナログ方式についてこれ以上触れない。

2.5 音声合成単位

任意のテキストを合成するために用いられる音声合成器の合成単位は、比較的短い音声単位、すなわち、

- 音素
- 音節 (CV、VCV、CVC など。なおＶは母音、Ｃは子音を表す)
- 1 ピッチの波形

が用いられる。通常、合成単位が小さいほど接続に伴う品質の低下が顕著となる。一方、合成単位が大きいほど、合成音声の品質は高くなるが、合成のために必要な蓄積が大きくなる。

短い合成単位を用いた場合の品質は、合成方式により大きく異なるが、一般には音声生成機構に近いモデルに基づく音声合成方式ほど、接続に伴う品質の低下

は小さい。

2.6 スペクトル包絡特性の推定

音素としての特徴は、一般にスペクトルの概形によって決まることから、分析合成手法においては、その概形を抽出するような分析手段を用いることが有効である。

現在、この目的によく用いられる分析手法として

1. 線形予測分析に基づく方法

2. ケプストラム分析に基づく方法

の大きく2つが用いられている。

線形予測分析は、基本的に音声をAR(Auto-Regressive, 自己回帰)過程に基づくものとして仮定するものであり、一方ケプストラムは、対数パワースペクトルのフーリエ係数表現である。一般に前者は生成機構寄りの、後者は音声波形の一般的特徴の表現手法であることから、それぞれ、パラメトリック分析、ノンパラメトリック分析、という分類もなされる。

ここでは、第4章で述べる分析編集合成システムでは、現在、生成機構の特徴を音声合成に用いることをその目的の1つとしており、それとの対応から、ここでは特に、線形予測分析について述べる。

2.6.1 線形予測分析

AR(Auto-Regressive, 自己回帰)モデルは次式で表される。

$$y(k) = -\sum_{i=1}^{p} a_i(k)y(k-i) + u(k) \tag{2.5}$$

ただし、入力$u(k)$は互いに無相関であるとする。

また、z領域で表し(ただしここでは$a_i(k)$はkについて定数であるとし、a_iで表す)これを変形すると

$$Y(z) = \frac{1}{1+\sum_{i=1}^{p} a_i(k)z^{-i}}U(z) \tag{2.6}$$

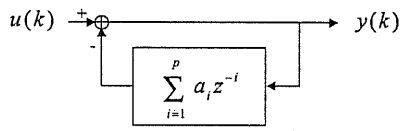


図 2.8: AR モデル

という形になる。従って全極モデルとも呼ばれる。

自然音声には零点の特性が現れることもあるが、全極のモデルを用いているのは、以下の理由による。

1. 人間の聴覚特性は、極に敏感で、零点については比較的鈍感である。

2. 比較的容易に問題を解くことができ、数理的な取り扱いが容易である。

線形予測分析とは、言い換えると出力 $y(k)$ が既知であるときに、$E(u(k)^2)$ を最小とする $a_i(k)$ を求めることであり、良い予測が行われたとき、$u(k)$ は白色となり、$\{a_i(k)\}$ はスペクトル包絡を表す。

特に有声音の場合、音声には声帯振動のハーモニック成分が含まれており、実際の分析においては、スペクトル包絡抽出のために、インパルス列を入力と仮定した分析が行われるが、このとき $u(k)$ の前提条件を厳密には満たさないため、実際の特性と分析結果との間には多少の誤差が生じる。また同様の理由により、$u(k)$ を白色として扱うことのできないような短時間分析には不向きである。一般的には音声の準定常性も考慮した上で、10ms 程度の短時間では定常であると仮定した時不変分析 (つまり $a_i(k)$ は k について定数であるとする) が行われる。

音源のピッチ成分 (インパルス列) は声道伝達特性に対しスペクトル的に平坦、つまり白色に近いため、ピッチ成分が $u(k)$ で表現され、$\{a_i\}$ がスペクトル包絡特性を表すように、実際の音声分析では、ピッチ成分によるスペクトルの微細構造を反映しない程度の比較的低い次数 (例えば 10 から 20 次程度) の分析を行う。分析結果より得られるスペクトル包絡の例を図 2.9 に示す。

この他、線形予測分析に関連する方法として、PARCOR(偏自己相関) 係数や LSP(線スペクトル対) 係数を用いる方法が知られている。

2.6.2 ケプストラムに基づく方法

ケプストラムはその定義により実係数による表現と複素係数の表現があるが、複素ケプストラム係数は、複素対数スペクトルの逆フーリエ変換により与えられる。このケプストラム係数により表現されるスペクトル包絡特性を表現するためには、

$$H(z) = \exp(\sum_{m=0}^{\infty} C_m z^{-m}) \tag{2.7}$$

の特性を持つフィルタが必要であるが、これを物理的には実現不可能である。そこで、この振幅特性を近似的に表現するフィルタである、LMA(対数振幅近似) フィ

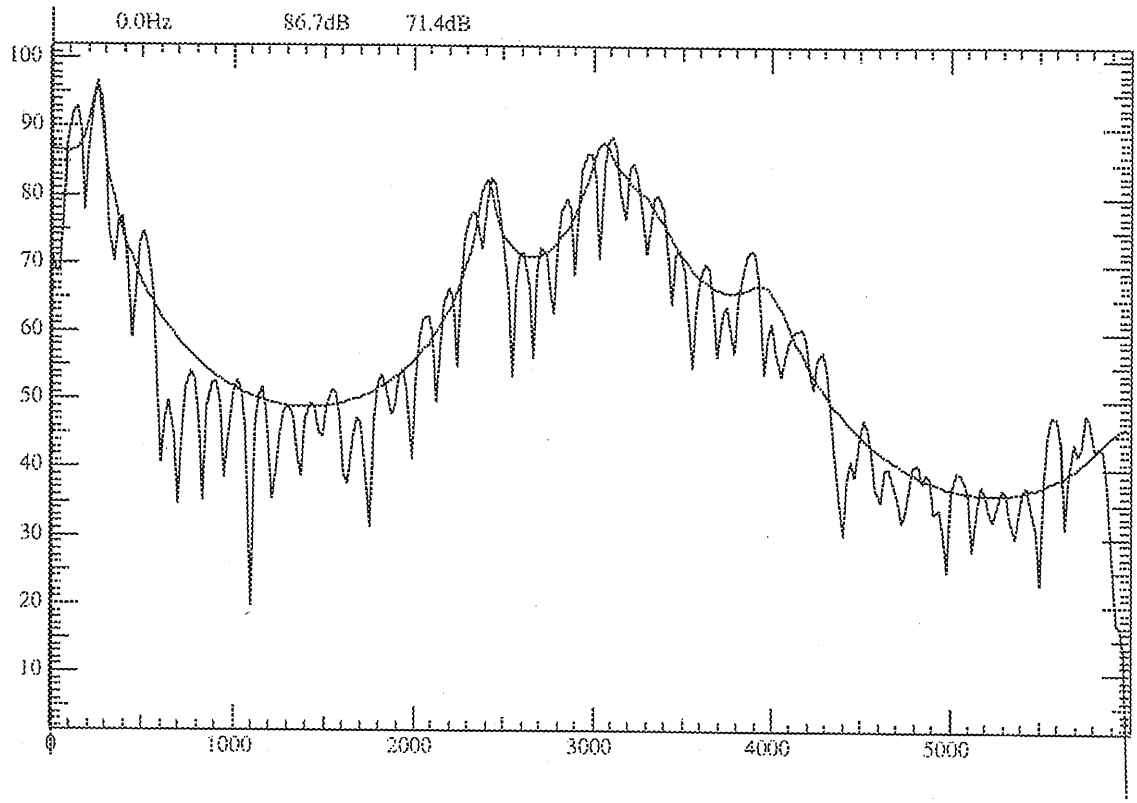


図 2.9: 線形予測分析によるスペクトル包絡の抽出 (男声 /isu/ 12 次分析)

ルタ [4] を用いるのが一般的である。

なお、実際にケプストラム係数はフーリエ変換を用いて計算すると、大きなバイアスを持つことが知られている。そこで、LPC 係数から計算されたケプストラム係数 (LPC ケプストラム) や、LMA 残差 (ここでは自然音声波形を入力した際の逆 LMA フィルタの出力) を小さくするような適応推定が行われる。

また、メルケプストラムに基づく方法もある。人間の聴覚特性を考慮した非直線周波数尺度の 1 つにメルスケールがあるが、メルケプストラム係数はメルスケール上で求めたケプストラム係数である。また、このメルケプストラムによるスペクトル特性を表現するフィルタとして、MLSA(メル尺度対数スペクトル近似) フィルタ [14] が知られている。近年の音声認識では、そのパラメータとしてメルケプストラムを用いるものが一般的であるため、認識との親和性という点では MLSA フィルタによる分析合成システムは有利である。しかし、合成音品質の点では必ずしも有利であるとはいえない。

2.7 音源特性と声道伝達特性の分離

線形予測分析等の入力に白色雑音を仮定したモデルから得られるスペクトル包絡は、音源特性の有色成分を含んだものであるため、このスペクトル包絡から得られるフォルマントは声道伝達特性とは厳密には一致しない。また、白色な入力を仮定することは、

1. 実際に分析を行ったときの推定誤差は、入力の仮定とは異なる。そのため、合成時にこの誤差を捨てることになる。

2. 短時間分析が難しい。分析区間長が、例えばその入力インパルス列の基本周期に比べ十分長ければ、入力が白色、という仮定を満たすが、そうでない場合にはその仮定には無理がある。しかし、分析区間長を長くしてしまうと、音声の定常性の仮定が満たされない。

という問題もある。

音声生成機構とのより直接的な対応を得るためには、音源と声道伝達特性を分離して扱うことが必要である。また、フォルマントとの対応性を考えた場合、分析次数をあまり上げたくないことから、より分析誤差の小さい分析手法を検討する必要がある。

以下ではARXモデルに基づく音源特性と声道伝達特性の分離手法について述べる。

2.7.1 ARXモデル

もし、入力波形が既知であるならば、既知の波形と白色雑音を入力とするモデルとしてARX (Auto-Regressive with Exogenous input) モデルを用いることができる。

ARXモデルは

$$y(k) + \sum_{i=1}^{p} a_i(k)y(k-i) = \sum_{j=0}^{q} b_j(k)g(k-j) + u(k) \tag{2.8}$$

で表される。

また、これはz領域において、

$$Y(z) = \frac{(\sum_{j=0}^{q} b_i z^{-j})G(z) + U(z)}{1 + \sum_{i=1}^{p} a_i z^{-i}} \tag{2.9}$$

と表すことができる。

ここで入力 $g(k)$ は音源波形であるとする。ARXモデルに基づく推定は、ARモデルに基づく推定とほぼ同じ方法で解くことができ、a_i, b_j を直接求めることができる、という特徴がある。また、音源特性を $u(k)$ に含める必要がなく、無相関であるという仮定を理想的には満たすことができる、という特徴がある。

しかし、実際にはこの音源波形を直接観測することはできない、という大きな問題がある。

2.7.2 声道伝達特性と音源特性の同時推定

音源波形の直接観測は不可能であるため、そこで、ARXモデルの入力を音源モデルとし、誤差を最小とする音源モデルパラメータを推定し、その際のARXモデルパラメータも分析結果として同時に用いる、という方法が考えられる。これにより、声道伝達特性と音源特性の同時推定が可能になる。

ただし、誤差の最小化問題は、一般に多数の局所的最適解を持つ傾向にあり、推定は容易ではない。

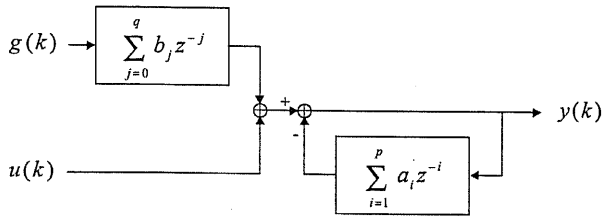


図 2.10: ARX モデル

2.7.3 音源波形モデル

声帯振動の観察等から、これまで様々な音源波形モデルが提案されてきた。近年用いられることが多いのはRK(Rosenbarg-Klatt)モデル[12]である。RKモデルはRosenbargのモデルにスペクトル傾斜を考慮しLPFを追加したもので、音源パラメータに、スペクトル傾斜に関するパラメータTL(spectral tilt)が追加されている。このモデルは、他のモデルよりも自由度が低いときの誤差が小さいという特徴があり、自動分析を目指した研究では用いられることが多い。

これに対し、FL(Fujisaki-Ljungqvist)モデル[15]がある。これは、声帯音源波形(体積流)を多項式近似により、より厳密に表現しようとしたもので6パラメータのモデルである。

以下では、本研究で用いたFLモデルについて説明する。

FLモデル

FL(Fujisaki-Ljungqvist)モデル(図2.11)は、このモデルは、音源波形の1周期を4つの区間に分けたもので、声帯音源波形$g(t)$の微係数$\dot{g}(t)$は以下のように表される。

$$\dot{g}(t) = \begin{cases} A - \dfrac{2A+R\alpha}{R}t + \dfrac{A+R\alpha}{R^2}t^2 & (0 < t \leq R) \\ \alpha(t-R) + \dfrac{3B-2F\alpha}{F^2}(t-R)^2 + \dfrac{2B-F\alpha}{F^3}(t-R)^3 & (R < t \leq W) \\ C - \dfrac{2(C-\beta)}{D}(t-W) + \dfrac{C-B}{D^2}(t-W)^2 & (W < t \leq W+D) \\ \beta & (W+D < t \leq T) \end{cases} \tag{2.10}$$

$$\alpha = \frac{4AR+6FB}{2R^2-F^2} \tag{2.11}$$

$$\beta = \frac{CD}{D-3(T-W)} \tag{2.12}$$

である。ただし、

T: 基本周期

W: 声門開放区間の長さ

S: 声門開放区間内でのパルスの非対称度$(= \dfrac{R-F}{R+F})$

D: 声門閉鎖時点から体積流が直線的なドリフトに移るまでの時間

A: 声門開放直後の微係数

B: 声門閉鎖直前の微係数

C: 声門閉鎖直後の微係数

である。このモデルの特徴としては、自由度が高いことと、それらが全て時間領域のパラメータであることが挙げられる。

実際に推定を行う場合、モデルの時間軸上の位置の推定に大きな計算量を要することが多い。そこで、あらかじめ基本周波数も含め、声門閉鎖時点の大まかな位置を、線形予測分析の残差等を用い決定しておくことも行われる。

2.8 ピッチ同期処理

自然音声波形の分析を行う場合には、まず時間軸上で分析区間を決める必要がある。二乗誤差最小化に基づくARXモデル推定では、共分散法による線形予測分析と同じ考え方に基づき、有限の分析区間を設定することが可能であるが、この分析区間により、分析結果は大きな影響を受ける。このため、分析区間を自然音声波形と独立に設定してしまうと、連続する2つの分析区間の間で、分析結果が大きく揺らぐという、という問題が生じる。

そこで、分析区間を自然音声波形の基本周期と同期させることで、隣接する分析区間の間の分析結果の変動を抑えることができると考えられる。

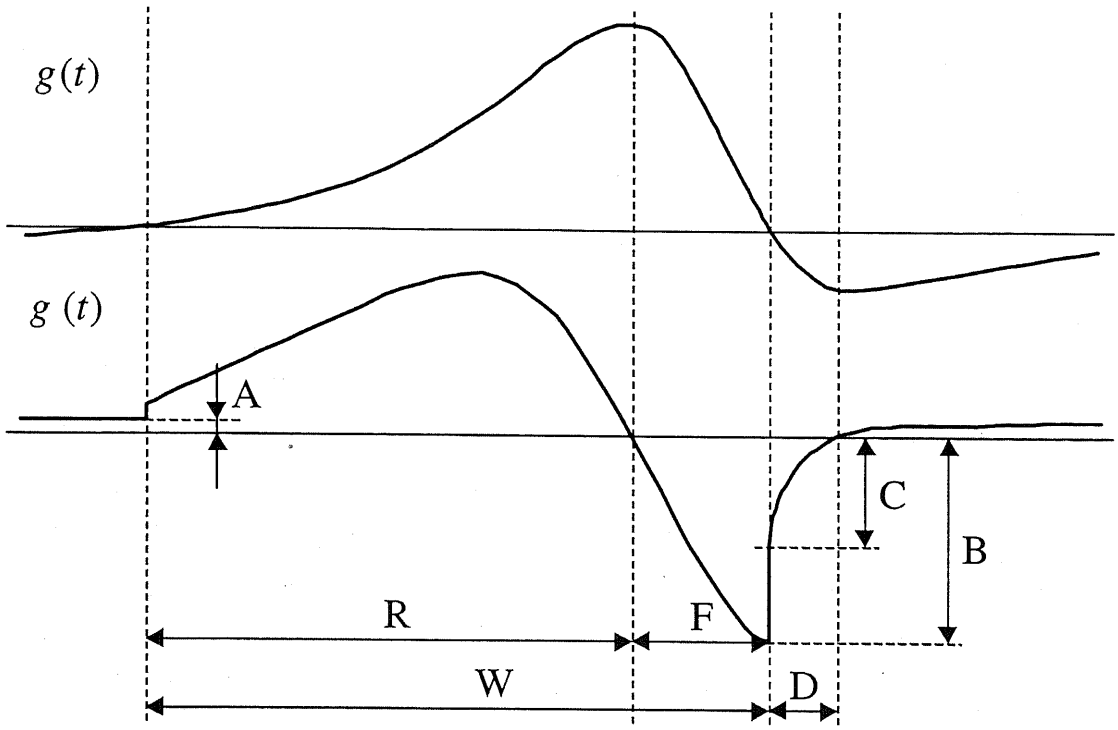
まず、分析区間の中心は、自然音声波形の一周期のうちで、大きなパワーを持つ部分とすることが望ましい。これは有声音の場合、通常声帯の閉鎖時点に相当する。そこで、線形予測分析等の残差波形と声門閉鎖時点との対応を考慮し、一周期のうち、線形予測分析の残差波形のピークを分析区間の中心になるようにする。

また、様々な基本周波数の音声に対応できるよう、分析区間の長さは自然音声波形の基本周期に比例した長さとすることが考えられる。

2.8.1 PSOLA(Pitch-Synchronous OverLap and Add)

波形接続を用いた代表的な音声変形技術がPSOLAである。全体的な音声品質においてその影響が大きい有声音についてその基本周期を単位に音声処理を行い、接続部分における不連続を緩和するために重ね合わせと加算処理が行われる。

このためにまず音声波形はある周期(基本的には有声音1周期)毎に単位として



$$g'(t)=\begin{cases} A-\frac{2A+R\alpha}{R}t+\frac{A+R\alpha}{R^2}t^2 & (0<t\leq R) \\ \alpha(t-R)+\frac{3B-2F\alpha}{F^2}(t-R)^2+\frac{2B-F\alpha}{F^3}(t-R)^3 & (R<t\leq W) \\ C-\frac{2(C-\beta)}{D}(t-W)+\frac{C-B}{D^2}(t-W)^2 & (W<t\leq W+D) \\ \beta & (W+D<t\leq T) \end{cases}$$

$$\alpha=\frac{4AR+6FB}{2R^2-F^2},\ \beta=\frac{CD}{D-3(T-W)}$$

図 2.11: FL(Fujisaki-Ljungqvist) model

中心に重みが掛けられた何らかの窓を掛け、その周期よりも長い範囲で切り出される。その切り出した波形の配置間隔を換え、足し合わせを行うことで音声の基本周期を制御することができる。またその波形単位での、間引き、補間を行うことで音声の持続時間長を制御することもできる。

切り出した波形に対し、周波数軸上での操作を何も行わず全て時間領域での処理によりPSOLAを行う方法は特にTD-PSOLA(Time Domain PSOLA)と呼ばれる。この方法の特徴は大きなピッチ変換を行わないとき、波形上での歪が小さく、品質の低下が少ないため、よく用いられる手法である。

一方、ピッチ変換後の歪を緩和するために、切り出した波形に対してDFT等を利用し周波数軸上における変換処理を行う手法はFD-PSOLA (Frequency Domain PSOLA)と呼ばれる。TD-PSOLAよりも大きくピッチを変換させたときに品質の低下が抑えられる反面、時間・周波数変換の際に歪が生じるため、ピッチを大きく変えない場合においては品質的には一般にTD-PSOLAより不利である。

また、PSOLAによる変形後の音声品質はピッチ同期分析の基準点(ピッチマーク)の精度に大きく依存する。ピッチ同期分析においては、誤差を小さくするために、分析フレームの中心付近に1周期音源波形のローカルピークが生じるように設定する方法が一般的である。ここでピッチマークを分析フレームの中心と定義するとき、ピッチマークの位置として通常

- 周期波形のローカルピーク
- 声門閉鎖点をピッチマークとする

のいずれかが用いられる。しかし、このどちらもロバストに抽出することは容易ではない。特に音声の過渡部において、両者の特長とも抽出が特に困難になる。ピッチマーク推定の誤りは、ピッチ変換時の品質を大きく低下させるため、これをロバストに推定する研究が行われているが、現状その精度には問題があり、多くのシステムでは自動推定後の手修正が行われている。

2.8.2 ピッチマークの自動推定

ピッチ同期分析において、分析の基準点(ピッチマーク)をどのように求めるかは問題である。近年ではPSOLAによる基本周波数の変換を行うためにピッチマークを付ける必要性もあり、ピッチマークの自動推定に関する様々な研究が行われている。

ピッチマークは主に、音声波形のローカルピークにマークをつける方法と、声門の閉鎖時点にマークを求める方法の2つがあるが、前者が音素や発話環境の差異による影響が大きいのに対し、後者はその影響が小さいと考えられる。

声門の閉鎖時点を推定する方法としては、線形予測分析やLMAフィルタの残差波形のピークを検出する方法がある。また、音声波形が声門閉鎖点で急峻に変化することに着目し、音声波形のwavelet変換のローカルピークを検出する方法[24]もある。しかし、実際にはこれらの方法によるピッチマークの抽出結果は安定しないことが多く、既に求められたピッチマークとの相関を用いること等も同時に行われる[25]。

2.9 コーパスベース音声合成

現在の音声合成システムの多くは、いずれも大規模な音声コーパスを直接的、あるいは間接的に利用した、統計的手法に基づいたものである。特に本節では音声認識分野で広く一般的に用いられているHMM(隠れマルコフモデル)に着目し、音声合成との関連について述べる。

2.9.1 隠れマルコフモデル(HMM)

隠れマルコフモデル(Hidden Markov Model: HMM)に基づく手法が、近年、音声情報を確率・統計的に扱うために、広く用いられている。ここでは、HMMに基づく手法の基本原理と、また、認識においてHMMを効率よく扱う手法である、Viterbiアルゴリズムについて、簡単に説明する。

2.9.2 HMMの基本原理

HMMは出力シンボルが確定しても、状態遷移先が不確定である、非決定性有限オートマトンとして定義される。音声認識ではleft-to-right型で、1つの初期状態と1つの最終状態がある構造を用いることが多い。ここでは、簡単のために図2.12のモデルを考える。

まず、π_iを状態q_iの初期確率、π_iをa_{ij}は状態q_iから状態q_jへの状態遷移確率、$b_{ij}(\mathbf{o})$は状態q_iから状態q_jへの遷移で、パターン$\mathbf{o}$が観測される確率と定義する。

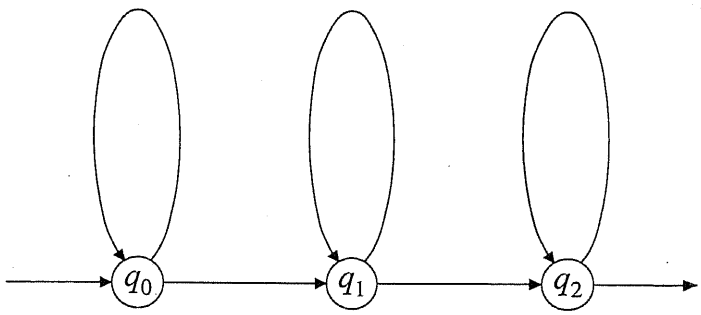


図 2.12: an example of HMM

従って、図4.1のモデルでは、

$$\pi_i = \begin{cases} 1 & (i=0) \\ 0 & (i=1,2) \end{cases} \tag{2.13}$$

であり、また、

$$a_{ij} = 0 (j < i, i+1 < j) \tag{2.14}$$

である。

また、$\mathbf{o}$を連続値として表す場合と、有限個のシンボルの組み合わせで表す場合があり、それぞれ連続HMM、離散HMMと呼ばれる。

例えば、$\mathbf{o}$がシンボルx_1, x_2のいずれかであるとし(従って離散HMMである)、a_{ij}を要素とする行列Aと、$b_{ij}(\mathbf{o})$がそれぞれ

$$\begin{aligned} A &= (a_{ij}) \\ &= \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \end{bmatrix} \end{aligned} \tag{2.15}$$

$$\begin{cases} b_{11}(x_1) = 0.4, & b_{11}(x_2) = 0.6 \\ b_{12}(x_1) = 0.7, & b_{12}(x_2) = 0.3 \\ b_{22}(x_1) = 0.8, & b_{22}(x_2) = 0.2 \\ b_{23}(x_1) = 0.3 & b_{23}(x_2) = 0.7 \\ b_{33}(x_1) = 0.8, & b_{33}(x_2) = 0.2 \end{cases} \tag{2.16}$$

であるとすると、観測系列$O = \{x_1 x_1 x_2\}$であるとき、考えられる状態遷移は$\{q_0 q_0 q_1 q_2\}$、$\{q_0 q_1 q_1 q_2\}$、$\{q_0 q_1 q_2 q_2\}$の3通りであるので、それぞれの状態遷移系列における出力確率を、それぞれ$P_1(\mathbf{O}|\lambda)$、$P_2(\mathbf{O}|\lambda)$、$P_3(\mathbf{O}|\lambda)$とすると、

$$\begin{cases} P_1(\mathbf{O}|\lambda) &= 0.4 \times 0.4 \times 0.6 \times 0.7 \times 0.5 \times 0.7 \\ &= 0.02352 \\ P_2(\mathbf{O}|\lambda) &= 0.6 \times 0.7 \times 0.5 \times 0.8 \times 0.5 \times 0.7 \\ &= 0.0588 \\ P_3(\mathbf{O}|\lambda) &= 0.6 \times 0.7 \times 0.5 \times 0.3 \times 1 \times 0.2 \\ &= 0.0126 \end{cases} \tag{2.17}$$

従って、モデルλの出力確率(尤度)$P(\mathbf{O}|\lambda)$は、

$$P(\mathbf{O}|\lambda) = P_1(\mathbf{O}|\lambda) + P_2(\mathbf{O}|\lambda) + P_3(\mathbf{O}|\lambda)$$

$$= 0.02352 + 0.0588 + 0.0126$$
$$= 0.09492 \tag{2.18}$$

となる。

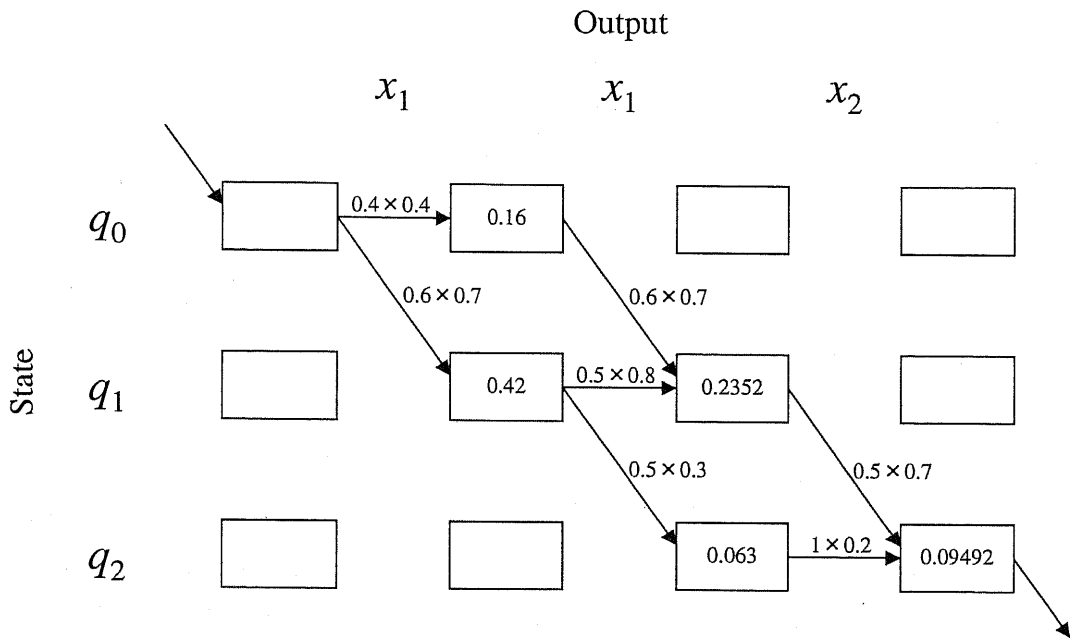


図 2.13: calculating likelihood of HMM

HMMによる認識とは、事前に用意された複数のHMMの中から、$\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T\}$であるとき、最大の尤度$P(\mathbf{O}|\lambda, T)$を与えるモデル$\lambda$が探すことである。

なお、HMMのパラメータは、与えられた観測系列の生成確率を最大にするような、モデルパラメータを繰り返し手法により推定する、Baum-Welch法 (EMアルゴリズム、forward-backwardアルゴリズムとも) により求めることができる。

2.9.3 Viterbiアルゴリズム

尤度 $P(\mathbf{O}|\lambda)$ を厳密に求めないで、近似的にモデル λ が系列 $\mathbf{O}$ を出力するときの、最も可能性の高い状態系列上での出現確率を用いる方法がある。前節の例では、状態遷移系列として、$\{q_0q_1q_1q_2\}$ を用いることになる。(図 2.14) この場合の出現確率 (尤度) は、各遷移での確率値を対数変換しておくことにより、加算と大小判定のみからなる演算により高速に求めることが可能である。すなわち、

$$f(j,t) = \begin{cases} \log \pi_j & (t=0) \\ \max\limits_i \{f(i,t-1) + \log a_{ij} + \log b_{ij}(\mathbf{o}_t)\} & \\ & (t=1,...,T) \end{cases} \tag{2.19}$$

を計算し、対数尤度として

$$L = \max_j f(j,T) \tag{2.20}$$

を求める。この方法を Viterbi アルゴリズムと呼ぶ。この方法は前節における尤度 $P(\mathbf{O}|\lambda)$ を用いた場合と、音声認識性能がほとんど変わらないことが実験的に確認されており、広く用いられている。

また、あらかじめ $f(j,t)$ を全て記憶しておき、尤度が最大となる (j,T) から逆に辿る事により、尤度最大の場合における、遷移系列を得ることができる。

2.9.4 音声認識技術を用いた蓄積の作成

音声波形データを合成に用いるためには、音素やその他の環境の情報を付与する作業 (ラベリング) や、蓄積単位に対応する区間に切り出す作業 (セグメンテーション) といった作業が必要である。近年、ディクテーションシステムが実用化されつつあり、もちろんそれを用いることも出来るが、一般に大量の音声波形データは、あらかじめ用意されたテキストの読み上げにより作成されたもので、あらかじめ発話内容がわかっていることが普通である。

そこで、認識用に学習された音素 HMM を発話内容に合わせそれぞれ HMM の初期状態と終了状態をそれぞれ接続し、1 つの大きな (例えば文に相当する程度の長さを持つ) HMM λ を作成する。この HMM において、入力音声に対し尤度 $P(\mathbf{Q}|\lambda)$ が最大となる状態遷移系列 $\mathbf{Q}$ を求め、元の各音素モデルの開始点・終了点と対応付けることにより音声の開始、終了時点を定めることが出来る。

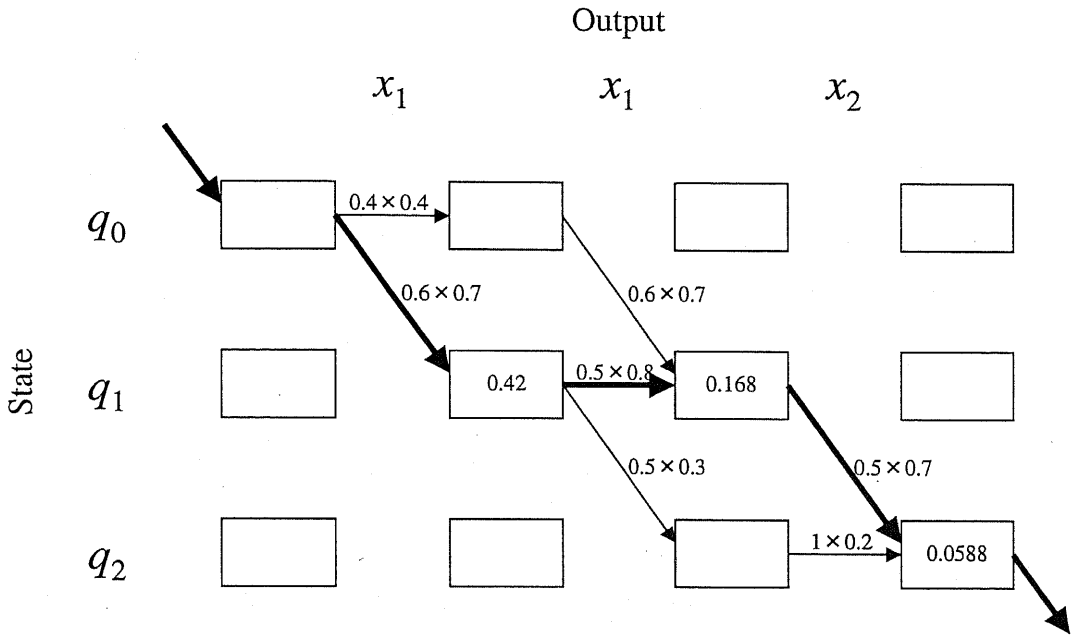


図 2.14: calculating based on Viterbi algorithm

この手法は、強制切り出し (forced alignment) と呼ばれ、音声データベースを作成する際に広く用いられている手法である。

なお、波形編集方式において TD-PSOLA 法によるピッチ変換を行う場合等には、この他に、ピッチマーク付けの作業を行うことが必要である。これは音声認識技術とは異なるものであるが、これについては、別に様々な自動化の研究が行われている [37]。

近年では、波形編集方式や分析合成方式による合成で必要となるデータベースの作成は、これらの手法により、ほぼ自動化されている。

2.9.5 統計的手法に基づく素片選択

音声合成において、蓄積されたデータをどう選択するかが、品質面において、大きな影響を及ぼす。

この選択基準を統計的に扱ったものとして、あらかじめ HMM の学習を行っておき、その HMM の尤度が最大となる波形やパラメータを選択する方法がある。以下ではその方法について述べる。

2.9.6 HMM の尤度最大化に基づく素片選択

ある話者について例えば triphone モデル (前後の音素の影響を考慮した音素モデル) を学習しておき、合成時にはその HMM において尤度が最大となる波形もしくはパラメータを選択する方法がある。

コンテキストの違いを考慮した triphone モデルは、数万～数十万個程度の規模となるが、それぞれに対応する音声波形を蓄積する方法 (CHATR 等の方法) では、蓄積波形のサイズが場合にもよるが数百メガバイト規模になってしまう。そこで、クラスタリングの結果等から、特徴ベクトル空間において、距離が近い音声素片は全て蓄積しない、等の方法により、蓄積数を減らす必要がある。

しかし、この手法により蓄積数を減らした場合、合成時に音素モデルに直接対応する音声素片はないので、何らかの方法で、音素モデルの特徴に近い音声素片を選択する必要があるが、コンテキスト等から決定した音声素片の候補について、合成したい音声に対応する HMM の尤度が最大になる、という基準で蓄積素片を選択する方法が考えられている。[38][39][40]

この方法では、合成システム構築の際に、HMM を学習したり、クラスタリング

を行う手間が増えるが、より少ない蓄積を用いて高い品質の合成音声が得られる、システムを構築することが可能である。

2.10 HMMによる合成パラメータの生成

分析編集方式におけるパラメータそのものを統計的に生成しようとする方法が提案されている。以下では、HMMに基づくパラメータの生成手法について述べる。

2.10.1 Viterbiアルゴリズムに基づくパラメータの生成

HMMを用いた合成器パラメータの生成において、単純な考え方は、与えられた状態系列に対して、尤度が最大となる音声パラメータ系列を生成することである。

そこで、連続型HMM λ が与えられたとき、λ から $P(O|\lambda,T)$ を最大にする長さTの出力ベクトルを生成することを考える。

この場合、Viterbiアルゴリズムに基づく方法として、

$$f(j,t) = \begin{cases} \log \pi_j & (t=0) \\ \max\limits_i \{ f(i,t-1) & \\ + \log a_{ij} + \max\limits_{\mathbf{o}} \log b_{ij}(\mathbf{o}_t) \} & \\ & (t=1,...,T) \end{cases} \tag{2.21}$$

による最大対数尤度をあらかじめ求めておき、その状態遷移経路を逆にたどることで、尤度が最大となる状態遷移系列を決定することができる。その後、各状態において、最大の出力確率を持つパラメータを出力することで、尤度が最大となるパラメータ系列が得られる。

2.10.2 動的特徴を考慮したパラメータ生成

前節における最大尤度の状態遷移系列からそれに対応するパラメータをそのまま得ると、そのパラメータは、1つの状態が継続している間は一定の値をとり、状態遷移毎に不連続にパラメータが変化することになる。従って、この場合、何らかの平滑化手法が必要になる。

そこで、この問題を厳密に取り扱うものとして、動的特徴をパラメータとして含む連続型の HMM から尤度最大という基準において最適なパラメータ系列を作成する手法がある。この手法について、簡単に説明する。

前節における $\mathbf{o}_t$ を、静的な特徴ベクトル $\mathbf{c}_t$ に加え、静的な特徴から計算される動的特徴ベクトル $\Delta\mathbf{c}_t$) および、$\Delta^2\mathbf{c}_t$ で構成されるとする。

ただし、

$$\begin{cases} \Delta c_t = \sum_{\tau=-L_1}^{L_1} w_1(\tau)\mathbf{c}_{t+\tau} \\ \Delta^2 c_t = \sum_{\tau=-L_2}^{L_2} w_2(\tau)\Delta\mathbf{c}_{t+\tau} \end{cases} \tag{2.22}$$

で表されるとする。ここで、$w_1(\tau)$、$w_2(\tau)$ は動的特徴量を計算するための重み係数である。

まず状態遷移系列が既知であると仮定する。この場合、与えられた状態遷移系列 Q に対して、

$$\frac{\partial logP(O|Q,\lambda,T)}{\partial C} = 0 \tag{2.23}$$

を解くことにより、最適なパラメータ系列
$\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_T\}$ が得られる。この方程式は線形方程式であり、容易に解くことが出来る。

ただしこの状態遷移系列 Q を前節の方法で定めると、(2.22) 式の条件が課せられたことにより、得られるパラメータ系列は最尤ではない。厳密には考えられる状態遷移系列すべてについて (2.23) 式を解き、さらにその中で最大の尤度になるものを得る必要がある。この際、異なる状態遷移系列に対する解を再帰的に計算するアルゴリズムを用いることで、準最適なパラメータ系列を効率的に求める方法もある [42]。

2.11 ノンパラメトリックな手法に基づく分析合成手法の研究動向

分析合成に関係する手法として、ノンパラメトリックな手法に基づく音声分析合成手法の研究動向についても論じるため、本節では、今日、高品質での音声変形が可能な代表的な手法として、sinusoidal modeling[5] と STRAIGHT[6][7][8] について述べる。

2.11.1 sinusoidal modeling

有声音においては、明確な調波特性が観測されるため、これを正弦波の重ね合わせで表現する手法が考えられる。これは一般にSinusoidal modelと呼ばれる。この手法では音声の雑音成分を表現できないため、さらに雑音モデルを加える手法が一般的である。

2.11.2 STRAIGHT

STRAIGHT(Speech Transformation and Representation of weighted spectrogram)は河原らによって提案された手法で、サウンドスペクトログラム上に生じる音声の周期性に由来した振動成分を除去する手法がメインとなっている。

STRAIGHTの基本となる考え方は、声帯音源による周期的な駆動の役割を、周波数・時間・振幅の三次元空間における時間周波数表現の曲面$S(w,t)$から、サンプリング操作により抽出することである、と解釈し直すことである。この解釈では、基本周期τ_0の周期信号$s(t)$から時間軸に対してτ_0毎に、また周波数軸に対して$f_0 = 1/\tau_0$毎に曲面$S(w,t)$の情報が得られることになる。

別の言い方をすると、有声音からはスペクトル包絡特性を表す曲面の部分的な情報が得られる、といういうことである。そして柔軟な変換を可能にするスペクトル分析の最終目標は、この部分的な情報を用いて$S(w,t)$平面を復元することである。

ところで、短時間フーリエ分析により与えられる時間周波数表現のことをスペクトログラムという。このスペクトログラムでは時間窓関数の選択が重要な問題となる。時間分解能と周波数分解能の積には不確定性の関係で規定される窓関数の形で決まる下限があるが、信号から振幅に関する情報をできるだけ詳しく取り出すために、時間分解能と周波数分解能の積が最小で、かつ時間分解能と信号の基本周期、周波数分解能と信号の基本周波数の比が等しくなるような時間窓を選択することである。これを満たす窓関数$w(t)$はすなわちGauss関数であり、$w(t)$とそのフーリエ変換$W(\omega)$は次式で与えられる。

$$w(t) = \frac{1}{\tau_0} e^{-\pi(t/\tau_0)^2} \tag{2.24}$$

$$W(\omega) = \frac{\tau_0}{\sqrt{2\pi}} e^{-\pi(\omega/\omega_0)^2} \tag{2.25}$$

音声の場合、時間と共に基本周期が変化するため、窓関数の時間長を基本周期に応じて適応的に変化させながら分析を行うことになる。

そして、得られたスペクトログラムにおける格子点上の振幅値から、1次関数補間によりスペクトル包絡特性を近似する。STRAIGHT では仮定 (制約) を最小にすることを狙い、できるだけ局所的な情報だけに基づき曲面の復元を行っている。具体的には、ある時間周波数の点における値は、近隣の 4 点の格子点の情報のみを用いて計算される。

時間軸方向および周波数軸方向の特性がそれぞれ 1 次関数として表される、区分的双一次曲面では、このような条件を満たす 1 つの解が定まり、STRAIGHT ではこれが補間で用いられている

ただし実際には、補間を行う際の計算に、格子点上の振幅値を直接的に用いると、雑音の影響を大きく受けるため実際には、平滑化関数を用いた、格子点近隣の重み付き平均値から曲面を推定している。

STRAIGHT の実現には音声の基本周波数の正確な抽出が必要であり、このために非対称時間窓を利用し基本波らしさを推定する TEMPO 法や、周波数と瞬時周波数との関係を利用した方法が考案された [46]。

2.12 まとめ

本章では、音声合成における波形生成手法の研究動向について、直接関連する音声学的知見、信号処理技法、蓄積選択手法、合成器制御手法等も併せて述べた。

第 3 章

子音波形に波形接続を用いる音声合成方式

3.1 はじめに

本章では柔軟な音声合成を実現するために重視すべき合成器構成、制御手法についてに論じる。

柔軟性という点では単にパラメトリックであれば良いと、ということには実際にはならない。一般に複雑な音声生成モデルを仮定するほど、各パラメータ間の依存関係が大きくなり、パラメータ制御が困難になる。また、モデル化に伴う誤差が生じるため、ある程度の品質低下が避けられない。

柔軟性と合成品質はトレードオフの関係にあると考えられるが、従来のノンパラメトリック、パラメトリックのどちらか、という考え方では結果も両極端なものとなり、分析合成方式のメリットである柔軟性を生かした実用的な音声合成システムの構築は困難である。

そこで本章ではセミパラメトリックとも言える、パラメトリックな部分のノンパラメトリックな部分の両方を備えた音声合成器の構成について検討を行い、その実現可能性について実験に基づいた議論を行う。

3.2 波形編集を併用したフォルマント音声合成

我々が従来から研究を行ってきた、全ての日本語の音素に対応したフォルマント合成システム [44] の構成を図 3.1 に示す。この合成器は音声生成過程との対応性を重視したもので、生成過程の異なる音声にそれぞれ対応する複数の調音フィルタから構成されており、各調音フィルタは二次線形系による共振回路とその逆回路である反共振回路を、それぞれ適当な個数、直列に接続し構成されている。また音源としては、声帯音源として声帯音源波形モデルが、摩擦音、破裂音の合成のために、それぞれ白色雑音源、インパルス音源が用意されている。

この合成システムでは、母音は声帯音源と母音用のフィルタ回路のみを用いて生成される。この過程は音声生成機構と対応したものであり、ARX モデルによりモデル化される。合成器パラメータは、ARX モデル推定誤差を最小化する声帯音源波形モデルパラメータを求めることにより自然音声波形から推定することが可能であり (GAR/GARMA 法 [17])、これにより、比較的高い品質の母音音声を得ることができる。またこの手法は母音に限らず、母音と同様の生成過程を持つ子音にも適用できる。我々は既にこの手法を用いたフォルマント分析合成システムの開発を行っており、実際に、/VrV/音声の合成についてその有効性を確認している

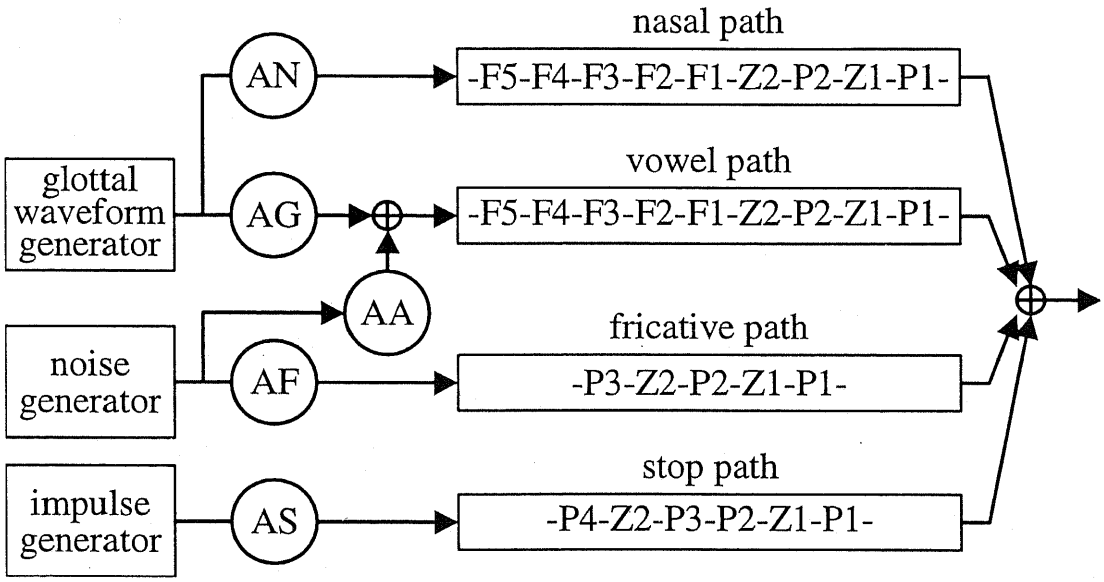


図 3.1: 複数の直列回路を持つターミナルアナログ合成システム [44]

[50]。

ところがこの回路構成では、同時に複数の回路を駆動する必要が有声子音等の合成時に生じる。その場合、自然音声波形からのモデルパラメータの推定は非常に困難なものとなる。また実際には、無声子音においても、音声の非定常性等の理由により、安定した分析結果を得ることができない場合が多い。

従来、そのような音素については、高い品質の合成音声を得るためにパラメータテンプレートを知識や経験に基づき作成しており、合成システム開発において、非常に大きな手間を要していた。

生成過程に対応する形での分析が困難な音素を合成する手法として有名なものに、Klatt による直列型回路と並列型回路のハイブリッド構成によるフォルマント合成器 [11] がある。しかし、並列型回路のパラメータの制御に関しては、フォルマント位置と強度に関する正確な知識が必要であり、この問題は解決されない。

そこで、この問題に対処するために、分析が困難な音素については、パラメータ推定の容易な、比較的簡単なモデルを用いて音声を表現する方法が考えられる。ところが、一般にそのようなモデルは生成機構との対応性が低く、品質の低下の小さいパラメータの操作が困難であるために、多くの蓄積を必要とし、結果として合成器の柔軟性が失われることになる。

しかし、特に日本語においては母音が音声区間のかなりの割合を占めることから、子音合成の柔軟性は母音ほど重要ではない可能性がある。そこで、テンプレート作成の容易性を考慮し、子音合成時の柔軟性が失われている場合の極端な場合として、母音合成用回路のみからなるフォルマント合成器と子音合成用の波形編集の合成器を並列に配置した構成の合成システムを提案する。

回路構成を図 3.2 に示す。提案する合成器では、母音型音声のみを声帯音源波形モデルと共振器・反共振器の直列接続によるフォルマント合成回路により合成し、それ以外の音素については、波形編集方式により自然音声波形素片を用いて合成する。無声子音合成のための自然波形重畳については既に検討が行われているが [43]、提案する合成器では、有声子音についても波形編集により合成する。このため、TD-PSOLA に基づくピッチ変換が必要に応じて行われる。

最終的に、これらの 2 つの方式の合成回路を切り替えて音声を合成することになるが、両者の切り替え、すなわち接続部分においても、波形レベルでの不連続を緩和するために、TD-PSOLA によるピッチ同期させた重ね合わせ処理が行われる。

ここでは、TD-PSOLA による処理の際の基準となるピッチマークに、線形予測分析の残差波形から推定される声門閉鎖点を用いる。フォルマント合成音声につ

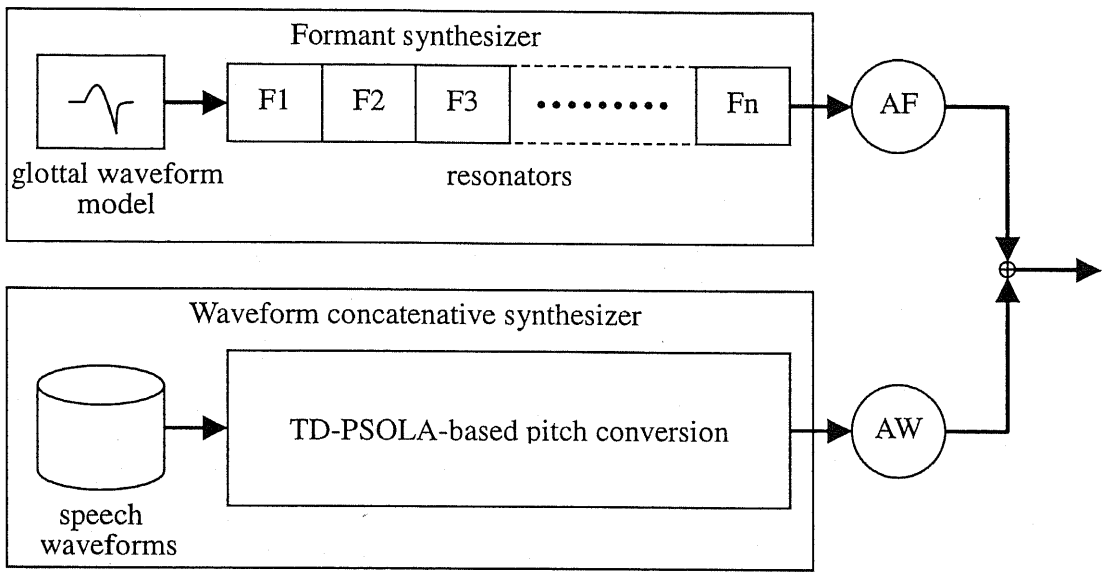


図 3.2: 波形編集を併用したフォルマント音声合成システムの構成

いても、声帯音源波形モデル上での声門閉鎖点を波形処理の基準点として用いており、全ての波形レベル処理はピッチ同期で行われる。

なお、フォルマント合成の音源として用いる声帯音源波形モデルは、我々の従来のフォルマント合成器同様、最大で6自由度のFL(Fujisaki-Ljungqvist)モデル[15]において、声帯開放時点での傾きと、声門閉鎖期間の2パラメータを固定した4自由度のモデルを用いている。

3.3 合成音声品質の予備的検討

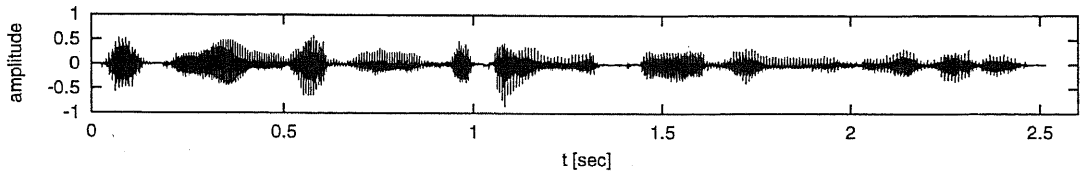
システムの構築に先立ち、合成システム出力の音声を模擬した合成音声を作成した。ここでは、波形編集方式による合成音声品質は充分に高いと仮定し、波形編集方式により合成する区間については、自然音声波形をそのまま用いた。また、フォルマント合成により生成する区間については、比較的容易に合成器パラメータの抽出ができる、母音のみを対象とした。すなわち、自然音声波形の母音区間をフォルマント合成音声で置き換えたものを、評価用合成音声とした。

3.3.1 評価用音声の作成

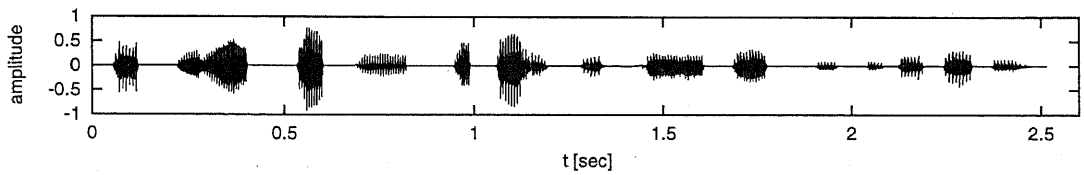
評価には、破裂音を多く含み、波形レベルでの接続の影響が大きいと考えられる、「爆音が銀世界の高原にひろがる」という文を用いた。破裂音を多く含む文を選択したのは、先述のようにターミナルアナログ方式では高品質な破裂音を合成することが難しいためである。

今回、サンプリング周波数は12kHzとした。使用したフォルマント合成回路は、声帯音源波形モデルにFL(Fujisaki-Ljungqvist)モデルを4自由度としたものを用いており、フォルマント数は6である。

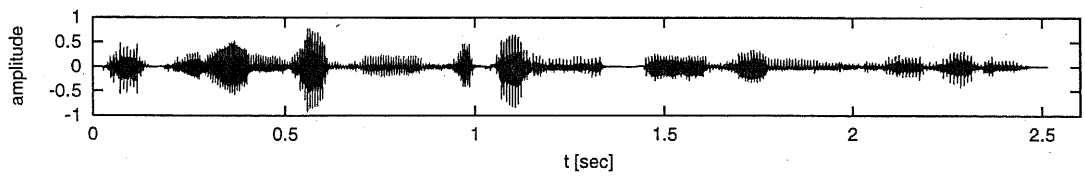
フォルマント合成で用いたVパラメータテンプレートは、自然音声と同一話者の離散発話による母音音声5個について、GAR法に基づく音源・声道伝達特性の分析を、独自に開発したフォルマント分析合成システム[50]を用いて行い、得られた定常部のパラメータ時系列から、1サンプルのみを用いて作成した。従って今回の実験ではC-V-C連鎖におけるVにおいて、そのスペクトル特性は基本周波数変動に伴う影響を除き一定である。一方、V-V連鎖については従来のフォルマント合成器で用いていたフォルマント接続規則に従い、パラメータを臨界制動2次



(a) 自然音声



(b) フォルマント合成による母音音声



(c) 評価用合成音声

図 3.3: 音声波形「爆音が銀世界の高原にひろがる」

系のステップ応答、あるいは直線により補間した。

また、置き換えの際に生じる自然音声とフォルマント合成音声の接続であるが、接続による波形レベルでの不連続を避けるために、波形編集音声、フォルマント合成音声をそれぞれ1ピッチの長さで、窓掛けし重ね合わせて接続した。

なお、自然音声波形における音素区間の切り出しは、波形の目視により行った。また、音源パラメータのうち声門閉鎖点は、自然音声波形の分析結果をそのまま用い、合成音声のピッチが自然音声と同じになるようにした。

3.3.2 評価実験

評価用の自然音声と合成音声を、録音室内でヘッドホンを用いて自由に聞かせ、合成音声の感想を自由に記述させた。被験者は13名である。

その結果、最も多かった指摘は、音がこもっている、あるいは高域成分が失われているというもので、これは9名により指摘された。次に多かったのが、接続に関する不自然さで、これは6名により指摘された。

前者の問題は、フォルマント合成音声の品質に起因したものと考えられる。この問題に対処するためには、モデルや分析方法の再検討が不可欠で、分析システムの改善が必要である。一方、後者の問題については、今回用いた合成音声が、簡単のために調音結合の影響を完全に無視したものであるため、調音結合の影響を考慮したテンプレートの作成や、パラメータ接続規則の修正により、比較的容易に改善できる可能性がある。

ところで、自然音声の音素とフォルマント合成により合成された音素との品質の差異を指摘したのは、13名中2名のみであった。このことから、両者の品質の差異は極端なものではなく、合成音声として受け入れられる水準にあると考えられる。

3.4 合成用テンプレートの作成

3.4.1 合成単位の検討

TTSシステム等での利用を目的とした音声合成器では、予め比較的短い音声単位のテンプレートを用意しておき、合成時にそれらを選択・接続して用いるのが一般的であるが、日本語では通常、子音の後に必ず母音が続き、母音の数が少な

く、また、フォルマント合成では母音区間内での接続がパラメータレベルで容易に行えることも考慮すると、蓄積の単位としてはCVやVCVが有効である。

我々の従来のシステムでは、蓄積単位としてCV単位が用いられており、CVテンプレートは先行母音によらず、各1つのみ用意されていた。実際には、各テンプレートは必要な全ての母音と接続可能であるように、実験に基づき予めパラメータが調整されており、合成時には子音部ではテンプレートをそのまま用いていたが、先行母音区間の後半では、子音との接続部分におけるスペクトルの不連続が緩和されるように、予め定められた接続規則に基づき、パラメータの操作が行われていた。これにより1つのCVテンプレートから、任意のVCV音声の合成を行っていた。

しかし、提案する合成システムのフォルマント合成部の回路構成は比較的簡単なものであるため、フォルマント合成により合成できる音素の制約が大きく、母音と子音の過渡部では、波形を利用する区間を比較的長くする必要が生じる。このために、波形部にも先行母音による調音結合の影響が大きく含まれてしまうため、先行母音に依存しないCVテンプレートを作ることは容易ではないと考えられる。従って提案システムでは、合成単位としてVCVを用いる。

なお、波形の利用はフォルマント合成のみによる合成システム構築と比較し、特に開発コストの面からは有効であるが、一方で先述したように合成の柔軟性が失われる。従って、波形の利用は、音声の分析が困難で高い品質での合成が容易ではない部分に限定することが望ましい。そこで、子音のうち、提案する合成器のフォルマント合成部の回路構成では生成機構に対応するモデル化ができない音声のみ波形編集方式により合成し、半母音等は波形編集ではなく、フォルマント合成により合成する。

以下では、波形編集方式により子音を合成するVCV音声のみを対象とし議論を行う。

3.4.2 自然音声波形からの子音波形の切り出し

テンプレート作成のために自然音声波形から子音部の波形を切り出す必要がある。まず、無声子音については基本的に自然音声中で声帯振動のない区間を子音部とした。ただし、声帯振動の終わりの部分や始まりの部分など、声帯振動が安定していない部分は分析が困難であり、また子音から母音への遷移部分も音素の知覚において重要と考えられるので、この部分はテンプレートの波形部に含め、合

成時にその特徴が失われないようにした。なおこの部分では、逆に品質が低下する危険性を考慮し、TD-PSOLA によるピッチ変換処理は行われない。

一方、有声子音の切り出しは無声子音と異なり声帯振動が持続していることが多く、子音区間を定義することが容易ではないが、本研究では、母音と子音、それぞれの遷移部において、サウンドスペクトログラム上で目視により急激に特性が変化している付近で声門閉鎖点を探し、その間を子音部とした。

実際には、接続時の波形レベルでのオーバラップ処理に用いるために、無声子音、有声子音とも、子音区間の前後で声帯振動の 1 周期分の長さだけ余計に切り出した。

上記の作業は、男性話者による VCV 孤立発話を DAT に録音し、12kHz サンプリング、16bit 線形量子化に変換した音声データについて行った。

3.4.3 VCV テンプレートの作成

VCV テンプレートは母音合成のためのフォルマント合成器パラメータと、子音波形から構成される。母音部のパラメータは、我々が開発したフォルマント分析合成システムを用いて行った。対象は、子音波形同様 12kHz サンプリング、16bit 量子化による音声データで、子音波形に用いた VCV 音声と同一の話者による孤立発話の母音 5 個である。これらをそれぞれ分析し、各 1 フレームの結果のみを用いた。すなわち、日本語のすべての母音は、計 5 個のパラメータベクトルから合成される。

フォルマント数は経験的に 6 とし、零点はないとした。すなわち、合成器モデルにおける ARX モデル次数は AR 係数について 12、MA 係数について 0 である。このモデル次数は、用いた母音音声において、実数軸上の極が生じないという条件を満たす。この条件は、実数軸上の極を認めると分析結果から得られた各フォルマントの時間軸上での連続性の特徴が悪化し、合成器制御が困難になる、という経験則に基づくものである。

実際には VCV 音声の前後の母音区間で、子音の調音結合の影響があり、母音区間の声道伝達特性が大きく変化する。この現象を表現するために、母音部でパラメータベクトル時系列を蓄積することも可能であるが、その場合、結果としてその区間では分析再合成を行うことになるためパラメータ制御が困難になり、合成器の柔軟性が低下する。従って、この調音結合の影響を規則により記述することが望ましい。調音結合の影響によるフォルマント遷移は臨界制動二次系のステッ

プ応答等により近似できると考えられるが、現時点ではルールを作成するのに充分な分析結果の蓄積がないため、本報告においては、簡単のために母音部での調音結合の影響は無視した。このため、母音と子音の接続点でスペクトル上において大きな不連続が生じ、品質が大きく低下する可能性がある。

提案システムによるVCV合成音声の例として、合成音声/aki/および/abi/の波形を図3.4に示す。

3.5 VCV音声の品質評価

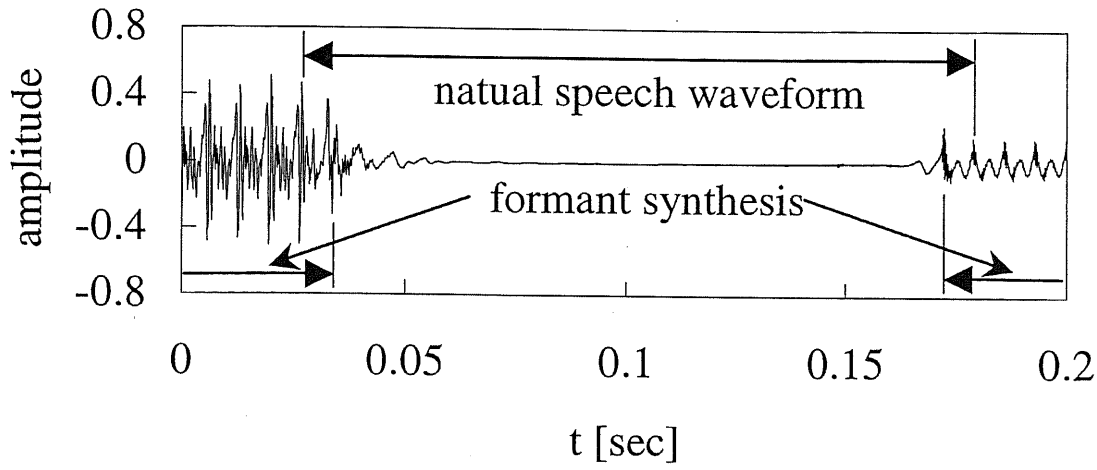
一般に、音声生成モデルに基づく方式であるフォルマント合成による合成音声の方が、波形編集方式による合成音声よりも品質が低いが、両者は品質の差異という形ではなく、全く声質の異なる音声として知覚される可能性がある。そのため、実際にVCV音声を合成した場合に、例えば、2人の話者が母音と子音をそれぞれ話しているように聞こえる、というような、極端な品質の低下が生じる危険性がある。

既に我々は先行研究において、ある文について、その母音部分を同一話者による別の孤立発話の分析結果から作成したフォルマント合成音声で全て置き換えた音声の品質について検討を行っており、自然音声波形とフォルマント合成による音声が接続可能であることを示唆する結果が得られている[49]。本報告では、これを確認するため、実際に提案システムによるVCV合成音声の明瞭度試験を行った。

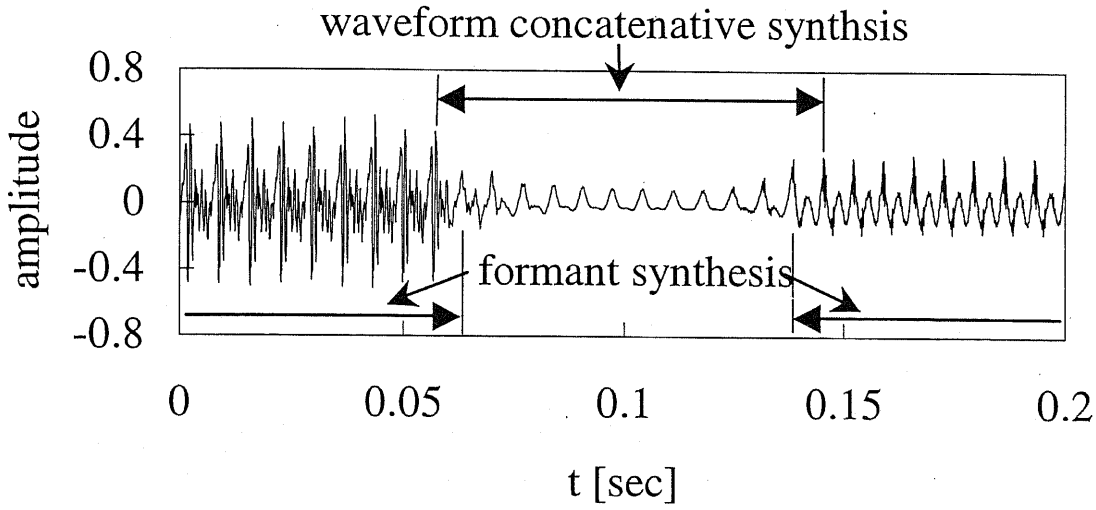
3.5.1 実験方法

対象は、日本語の母音のうちaまたはiと、提案システムにおいて波形を用いて合成する子音を組み合わせてできるVCV音声34個である。これらは、前節で作成されたVCVテンプレートを用いて作成されたもので、サンプリング周波数は12kHzである。声帯振動の基本周波数については、フレーズ成分に相当する成分のみ与えた。

また、我々の従来の合成システムを用いて、全ての音素をフォルマント合成で合成したVCV音声(女声)も同様に34個作成した。サンプリング周波数は同じく12kHzであるが、提案手法による合成音声の母音のフォルマント数が6であるのに対し、従来のフォルマント合成音声の母音のフォルマント数は5である。



(a) 合成音声/aki/



(b) 合成音声/abi/

図 3.4: VCV 合成音声例

これらを混合しランダムに並べ、録音室内で、ヘッドフォンを用い片耳にのみ提示し、その内容を仮名で書き取らせた。合成音声は5秒間隔で提示した。被験者は日本語の母語話者13名である。

なお提案手法による合成音声が男声であるのに対し、我々の従来のフォルマント合成システムは女声を対象としたものであるため、その結果を単純に比較することはできない。従来のシステムによる合成音声についても評価を行ったのは、子音合成のために相当の時間を要して修正を行ったテンプレートを用いたフォルマント合成音声の明瞭度を、参考として調べるためである。

3.5.2 実験結果

提案手法による音声の明瞭度を図3.5に示す。まず結果より、幾つか極端に明瞭度が低い音声があることがわかる。この原因として、母音部において調音結合の影響を無視しているために、母音と子音の接続部において、スペクトル上の大きな不連続が生じた影響が考えられる。また、無声子音よりも有声子音の方が、明瞭度の低い音声が多い。この原因として、まず、TD-PSOLAによる品質低下が考えられる。さらに、母音部で調音結合を無視した影響が、声帯振動が持続している有声子音でより顕著に表れた可能性も考えられる。

明瞭度が低いVCV音声のうちg音を含むもの以外では、同じ子音を含む別のVCV音声で明瞭度が高いものがある。そこで蓄積する波形を選ぶことにより、あるいは、複数の波形を蓄積しておき適当なものを選択することにより、容易に明瞭度は改善されると考えられる。一方g音を含むVCV音声については、すべて明瞭度が低い。その原因が、TD-PSOLAにあるのか、スペクトルの不連続性にあるのか、あるいは全く別の原因であるのかについて、詳細に検討する必要がある。

なお、従来のシステムにより全ての音素をフォルマント合成により合成したVCV音声の明瞭度は、平均で81.6%であり、提案手法によるものよりも高い値となっている。これは提案手法による音声のうち、極端に明瞭度の低いものが多いためで、十分に高い明瞭度を有する音声は提案手法によっても得られている。なお、提案手法によるVCVテンプレートセットは、従来のCVテンプレートセットと比べ、短時間で作成されている。

本実験において明瞭度が極端に低い音声については、VCVテンプレートを修正し、再度評価する必要があるが、テンプレートの修正により少なくとも80%程度の明瞭度は期待できることから、提案手法による音声合成は充分可能であると考

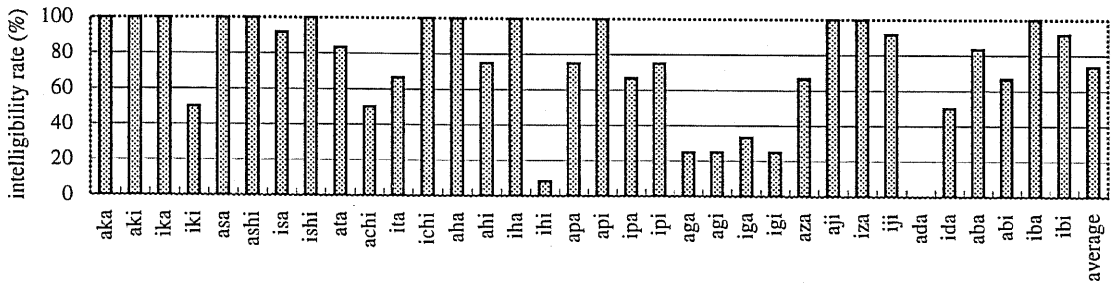


図 3.5: 提案手法による VCV 音声の明瞭度試験結果

えられる。

3.6 母音パラメータのみの変更による声質への影響の検討

波形の利用により、合成器の柔軟性が失われる。しかし提案手法は、日本語の音声合成においては、母音と比較し子音の柔軟性はそれほど重要ではない、という仮定に基づくものであり、その仮定についてもさらに検討する必要がある。

そこで、提案手法によるVCVテンプレートにおいて、そのVCVテンプレートにおける母音パラメータを、別の話者による母音から作成された母音パラメータに変更した際の、合成音声の品質について調べた。

具体的には波形テンプレート作成に用いた話者とは異なるある1名の話者による、テンプレート作成に用いた音声と同じ条件で収録された孤立発話の母音音声から、同じ条件で分析を行いフォルマント合成のパラメータベクトルを各母音について1つ抽出し、フォルマント合成のパラメータをそのベクトルで置き換えた合成音声と、元のテンプレートによる合成音声との比較実験を行った。

3.6.1 実験方法

本実験は2段階で行った。まず、母音パラメータを変えたことにより、その変化が知覚できるかどうかについて調べた。元になるVCVテンプレートから合成した音声(A)と、母音パラメータを変更し合成した音声(B)をそれぞれ34個ずつ作成し、同じ内容のVCV音声について、(i)AとBのペアを34組(AとBの順番はランダムである)、(ii)Aを2度繰り返すものを17組、(iii)Bを2度繰り返すものを17組それぞれ用意し、それらをランダムに配置した。これを被験者に各1度聞かせ、各組の中の2つのVCV音声が同じものであるか、異なるものであるかを判断させた。

次に、合成音声の品質の差異を評価させた。同じVCVのAとBの組を合計34個ランダムに提示した。各組内でのAのBの提示順もランダムである。各組について、2つのVCV音声の前者と後者についてどちらがどの程度よいのかを+2〜−2の5段階で選択させた。

被験者は日本語の母語話者11名で、録音室内でヘッドフォンを用いて両耳提示

した。この際、各組の2つの音声は1秒間隔で提示した。

3.6.2 実験結果

識別率を図3.6に示す。ただし、ここで識別率とは(i)の結果のみから計算した値であり、(ii)および(iii)の結果は含まれていない。なお、(ii)および(iii)の誤答は平均10%であった。結果から、多くのVCV音声で知覚的に異なる声質の合成音声となっていることが判る。

また、品質の評価については図3.7に示す。ただし、図での音声の順序は図3.6と同じで、また正の値は元のVCVテンプレートを用いた音声の方が品質的に高いことを示す。結果から、全体としては母音の変更により品質の低下が生じているものの、その平均値は0.5以下であり、全体としては大きな品質低下はないと言える。また、識別率と品質の低下との間の相関は、図より比較的小さいことがわかる。すなわち、声質の変化を品質の低下という形で知覚している可能性は低い。

以上の結果は、母音パラメータの変換のみで、大きく品質が低下することなく声質の変換が可能である、ということを示唆していると考えられる。

3.7 まとめ

本章ではフォルマント合成と波形接続方式のハイブリッド構成による音声合成回路構成を提案し、その構成において、異なる合成方式を組み合わせたことによる品質の破綻が生じないことを実験的に確認した。

そして子音波形生成に蓄積波形を用いた場合でも、VCVテンプレート作成の適切に行うことで、フォルマント合成の最大の特徴であるその柔軟性が大きく失われないことを、VCV合成音声の聴取実験により確認した。

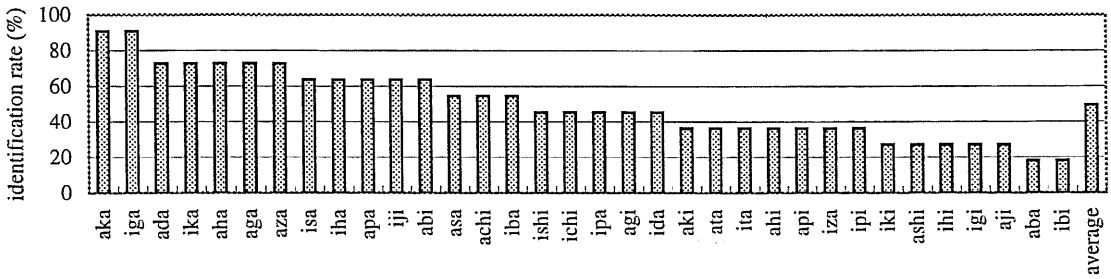


図 3.6: 元の VCV 音声と母音パラメータを変更した VCV 音声の識別率

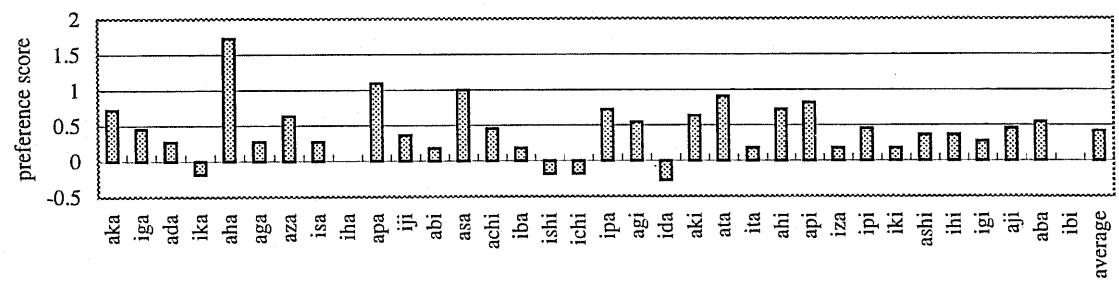


図 3.7: 元の VCV 音声と母音パラメータを変更した VCV 音声の品質の差

第 4 章

AR-HMMモデリングに基づく母音分析

4.1 はじめに

音声生成過程を音源と調音フィルタに分解して考えるソース・フィルタモデルに基づく音声合成において、両者の特性を明確に捉えることができれば、それぞれを独立に制御することによる柔軟な音声合成の実現が期待される。

従来、ソース・フィルタモデルに基づく多くの音声分析合成システムでは、線形予測分析等の手法により、自然音声波形を白色化するフィルタパラメータを求め、合成時には、その逆特性を調音フィルタの特性として与えている。一方、音源には、簡単のためにインパルス列と白色雑音源を組み合わせたものが用いられることが多い。

しかしこのモデルで、音源と調音フィルタを制御することは実際には容易ではない。なぜならば、モデル上の音源は実際の音声生成過程における音源の特徴の一部しか表現しておらず、調音フィルタにも生成過程における音源由来の成分が含まれているため、結果的に、音源と調音フィルタを独立に制御することが出来ないためである。このことを無視して独立に制御した場合、例えば音源の基本周波数を大きく変化させた際に極端な品質低下が起きるといったような問題が生じる。

従って、生成機構における音源と声道伝達特性という形で自然音声の特徴を分離することができれば、もちろん両者は相互に関連しており、完全に独立には制御できないせよ、従来よりも独立に制御した際の品質の低下を抑えることができると期待される。

本研究では、声道伝達特性が共振特性のみの積で表現されるという仮定に基づき、AR-HMM モデルに基づくパラメータ推定において、得られた AR パラメータから実極成分を取り除くことを繰り返すことで、音声合成に適した AR パラメータを推定する手法について検討を行った。

4.2 音源特性と声道伝達特性の分離

先述のように、自然音声波形を生成機構と対応した音源特性と声道伝達特性に分離することは、柔軟な音声合成に有効であると考えられる。しかし、声帯音源波形の直接観測は不可能であることから、自然音声波形を正しく音源特性・声道伝達特性に分離することは容易ではない。

これを実現する方法として、まず、比較的容易な手法として、線形予測分析により得られた極の分類による分離手法が考えられる。線形予測分析において、分析

次数を適当に制御することによりスペクトル包絡を表現するパラメータを推定することができるが、この際、フォルマントに対応すると考えられる共振特性に相当する複素共役な極の特性だけでなく、実数の極がしばしば分析結果に見られる。これは、主にスペクトル包絡の傾斜成分を表現するもので、主に音源特性に由来すると考えられる。一方、複素共役な極は声道伝達特性に由来すると考えられることから、実極と複素共役な極を分けることにより音源特性と声道伝達特性のおおよその分離は可能である。しかし、線形予測分析により得られる実極は、実際の声帯振動過程を考慮した場合、音源特性を表現するには精度的に不十分であると考えられ、これは結果的に音源・声道伝達特性双方の推定値の精度を低下させる。ところが、精度を上げるために分析次数を上げると、線形予測係数がスペクトル包絡特性ではなく、声帯振動の調波成分であるスペクトルの微細構造を表現するようになる。この場合、得られた極の特性を音源に由来するものか声道伝達特性に由来するものかを分類することが出来なくなり、音源・声道伝達特性の分離が不可能になる。

また、音源として声帯音源波形モデルを仮定する方法が試みられている。声帯音源波形モデルとしては、声帯音源波形上で声帯振動1周期を幾つかの区間に分け、それを多項式近似するRosenberg-Klattモデル[12]やFujisaki-Ljungqvistモデル[15]等があり、これを入力とするようなARXモデルの誤差が最小となる音源パラメータを推定することで音源パラメータと声道伝達特性に相当するARモデルパラメータを推定する手法が試みられている[15][17][18]。声帯音源波形モデルが実際の声帯音源波形をよく模擬すれば、この方法により精度よく、声道伝達特性・音源特性の分離が可能であると考えられる。しかし、モデルパラメータの推定問題は通常、非線形問題であり、一般にパラメータ推定に逐次近似が用いられる。さらに、実際のパラメータ推定問題では、局所的最適解が大量にあることが多く、かつ、誤差最小化基準による最適解が合成の観点からの最適解である保証もないため、実際の分析では、逐次近似推定の初期値として求めたい解に近い値を設定する必要が生じる。加えて、声門閉鎖時点に相当するパラメータが分析結果に大きく影響するために、線形予測分析の残差波形等から事前に声門閉鎖時点を推定しておく手法が良く用いられるが、この正確な自動推定は比較的困難である。これらのことから、声帯音源波形モデルとARXモデルを用いる手法での自動分析は容易ではない。

これに対し、提案する手法は、AR-HMMモデルに基づくものである。AR-HMMモデルは、線形予測分析のように残差波形の特徴として、フレーム内において定

常的なガウス雑音をモデル化するのではなく、HMMに基づき残差波形の特徴を記述するもので、より複雑な残差波形の特徴を表すことが可能である。つまり、分析フレーム全体から見たときに白色ではないような残差波形を仮定することができる。提案手法は、推定されたARパラメータのうち、音源に由来すると考えられるものを取り除き、それを残差波形で表現するような形に分析結果を誘導する反復推定手法であり、AR過程で声道伝達特性を直接的に表そうとするものである。そして提案手法の最大の特徴は、声帯音源波形モデルのように音源特徴に対する事前知識は不要であり、自動分析が容易であることである。

4.3 AR-HMMモデル

本節では提案手法において用いられる、AR-HMMモデルについて述べる。AR-HMMモデルは佐宗らにより導入されたモデルで、AR過程の音源モデルとしてHMMを用いる手法であり、そのモデル構造は図4.1に示す通りである。またHMMと残差波形との関係について例を図4.2に示す。

佐宗らはAR-HMMモデルを高基本周波数音声分析に用いており、その有効性が示されている[51]。

4.3.1 HMMに基づく音源モデル

音源モデルとして用いられるHMMは、各時刻における残差波形の振幅値を出力とするようなモデルである。各状態における生成確率分布には正規分布を仮定しており、また音源の周期性を表すために、図4.1に示すようなループ状の構造となっている。つまり、残差波形として周期的に変化するガウス雑音を認める、ということである。

このモデルを用いることで周期的な音源波形の影響をARパラメータの推定から除去することが期待される。

4.3.2 パラメータ推定手法

パラメータ推定の手順は次の通りである。ただし、以下では実際の計算アルゴリズムについてのみ述べる。

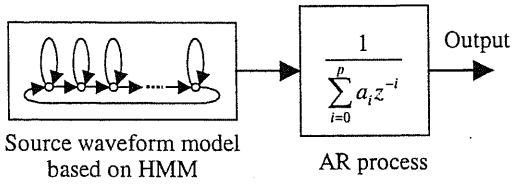


図 4.1: AR-HMM モデル

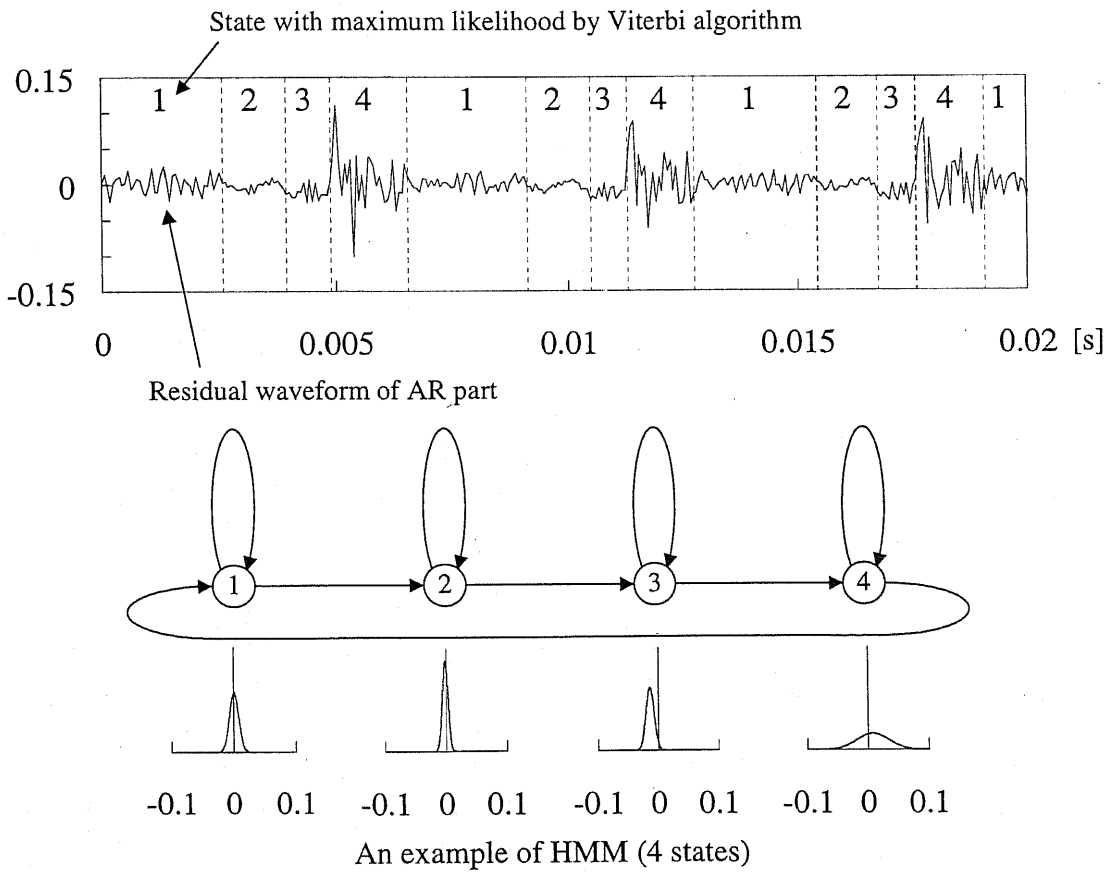


図 4.2: HMM の例 (状態数 4)

(1) まず音源の生成確率分布についてHMMの各状態sについて確率分布を表す正規分布の平均値$\tilde{\mu}_s$、分散$\tilde{\sigma}_s^2$、HMMの各状態遷移確率についてそれぞれ初期値を設定する。また、入力に対する状態遷移系列$(s_p, s_{p+1}, \cdots s_{N-1})$についても初期値を設定する。

(2) 状態遷移系列において予測誤差系列の尤度が最大となるARモデルパラメータと予測誤差系列を求める。ここでARモデルパラメータの推定値$\hat{\theta} = [\hat{a}_1 \hat{a}_2 \cdots \hat{a}_{p-1}]^T$は

$$\hat{\theta} = -[\Omega^T \tilde{\tilde{\Sigma}}_p^{-1} \Omega]^{-1} \Omega^T \tilde{\tilde{\Sigma}}_p^{-1} (\mathbf{y}_p - \tilde{\mathbf{m}}_p) \tag{4.1}$$

により与えられる。ただし、

$$\begin{aligned} \tilde{\mathbf{m}}_p &= [\tilde{\mu}_{s_p} \tilde{\mu}_{s_{p+1}} \cdots \tilde{\mu}_{s_{N-1}}]^T \\ \tilde{\tilde{\Sigma}}_p &= diag(\tilde{\sigma}_{s_p}^2, \tilde{\sigma}_{s_{p+1}}^2, \cdots, \tilde{\sigma}_{s_{N-1}}^2) \\ \Omega &= [\mathbf{y}_{p-1} \mathbf{y}_{p-2} \cdots \mathbf{y}_0] \\ \mathbf{y}_p &= [y_p y_{p+1} \cdots y_{N-1}]^T \end{aligned}$$

である。

(3) (2)で得られた尤度が収束していれば終了し、そうでなければ(4)に進む。

(4) Baum-Welchアルゴリズムにより予測誤差系列からHMMを推定し、HMMの各パラメータ(各状態の平均値、分散、および各状態遷移確率)を更新する。

(5) 予測誤差系列に対して(4)で推定したHMMで最尤となる状態遷移系列をViterbiアルゴリズムにより求め、その結果で状態遷移系列を更新する。

(6) (2)へ戻る。

なお、HMMの初期値であるが、本稿では簡単のために、線形予測残差波形全体の平均値および分散を全ての状態の確率分布として設定する。また、状態遷移確率は、音源波形の1周期によりHMMが1周するような確率を、基本周波数のおおよその値から計算し、各状態について等しく設定する。例えば、16kHzサンプリングによる男声話者に対し、HMMの状態数を8としたとき、同じ状態への遷移確率を0.95、次の状態への遷移確率を0.05程度に設定する。

上記アルゴリズムにより非定常的な特性を最も良く表すHMMを求めることで、分析フレーム内で非定常的だが周期的な音源に由来する特性と、分析フレーム内で定常的ば特性の分離が期待される。

4.3.3 声道伝達特性における線形予測分析との比較

従来、線形予測分析により声道伝達特性推定を行う場合、各フォルマントが共振特性を有する2次系により表現され、また音源に由来するスペクトル傾斜成分が数個の極により表現されると考え、経験的に、例えば、観測されるフォルマント数の2倍に2ないし4程度を足した数を、線形予測分析の分析次数として設定していた。つまりこれは、声帯音源特性は数個の極で表される、というモデルであるが、先述のように、音源特徴を記述するのに充分なモデルではなく、結果的に声道伝達特性、音源特性双方の推定値の精度が低下する。

一方、AR-HMM モデルに基づく分析においては、容易に残差波形を記述するHMM の自由度を上げることができる。しかし、その自由度により得られるARパラメータの結果が大きく影響を受ける可能性があり、その決定には注意を要する。また、HMM の学習過程はBaum-Welch アルゴリズムによるため、局所的最適解への収束しか保証されておらず、分析の初期値決定が分析結果に大きな影響を及ぼす。

なお、上記のアルゴリズムにおいては、HMM における各状態の出力確率分布が等しい状態で、まずARパラメータを推定するが、それは線形予測分析の残差波形に等しいため、分析結果としては、線形予測分析の結果の近くにある局所解に収束すると考えられる。

4.4 AR-HMMモデルに基づく実極除去による声道伝達特性の推定

先述のようにAR-HMM モデルの特徴の1つとして、高い自由度の音源モデルを扱えることがある。従来の線形予測分析のように数個の極による音源モデルをHMM に置き換えることによって、より精密なモデル化が可能である。しかしながら、AR-HMM モデル分析により得られるARモデルは依然として音源に由来する成分と声道伝達特性に由来する成分の両方が含まれており、これを分離しなければならない。しかし、ARモデル上で両者を分離する方法は、先述の線形予測分析の極分類と同じであり、その分析結果の精度はあまり高くない可能性がある。

そこで、ARモデル上での分離を行わない方法として、声道伝達特性はフォルマントに対応する共振特性のみで表現される、という仮定を置き、自然音声波形か

ら、声道伝達特性に対応する共振特性のみからなるARモデルパラメータと音源特性を表現するHMMへの分離をAR-HMMモデル推定を繰り返すことで行う分析手法を提案する。

4.4.1 分析手順

分析の手順は以下の通りである。

(1) 前節による手順に従いAR-HMM分析を行う。

(2) ARパラメータの極配置を調べる。得られたARパラメータに実極がなければ終了。

(3) ARモデルから実極の特性を除去する。

(4) 分析対象の音声に対する推定声道伝達特性の逆フィルタ波形で音源HMMを再学習する。

(5) AR次数を除去した極の数だけ下げ、AR-HMM分析を行う。

(6) (2)に戻る。

ここで(1)におけるAR-HMMモデル分析と(5)におけるAR-HMMモデル分析はその目的が異なる。(1)では線形予測分析同様のスペクトル包絡特性の推定を目的としているが、(5)では実極除去後の局所的な最適解探索がその目的である。

従来の線形予測分析同様、フォルマントの個数に対し余裕を持ったARモデル次数を初期値として設定し、本手法を繰り返すことで、ARパラメータの徐々に次数は下がり、フォルマントの個数に対応した次数となる。しかし、これを実現するためは、音源モデルが実極の特性を表す必要があり、HMMについてそれを表現するのに充分な数の状態数を設定する必要がある。

4.5 評価実験

線形予測分析とAR-HMMモデルに基づく反復推定を行わない分析、また提案手法を用いて、母音音声に対する音源・声道伝達特性の分離を行い、それぞれについて、得られたパラメータの精度に関して検討を行った。

対象はATR音声データベースの自由発話文のうち、50文中に含まれる全ての母音の中心部分である。まず、20kHzサンプリングのデータを16kHzにダウンサンプリングしたものを付属のラベルデータに含まれる母音中心の情報を用い、そこを中心とする528サンプルを切り出し、それを分析対象とした。各母音の個数を表4.1に示す。

分析であるが各サンプルについて、線形予測分析(LP)、反復推定を行わないAR-HMMモデル分析(AR-HMM)、提案手法(iterative AR-HMM)でそれぞれ分析を行い、声道伝達特性とその逆フィルタ波形(推定音源波形)を推定した。

この際、線形予測分析および反復推定を行わないAR-HMM分析においては、提案手法同様に極の分類を行い、複素共役な極のみを含むARパラメータと、そのARパラメータの逆フィルタ波形を求め、それぞれ推定声道伝達特性、推定音源波形とした。

予備的な実験の結果から、AR-HMMモデルにおけるHMM状態数については8、従来法におけるAR次数、および提案手法におけるAR次数の初期値は16とした。また$1-0.97z^{-1}$によるプリエンファシス特性を行った。

そして、得られた声道伝達特性を表すARパラメータについては、32次のLPCケプストラム係数を、推定音源波形について32次のDFTケプストラム係数を求めた。

分析の結果であるが、まず、得られた各ケプストラム(0次項は含まない)について中心を求め、中心との間とユークリッド距離をそれぞれ計算した。声道伝達特性、音源特性それぞれにおける各サンプルと中心との間の平均二乗距離の平方根の大きさを、それぞれ図4.3、図4.4に示す。まず図4.3より、声道伝達特性について、提案手法により得られる特性は、線形予測分析よりもその変動が小さいことがわかる。また図4.4より、音源特性についても/e/を除き提案手法において変動は最も小さく、/e/についても線形予測分析による結果と比較し、その変動の差は僅かである。声道伝達特性・音源特性ともに変動が小さくなっていることは、一方のパラメータ変動を抑えるためにもう片方の特性が犠牲になっていないことを示している。もちろん同じ母音でも特徴の変動があるため、平均距離の値はある値以下にはならないと考えられるが、変動を比較した際にその値が小さいということは、音声自体の変動以外の要因による変動が比較的に小さいことを示している。ソース・フィルタモデルに基づく音声合成を目的した分析の場合、安定した分離特性を有することが、最も重要であると考えられる。その点で、従来の線形予測分析と比較し提案手法が優れていることがわかる。

表 4.1: 分析に用いた母音サンプルの数

母音	a	i	u	e	o
個数	325	223	197	185	279

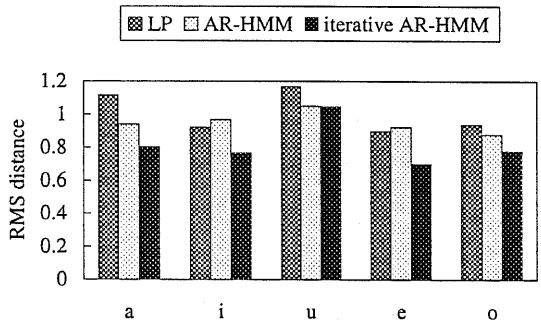


図 4.3: 推定声道伝達特性に対する 32 次 LPC ケプストラム距離空間における各サンプルとケプストラム中心間の平均二乗距離の平方根

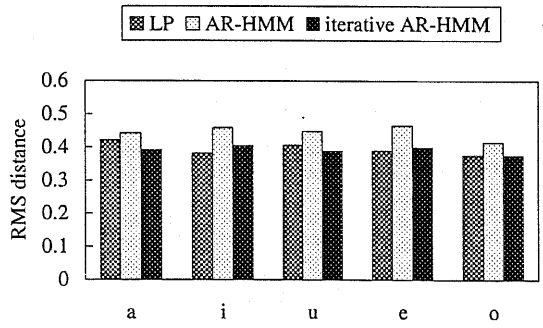


図 4.4: 推定音源波形に対する 32 次 DFT ケプストラム距離空間における各サンプルとケプストラム中心間の平均二乗距離の平方根

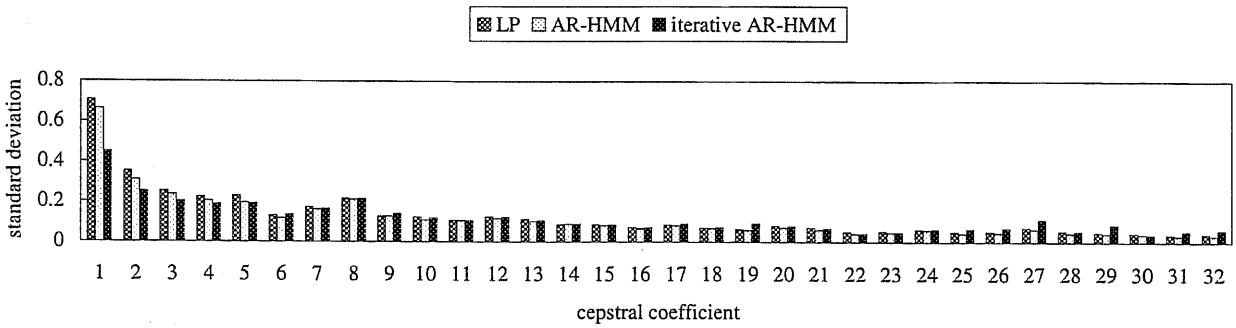


図 4.5: 母音 a の推定声道伝達特性に対する各ケプストラム係数の標準偏差

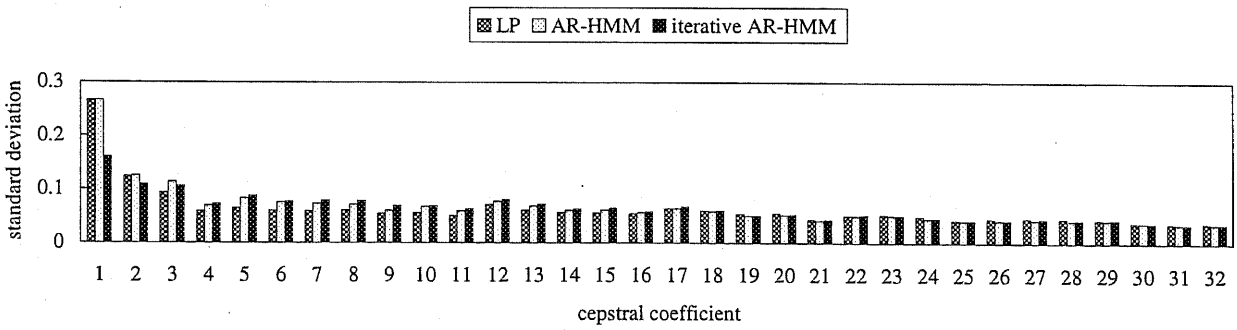


図 4.6: 母音 a の推定音源波形に対する各ケプストラム係数の標準偏差

一方、反復推定を行わないAR-HMM分析では線形予測分析よりも逆に変動が大きくなってしまっている場合が多い。これはモデルの自由度が高いために、逆に推定特性が変動してしまったと考えられる。しかし、反復推定によりAR特性から実極を取り除くことで分析結果の変動が小さくなる。このことは、ARモデルにおいて実極特性(主にスペクトル傾斜成分に対応すると考えられる)が支配的で、声道伝達特性・音源特性の推定値に大きな影響を及ぼすことを示唆している。

この点についてさらに調べるために、ARフィルタにより表されるLPCケプストラム係数について、次数毎に標準偏差を求めた結果を図4.5に、また推定音源波形のDFTケプストラムについて同様に次数毎に標準偏差を求めた結果を図4.6に示す。

これらの図より、提案手法により、主に1次のケプストラム係数の変動が抑えられていることが判る。1次のケプストラム係数はスペクトル傾斜に大きく関係していると考えらることから、提案手法により、線形予測分析よりスペクトル傾斜特性の変動が小さく推定されていると考えられる。

従来手法において変動が大きい理由の1つとしては、ARモデル次数の値を固定していることから音声サンプルによってその極配置が大きく変わってしまうことが考えられる。分析において、ARモデル次数を音声サンプル毎に適切に制御することができれば、その変動を多少は抑えることはできると考えられる。しかし、先述のようにモデルの制約から誤差を小さくすることが出来ず、また、AR次数制御には何らかの知識や、分析結果の蓄積等が必要で自動分析は遥かに困難なものとなる、と考えられる。

4.6 まとめ

本章ではAR-HMMモデリングを導入し、自然音声分析のための反復推定法を提案した。

そして、自動分析を目的とし特にモデル次数、または次数の初期値を固定した場合に、従来手法によるモデル誤差最小化基準の下では、得られる声道伝達特性・音源特性のスペクトル傾斜特性が大きく変動してしまうのに対し、提案手法による実極を含まないようなAR-HMMモデル推定ではそれが抑えられ、声道伝達特性・音源特性の分離に有効であることを示した。

第 5 章

頑健な高精度音声分析

5.1 はじめに

本章ではより誤差の少ないAR-HMMモデリングに基づく音声自動分析手法の実現に関して2つの提案を行う。

先に述べたAR-HMMモデル推定ではHMMによりモデル化される音源特性が真の音源特性を正確に表す場合に、AR過程のパラメータが推定される、というものであったが、音源のモデル化に伴う誤差を小さくするためにはHMMの状態数を大きく設定する必要がある。

しかしHMMの状態数を増やすと、実効的な1状態当たりの学習データ量が減少し、各状態における出力確率関数の分散値が極端に小さくなる。この結果、出力確率が極端に大きな値となる場合が生じ、HMM推定が不安定なものと成りやすい。分析フレーム長をより長く設定することで学習データ量は増えるが、AR過程の時不変性を仮定しているためそれには限界がある。

そこで本節では頑健な分析を実現するために、

- HMM推定過程における安定性の改善
- 初期HMMの設定手法

の2つについて注目し、それぞれ検討を行う。

5.2 逐次状態分割に基づくHMMの最適化

AR-HMMモデリングはインパルス駆動による合成音声のように音源の特徴をHMMでうまくモデル化できる場合、有効であると考えられる。

しかし、例えば声帯音源波形モデル駆動による合成音声や自然音声を対象とする場合、HMMにより音源波形を精密にモデル化するためには、HMMの状態数を大きく設定しなければならない。しかし、適切な初期HMMが与えられない場合、状態数の多いHMMの学習は容易ではない。AR-HMMモデリングにおけるHMMでは声帯振動の1周期とHMMの最尤状態遷移の1周期との対応関係が重要であるが、これを実現するために、HMMの状態数を逐次増加させる手法を導入する。具体的には、音声認識における音響モデル学習等で用いられた手法[52]等を参考に、ある1つの状態について、時間軸方向に2つの状態に分割する手法を用いる。

具体的な手順は次のとおりである。

ここでは、HMMの状態$S_k^{(n)}$を状態$S_{k1}^{(n+1)}, S_{k2}^{(n+1)}$に分割するとする。まず、出力確率分布については状態$S_{k1}^{(n+1)}, S_{k2}^{(n+1)}$共に、分割前の状態$S_k^{(n)}$の分布と等しい分布を設定する。また状態遷移確率については、状態$S_k^{(n)}$の自己遷移確率をpとするとき、分割後の状態$S_{k1}^{(n+1)}$における自己遷移確率をp、$S_{k1}^{(n+1)}$から$S_{k2}^{(n+1)}$への遷移確率を$1-p$、状態$S_{k1}^{(n+1)}$から$S_{k1}^{(n+1)}, S_{k2}^{(n+1)}$以外への状態遷移確率は0に設定する。状態$S_{k2}^{(n+1)}$については、$S_k^{(n)}$からの遷移確率を自己遷移確率も含めそのまま等しい値を設定する。最後に、以上を初期値としてBaum-Welchアルゴリズムにより HMMの再学習を行う。

分割する状態の決定方法であるが、HMM学習過程において1状態に割り当てられるサンプル数が実質的に極端に少ない状態が生じると、分散の推定値も極端に小さくなり、結果的に分析が不安定となりやすいことを考慮し、最尤状態遷移系列において割り当てられた数が多い状態を分割することとする。

5.3 予備的なAR-HMMモデル推定結果を利用した初期HMMの推定

真の音源分布と仮説分布と間の誤差が大きいと、AR-HMMモデル推定によるARパラメータ推定値は、真の極配置とは異なる局所解に収束する可能性が高くなる。従ってより安定した分析のためには、精密な分析を行う前に音源の仮説分布を適切に定めることが重要となる。そこで、まず音源の影響分を考慮した分析次数を設定した線形予測分析によりフォルマントの粗い推定を行い、その結果に基づきフォルマントの影響を除去した音声波形を用いて初期仮説を決定する方法がまず考えられる。

そこで音源の仮説分布を決定するために、線形予測分析の分析結果を利用を検討する。フォルマント1つが複素共役な極対により表され、また、音源のスペクトル傾斜分を表現するのに数個の極が必要である、との仮定に基づき、観測されるフォルマント数を2倍した値に2ないし4を足した数を線形予測分析の分析次数として設定し、得られた極の中から、共振特性を記述するものを選択し、各フォルマントのパラメータを推定する手法があるが、この結果をHMMの初期モデル学習に利用する方法がまず考えられる。

ところで、AR-HMMモデル推定において、HMMの状態数を比較的少なく設定した場合においても、実験的には、線形予測分析よりも安定した結果が得られる

場合が多い。HMM の状態数が少ない場合、声門閉鎖の付近において予測誤差の分散が大きくなり、それ以外の点では分散が小さくなるような推定結果となる傾向があり、これにより線形予測フィルタの推定において音源の仮説分布において分散が大きい部分の誤差が相対的に軽く評価されるため、声門閉鎖の影響が小さいフィルタ係数が推定されると考えられる。

以上より、まず比較的状態数の少ない HMM を用い、フォルマントの表現に必要な次数より若干高い AR 次数を設定した AR-HMM モデル推定を行い、推定された AR 過程の極からフォルマントに対応する共振特性の積を求め、音声波形に対するその逆フィルタ波形を用いて HMM を再学習し、これをより精密な AR-HMM モデル推定のための初期 HMM として利用する手法が考えられる。

5.4 フォルマント合成音声に対する分析実験

5.4.1 分析手順

分析対象とした音声は 16KHz サンプリング、16bit 量子化によるフォルマント合成音声で、基本周波数を 100.0Hz から 251.2Hz まで対数周波数軸上で等間隔になるよう 9 段階に変化させた男声母音/a/,/i/,/u/,/e/,/o/計 45 サンプルである。各母音のフォルマント数は 6 であり、駆動音源として FL モデル [15] を利用した。また分析に先立ち $1-0.97z^{-1}$ のプリエンファシスを行った。分析フレーム長は 272 サンプルである。

分析の手順は次の通りである。

(1) AR 次数を 16、HMM の状態数を 4 とし、AR-HMM モデル推定を行う。

(2) 得られた AR フィルタの極のうち、フォルマントの特徴を直接的に表現しないと考えられる極を除去する。ここでは実極および帯域幅が 4kHz 以上となる共振特性を有する複素極対を除去する。この際、フォルマントに対応する極配置を維持するために分析における AR 次数を下げる。そして音声波形に対する逆フィルタ波形を求め、それを学習データとして HMM を再学習する。

(3) Viterbi アルゴリズムにより最尤状態遷移系列を求め、割り当てられた数が多い状態を、前節で述べた手法により分割する。これを繰り返すことで、最終的に HMM の状態数を 16 にする。

(4) 既に推定されたHMMを初期HMMとして、AR-HMMモデル推定を再度行う。(この結果から得られたフォルマント推定結果をCとする)

また比較のために、状態分割を用いずに初期モデルの時点から16状態のHMMを用いたAR-HMMモデル推定と(この結果から得られたフォルマント推定結果をAとする)、上記において(2)の操作を省いた(AR次数を16のままとした)AR-HMMモデル推定を行った(この結果から得られたフォルマント推定結果をBとする)。

分析においては文献[51]の方法と同様に、初期仮説および状態遷移確率の初期値を乱数を用いて決定しているため、この影響を考慮しそれぞれの分析を各4回ずつ行い、さらに6個のフォルマントが抽出されない場合が生じることも考慮し、各サンプルについて、6個のフォルマントが推定されたもののうち、最も誤差の小さいものを最終的な分析結果として選択した。

5.4.2 実験結果

対数周波数軸上における推定されたフォルマントの中心周波数・帯域幅の平均二乗予測誤差を図5.1に示す。図5.1におけるAとBの比較から、全体的にはHMMの状態の逐次分割により、より誤差の小さい結果が得られていることが判る。一方、HMM状態数の少ないAR-HMMモデル推定結果を初期HMMの推定に用いる手法については、低次フォルマントの推定精度が改善される反面、逆に高次フォルマントの推定精度が低下する傾向が見られる。これは音源由来のスペクトル傾斜特性を取り除いたことにより、誤差波形とHMMから定めた音源仮説との誤差が、特に高域で強調された結果であると考えられる。改善のためには、音源仮説と真の音源との間で生じる誤差を減らすためにHMMの状態数をさらに増やす必要があるが、不適切な局所解に収束する可能性も大きくなるため、これについては更なる検討が必要である。

5.5 まとめ

音源に逆フィルタ波形を用いるフォルマント合成による分析合成を実現するため、AR-HMMモデリングに基づく音声分析手法について検討を行った。AR-HMMモデリングでは音源の仮説分布を適切に設定することが重要であるが、本節ではこれを実現するために、HMM状態の逐次分割手法と、予備的なAR-HMM分析結

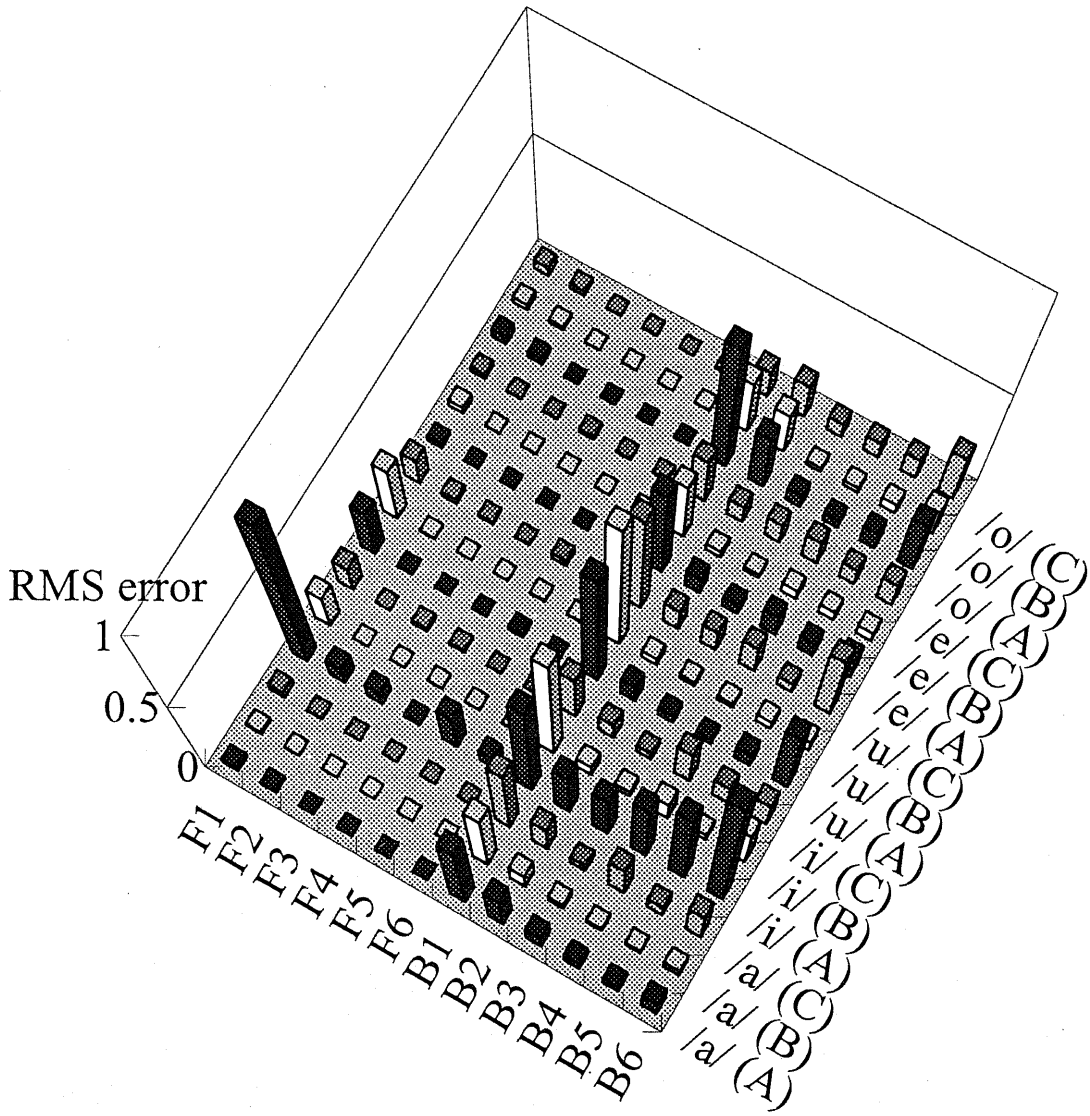


図 5.1: 対数周波数軸上における推定されたフォルマント周波数・帯域幅の平均二乗誤差

(Fn:第 n フォルマントの中心周波数,Bn:第 n フォルマントの帯域幅)

果を利用した初期 HMM の推定手法を導入し、フォルマント合成による男声母音音声に対する分析実験を行った。

第 6 章

フォルマントの高精度分析に基づく逆フィルタ波形駆動フォルマント合成

6.1 はじめに

前章で述べた音声分析手法は、モデルが実際の音声生成機構に厳密に対応しているならば、その音声のフォルマントを、AR過程における共振特性と完全に対応付けて取り扱うことができる。

AR過程における各共振特性をフォルマントと定義し、フォルマント合成は音声合成用ARフィルタの複素共役な極を制御し、音声を合成する手法、と考えれば、音声合成のためのフォルマント推定手法として、前章で述べた音声分析手法を用いることができる。

しかし、自然音声波形に対して、提案手法による音声のモデリングが妥当である、ということは何ら保障されておらず、その結果、推定される共振特性についても、それが生成機構の伝達特性に必ずしも対応したものではなくなり、推定結果の妥当性を別途判断する必要が生じる。

本章では、合成回路のモデル化をまず行い、そのパラメータとして前章における提案手法が妥当なものであるかを、主観評価実験に基づき実験的に調べ、議論を行う。

6.2 逆フィルタ波形駆動フォルマント合成

従来のフォルマント合成システムでは、駆動音源として声帯音源波形モデルが用いられることが多かった。インパルス駆動等の極端に単純なモデルと比較し、高い合成音品質を得ることが出来る反面、音源パラメータの推定が困難であり、分析合成系の実現は容易ではなかった。

そこで本節では、本論文において既に提案したAR-HMMモデリングに基づく合成フィルタの推定値を用い、フォルマント合成による分析合成系を提案する。

提案手法は、音源のモデル化は基本的に行わず、推定された合成フィルタから逆フィルタを求め、自然音声波形に対する逆フィルタ波形を駆動音源として用いるものである。このモデルでは合成フィルタを全く制御しなければ、原理的には品質の低下は生じないことになる。

実際には、柔軟性を得るために、合成フィルタ係数は制御することになるため、ある程度の品質の低下は生じる。また、特に音源の基本周波数制御のために別途ピッチ変換技術を音源部に適用する必要が生じる。ここではTD-PSOLAによるピッチ変換を考える。品質と蓄積波形数はトレードオフの関係にあると考えられ

るが、特に以下では、音韻を超えるような極端な制御を行った場合について議論を行う。

6.3 合成フィルタパラメータの非線形制御による母音合成

推定された共振特性の積が生成過程における声道伝達特性と正確に対応するとき、自然音声に対するその逆フィルタ波形は生成過程における音源と放射特性を表すことになる。両者の制御を行わないならば、その逆フィルタ波形を駆動音源とするフォルマント合成により、合成フィルタ制御時に品質低下の小さい音声合成が実現されると考えられる。実際の音源・声道伝達特性・放射特性は相互に関係し合っており、厳密には独立の制御は行うことが出来ないものの、生成機構を考慮すると、ある程度の独立性は期待できるからである。

もしこれが実現できれば、フィルタ制御により音韻を制御でき、かつ音源が分離されたことでピッチ変換による波形上での歪の影響が小さくなるため、少ない波形蓄積でも高品質かつ柔軟な音声合成システムを構成できることになる。

6.4 評価実験

提案手法による高品質かつ柔軟な音声合成システム構築の可能性を評価するために、前章で提案した音声分析手法に基づき、母音音声の分析・合成を行い、聴取実験に基づく合成音声の評価を行った。

6.4.1 合成音声の生成

実験試料として日本語音声データベースのうち、話者 MHT による孤立母音発話のデータを用いた。元のデータは 20kHz サンプリング、16 ビット量子化によるものであるが、これらを 16kHz にダウンサンプリングして利用した。

まず、音声分析を行い共振特性を推定した。分析条件は前章の手順と同じく初めに $1-0.97z^{-1}$ のプリエンファシスを行ったうえで、AR 係数 16 次、HMM 状態数 4 の AR-HMM モデル推定をまず行い、この結果から得られた予測誤差波形を用いて逐次状態分割により 16 状態の HMM を作成し、それを初期 HMM とする AR

係数16次、HMM状態数16の分析を行う。この際のフレーム長は336サンプルである。そしてフレームシフト長160サンプルとするオーバラップ分析によりこの操作を繰り返し、フィルタ係数の時系列を推定した。

そして、得られたフィルタ係数の時系列から、160サンプル周期で変化する逆フィルタを構成し、自然音際波形に対するその出力を音源波形として得た。そして音声の生成であるが、推定された音源波形を駆動音源とし、母音開始時点をそろえた上で、別種の母音のフィルタ係数時系列の推定値を用いて音声の合成を行った。

また、比較のために、AR-HMMではなく線形予測分析を用い、同様の手順で分析・合成を行った。

6.4.2 明瞭度試験

まず、分離の妥当性を評価するために、AR-HMMモデリングに基づく分析合成音25種類(音源波形5種、合成フィルタ係数時系列5セットの組み合わせ)を用いて、提示音声の書き取りによる音韻明瞭度試験を行った。比較対象用の線形予測分析に基づく分析合成音25種を加えた50種に、さらにダミーの音声計10個を加え、合計で60個の合成音声をランダムに提示した。

提示はヘッドフォンを用いた両耳提示で、16kHzサンプリング16ビット量子化である。なお、提示間隔は2秒とした。被験者は日本語の母語話者9名である。

実験の結果であるが、AR-HMMモデリングに基づく分析再合成音、線形予測分析に基づく分析再合成音共に100%の音韻明瞭度が得られた。これにより、音源・フィルタ特性分離の妥当性が示された。

6.4.3 自然音声との比較実験

明瞭度試験で用いた50種の分析再合成音それぞれについて、同じ種類の母音5種(自然音声)と組み合わせ、間隔1秒で2つの母音を提示し、総合的な品質について2種の母音のどちらが良いか、同程度の選択肢も含めて3段階で評価させた。2種の音声の提示順、および各セットの提示順はランダムである。

提示方法であるが、何度も繰り返し聞くことを認めた他は、先の明瞭度試験と同様の条件とした。また、被験者も同様に9名である。

実験結果を表6.1,6.2に示す。実験結果より、逆フィルタ波形を求めた母音と同種の母音を合成した場合、合成音と自然音声の間の相違がないことが示されてい

表 6.1: 分析合成音性と自然音声との比較実験の結果

	選択数					
	線形予測分析			提案手法		
合成音	a	b	c	a	b	c
aa	0	9	0	0	9	0
ai	9	0	0	8	1	0
au	9	0	0	9	0	0
ae	3	6	0	2	7	0
ao	7	2	0	7	1	1
ia	9	0	0	9	0	0
ii	0	9	0	0	9	0
iu	8	1	0	9	0	0
ie	9	0	0	7	2	0
io	9	0	0	9	0	0
ua	8	1	0	8	1	0
ui	8	1	0	7	2	0
uu	1	8	0	0	9	0
ue	8	0	1	8	1	0
uo	6	3	0	7	2	0

a: 自然音声の方が品質的に優れている

b: 品質的に自然音声と合成音声は同等である

c: 合成音声の方が品質的に優れている

(表において V_1V_2(V:母音) は母音 V_2 から抽出した逆フィルタ波形を用いて母音 V_1 を合成したものを表す)

表 6.2: 分析合成音性と自然音声との比較実験の結果 (続き)

	選択数					
	線形予測分析			提案手法		
合成音	a	b	c	a	b	c
ea	7	2	0	1	8	0
ei	9	0	0	8	1	0
eu	9	0	0	9	0	0
ee	0	9	0	0	9	0
eo	9	0	0	9	0	0
oa	9	0	0	5	3	1
oi	9	0	0	5	4	0
ou	9	0	0	8	1	0
oe	7	2	0	5	4	0
oo	0	9	0	0	9	0

a: 自然音声の方が品質的に優れている

b: 品質的に自然音声と合成音声は同等である

c: 合成音声の方が品質的に優れている

(表において V_1V_2(V:母音) は母音 V_2 から抽出した逆フィルタ波形を用いて母音 V_1 を合成したものを表す)

る。原理的には自然音声に対する逆フィルタリング後に、同じ係数でフィルタリングを行うことになるため当然の結果であるが、これは実験に用いた分析合成系が適当に構成されていることを示している。

音声生成機構とは異なり、フィルタ操作により全体のバランスが崩れているため、自然音声と比較し、品質の低下することは避けられないが、提案手法においてより、自然音声と同等の品質である、との回答が多いことが示されている。例えば自然音声/e/から得た逆フィルタ波形を用いて生成した合成音声/a/や、その逆の/a/の逆フィルタを用いた合成音声/e/については自然音声と同等の品質との回答が多い。この結果から、実際の合成システムにおいて、音韻を超えた範囲で蓄積波形数の削減が可能な場合が存在しうることが示された。

6.4.4 TD-PSOLA に基づくピッチ変換の影響比較

同じ種類の母音の自然音声波形と分析再合成音について、TD-PSOLA 法に基づきそれぞれ音声波形に対するピッチ変換 (従来手法) と、音源波形に対するピッチ変換 (提案手法) を行った。これらをセットとし先の実験同様の比較実験を行った。ただし、ピッチ変換により影響だけを見るために、用いた分析再合成音は、音源とフィルタ係数時系列は同じ母音データから推定したものをセットとしたもののみ、AR-HMM モデリングに基づく手法、線形予測分析に基づく手法それぞれについて 5 種ずつとした。

評価対象はピッチ変換率をそれぞれ 0.50, 0.72, 1.41, 2.00, 2.83 の 5 種類とする、合計 50 セットである。その他は被験者数も含め、先の自然音声との比較実験と同様の条件とした。

実験結果を表 6.3 に示す。全体的には従来手法と提案手法による差がない、との回答が多いが、線形予測分析に基づく分析結果を用いた場合よりも、AR-HMM モデリングに基づく前章で提案した分析手法を用いた場合の方が、品質の低下が少ないことが判る。これは前章での提案手法による分析が、線形予測分析よりも音声の特徴をより適切に表しているため、と考えられる。

また、ピッチ変換率が 100%とから離れるほど、比較において提案手法の結果が良くなっている傾向があることがわかる。これはピッチ変換率が 100%から離れるほど TD-PSOLA では波形の歪が大きくなるが、音声波形に対する TD-PSOLA よりも逆フィルタ波形に対する TD-PSOLA の方がその影響が小さいことを示唆した結果であると考えられる。

表 6.3: TD-PSOLA によるピッチ変換の影響に関する比較実験結果

		選択数					
		線形予測分析			提案手法		
ピッチ変換率 (%)	合成音	a	b	c	a	b	c
50	a	0	8	1	1	8	0
	i	3	6	0	0	9	0
	u	3	5	1	0	8	1
	o	0	6	3	3	6	0
71	a	2	7	0	1	8	0
	i	6	3	0	1	8	0
	u	0	9	0	0	8	1
	e	1	6	2	3	6	0
	o	1	5	3	1	5	3
141	a	3	4	2	1	6	2
	i	7	2	0	5	4	0
	u	4	5	0	3	4	2
	e	3	4	2	3	4	2
	o	5	3	1	0	9	0
200	a	4	4	1	0	9	0
	i	7	2	0	7	2	0
	u	3	6	0	1	5	3
	e	3	6	0	2	2	5
	o	7	1	1	0	5	4
283	a	2	5	2	2	5	2
	i	7	1	1	3	6	0
	u	4	4	1	1	8	0
	e	1	5	3	1	5	3
	o	5	2	2	1	7	1

a: 従来手法の方が品質的に優れている

b: 品質的に従来手法と提案手法は同等である

c: 提案手法の方が品質的に優れている

6.5 まとめ

蓄積波形駆動によるフォルマント合成システムを提案した。そして、分析合成音の聴取実験を行い、駆動音源波形を取得した自然音声波形と合成フィルタ係数を抽出した自然音声波形の音韻が異なる場合においても母音については100%の音韻明瞭度が得られることを実験により確認した。

また提案システムにおいて、駆動音源波形と合成フィルタの音韻を超えた形での組み合わせが可能であり、品質の低下を抑えつつ、大幅な蓄積音源波形数削減の可能性があることを比較聴取実験により確認した。

さらに、TD-PSOLAによるピッチ変換の影響について調べ、駆動音源波形に対するTD-PSOLAにより、少なくとも従来の音声波形に対する直接的なピッチ変換と同様の、組み合わせによっては従来法よりも品質低下を抑えたピッチ変換が実現されることを比較聴取実験により確認した。

第 7 章

結論

本論文では、高品質かつ柔軟な音声合成の実現に関し、特にパラメトリックな手法に基づく分析合成方式による合成音声品質の改善のために、ノンパラメトリックな手法の導入についての議論を行った。

第2章において、まず音声合成技術のうち波形生成手法を中心に、それに関係する知見・手法等も含めた研究の動向について述べた。計算機で利用可能な記憶容量が増えたことで、いわゆるコーパスベースの音声合成手法が主流となったが、単純な波形接続方式に限らず、分析合成方式におけるコーパスベースの統計的手法についても紹介した。また、パラメトリックな分析手法に関連する信号処理手法、ノンパラメトリックな手法に基づく高品質な分析合成手法等に紹介した。

第3章においてパラメトリックな手法に基づく音声合成システムにおいて全体としての品質性能を向上させるために、ノンパラメトリックな手法を導入する手法について論じた。従来のフォルマント合成において、子音はパラメータの制御がほとんど不可能で、かつ高品質な音声合成が困難である。この場合、パラメトリックのメリットが既に失われてしまっていることから、波形というノンパラメトリックな特徴を導入することで、全体としての性能向上を達成する、という方法について検討を行った。

第4章ではAR-HMMモデリングを導入し、AR-HMMモデル推定の反復推定よる自然音声分析手法を提案した。まず自然発話中に含まれる母音音声に対し、線形予測分析、AR-HMM分析、提案手法でそれぞれ分析を行い、各母音ごとに、推定された音源特性・フィルタ特性の32次ケプストラム空間におけるパラメータの広がりについて調べた。その結果、提案手法により、分布の小さいパラメータが得られることが確認された。

第5章では第4章で述べた手法に基づき、高精度でかつ頑健な音声分析を実現するための2つの手法の提案を行った。1つがHMM状態数の逐次状態分割手法であり、もう1つがAR-HMMモデリングにおける初期HMMの推定手法である。その特性が判っている、ということからフォルマント合成による合成音声波形に対して分析を行い、提案手法が有効であることを客観評価に基づき示した。

第6章ではAR-HMMモデリングに基づく音声分析手法を前提とした逆フィルタ波形駆動フォルマント合成を提案した。また音声合成においては、最終的な評価は合成音品質が基準となる。このため、客観評価だけでなく主観評価が重視される。第4章および第5章で提案した音声分析手法による分析の妥当性を評価するため、推定フォルマントに対する逆フィルタ波形により駆動されるフォルマント合成器を構築し、それを用いた分析再合成音に対する評価を行った。この結果、

音韻レベルでの音声変形、というノンパラメトリックな手法では困難な複雑な音声変形処理を、提案した音声合成手法により高品質で実現できることが示された。

今後の課題としては、蓄積波形駆動のフォルマント合成と波形接続合成のハイブリッド合成システムの実用レベルでの実装や、そのためのロバストな音声分析に基づく統計的なフォルマントパラメータモデルの作成、また柔軟かつ高品質な音声合成を実現する音声合成システムの規模に関する検討などが挙げられる。

謝辞

本研究を進めるにあたり、6年間にわたり親身に御指導をして下さいました広瀬啓吉教授に深く感謝致します。

研究に関する議論、論文執筆等に関しましてご指導を頂きました峯松信明助教授に感謝いたします。また、研究活動を様々な面で支えて下さいました、高橋登技官に感謝致します。そして、お忙しい中、聴取実験にご協力頂き、さらに、研究に限らずいろいろとお世話になりました広瀬研究室、峯松研究室の学生の皆様に感謝致します。

ありがとうございました。

参考文献

[1] 広瀬啓吉, "音声の出力に関する研究の現状と将来," 音響学会誌 52, 11, pp. 857-861 (1996).

[2] 古井貞熙, "ディジタル音声処理," 東海大学出版会 (1985).

[3] 今井聖, "音声信号処理," 森北出版 (1996).

[4] 今井聖, "対数振幅近似 (LMA) フィルタ," 信学論 (A), vol. J63-A, pp. 886-893, 1980.

[5] McAulay, R. J., and Quatieri, T. F., "Speech Analysis/Synthesis Based On a Sinusoidal Representation," IEEE Trans. on ASSP, ASSP-34, 4, pp. 744-754, 1986.

[6] 河原英紀, 増田郁代, "音声分析・変換・合成法 STRAIGHT のスペクトル近似特性の評価と改良について," 信学技報, SP96-97, pp.19–24, Jan. 1997.

[7] 河原英紀, "聴覚の情景分析が生んだ高品質 VOCODER: STRAIGHT," 日本音響学会誌, 54, 7, pp. 521-526, Jul. 1998.

[8] Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A., "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, vol. 27, pp. 187-207, 1999.

[9] Charpentier, F.J. and Stella, M.G., "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," Proc. IEEE-ICASSP86, 2015-2018 (1986).

[10] Valbret, H., Moulines, E. and Tubach, J. P., "Voice transformation using PSOLA technique," Speech Communication, 11, pp. 175-187 (1992).

[11] D. Klatt, "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. 67, pp. 971-995 (1980).

[12] D. Klatt and L. Klatt, "Analysis, sysnthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am 87(2), pp. 820-857 (1990).

[13] 浅野太郎, "テキストからの音声合成におけるターミナルアナログ方式による音声波形生成処理に関する研究," 東京大学大学院工学系研究科電子工学専攻修士論文 (1993).

[14] 今井聖, 住田一男, 古市千枝子, "音声合成のためのメル対数スペクトル近似 (MLSA) フィルタ," 信学論 (A), vol. J66-A, pp. 122-129, 1983.

[15] H. Fujisaki and M. Ljungqvist, "Proposal and Evaluation of Models for the Glottal Source Waveform", Proc ICASSP, 31.2, pp. 1605-1608 (1986).

[16] M. Ljungqvist, "Speech analysis-synthesis based on modeling of voice source and vocal-tract characteristics," 東京大学大学院工学系研究科電気工学専攻学位論文 (1986).

[17] 藤崎博也, マッツ ユンクヴィスト, "声帯音源波形の新しいモデルとその音声分析への応用," 信学論 (D-II), J72-D-II, 8, pp. 1109-1117, Aug. 1989.

[18] W. Ding, H. Kasuya and S. Adachi, "Simultaneous Estimation of Vocal Tract and Voice Source Parameters Based on an ARX Model," IEICE Trans. IS vol. E78-D, 6, pp. 738-743 (1995).

[19] W. Zhu and H. Kasuya, "A speech analysis-synthesis-editing system based on the ARX speech production model," J. Acoust. Soc. Jpn. (E) 19, 3, pp 223-230 (1998).

[20] K. Funaki, Y. Miyanaga and K. Tochinai, "A time varying ARMAX speech modeling with phase compensation using glottal source model," Proc. ICASSP 97, 2, pp. 1229-1302 (1997).

[21] 舟木慶一, 宮永喜一, 栃内香次, "Glotal-ARMAX 音声分析法におけるマルチレート処理の導入について," 音響学会誌 55, 2, pp. 75-82 (1999).

[22] 片山徹, “システム同定入門,” 朝倉書店 (1994).

[23] 福島雅夫, “数理計画入門,” 朝倉書店 (1996).

[24] 阪本正治, 斉藤隆, 鈴木和洋, 橋本泰秀, 小林メイ, “波形重畳法を用いた日本語テキスト音声合成システムについて,” 信学技報, SP95-6, pp.39-45 (1995).

[25] 村松茂樹, “ピッチ同期処理に基づく波形編集型音声合成の研究,” 東京大学大学院工学系研究科電子情報工学専攻修士論文 (1999).

[26] 古園貴則, “ホルマント合成音声の高品質化の研究,” 東京大学工学部卒業論文 (1995).

[27] 柳瀬隆史, “波形重畳を併用したターミナルアナログ音声合成,” 東京大学工学部卒業論文 (1997).

[28] 長谷川澄志, “波形接続を併用した柔軟な構成のターミナルアナログ音声合成システム,” 東京大学大学院工学系研究科電子情報工学専攻修士論文 (1998).

[29] Y. Miyanaga, N. Miki and N. Nagai, “Adaptive identification of a time-varying ARMA speech model,” IEEE Trans. ASSP-34, 3, pp. 423-433 (1986).

[30] 豊島拓史, 三木信宏, 永井信夫, “スペクトル傾き補正を組み入れた適応的ホルマント推定,” 信学論 (A), J-74A, 6, pp.813-821 (1991).

[31] 匂坂芳典, “コーパスベース音声合成,” 信号処理, vol. 2, no. 6, (1998-11).

[32] 小林隆夫, “HMM に基づく音声合成,” 日本音響学会講演論文集, pp. 201-204, (1999-9).

[33] 古井貞熙, “音声情報処理,” 森北出版, 1998.

[34] “ CHATR speech synthesis,”
http://www.itl.atr.co.jp/chatr/

[35] N. Iwahashi, N. Kaiki, Y.Sagisaka, “Speech segment selection for concatenative synthesis based on spectoral distortion,” IEICE Trans. vol. E76-A, no. 11. pp. 1942-1948 (1993).

[36] A. W. Black, N. Campbell, "Optimising selection of units from speech database for concatenative synthesis," Proc. EUROSPEECH '95, pp. 581-584, (1995-9).

[37] 峯松信明, 中川聖一, "残差波形からのピッチパルス駆動点検出におけるエラー低減の試み," 音講論, 2-7-8, pp. 231-232 (1997-3).

[38] R. E. Donovan and P. C. Woodland, "Improvements in an HMM-based speech synthesiser," Proc. EUROSPEECH '95, pp. 573-576 (1995-9).

[39] X. D. Huang, A. Acero, J. Adcock, H. W. Hon, J. Goldsmith, J. Liu and M. Plume, "WHISLTLER: A trainable text-to-speech synthesis," Proc. of ICSLP, pp. 2387-2390, 1996.

[40] X. Huang, A. Acero, H. Hon, Y. Ju, J. Liu, S. Meredith and M. Plumpe, " Recent improvements on Microsoft's trainable text-to-speech system - WHISTLER," Proc. ICASSP '97, pp. 959-962, 1997.

[41] 益子貴史, 徳田恵一, 小林隆夫, 今井聖, "動的特徴を用いたHMMに基づく音声合成," 信学論 (D-II), vol. J97-D-II, no. 12, pp. 2184-2190, Dec. 1996.

[42] 徳田恵一, 益子貴史, 小林隆夫, 今井聖, "動的特徴を用いたHMMからの音声パラメータ生成アルゴリズム," 音響学会誌, vol. 53, no. 3, pp. 192-200, Mar, 1997.

[43] 長谷川澄志, 広瀬啓吉, "柔軟な構成のターミナルアナログ音声合成システムとそれによる音声合成実験," 音講論集, 1-7-8, pp. 195-196, Mar. 1998.

[44] K. Hirose and H. Fujisaki, "A System for the synthesis of high-quality speech from texts on general weather conditions," IEICE Trans. Fundamentals, vol. E76-A, no. 11, pp. 1971-1980, Nov. 1993.

[45] H. Fujisaki and M. Ljungqvist, "Proposal and Evaluation of Models for the Glottal Source Waveform," in Proc ICASSP, 31.2, pp. 1605-1608, Apr. 1986.

[46] 阿竹義徳, 入野俊夫, 河原英紀, 陸金林, 中村哲, 鹿野清宏, "調波成分の瞬時周波数を用いた基本周波数推定方法信学論 (D-II), Vol. J83-D-II, No.11, pp. 2077–2086, Nov. 2000.

[47] 平井啓之, 橋本誠, 大西宏樹, "STRAIGHTを用いたテキスト音声合成の開発と評価," 音響学会2001年秋季研究発表会講演論文集, 1-P-7, pp. 369-370, Oct. 2001.

[48] 河原英紀, 阿竹義徳, "音声の群遅延特性に基づく声門閉止等のイベント抽出について," 信学技報, SP99-171, pp.33–40, March 2000.

[49] 西澤信行, 峯松信明, 広瀬啓吉, "波形編集とターミナルアナログを併用した音声合成の検討," 音講論集, 2-6-17, pp. 315-316, Mar. 2001.

[50] Nishizawa, N., Minematsu, N., and Hirose, K., "Development of a formant-based analysis-synthesis system and generation of high quality liquid sounds of Japanese," Proc. ICSLP, Beijing, vol. I, pp. 725-728, Oct. 2000.

[51] 佐宗晃, 田中和世, "HMMによる音源のモデリングと高基本周波数に頑健な声道特性抽出," 信学論 (D-II), J84-D-II, 9, pp. 1960-1969, Sep. 2001.

[52] 鷹見淳一, 嵯峨山茂樹, "逐次状態分割法による隠れマルコフ網の自動生成," 信学論 (D-II), J76-D-II, 10, pp. 2155-2164 (1993-10).

発表文献

[1] 西澤信行, 広瀬啓吉, "基本周波数の影響を考慮したフォルマント音声合成," 日本音響学会平成12年度春季研究発表会講演論文集, Vol. I, 1-7-10, pp. 215-216 (2000-3).

[2] 西澤信行, 峯松信明, 広瀬啓吉, "フォルマント分析合成システムの開発と流音合成," 電子情報通信学会技術研究報告, SP2000-31, pp. 33-40 (2000-7).

[3] 西澤信行, 峯松信明, 広瀬啓吉, "ターミナルアナログ合成による高品質な流音の生成," 日本音響学会平成12年度秋季研究発表会講演論文集, Vol. I, 1-Q-4, pp. 237-238 (2000-9).

[4] Nobuyuki Nishizawa, Nobuaki Minematsu, Keikichi Hirose, "Development of a Formant-based Analysis-Synthesis System and Generation of High Quality Liquid Sounds of Japanese," Proceedings 6th International Conference on Spoken Language Processing, Beijing, Vol. I, pp. 725-728 (2000-10).

[5] 西澤信行, 峯松信明, 広瀬啓吉, "波形編集とターミナルアナログを併用した音声合成の検討," 日本音響学会2001年春季研究発表会講演論文集, Vol. I, 2-6-17, pp. 315-316 (2001-3).

[6] 西澤信行, 峯松信明, 広瀬啓吉, "波形編集を併用したフォルマント音声合成－VCV音に関する検討," 電子情報通信学会技術研究報告, SP2001-20, pp. 35-42 (2001-5).

[7] 西澤信行, 峯松信明, 広瀬啓吉, "自然音声波形を併用したハイブリッド型フォルマント音声合成システムにおける子音波形テンプレート削減の検討," 日本音響学会2001年秋季研究発表会講演論文集, Vol. I, 1-2-16, pp. 237-238 (2001-10).

[8] 西澤信行, 峯松信明, 広瀬啓吉, "HMMによる音源モデルを用いたフォルマント合成パラメータ推定," 日本音響学会2002年春季研究発表会講演論文集, Vol. I, 1-P-3, pp. 357-358 (2002-3).

[9] 西澤信行, 広瀬啓吉, 峯松信明, "音声合成のための AR-HMM モデルに基づく音声分析手法の検討," 電子情報通信学会技術研究報告, SP2002-63, pp. 35-40 (2002-07).

[10] Nobuyuki Nishizawa, Keikichi Hirose, Nobuaki Minematsu, "Separation of Voiced Source Characteristics and Vocal Tract Transfer Function Characteristics for Speech Sounds by Iterative Analysis Based on AR-HMM model," 7th International Conference on Spoken Language Processing ICSLP-2002, Denver, Colorado, Vol. 3, pp. 1721-1724 (2002-09).

[11] 西澤信行, 広瀬啓吉, 峯松信明, "柔軟な音声合成のための AR-HMM モデルに基づく AR フィルタ係数の推定, " 日本音響学会 2002 年秋季研究発表会講演論文集, Vol. I, 2-1-2, pp. 281-282 (2002-09).

[12] 西澤信行, 広瀬啓吉, 峯松信明, "音声合成のための AR-HMM モデリングに基づく音声自動分析, " 日本音響学会 2003 年春秋季研究発表会講演論文集 (2003-03 発表予定).