論文の内容の要旨

論文題目　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

氏名　西澤 信行

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った。

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しまい、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、声帯音源波形を数式により表現したモデル(声帯音源波形モデル)をその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。
特に母音型音声については、声帯音源波形モデルとARX(Auto-regressive with exogenous 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って、フォルマント合成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となる。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、以降、本論文においては主に母音型音声に対し議論を行った。
一方、フォルマントの制御方法に関する問題に対し、将来的にはパラメータ制御に統計的モデルを利用することが考えられる。特に近年、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。この手法では、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のパラメータとして用いることが一般的であるが、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、大量の分析結果に基づく高精度なモデルを作ることは困難である。
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、AR-HMM モデリング手法を用い、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 で表現されるモデルで表すことで、より精密に音声のモデル化を行う手法である。佐宗らによると、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がモデル上でより精密に表され、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。ここで用いられるループ状の HMM は周期性を有し、かつ振動周

期に揺らぎが存在する声帯音源波形の表現にも適したものであると考えられるが、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 モデル推定は AR 部と HMM 部の反復推定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さらに、モデルにおける AR 部に実極が現れなくなるまで AR 次数が減らされ、一方でその分の特徴が HMM において表現されるように分析が誘導される。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、生成過程における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、母音毎に、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を求めた。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について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ている、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。以上より、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。

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も重視される。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が用いられており、スペクトル包絡に対する非線形な制御が比較的容易である。実験の結果、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。