

学 位 論 文 題 名

Speech Analysis Methods Based on Accurate Models

(高精度なモデルに基づく音声分析法)

学位論文内容の要旨

近年、パーソナルコンピュータや LSI の低価格化により、携帯電話やカーナビを始めとする多彩な音声応答関連の製品が販売されている。音声のデジタル信号処理技術は、1970 年代より研究されている歴史のある研究分野であるが、現在市販されている音声応答製品は、音質の面で十分な品質を実現しているとは言い難い。音質改善の 1 つの方向性として、モデルの高精度化を挙げることができる。本研究は、音声の生成モデルを高精度化した音声分析法の提案である。

本研究では 2 種類の高精度音声生成モデルを導入し、各モデルに基づく音声分析法の提案を行なっている。2 種類のモデルとは、音声の声道特性を ARMAX モデル、入力を白色 GAUSS 過程と声帯音源波モデル励振音源とする Glottal-ARMAX (Auto Regressive and Moving Average eXogenous) モデルと、時変複素 AR(TV-CAR: Time-Varying Complex AR) モデルである。

(1) Glottal-ARMAX モデルに基づく音声分析

音声信号の音源特性と声道特性を分離することは音声研究の目標の 1 つであり、1970 年代より様々な研究がなされて来た。音源と声道特性の分離は次の 2 種類のアプローチに分類できる。1 つは glottal-inverse filtering と呼ばれるアプローチであり、音源の影響を除去した声道フィルタを推定し、その逆フィルタリングにより音源信号を推定する方法である。もう 1 つは、音源モデルを仮定し、音源モデルパラメータと声道モデルパラメータを同時推定するアプローチである。前者は比較的安定した声帯音源が抽出できるが、音源をいかに正確に音声から除去できるかが問題になる。後者の方法として、藤崎らによる GARMA 分析、粕谷らによる ARX 分析が提案されている。これらの分析法において仮定されている音声生成モデルでは、音源を音源モデルにより生成される声帯音源波のみと仮定している。しかし、実際の有声音源には雑音成分も含まれていることが、最近の音声生成の研究により明らかにされている。そこで、本論文では、音声の生成過程をより一般化かつより高精度化する Glottal-ARMAX モデルを導入し、そのパラメータ推定アルゴリズムである Glottal-ARMAX 分析の提案を行う。Glottal-ARMAX モデルは、2 つの MA モデルと 1 つの AR モデルからなる声道 ARMAX モデルと、声帯音源波モデルで生成される外生入力 (X 入力) と白色 GAUSS 過程の 2 種類の入力により構成される。

本方式において、ARMAX モデル同定は、MIS(Model Identification Method) と呼ばれる ARMA 分析法を X 入力既知の ARMAX モデル同定に拡張した方式で実現する。また、非線形最適化問題となる音源モデルと声道モデルの同時推定は、遺伝的アルゴリズム (GA) と焼きなまし法 (SA) を組み合わせたハイブリッド探索により実現する。声帯音源波モデルとして Rosenberg-Klatt モデルを採用する。ハイブリッド探索では、GA の各個体の遺伝子を Rosenberg-Klatt モデルのモデルパラメータとし、適応度を ARMAX モデル同定の 1 ピッチ分の推定誤差の 2 乗和の逆数として、GA による大域探索を行なう。GA の局所推定能力の欠如を補うため、GA の各世代の最適な個体に対し SA による局所探索を行なう。さらに、演算量低減と推定精度向上のため、QMF フィルタバンクや Haar 基底などの直線位相 2 チャンネルフィルタバンクを用いたマルチレート処理を導入する。マルチレート化により ARMAX 同定次数を小さくできるため、演算量が低減できる。また、マルチレート化により周波数分解能が向上するため、スペクトルの局所的特徴に関連する声帯音源パラメータ、すなわち声門開口比 OQ の推定精度が向上する。合成音声と実音声を用いた実験により、マルチレート化による精度の向上と演算量の低減を示す。

(2) 時変複素 AR (TV-CAR) モデルに基づく音声分析

1960 年代初頭に板倉らと B.S.Atal らにより個別に提案された線形予測分析 (LPC) 分析は、その適度なロバスト性と演算量のため、音声処理、特に低ビットレート音声符号化における特徴抽出法の主流となっている。しかし、音声信号は非定常信号であるにも関わらず、LPC 分析には音声の動的特徴を抽出できないという重大な欠点がある。一方、複素信号である解析信号をターゲット信号にする複素 LPC 分析法が、Kay や島村らにより個別に提案され、解析信号の性質によるスペクトル推定精度の向上や係数の量子化精度の向上などの利点が報告されている。しかし、これらの分析法も動的特徴を抽出できない時不変分析法である。本論文では、解析信号に対し時変複素 AR モデル同定を行う音声分析法の提案を行う。本方式では、AR 係数を複素基底関数展開で表現し、最小 2 乗誤差規範 (MMSE) に基づく時変複素 AR モデル同定を行う。複素 LDU 分解による解法を導出するとともに、合成音声と実音声を用いて、従来方式との比較と基底関数による推定スペクトルの違いを示す。さらに、MMSE 規範を Huber's Robust M-推定に拡張した、ロバスト時変複素 AR 分析の構築も行なう。ロバスト推定アルゴリズムとして、Newton-type と重みづけ最小 2 乗法 (WLS) の 2 種類のアルゴリズムを導出する。ハイピッチの合成信号と実音声で分析を行なうことにより、TV-CAR 分析のロバスト性を示す。

最後に、Glottal-ARMAX 分析において、Haar 基底を用いたマルチレート化と低域のみを用いる分析が、十分有効であることを明らかにした。

また、時変複素 AR 分析において、基底関数としては、母音のような直線的な遷移区間では一次変動関数が、半母音や子音のような非直線的な遷移区間では複素 Fourier 基底が適していることを明らかにした。また、ロバスト分析との比較により、MMSE による時変複素 AR (TV-CAR) 分析が、非 GAUSS 性である声帯音源特性に対し十分にロバストであることを明らかにした。

学位論文審査の要旨

主査	教授	栃内香次
副査	教授	青木由直
副査	教授	北島秀夫
副査	教授	宮永喜一

学位論文題名

Speech Analysis Methods Based on Accurate Models

(高精度なモデルに基づく音声分析法)

近年、LSI の低価格化により、携帯電話やカーナビゲーションシステムなどを始めとする、音声ディジタル信号処理技術を応用した多彩な音声応答関連製品が出現している。しかしながら、現在市販されている製品は、音質面で十分な品質を実現しているとは言い難く、その改善が求められている。これらの機器の音質改善を実現するために、用いられる音声生成モデルの高精度化が必要とされている。

音声のディジタル信号処理の基本的な手法として、線形予測分析 (LPC) が1960年代初頭に提案され、広く応用されてきた。しかし、LPC 分析で用いられる音声生成モデルでは、複雑で非定常な音声信号を完全には表現できないという問題があり、さまざまな改善手法が研究されている。

本論文はこの問題に対し、音声信号処理のさまざまな応用における高品質化を目標として、2種類の高精度音声生成モデルを検討し、それぞれについて分析法の提案を行ったものである。その一つは、音声の声道特性を ARMAX (Auto Regressive and Moving Average eXogeneous) モデル、入力を白色 GAUSS 過程と声帯音源波モデルとする Glottal-ARMAX モデルで、もう一つは、時変複素 AR (TV-CAR: Time-Varying Complex AR) モデルである。

始めに、Glottal-ARMAX モデルに基づく音声分析法の検討を行っている。音声信号の音源特性と声道特性を分離することは音声信号処理における主要な研究目標の1つであり、これまで多くの研究がある。本論文では、それらを基礎とし、音声生成過程をより一般化かつ高精度化し、2つの MA モデルと1つの AR モデルからなる声道 ARMAX モデルと、声帯音源波モデルで生成される外生入力と白色 GAUSS 過程の2種類の入力により構成される Glottal-ARMAX モデルを提案している。

ここで、遺伝的アルゴリズム (GA) と焼きなまし法 (SA) を組み合わせたハイブリッド

探索により、音源モデルと声道モデルの同時推定を実現する手法を提案している。さらに、マルチレート処理により、ARMAX モデル同定次数を小さくできるとともに、声帯音源パラメータの推定精度が向上することを明らかにしている。そして、合成音声と実音声を用いた実験を行って、提案手法の有効性を実証している。

次に、TV-CAR モデルに基づく音声分析法の検討を行っている。LPC 分析は、その適度なロバスト性と演算量のため、特に低ビットレート音声符号化における特徴抽出法の主流となっているが、音声の動的特徴を抽出できないという問題点がある。

この点に対し、本論文では、AR 係数を複素基底関数展開で表現し、最小 2 乗誤差規範 (MMSE) に基づく時変複素 AR モデル同定を行う分析法を提案し、複素 LDU 分解による解法を導出するとともに、従来方式との比較を行っている。さらに、2 種類の推定アルゴリズムを導出し、合成信号と実音声を用いた分析実験により、時変複素分析のロバスト性を実証している。

最後に、Glottal-ARMAX 分析において、Haar 基底を用いたマルチレート化と低域のみを用いる処理が有効であることを示し、時変複素 AR 分析に用いる基底関数として、母音のような直線的な遷移区間では一次変動関数が、半母音や子音のような非直線的な遷移区間では複素 Fourier 基底が適していることを明らかにしている。また、MMSE による時変複素 AR 分析が声帯音源特性に対し十分にロバストであることを明らかにしている。

これを要するに、著者は、ディジタル音声信号処理の基本である音声の高精度モデル化について、従来手法の問題点の検討に立って新しい分析法を提案し、実音声と合成音声を用いた実験により提案手法の有効性を確認し、今後ますます普及することが予想される携帯電話等の小型ディジタル音声機器の高品質化の可能性を明らかにしたもので、音声情報処理工学ならびに情報メディア工学の発展に寄与するところ大である。

よって著者は、北海道大学博士（工学）の学位を授与される資格あるものと認める。