早稲田大学大学院理工学研究科

博士論文審査報告書

論　文　題　目

高品質音声合成のためのスペクトル包絡の推定及び変換に関する研究
Studies on Spectral Envelope Estimation and Conversion for High Quality Speech Synthesis

申　請　者

望月	亮
Ryo	Mochizuki

情報・ネットワーク専攻
知覚情報システム研究

２００６年　２月

近年，コーパスベースの音声合成方式によって，音質の良い音声の合成が可能となった．特に大規模な音声コーパスを用い，韻律変換をまったく行わない波形接続合成方式では，読み上げ口調の音声に限れば自然発声と比較してほとんど遜色の無い合成が可能になっている．一方，音質の改善が進むにつれ，最近では感情や態度，話者性，発話口調等を自由に制御するための技術が求められるようになってきている．例えば音声合成を音声対話システムへ応用する場合，ユーザとシステムとの自然なやり取りを実現するためには単なる読み上げ口調ではなく，システムの発話意図や態度を表現する多彩な声質の制御が必要とされている．

音声合成によって多彩な発話を実現する手段としては，(1)発話スタイルや話者ごとに音声を録音しておき合成時にはこれを編集する，(2)少量データの学習によって合成素片生成モデルを適応し合成素片を変形させる，等のアプローチが考えられる．前者は高音質を実現するという意味では有効であるが，現在の波形接続合成方式では録音やラベル情報の付加に膨大な人手の作業が発生するため，発話スタイルや話者ごとにデータベースを構築するのは現実的な方法とは言いがたい．一方，後者においては，現時点では十分な適応・変換方法が存在しないため変換処理を施すと音質劣化が目立ったり変換自体が不十分だったりといった問題があるものの，この問題は今後検討が進むにつれて改善されることが期待される．

現在，高音質な合成を実現している波形接続合成方式は，合成時に元となる音声データを一切加工・変形しない方式であり，このことに依存した形で高音質を実現している．しかし，発話の多様化を目指すためには，音声信号処理による加工・変形が可能な方式を採用する必要がある．PSOLA（Pitch Synchronous OverLap Add）法は波形接続合成より変換に対する自由度が高く，変換率が低い場合は高音質な韻律変換が可能であり，従来の線形予測を代表とするパラメトリックな方式よりも格段に音質が良いという長所を持つ．本研究では，この高音質な音声合成が期待できる PSOLA 法に注目し，音質の改善，及び多彩な発話表現の実現に向けたスペクトル包絡の抽出，補正，及び変換に関する要素技術を提案・検討したものである．

以下に本論文の概要とその評価について述べる．

第１章は序章であり，本研究の目的と，その背景について述べている．

第２章では，歪の少ないスペクトル包絡の推定を目的とし，ピッチ同期で短時間波形を抽出する方法について提案している．PSOLA 法は短時間窓を利用して基本周期の影響を含まない短時間波形を抽出し，この短時間波形を所望する基本周期で再配列することによって F0 変換を行うことができる．しかし，安定したピッチ同期分析が行えない場合，波形抽出位置がふらつき，韻律変換処理によって音質劣化を引き起こす．従来，短時間波形の抽出は基本周期の 2

倍の窓長を持つハニング窓で抽出するのが一般的であったが，先行研究ではどの位置を窓関数の中心に設定するのが音質として良いのか明確な回答を持っていなかった．本研究ではこの問題に対し，変形自己相関によって線形予測残差波形のピーク抽出を行い，このピーク位置を短時間波形抽出の基準位置（ピッチマーク）として波形抽出する方法を提案している．また，提案方法によって決定したピッチマークを基準に，どの程度遅延した位置にスペクトル歪が最小となる波形抽出位置が存在するのか，音声信号モデルを用いて最適な波形抽出位置を実験的に調査している．これらの提案・検討により，従来手法に比べ波形の揺らぎに影響されることなく安定して品質の高い短時間波形を切り出すことに成功しており，その成果は高く評価できる．

第３章では，ピッチ同期で抽出した短時間波形の低域におけるスペクトル包絡を，スペクトル傾斜と F0 変換率に応じて動的に再構築する方法を提案している．PSOLA 法によって韻律変換を行う場合，抽出した短時間波形をそのまま利用すると変換音声に著しい音質劣化が生じる場合がある．この音質劣化は原音声から抽出した短時間波形のスペクトル包絡が韻律変換後本来あるべき形状から外れるためであるが,この原因として著者は PSOLA 法では元の F0 より低域において信頼できるスペクトル情報が得られないという問題が存在することを初めて指摘した．本来，周波数分析によって求められるスペクトルは，F0 の整数倍にあたる高調波のみで構成される線スペクトルとなるのが理想であるが，実際は短時間波形抽出に用いる窓関数の漏れが隣接する高調波間で重畳され，滑らかなスペクトル包絡が形成される．しかし F0 より低い帯域においては，F0 における窓関数の漏れの影響が観測されるのみで，正しいスペクトル包絡情報が観測できない．この低域スペクトルの問題により， F0 を低い方へ変換した場合に音質劣化が顕著になっているとしている．この問題に対処するため，F0 変換を行ってもスペクトル傾斜は保存されるという仮定に基づいて，動的に低域におけるスペクトル包絡を再構築し，音質劣化を軽減する方法を提案している．提案方法は，F0 を低い方へ変換する音声変換において，従来手法に比べ格段に高いプリファレンスを与えており，高く評価できる．

第４章では，韻律特徴量を利用し，統計的な手法によってスペクトル特徴量をターゲットの環境にあったスペクトル特徴量へ変換する方法について提案している．音声合成によって多様な発声を実現するためには，音声収録時の発話から，ターゲットの発話へ変換するための適応技術が必要となる．話者の発話スタイルや話者性を決定づける要因としては，イントネーションやアクセントなど韻律的な特徴が重要であるが，それに劣らず，声質を決定するスペクトル包絡に関しても精度の良い変換が強く望まれる．この適応・変換を実現するために，今まで統計的な手法を用いた様々な方法が検討されているが，従来方法のほとんどの研究では変換元となるスペクトルとターゲットのスペクトルとの 1 対 1 の対応学習によって変換が行われていた．しかし，スペクトル変換を音

声合成へ応用した場合を考えると，変換関数の入力にはスペクトル以外にも韻律や音素系列などのコンテキスト情報を利用することが可能である．特にスペクトルは韻律特徴量との間にある程度の相関があるため，変換モデルに韻律情報を考慮することで変換精度の改善が期待できる．この点に着目し，本研究では韻律情報を加味した上でスペクトルを変換する統計的手法を提案し，声質変換に応用することを試みている．比較実験の結果，スペクトル変換時に韻律情報を組み入れることで，それを組み入れないときに比べ高い品質を実現できることを示しており，着実な実験結果が評価できる。

第５章は結論であり，本論文のまとめと今後の展望について述べている．

以上を要するに，本研究では声質の変換を対象とする音声合成において実用化の観点で有望視される PSOLA を対象として，その短時間基本波の抽出方式，韻律の変形時におけるスペクトル再構成方式，スペクトルの変換方式について新たな手法を提案することで，従来にない柔軟性の高い高品質な音声合成を可能にしたものであり，その工学的価値は高い．よって，本論文は，博士（工学）の学位にふさわしいものと認める．

２００６年２月

審査員

（主査）	早稲田大学教授	工学博士（早稲田大学）	小林　哲則
	早稲田大学教授	工学博士（早稲田大学）	白井　克彦
	早稲田大学教授	工学博士（早稲田大学）	誉田　雅彰
	早稲田大学教授	工学博士（早稲田大学）	匂坂　芳典