報告番号 ※乙第 4469 号

主論文の要旨

題名 複合音声単位の選択的使用に基づく
規則による音声合成の研究

氏名 武田一哉

主論文の要旨

報告番号	※乙第　　号	氏名	武田一哉

機械と人間のコミュニケーション手段として、人間にとってもっとも自然な情報伝達手段である音声を利用することが強く望まれており、音声を自動生成する音声合成と、音声入力を自動認識する音声理解とが、機械と人間とのインタフェースを飛躍的に高度化する技術として注目を集めている。音声合成の分野では、現在特定の単語や文のみを合成する、録音・編集型音声合成が実用化され普及しているが、任意の文章を生成する「規則による音声合成」の品質は未だ十分でなく、広く普及するには至っていない。これは（1）同一のかなや文字に対し、様々な音響的バリエーションを生成する音声生成のメカニズムを、単純な規則で制御することが困難であること。（2）開発された制御規則の評価が、受聴による比較に頼らざるを得ないこと、等の理由から、近年のデバイス技術の進歩に伴う「大容量化」「高速化」を、合成音声の品質向上に直接結びつけ難いためである。

本論文ではこの問題を解決するために、複雑な制御規則を用いるのではなく、大量の音声データに含まれる様々な音声区間を直接用いて、高品質な合成音声を生成する方式を提案した。提案した方式は、予め蓄積された大規模な音声データの中から、入力文字系列に最も良く対応する音声区間（複合音声単位）を選択し、それらを接続して音声を生成することを特徴としている。この方式を用いることで、音声データの蓄積量に比例して高品質な合成音声を得ることができ、従来よりも高品質な合成音声が実現可能なことを、主観評価および客観評価によって確かめた。

第1章、序論では本論文の目的について述べた。本論文の目的は、複合音声単位を選択的に用いる音声合成法を提案し、高品質な規則による音声合成を実現することである。本章では複合音声単位が以下の3つの特徴を持つ新しい音声処理単位として提案された。（1）従来の、音韻、音節等の固定長の処理単位に対し、任意の長さの音韻連鎖を処理単位とし、それらの組み合わせで音声を合成すること。（2）処理単位毎に複数の中から最適な音声区間（音声単位素片）を選択・接続して音声を合成すること。（3）従来音声単位素片毎に固定されていた接続境界を、隣接音声単位素片に応じて適応的に変化させること。さらに提案する方式の利点として、音声素片の選択・接続アルゴリズムと、単位素片抽出用の音声データベースが独立しているため、データベースを大規模化することにより、システムの複雑さを増すことなく高品質な音声合成を実現することができる、ことを示した。

第2章では、複合音声単位を選択的に用いる規則合成方式に基づく音声合成システムの、構成要素と合成手順を示した。本章では提案する手法が従来の音声合成方式を含む自由度の高い方式であることを示し、本方式に基づき規則による音声合成システムを構築するための課題として、（1）単位素片抽出用音声データベースの構築（2）選択的な単位素片使用法の確立（3）高精度な韻律制御規則の作成、の3点を指摘した。

第3章では、単位素片抽出用音声データベースの構築について述べた。本章では、大規模発声の収集と、詳細な音声記述の2点を技術的な課題としてとらえ、以下の3つの特徴を持つ音声データベースの構築を行なった。（1）音声データ記述用の音韻ラベルが、仮名表記の再現性を保ちつつ、音声事象レベルの詳細な発声記述が可能となる様、階層的に構成されてい

主論文の要旨

報告番号	※乙第　　号	氏名	武田一哉

る。（2）音韻ラベル付けは視察に基づき行なわれ、ラベル付けの精度は、セグメント境界の誤差が5 ms以内、ラベル記述誤りの割合が0.5％以下である。（3）収集したデータを各種の目的に共用するため、音声データベース管理システムが構築されており、簡易検索言語を用いて希望の音声・言語環境に容易にアクセスすることができる。

さらに、設計した階層的音韻ラベルに基づき、音声単位素片の伸縮や接続点の制御を、高精度に処理可能になることを示した。最後に、本分野における今後の課題として、日本語に含まれる音韻系列の分析を大規模なテキストに基づいて行ないデータベースの規模拡大に反映させること、を指摘した。

第4章では、複合音声単位を選択的に用いる規則による音声合成方式における、選択的な単位素片使用法を検討した。まず単位素片選択手法について検討を行なった。単位素片選択基準として（1）単位素片間の接続特性（2）単位素片の抽出・使用環境の類似性、の2つが重要であることを指摘し、これらの基準に基づき、使用単位素片の候補を逐次絞り込む選択手法を提案した。本選択手法により得られた合成音声の音節明瞭度は90％以上と、高い合成音声品質が確認された。

さらに明瞭度試験の結果から、単位素片系列と合成音声品質との関係を定量的に分析し、以下の結果を得た。（1）合成単位の長さと音声品質との関係から、単音節を合成単位に用いた場合、品質の劣化が顕著である。（2）単位素片の抽出・使用環境の類似度が低下するに従い音節の挿入・付加といった聞き取り誤りが増加する。（3）単位素片間を、無声摩擦音と母音の間、母音の定常部、無声破裂音と母音の間、等で接続した場合には、接続点周辺においても、聞き取り誤りが起こる割合が10％未満と少ない（その他の環境で接続した場合には20％以上）こと。

次に、単位素片系列に応じて適応的に接続境界を制御する単位素片接続方式を提案した。本接続方式により、接続歪みの値が視察による音韻境界において接続を行なった場合の、10％程度に低減され、主観評価においても75％以上の割合で選好された。

最後に提案した選択・接続手法により得られた合成音声と、抽出環境を固定したCV音節を単位とする合成音声との比較実験を行なった。実験の結果、自然性、明瞭性の両者において、選択された単位による合成音がより高い品質を示し、選択的な単位使用に基づく合成方式の有効性が確認された。さらに本分野における今後の課題として、定量的な単位素片抽出基準の確立を指摘した。

第5章では、本方式に基づく規則による音声合成システムの韻律制御規則について検討した。本章では制御対象として音韻継続時間長を取り上げ、数量化理論の枠組みの上で、種々の要因と単語発声中の音韻継続時間長との関係の分析を行なった。分析により、隣接音韻間の継続時間長の補償傾向、単語モーラ長、等の要因が音韻継続時間長に与える影響を定量的に明らかにした上で、線形重回帰モデルにより音韻継続時間長をモデル化した。モデル化により得られた音韻継続時間長制御規則の精度を、単語音声を用いた実験により確認したところ、得られた予測誤差は従来法の20.5 msに対して、18.3 msであり本分析・モデル化の有効性が

主論文の要旨

報告番号	※乙第　　号	氏名	武田一哉

確認された。

次に、本モデルを文発声中の音韻の継続時間長の予測に用い、生じた誤差を分析した。本分析は、単語レベルでの音韻継続時間長の変動を予め単語レベルでの継続時間長モデルにより正規化し、文レベルに固有の音韻継続時間長の変動現象を明らかにすることを目的として行なった。分析の結果、（1）呼気段落末における音韻継続時間長の伸長が単語末の伸長を大きく上回ること。（2）文末の効果として、音韻継続時間長の短縮が朗読文に存在する反面、会話文の読み上げではそれが顕著でないこと。等の文レベルでの音韻継続時間長の変動現象が明らかになった。

さらに、単語レベルでの予測モデルをこれら文レベルでの変動要因により補正した結果、文発声内の音韻継続時間長も単語発声内の場合と同程度の誤差範囲で予測可能となり、精度の高いモデル化が可能なことが明らかになった。最後に本分野における今後の研究課題として、会話文の分析による、自然な対話音声が生成可能な韻律制御規則の作成、を指摘した。

第6章では本論文をまとめた。本論文では、複合音声単位を選択的に用いる規則による音声合成法を提案し、本方法による音声合成を実現するために、単位抽出用音声データベースの構成法、適応的な単位選択アルゴリズム、統計的モデルに基づく韻律制御方式を提案し、提案した各手法の有効性を主観評価及び客観評価により確認した。