審 査 の 結 果 の 要 旨

氏　　名　　齋藤　佑樹

音声合成とは、コンピュータを用いて音声を人工的に生成・変換・加工する技術である。特に、テキストから音声を生成する技術をテキスト音声合成といい、入力された音声の言語情報を保持しつつ、非言語・パラ言語情報を変換する技術を音声変換という。音声合成は、人間とロボットの間での音声コミュニケーションを仲介するマンマシンインターフェースのみならず、音声バーチャルリアリティやエンターテインメント応用などを通じて、物理的制約を超えた音声表現を実現する技術である。本論文ではこれらの背景を踏まえ、多様な話者の音声を高品質に合成でき、かつ合成された音声の話者性を直感的に制御可能な音声合成技術の確立を目指している。特に本論文では、（１）音声敵対過程を統合した高品質な音声合成法、（２）音声認識過程を統合した多様な音声合成法、そして（３）人間の音声知覚を導入した解釈性の高い話者表現学習法を提案している。本論文は全 7 章から構成されている。

第 1 章「Introduction（序章）」では、本論文の主題である統計的音声合成の概要および問題点について述べている。音響モデルとして deep neural network (DNN)を用いる DNN 音声合成は、大規模音声データベースの公開に恩恵を受け、入力特徴量から音声パラメータへの複雑な写像を学習可能な手法として広く研究されている。しかしながら、従来技術では、DNN 学習時の統計処理に起因する合成音声パラメータの過剰な平滑化により、合成音声の品質が著しく劣化し、また事前学習データ以外の表現が困難な DNN の特性より、合成音声の話者性の多様性を高めることは困難であることが指摘されている。

第 2 章「Statistical Speech Synthesis Using DNNs（DNN を用いた統計的音声合成）」においては、従来の統計的音声合成の詳細について述べている。音声分析合成系の構築、DNN による音声パラメータのモデリング等について解説がなされ、従来技術の問題点に触れている。

第 3 章「Vocoder-based Statistical Speech Synthesis Using GANs（GAN を用いたボコーダに基づく統計的音声合成）」および第 4 章「Vocoder-free Statistical Speech Synthesis Using GANs（GAN を用いたボコーダフリー統計的音声合成）」においては、音声敵対過程を利用した音声合成を提案している。音声なりすまし検出と敵対する高品質な音声合成法では、generative adversarial network (GAN)を音声合成に導入し、高品質な音声を合成可能な DNN 音響モデルを学習する。この提案法では、DNN 音響モデルと、自然音声と合

成音声を識別する discriminator を交互に学習する。GAN の学習の目的関数は生成データ分布と実データ分布の擬距離最小化とみなされるため、この提案法は過剰な平滑化を定量的に補償しており、定性的な統計的差異補償により合成音声品質を改善する従来法の理論拡張と解釈できる。本論文では、第 3 章にて、ボコーダ分析合成に基づくテキスト音声合成および音声変換のための GAN に基づく音声合成法を提案し、その有効性を評価している。その後、第 4 章において、より先進的な短時間フーリエ変換スペクトルを用いたボコーダフリーなテキスト音声合成に向けて、この提案法を拡張している。実験的評価の結果から、この提案法によって合成音声の品質が有意に改善することが示されている。

第 5 章「VAE-based Multi-speaker Acoustic Modeling Using Speech Recognition Process（音声認識過程を用いた変分オートエンコーダに基づく多話者音響モデリング）」では、音声認識過程を統合した多様な音声合成法について提案を行っている。ここでは、音声から発話内容と発話者を認識する DNN を導入し、高品質かつ多様な話者性を持つ DNN 音響モデルを学習する。さらに、多様な話者性を持つ音声を合成するために、話者認識由来の連続的な話者表現を変分オートエンコーダの潜在表現として用いることも提案している。実験的評価の結果から、この提案法によって変換音声の有意な品質改善を達成しつつ、DNN 学習に用いていない未知話者の音声も合成可能であることが示されている。

第 6 章「Perceptual-similarity-aware Deep Speaker Representation Learning（知覚類似度を導入した深層話者表現学習）」においては、人間の音声知覚特性を導入した解釈性の高い話者表現学習法を提案している。ここでは、合成音声の話者性を直感的に制御可能な音声合成を実現するために、人間の主観的な音声知覚を導入して話者表現を学習する。本提案法では、まず多数話者間の知覚的な類似度のスコアを収集し、その後このスコアを利用して聞き手の話者知覚を高精度に再現する話者表現を学習する。このように学習された話者表現は、聞き手の話者知覚を強く反映するため、合成音声の話者性をより直感的に制御できるのみならず、未知話者の音声合成時の品質劣化に対する頑健性の向上が期待できる。実験的評価の結果から、この提案法によって学習された話者表現の導入により、合成音声の品質と制御性が向上することが示されている。

第 7 章「Conclusion（結論）」では、本研究の成果がまとめられている。

以上、本論文は要するに、従来の統計的音声合成では陽に考慮されてこなかった人間による三種の音声情報処理能力（敵対過程、認識過程、知覚過程）を導入した統一的な統計的音声合成理論を提唱し、その有効性を音声合成評価実験で検証・確認したものである。本論文で提案された理論を導入する事により、従来では不可能とされてきた高品質かつ解釈性の高い音声合成処理が可能となった。これは音声バーチャルリアリティや身体能力拡張システムの基礎を与えるものであり、システム情報学に対する貢献が大きいと判断される。よって本論文は博士（情報理工学）の学位請求論文として合格と認められる。