論 文 の 内 容 の 要 旨

論文題目

Statistical Speech Synthesis Based on
Human's Speech Information Processing Abilities
(人間の音声情報処理能力に基づく統計的音声合成)

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す．

本論文の主題である統計的音声合成は，統計的機械学習の枠組みに基づき，入力特徴量から合成音声パラメータを生成する音響モデルを学習する手法である．特に，音響モデルとしてdeep neural network (DNN)を用いるDNN音声合成は，計算機性能の向上や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する．また，合成可能な話者はDNN学習時に用いられたものに限定されるため,合成音声の話者性の多様性を高めることは困難である．さらに，合成音声の話者性を制御するために用いる従来の話者表現は，人間の主観的な話者知覚を無視しているため，多様な話者性の音声を合成可能なDNN音声合成には不適切である．

本論文では，これらの問題点を解消するために，(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する．これらの提案法は，音声合成を改善するために，人間の音声情報処理能力（敵対過程，認識過程，知覚過程）を活用するという着想に基づく．

(1)音声敵対過程を統合した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．この提案法は，人

間の自然音声と合成音声を識別する能力（即ち，音声合成に対する敵対過程）を活用してDNN音響モデルを学習するという着想に基づく．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化するパラメトリックな統計的差異（音声パラメータ分布の2次モーメントなど）の補償により合成音声の品質を改善する従来法の理論拡張と解釈できる．また，この提案法におけるdiscriminatorは，音声合成技術の悪用による音声なりすましを防ぐためのanti-spoofingとして解釈できる．したがって，anti-spoofingの知見を導入することで，より効率的なDNN音響モデル学習が実現できる．本論文では，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，その有効性を評価する．その後，短時間フーリエ変換スペクトルを用いたより先進的なテキスト音声合成に向けてこの提案法を拡張する．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を示す．

(2)音声認識過程を統合した多様な音声合成法では，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音声を合成可能なDNN音響モデルを学習する．この提案法は，人間の音声認識と話者認識（即ち，音声合成の逆過程）の能力を活用してDNN音響モデルを学習するという着想に基づく．この提案法では，発話内容が明瞭で高品質な音声を合成するために，音声認識を用いて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用いる．本論文では，変分オートエンコーダを用いた音声変換においてこの提案法の有効性を評価する．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になることを示す．

(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．この提案法は，人間の知覚を計算資源とみなし，解釈しやすい話者表現の学習に活用するという着想に基づく．この提案法では，まず，多数の評価者（音声の聞き手）による大規模主観評価により，多数話者間の知覚的な類似度のスコアを収集する．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．本論文では，この提案法の有効性を(2)音声認識過程を統合した多様な音声合成法において評価する．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を示す．