ポスター番号

P-09-47

戦略的情報通信研究開発推進事業(SCOPE)_若手ICT研究者等育成型研究開発

声の個人性が保持された音声翻訳を実現するための大規模音声データと深層学習に基づく多言語音声合成技術に関する研究開発

橋本 佳[1] [1]名古屋工業大学 hashimoto.kei@nitech.ac.jp

1. 背景と目的

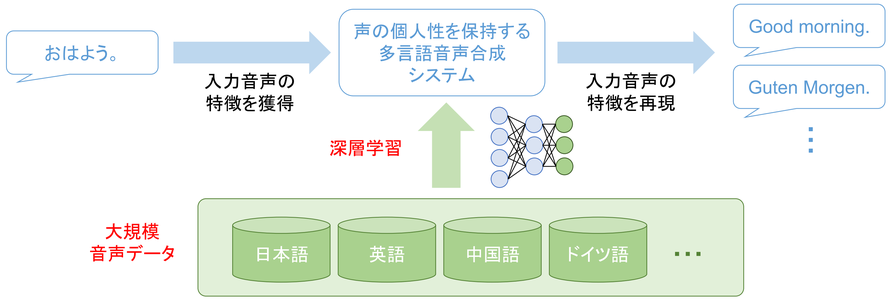
- 従来の音声翻訳の問題点
  1. 翻訳先の言語によって異なる人物の声で音声が出力される
  2. 誰が使用しても翻訳先の言語が同じなら同じ声の音声が出力される
- 本研究開発の目的
  1. 話者・言語が混在する音声データから音声合成用モデルを学習する方法を確立する
  2. 音声を入力した話者の声であらゆる言語の音声を合成可能な多言語音声合成システムを構築する方法を確立する

- 入力音声の声を翻訳先の言語で再現することで声の個人性を保持した音声翻訳システムを構築
- 自然なグローバルコミュニケーションを実現

2. アプローチ

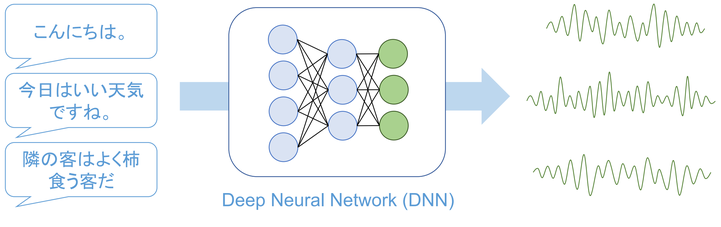
- 大規模音声データと深層学習に基づく手法
  - 言語や話者によらない共通した特徴をモデル化
  - 言語を超えて現れる話者の個人性を表す特徴をモデル化

→ 音声における言語と話者の特徴を分離



3. 深層学習に基づく音声合成の基本

- 深層学習に基づきテキストから音声への変換をモデル化
  - テキストと音声のペアデータを利用
  - 学習に用いた音声の特徴をモデル化
  - 基本は単一言語・単一話者・読み上げ調



- 本研究開発では複数話者・複数言語へと発展

4. 研究開発の内容および成果

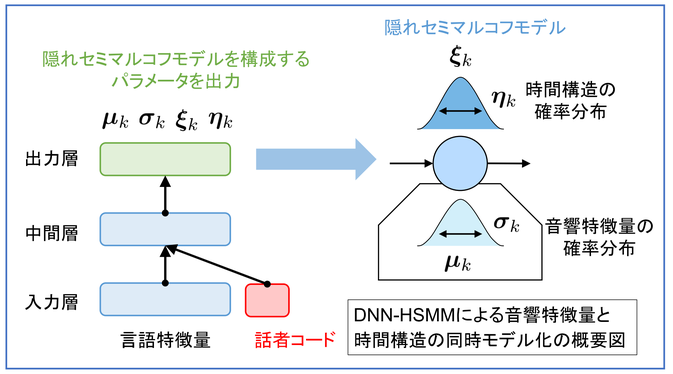
- 【研究開発項目1】 同一言語の複数話者・発話スタイルが混在する音声データを用いた音声合成用モデルの学習
  - 話者コード・発話スタイルコードを用いたDeep Neural Network (DNN)による音声合成用モデルの構築
    → MOS値を0.9ポイント改善

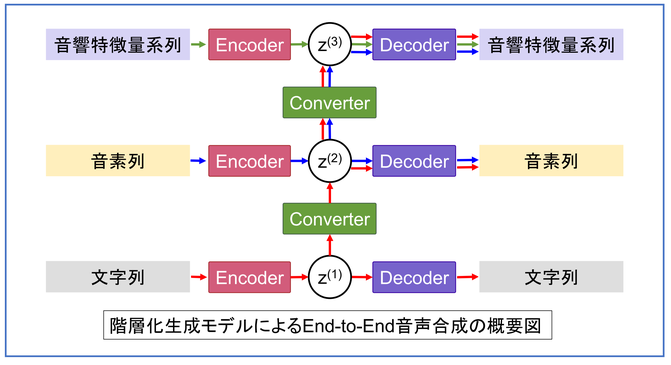
- 【研究開発項目2】 同一言語において指定した人物の声の特徴を再現する話者適応
  - 音響特徴量と時間構造を同時モデル化する隠れセミマルコフモデルの構造を導入したDNN (DNN-HSMM) による話者適応の実現
    → MOS値を0.6ポイント，DMOS値を1.1ポイント改善
  - DNNを用いた音声波形モデルに基づくボコーダのためのノイズシェーピング手法を提案
    → MOS値を0.7ポイント，DMOS値を1.0ポイント改善

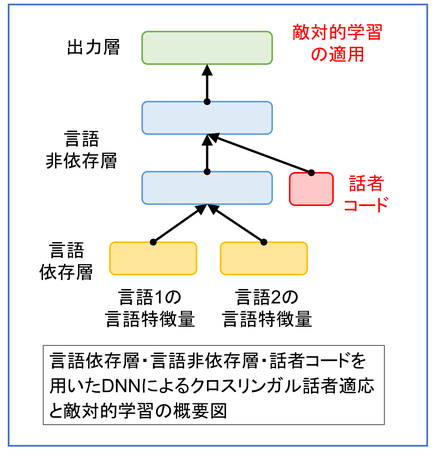
- 【研究開発項目3】 複数の言語が混在する音声データを用いた音声合成用モデルの学習
  - 言語依存層・言語非依存層・話者コードを用いたDNNの提案
    → 複数言語・複数話者の同時モデル化
  - テキストと発音の対応関係が未知の言語のための階層化生成モデルによるEnd-to-End音声合成の提案
    → MOS値を1.3ポイント改善，テキストの読み間違いを改善

- 【研究開発項目4】 指定した人物の声の特徴をその人物が話すことができない言語で再現するクロスリンガル話者適応技術
  - 言語依存層・言語非依存層・話者コードを用いたDNNによるクロスリンガル話者適応の実現
    → MOS値を0.8ポイント，DMOS値を0.5ポイント改善
  - 言語依存層・言語非依存層・話者コードを用いたDNNにおける敵対的学習を提案
    → DMOS値を0.25ポイント改善

MOS値：合成音声の自然性（音質）に関する評価値
DMOS値：合成音声の声質の再現性に関する評価値







5. 今後の研究開発成果の展開および波及効果創出への取り組み

- 個人性が保持された新たな音声翻訳サービスへと展開
  - より自然なグローバルコミュニケーションを実現
  - 本人の声のまま多言語講演，本人の声のまま映画音声の吹き替えなど多方面のサービスへと波及
- 発話スタイルを考慮したグローバルコミュニケーションへの発展
  - 現状では読み上げ調を対象としていたが異なる言語で感情を再現する音声合成技術への発展
- 「なりすまし」の問題への対処
  - 生身の人間の音声か合成音かを識別する技術と個人性を再現する音声合成技術の両方面から考える必要がある

ICTイノベーションフォーラム2020
戦略的情報通信研究開発推進事業 (SCOPE)