異なる言語において入力音声の特徴を再現する音声翻訳のための多言語音声合成

Multi-lingual speech synthesis for speech translation transferring voice characteristics of input speech in different languages.

橋本　佳(Kei HASHIMOTO, Ph. D.)

名古屋工業大学 准教授

(Nagoya Institute of Technology, Associate Professor)

日本音響学会、電子情報通信学会、情報処理学会、IEEE

受賞：日本音響学会粟屋清学術奨励賞（2015 年）

電子情報通信学会・日本音響学会音声研究会研究奨励賞（2017 年）

日本音響学会東海支部学会活動貢献賞（2022 年）

著書：音響学入門ペディア（担当；統計的音声合成のしくみを教えてください）、日本音響学会編, コロナ社（2017 年）人工知能学辞典新版（担当：音声合成（HMM 合成方式））、 人工知能学会編, 共立出版（2017 年）

研究専門分野：音声情報処理

あらまし

近年、言語の壁を超えて音声によるコミュニケーションを実現する、「音声翻訳」技術が注目を集めている。今後、音声翻訳技術は世界中で広く普及していき、将来的には誰もが音声翻訳システムを気軽に利用して、異なる言語の人々と自在に会話する日が来ると考えられる。しかしながら、現在の音声翻訳システムでは、翻訳先の言語ごとに予め定められた人物（話者）のナレーション調の音声が出力される、という制限がある。そこで本研究では、入力音声と異なる言語において入力音声の話者や感情などの特徴を再現する多言語音声合成技術を確立することを目的とする。任意の言語・話者・感情の組み合わせの合成音声を生成可能とするためには、言語・話者・感情それぞれに依存する音声の特徴を分離し、さらにそれらを学習データにない組み合わせで音声を合成可能にする枠組みが必要である。本研究では、言語・話者・感情に依存する音声の特徴を分離し、異なる言語において入力音声の話者・感情を再現することを可能とする深層学習に基づく多言語音声合成技術の開発に取り組んだ。

1. 研究の目的

本研究の目的は、入力音声と異なる言語において入力音声の話者・感情を再現する多言語音声合成技術を確立することである。本研究では、音声に含まれる言語・話者・感情の 3 つの要素に注目する。任意の言語・話者・感情の組み合わせの合成音声を生成可能とするために、言語・話者・感情に依存する音声の特徴を分離し、異なる言語において入力音声の話者・感情を再現することを可能とする深層学習に基づく多言語音声合成の研究開発に取り組む。このような多言語音声合成技術を開発し、音声翻訳システムへと応用することで、自分の声のまま感情も伝えることができる自然な異言語間コミュニケーションの実現を目指す（図 1）。

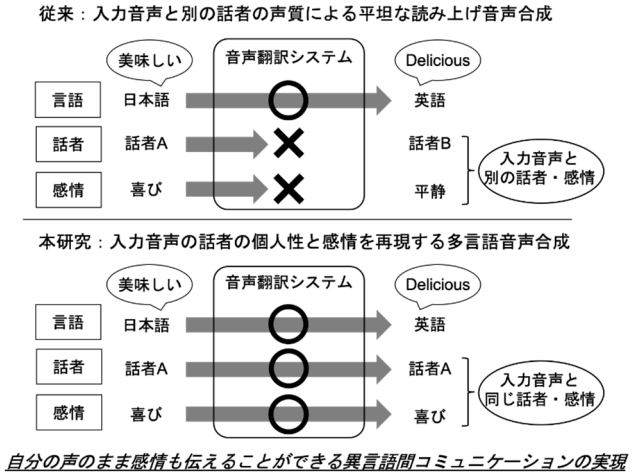

図 1 本研究の目的および概要

2. 研究の背景

近年、深層学習の発展に伴い、音声情報処理技術の性能は大きく向上している。入力テキストを音声へと変換する音声合成技術においても深層学習は適用されており[1]、条件によっては人間と聞き分けることができないほど高品質な合成音声の生成を実現している[2]。このような音声合成技術の応用例の一つに「音声翻訳」技術がある。音声翻訳技術は言語の壁を超えて音声によるコミュニケーションを実現する技術として近年注目を集めている。今後、音声翻訳技術は世界中で広く普及していき、将来的には誰もが音声翻訳システムを気軽に利用して、異なる言語の人々と自在に会

異なる言語において入力音声の特徴を再現する音声翻訳のための多言語音声合成

Multi-lingual speech synthesis for speech translation transferring voice characteristics of input speech in different languages.

話する日が来ると考えられる。しかしながら、現在の音声翻訳システムでは、翻訳先の言語ごとに予め定められた人物（話者）のナレーション調の音声が出力される、という制限がある。このため、問題点１：入力音声が同じでも翻訳先の言語によって異なる声の音声が出力される、問題点２：誰が音声を入力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、問題点３：どのような音声を入力してもナレーション調の音声が出力される、といった問題がある（図 2）。母国語の異なる複数人での会話を想定した時、音声翻訳システムに入力された音声は、複数の言語へと翻訳されることとなるが、現在の音声翻訳システムでは、上記の問題によって誰が話しているのかが声から判断することができない。また、楽しそうにしている、悲しそうにしている、怒っているなどの感情を伝えることもできず、コミュニケーションに大きな支障をきたすこととなる。

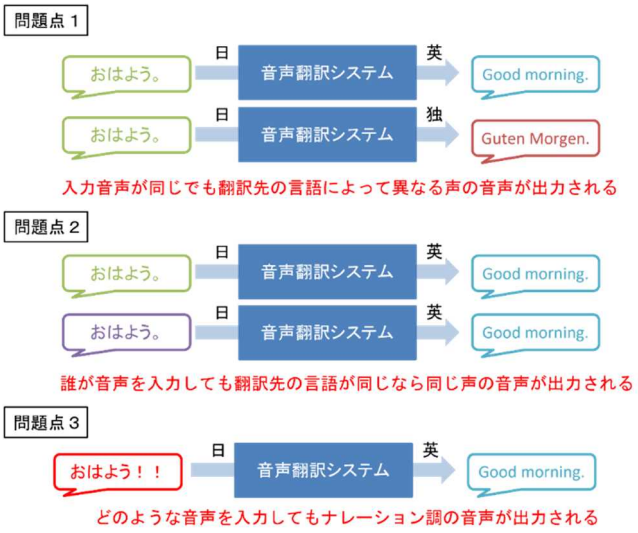

図 2　音声翻訳における問題点

これらの問題を解決するために、入力音声の話者に注目し、異なる言語でその話者の声質を再現するクロスリンガル話者適応と呼ばれる技術の開発が進められている。音声翻訳システムの利用を想定した場合、ユーザは自分の母国語で音声を入力すると、クロスリンガル話者適応技術によって自分の声質のまま、翻訳された音声が出力されることとなる。異なる言語で声質を再現するためには、音声に含まれる特徴のうち、言語に依存する部分と声質（話者性）に依存する部分を分離しながらその特徴を学習する枠組みが必要であり、非常に困難な問題である。この技術によって自分の声質のまま異なる言語の音声を高品質に生成することで、上記の問題点１，２が解決されることとなる。しかしながら、問題点３を解決するためには、話者の再現だけでなく入力音声の感情や話し方などを再現する必要がある。

3. 研究の方法

本研究では、言語・話者・感情の 3 つの要素に注目する。任意の言語・話者・感情の組み合わせの合成音声を生成可能とするためには、言語・話者・感情それぞれに依存する音声の特徴を分離し、さらにそれらを学習データにない組み合わせで音声を合成可能にする枠組みが必要である。そこで、言語・話者・感情に依存する音声の特徴を分離し、異なる言語において入力音声の話者・感情を再現することを可能とする深層学習に基づく多言語音声合成の研究開発に取り組んだ。

3.1 Variational Autoencoder に基づく多言語音声合成

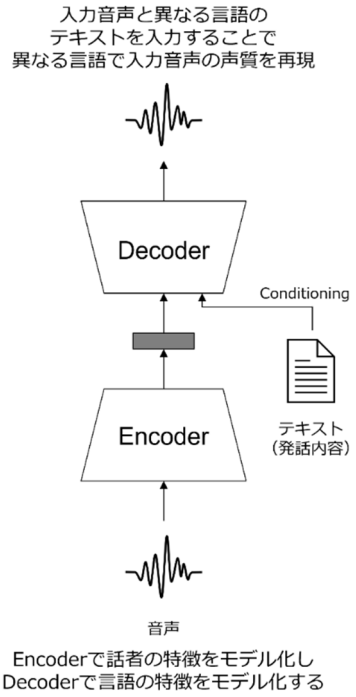

図 3 VAE に基づく多言語音声合成の概要

異なる言語において入力音声の特徴を再現する音声翻訳のための多言語音声合成

Multi-lingual speech synthesis for speech translation transferring voice characteristics of input speech in different languages.

入力音声の特徴を出力音声において再現するための枠組みとして、近年注目を集めている Variational Autoencoder (VAE)の構造を導入した多言語音声合成を提案した(図3)。VAEはEncoder(符号化器)とDecoder(復号化器)から構成され、音声合成においては、入力音声の特徴を表す潜在変数をEncoderによって獲得し、獲得した潜在変数とテキストをDecoderへと入力することで入力音声の特徴を再現した音声を生成する。提案法では、Decoderに言語依存層を導入することで多言語へと対応する。複数言語・複数話者のテキストと音声のペアデータを学習データとして用いて、VAEに基づく多言語音声合成モデルを学習する。これによって、言語依存の特徴は言語依存層においてモデル化され、話者依存の特徴はEncoderにおいてモデル化され、音声に含まれる特徴のうち言語と話者に依存する特徴を分離することが可能となる。このようにして学習されたVAEに基づく多言語音声合成モデルは、ある言語の音声をEncoderへと入力し、Encoderへ入力した音声と異なる言語のテキストをDecoderへと入力することで、入力音声と異なる言語において入力音声の話者を再現する合成音声が生成される。

3.2 敵対的学習に基づく多言語音声合成

入力音声の話者の特徴を、より再現するためにVAEに基づく多言語音声合成において敵対的学習の枠組みを導入した敵対的学習に基づく多言語音声合成を提案した[3]。テキストから音声を予測する多言語音声合成モデルをGenerator(生成器)とし、人間が発声した音声かモデルが生成した音声かを識別するDiscriminator(識別器)を導入する。敵対的学習では、入力された音声が人間の発声した音声かモデルが生成した音声かをDiscriminatorが識別できなくなるようにGeneratorの学習を行うことで、より人間の発声に近い音声をモデルから生成可能とする。さらに、より話者の特徴を学習するために、話者情報をDiscriminatorの入力に加える手法についても検討した。実験結果から、敵対的学習を導入することで、異なる言語においてより話者の特徴を再現することが示された。

3.3 異なる話者で感情を再現するためのモデル構造の検討

従来の感情音声合成では、ある特定の話者の様々な感情音声を収録し、テキストと感情音声のペアデータを用い ることで、感情音声合成モデルの学習が行われる。本研究では、ある特定の話者には感情音声のペアデータが存在するが、別の話者には感情音声のペアデータは存在せず、ナレーション調音声のペアデータのみが存在するという状況を想定する。ナレーション調音声のデータしか持たない話者の声質で感情音声を生成するために、感情と話者の特徴を分離して学習するためのモデル構造について検討を行った。提案法では、ナレーション調音声を用いた話者の変換とナレーション調音声から感情音声への2段階の変換を実現するようなモデル構造を導入することで、感情音声を収録していない話者の声質のまま感情音声を生成した。

3.4 顔画像を補助特徴として利用した複数話者音声合成

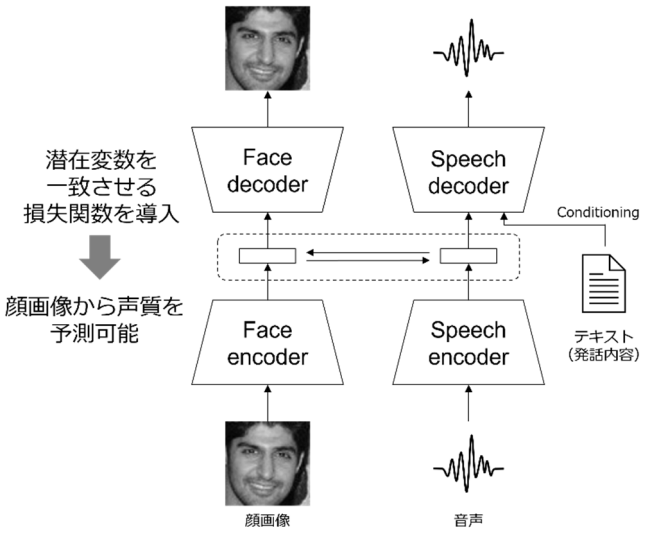

図4 顔画像を補助特徴として利用した
複数話者音声合成の概要

入力音声のみから音声の特徴を獲得するのではなく、顔画像から音声の特徴を獲得する枠組みを提案した。人間は顔から声の特徴を予測するため、顔画像と声質の間には何らかの相関があると考えられている。そこで、顔画像・テキスト・音声の関係を深層学習によっ

てモデル化する手法を提案した(図 4)[4]。提案法は、顔画像と音声をそれぞれ VAE によってもモデル化する。テキストは音声をモデル化した VAE の Decoder の入力に用いる。顔画像をモデル化した VAE の潜在変数と、音声をモデル化した VAE の潜在変数の間で同一人物の顔画像と音声の場合は共通する潜在変数によって表すという制約を加えることで、顔画像と音声の関係をモデル化することが可能となった。提案法では、顔画像を入力することで、その顔画像から予測される声質の合成音声を生成する。評価においては、顔画像を見せながら、実際にその人物の声質の合成音声と顔画像から予測した声質の合成音声を比較した。実験結果から、顔画像に対して違和感のない音声の合成を生成可能であることを示した。

４．将来展望

現在の音声翻訳システムでは、翻訳先の言語ごとに予め定められた人物のナレーション調の音声が出力される、という制限がある。本研究では、入力音声の話者・感情を翻訳先の言語で再現する多言語音声合成を開発することで上記の問題を解決することを目指した。音声を入力した人物の声のまま、感情を込めて音声翻訳が行われることで、より自然で円滑なグローバルコミュニケーションを実現することができると考える。また、音声翻訳システムは、日常の会話以外にも様々な場面で利用されると考えられる。例えば、著名人の講演やインターネット上での動画配信を様々な言語に翻訳するといった状況を想定すると、声質自体がその人物のブランドイメージの一部であると考えられるため、異なる人物の声に翻訳されることは望ましくない。また、講演や演説の多くはニュースのような単調な読み上げ音声ではなく、感情を込めることで多くを伝えようとしている。そのような状況においては、その人物の個人性や感情が保持された声で出力されることが必須である。これまで、音声翻訳システムは、いかに正しく翻訳されるかといった翻訳性能に重点が置かれてきたが、今後は利用者の個人性や感情などの表現を再現することが重要な課題になると考えられる。

おわりに

本研究では、異なる言語において入力音声の話者・感情を再現することを可能とする深層学習に基づく多言語音声合成技術の開発に取り組んだ。音声から言語・話者・感情に依存する特徴を分離し、これらの特徴をモデル化する手法を提案した。今後は、任意の言語・話者・感情の組み合わせの音声を、より高品質に生成可能にする手法やより細かなニュアンスを異なる言語で再現する手法の開発に取り組む必要がある。これらが実現されることで、言語の壁を超えた、音声によるグローバルコミュニケーションが実現されることが期待される。

参考文献

[1] 橋本佳、 高木信二、 “深層学習に基づく統計的音声合成、” 日本音響学会誌、 vol.73, no.1, pp.55-62, 2017.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, RJ Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions,” ICASSP 2018, pp.4779-4783, 2018.

[3] 大谷眞史、 佐藤優介、 高木信二、 橋本佳、 大浦圭一郎、 南角吉彦、 徳田恵一、 “音声合成における敵対的生成ネットワークを用いた複数言語・複数話者モデリング、” 日本音響学会 2020 年秋季研究発表会講演論文集、 pp. 695-696, 2020.

[4] 平光啓祐、 橋本佳、 南角吉彦、 徳田恵一、 “深層学 に基づく音声合成における顔画像情報を用いたクロスモーダル話者適応、” 日本音響学会 2022 年春季研究発表会講演論文集、pp. 905-906, 2022.

関連文献

藤本崇人、 橋本佳、 南角吉彦、 徳田恵一、 “隠れセミマルコフモデルによる構造化アテンションを用いた自己回帰型 VAE に基づく sequence-to-sequence 音声合成、” 日本音響学会 2021 年秋季

異なる言語において入力音声の特徴を再現する音声翻訳のための多言語音声合成

Multi-lingual speech synthesis for speech translation transferring voice characteristics of input speech in different languages.

研究発表会講演論文集、 pp. 915-918, 2021.
高木信二、 牛田光一、 橋本佳、 南角吉彦、 徳田恵一、 “因子分析に基づく HSMM を利用した構造化アテンション音声合成、” 日本音響学会 2021 年秋季研究発表会講演論文集、 pp. 871-874, 2021.
藤本崇人、 橋本佳、 南角吉彦、 徳田恵一、 “学習時と合成時の一貫性を考慮した VAE に基づく自己回帰型 sequence-to-sequence 音声合成、” 日本音響学会 2021 年春季研究発表会講演論文集、 2021.
角谷健太、 吉村建慶、 高木信二、 橋本佳、 大浦圭一郎、 南角吉彦、 徳田恵一、 “隠れセミマルコフモデルに基づく構造化アテンションを用いた Sequence-to-Sequence 音声合成、” 日本音響学会 2021 年春季研究発表会講演論文集、 pp. 943-946, 2021.
岩田康平、 高木信二、 橋本佳、 南角吉彦、 徳田恵一、 “勾配ブースティング決定木を用いた音声合成手法の検討、” 日本音響学会 2021 年春季研究発表会講演論文集、 pp. 813-814, 2021.

この研究は、令和元年度ＳＣＡＴ研究助成の対象として採用され、令和２～令和３年度に実施されたものです。