TSUBAME 共同利用　令和7年度　産業利用　成果報告書

利用課題名　感情豊かな音声合成技術の開発
英文：Construction of an emotion-rich speech synthesis model

利用課題責任者
大嶽匡俊
Masatoshi Otake

所属
株式会社 DubGuild
DubGuild Inc.
https://dubguild.com/

邦文抄録（300 字程度）
近年、音声基盤モデルの研究が急速に進展している。音声には、(1) テキスト化可能な言語情報、(2) 感情や演技など話者が意図的に変化させられるパラ言語情報、(3) 年齢・性別など話者属性に関わる非言語情報が含まれる。このうちパラ言語情報は、感情や意図を豊かに伝えるうえで重要である一方、一貫した定義やラベリングが難しく、深層学習での扱いが容易ではなかった。本研究では、パラ言語情報を効果的に表現できる音声モデルの構築を行った。あわせて、チューニング手法の探索を通じて、感情豊かな対話システムや吹き替え生成に資する基盤技術を確立する。

英文抄録（100 words 程度）
In recent years, research on speech foundation models has advanced rapidly. Speech contains (1) linguistic information that can be transcribed into text, (2) paralinguistic information, such as emotion and acting style, which speakers can intentionally vary, and (3) non-linguistic information related to speaker attributes, such as age and gender. Among these, paralinguistic information is particularly important for conveying emotion and intention, but it has been difficult to define consistently and annotate systematically, making it challenging to handle in deep learning. This study develops a speech model that can effectively represent paralinguistic information and establishes core technologies for expressive dialogue systems and dubbing generation.

Keywords:
Speech Synthesis
Language Model
High Performance Computing
Speech Analysis

背景と目的

近年、深層学習技術の飛躍的な発展に伴い、音声から取得できる情報のうち、話者の感情、意図、態度、話し方のニュアンスといった「パラ言語情報」を精緻に捉え、活用することの重要性が急速に高まっている。こうした情報は、人と AI との自然な対話、感情豊かな音声インタラクション、コンテンツ制作における表現力の向上に不可欠である。一方で、従来の音声モデルの多くは、音素列や文字列といった言語情報の処理を主眼として設計されており、感情表現や話者ごとの韻律変動を含む複雑なパラ言語情報を十分に扱い切れていない。とくにパラ言語情報の制御可能な生成への接続は依然として十分に確立されていない。

本研究では、多様な感情表現を自然に扱いうる深層学習音声モデルの構築を目的とした。特に感情音声合成に重点を置き、日本語の多話者・情動音声において課題となるアラインメント学習の不安定性に着目し、その改善に有効なモデル設計を検証した。これにより、感情豊かで自然な音声合成を可能にする基盤技術を確立し、将来的には対話 AI や音声インタフェース、吹き替え生成の品質向上に資することを目指した。

概要

本研究では、Flow Matching ベースの TTS モデルである Matcha TTS[1]を出発点として、日本語の多話者・情動音声に対応可能な zero-shot TTS モデルの構築を試みた。Matcha TTS では、テキスト系列からメルスペクトログラムの概形を予測し、その尤度に基づいて MAS によりアラインメントを推定する構成が採られている。しかし、多話者・情動音声ではメルスペクトログラムの変動が大きいため、この前提が崩れ、アラインメント学習が不安定になる。

この問題に対処するため、GlowTTS[2]に倣い、可逆な Glow モジュールを導入した。具体的には、観測された音響特徴を Glow モジュールによってより単純な潜在空間へ写像し、その空間上で尤度を定義して MAS によるアラインメント計算を行う構成とした。これにより、直接メル空間で整列を取る場合に比べ、アラインメント推定に必要な尤度構造を単純化し、情動音声特有の大きな変動を吸収しながら学習を進めることを狙った。

具体的な実装について

テキスト入力はエンコーダにより、メルの概形を予測したうえでメルの空間で尤度を計算することでその尤度を最大化するアラインメントを予測する。L_{prior}は尤度最大化に基づくエンコーダの損失を指す。

$$
\begin{aligned}
x &= \mathrm{Encoder}(w) \\
a_{i,j} &= \log \mathcal{N}(y_j; x_i, I) \\
A &= \mathrm{MAS}(a) \\
\hat{x} &= \mathrm{LengthRegulator}(x, A)
\end{aligned}
$$

$$
\mathcal{L}_{\text{prior}} = \sum_{j=1}^{T} \log \mathcal{N}(y_j; \hat{x}_j, I)
$$

ただし、多話者・情動音声においてはそもそもメルの概形を予測するのが難しいため、アラインメントを計算するための前提条件が成立せず、アラインメントの学習が成功しないことが分かった。

そのため、GlowTTS に倣い、可逆な Glow モジュールを導入する。

$$
\begin{aligned}
x &= \mathrm{Encoder}(w) \\
z &= f_\theta^{-1}(y) \\
a_{i,j} &= \log \mathcal{N}(z_j; x, I) \\
A &= \mathrm{MAS}(a)
\end{aligned}
$$

$$
\mathcal{L}_{\text{prior}} = \sum_{j=1}^{T} \log \mathcal{N}(y_j; \hat{x}_j, I) + \log \left| \det \frac{\partial f_\theta^{-1}(y)}{\partial y} \right|
$$

なお、本研究で用いた記号は以下の通りである。
xはテキスト潜在表現、y はメルスペクトログラム、$a_{i,j}$はアテンションスコア、A はアラインメント、$\hat{x}$ は長さ調整後の表現、w は音素列を表す。テキスト入力 wから エンコーダにより潜在表現 xを得て、尤度に基づきアラインメント Aを推定する。

結果および考察

実験の結果、Glow モジュールを導入しない構成では、アテンションスコアの分布が不安定となり、アラインメント行列としても不自然な結果が観察された。すなわち、音素列と時間軸との対応関係が適切に学習されず、意味のあるアラインメントが形成されていないことが確認された。これに対し、Glow モジュールを導入した構成では、アテンションスコアが安定した値を示し、音素列と時間軸との対応として妥当な、対角線状に近いアラインメントが得られた。

この結果は、多話者・情動音声のように音響変動の大きい条件下では、アラインメント計算をそのままメル空間で行うことが本質的に難しいことを示唆している。一方で、Glow モジュールにより音響特徴をより単純な潜在空間へ写像することで、アラインメント学習に必要な尤度構造が扱いやすくなり、学習の安定化につながったと考えられる。

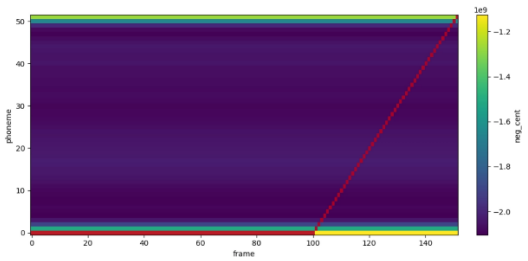

図 1 Glow 導入前のアライメント

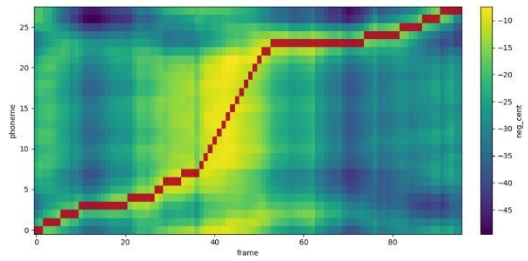

図 2 Glow 導入後のアライメント

まとめ

本研究では、感情や意図といったパラ言語情報を豊かに含む音声を対象として、深層学習に基づく感情音声合成モデルの構築に取り組んだ。特に、Flow Matching ベースの Matcha TTS を日本語の多話者・情動音声へ拡張する際に課題となるアラインメント学習の不安定性に着目し、GlowTTS に倣った可逆 Glow モジュールを導入した。

その結果、従来構成では困難であった意味のあるアラインメント学習が可能となり、アテンションスコアの安定化と、対角的で妥当な整列構造の獲得を確認した。これにより、感情音声合成において重要となる学習安定性の改善が達成され、多様な感情を自然に扱いうる音声モデルの実現に向けた有望な足掛かりが得られた。

今後の課題

今後の課題として、第一に、アラインメントの改善が実際の生成音声品質にどの程度寄与するかを、客観評価および主観評価の両面から定量的に検証する必要がある。具体的には、自然性、感情表現の再現性、話者性保持、発話明瞭性などの観点から評価を行い、既存手法との比較を明確にすることが重要である。

第二に、感情表現の制御性向上が課題として挙げられる。現段階では、感情を含む多様な音声データに対して安定した学習基盤を得ることに重点を置いたが、今後は感情カテゴリや連続的な感情表現を明示的に制御できるモデルへ発展させる必要がある。

第三に、zero-shot 条件下での汎化性能向上も重要である。多様な話者、録音条件、発話スタイルに対して頑健に感情表現を再現するためには、さらなるデータ整備とモデル改善が必要である。とくに、話者性と感情表現を適切に分離しつつ制御する学習枠組みの検討が今後の重要なテーマとなる。

参考文献

[1] Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, Gustav Eje Henter.
“Matcha-TTS: A fast TTS architecture with conditional flow matching”,
IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024.
https://arxiv.org/abs/2309.03199

[2] Jaehyeon Kim, Sungwon Kim, Jungil Kong, Sungroh Yoon.
“Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search”,
Advances in Neural Information Processing Systems (NeurIPS), 2020.
https://arxiv.org/abs/2005.11129