E-017

ディープラーニングを用いた潜在表現による感情を表現する音声合成の実現

Speech Synthesis to Express Emotions through Latent Expression with DeepLearning

鎌田凌輔† 土屋誠司‡ 渡部広一‡

Ryosuke Kamada Seiji Tsuchiya Hirokazu Watabe

1. はじめに

近年，AI による音声技術を活用したアプリケーションが広く活用されている．AI による音声技術は，スマートスピーカやバーチャルキャラクタの声などに歌声合成に活用されている．これらには，音声合成，音声変換，音声認識といった AI 音声技術が利用されており，音声合成技術は機械学習が取り入れられる前からも研究は進んでいた．音声合成技術に機械学習が取り入れられる事によって，合成された声は以前よりも自然さと明瞭さが向上した．

文字から直接音声を合成する技術である End-to-end Text-To-Speech モデルは近年急速に発展し，人間の声に近い品質となっている．音声合成で生成された音声が表現豊かな音声を実現することで，より人間に近い印象を与えることができると考えられる．理由として，人間は声や表情，態度で感情表現をすることで他者とのコミュニケーションを図ることが多く，特に会話時に感情表現が見受けられるからである．そこで，本研究の目的は，音声合成によって音声を生成する際に，感情を表現する音声を合成することである．音声合成によって生成された音声が，聞き手に意図や感情を想起させるような韻律で発話を行うことができれば，聞き手が理解しやすい会話を行うことができると考える．また，提案手法として，教師なし学習である Variational Autoencoder(VAE)[1]を使用して，感情ラベルありの音声を学習を行う．なお，本研究で採用する感情はエクマンの基本感情[2]を参考にする．

2. 関連研究

2.1. 音声合成

音声合成とは任意の文章から対応する音声を人工的に生成する技術のことである．音声合成を実現する基本要素として，テキスト解析，言語特徴量から音響特徴量へ変換する音響モデル，音声波形を生成するボコーダーの３つのモジュールがある．

2.2. エクマンの基本感情

エクマンの基本感情[2]とは，「喜び」，「悲しみ」，「怒り」，「驚き」，「嫌悪」，「恐れ」のことで．人間の感情はこれらを組み合わせて表現されているとされる．エクマンは基本感情が全人類に普遍的であり生物学的基盤を持つと結論した．

2.3. VAE(Variational Autoencoder)

VAE(Variational Autoencoder)[1]とは教師なし学習である Autoencoder の一種であり，Kingma らによって最初に定義された．損失関数は入力と出力の誤差である．Autoencoder と VAE の異なる点は，潜在変数に確率分布を用いることである．具体的に，VAE では観測変数 x と潜在変数 z のデータ生成過程を，確率密度分布のモデルパラメータ θ を用いて z~p(z),x~p_θ (x | z)のように定める．真の事後分布 p_θ (x | z)を近似した分布 q_ϕ (z|x)（ϕ はモデルパラメータ）を用いて，周辺尤度の下界 L(θ,ϕ,x)が以下のように求まる．

$$
\begin{aligned}
log p\theta(x) &= D_{KL}(p_\phi(z|x)||p_\theta(z|x)) + \mathcal{L}(\theta,\varphi,x) \\
&\geq \mathcal{L}(\theta,\varphi;x) \\
&= -D_{KL}(p_\phi(z|x)||p(z)) + E_{q_\phi}(z|x)[\log p_\theta(x|z)]
\end{aligned}
$$

学習の際は，下界 L(θ,ϕ,x)が最大になるようなパラメータ θ，ϕ について最適化を行う．

† 同志社大学大学院理工学研究科

‡ 同志社大学理工学部インテリジェント情報学科

3. 先行研究

阿久澤らの研究に「変分自己符号化器を用いた表現の多様性のモデル化による表現豊かな音声合成」[3]がある．この研究では，機械学習の自己回帰生成モデルと VAE を用いて，感情ラベルなしの音声データから音声合成モデルを訓練し，表現豊かな音声合成の生成を行っている．本研究では，VAE を用いたモデルの学習を参考にしている．

4. 既存システム

4.1. Tacotron2

Tacotron2[4]は 2017 年に Google で開発された TTS(Text To Speech)アルゴリズムであるが，現在でも最も肉声に近い高品質な音声合成モデルだと言われている．Tacotron2 ではテキストを音声に変換する際にテキストからメルスペクトログラムに変換，メルスペクトログラムを音声に変換という大きく 2 段階の処理で TTS を実現している．

4.2. 音声コーパス

本研究では日本声優統計学会[5]が提供している「声優統計コーパス」を使用する．声優統計コーパスは，様々な発音がバランス良く含まれている日本語の台本（音素バランス分）と，それらをプロの女性声優 3 名が読み上げた音声データの総称である．含まれる音声の種類には 3 種類の感情があり，普通，喜び，怒りがある．

5. 提案手法

5.1. Style Encoder と Tacotron2

本研究では End-to-end Text-To-Speech モデルである Tacotron2 に VAE を導入し，Tacotron2 へ入力する潜在変数を表現するため，Style Encoder を提案する．そのため，感情表現を行うために学習音声データをメルスペクトログラムに変換し VAE の入力データとする．更に，本研究で使用するデータには感情ラベルを付加して学習させるため，CVAE(Conditional Variational Autoencoder)を使用する．CVAE とは VAE に対してデータの状態を付加させて学習を行うことで，データ生成を行う時にデータの状態を指定することができるオートエンコーダである．Tacotron2 では，潜在変数とテキストエンコーダの状態を合わせたエンコーダの状態を目標とする音声に変換する．

5.2. 学習

Tacotron2 で日本語の文章を学習できるよう，以下に示す表のように入力テキストをローマ字変換を行う．更に，Style Encoder の学習の際には感情ラベルを含めて学習を行うため，テキスト文の末に感情ラベルを追加する．

表 1 音声コーパスに含まれている text

meian/meian_0000.wav|この前探った時は、途中に瘢痕の隆起があったので、ついそこが行きどまりだとばかり思って、ああ言ったんですが、|kono mae sagut ta toki wa 、 tochu- ni hankon no ryu-ki ga at ta node 、 tsui soko ga yukidomari da to bakari omot te 、 a- yut ta n desu ga 、|8.77

表 2 感情ラベルを追加した text

meian/meian_0000.wav|konomaesaguttatokiwa,tochu-nihankonnoryu-kigaattanode,tsuisokogayukidomaridatobakariomotte,a-yuttandesuga,|0

Copyright © 2022 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

5.3. 推論

推論を行う際の入力には，テキスト文と潜在変数 Z を入力とし，出力として音声波形を取得する．感情を指定する際には，感情ラベルごとに潜在変数 z の平均をとった値を，それぞれの感情に対する値とする．

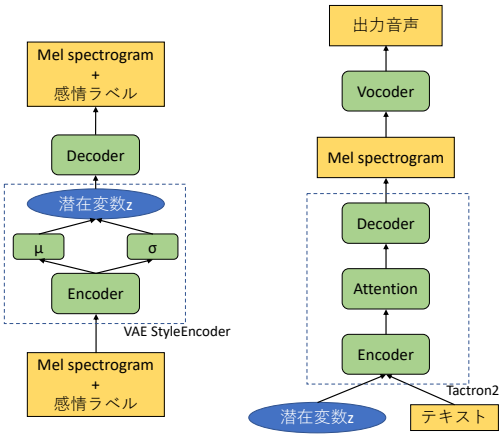

図 1　推論の際のフロー

6. 評価

6.1. 評価手法

感情を含んだ音声合成の評価には，大学生や社会人など 5 名の協力を得た．各人に対して，音声合成をした文がどのような感情に聞こえたかを自由選択方式により感情を選択してもらった．また意図した感情の合成音声が，どれだけ意図した感情に聞こえたかを 5 段階で採点してもらった．これによって感情の正答率と，どれだけ自然に感情を読み取れたかを評価する．

評価セットとして「おはよう」「こんにちは」「ありがとう」「お疲れ様です」「かしこまりました」のあいさつ文，応答文の合計 5 文を用いる．それぞれ「感情なし」，「喜び」，「怒り」の感情が含まれた音声合成に対し，どのように聞こえたかの回答を「喜び」「悲しみ」「怒り」「恐れ」「驚き」「嫌悪」「感情なし」「該当なし」の 8 個の感情の中からを選ばせた．

6.2. 評価結果

表 3 に，評価結果を示す．この表は，縦軸が付与した感情をを示しており，横軸は被験者が選択した感情である．各セルには，計５名の被験者が選択した割合を示している．また点数は，どれだけ意図した感情が聞こえたかの評価点の５名の平均であり，得点が高いほど，意図した感情が感じられている．

表 3　感情ごとの選択割合と点数

		回答した感情								
		喜び	悲しみ	怒り	恐れ	驚き	嫌悪	感情なし	該当なし	点数 (5点満点)
付与した感情	喜び	96%	0%	0%	0%	0%	0%	4%	0%	4.60
	怒り	0%	0%	84%	0%	0%	16%	0%	0%	4.12
	感情なし	4%	0%	8%	0%	0%	0%	84%	4%	4.04

7. 考察

島部さんの研究[6]では，「喜び」が 78%，「怒り」が 82%という結果であったが，本研究で生成した音声合成の方が，評価を上回っているため，提案システムの有効性を示すことができている．結果として，「喜び」の感情を含んだ音声合成が，一番被験者に感情が伝わった．この結果は「喜び」の音声のほうが他の感情を含んだ音声よりも識別しやすいと考える．また，感情を付与することはできているが，Tacotron2 単体で生成された音声よりも本研究の VAE を導入した Tacotron2 の方が合成した音声の人間の声のような自然さが欠けているように感じる．この問題に対して考えられることとして，学習させる音声データの数が足りてないことや，ハイパラメータの設定，入力にテキスト文と潜在変数 z を入力することによる学習の際のノイズとして認識されたことが挙げられる．

8. おわりに

本研究で示した，End-to-end Text-To-Speech モデルは VAE を導入し，音声の連続空間における話し方の潜在表現を教師なしで学習することで，合成音声の感情表現を制御することができた．しかし，潜在変数 z をテキストと合わせて学習することによって，感情表現を制御できたが，不自然さが残る音声になることもあった．

また，本研究の提案システムでは，学習の際に音声データの感情ラベルも必要なため，他の音声を学習する際に，感情ラベルがなければ学習させることができないという問題があるため，より人間のような感情表現をするという表現の豊かさについては課題として残された．

そのため将来の研究の方向性としては，感情ラベルなしでの学習を行う．例えば提案システムである Style Encoder を学習させる際に，潜在変数 z の値をクラスタリングすることができれば，感情ラベルなしの音声データであっても感情ごとに，合成した音声を表現することができると考えられる．Tacotron2 と同レベルの音声クオリティを実現するために，ハイパーパラメータの調整やシステムの見直しを行う必要がある．

よって VAE を導入することで音声の感情表現を行うことはできたが，このテーマはこれからも研究が必要だと考える．

謝辞

本研究の一部は，JSPS 科研費 16K00311 の助成を受けた

参考文献

[1] Diederik P Kingma and Max Welling, “Autoencoding variational bayes,” in Proc. 2nd International Conference on Learning Representations, 2014.
[2] ポール・エクマン，W．フリーセン：“表情分析入門-表情に隠された意味をさぐる”，第 7 版，誠信書房，2000
[3] 阿久澤圭 et al.，“変分自己符号化器を用いた表現の多様性のモデル化による表現豊かな音声合成”，2018 年度人工知能学会全国大会，2018
[4] Shen et al. 2018 natural tts synthesis by conditioning wavenet on mel spectrogram predictions, ICASSP 2018.
[5] 日本声優統計学会，“声優統計コーパス”，2020
[6] 島部貴由，“会話ロボットのための感情を表現する音声合成”同志社大学修士論文，2013
[7] 大浦圭一郎，酒向慎司，徳田恵一：日本語テキスト音声合成システム Open JTalk，日本音響学会春季構論集，Vol.1, No.2-7-6, pp.343-344(2010)
[8] Hochreiter, S. and Schmidhuber, J Long Short-Term Memory, Neural Computation, Vol. 9, No. 8, pp. 1735–1780 (1997)

Copyright © 2022 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.