7S-08

GANによる音声感情を反映させた フォント自動生成システムの改良

山本　悠†　　鈴木　祐太‡　　土屋　奎太‡　　陳　キュウ†

工学院大学情報学部情報通信工学科†　　工学院大学大学院工学研究科電気・電子工学専攻‡

1. はじめに

近年、PCやスマートフォンの所持は一般化しつつあり、SNSやメディアなどで文字によるコミュニケーションをとることは日常的なこととなっている。これらのシーンでは、文字だけでは感情が伝わりづらいため、感情表現のツールとしてフォントやデザインを工夫することが行われている。しかし、一般のユーザーが感情表現を行いたい場合、フォントを自分で作成することは困難であると言える。

専門性が要求されると考えられるフォントの決定に注目し、顔画像の感情成分に合わせた文字の形状変化で表現支援を行う先行研究がある[1]。本研究では[1]の手法に基づいた、リアルタイムで入力された音声に含まれる音声感情表現フォント生成システムの問題点の改良を目的とする。

2. 概要

2.1. 音声からの感情検出

本研究のシステムでは、入力音声から感情を抽出する手法として、Empath社のWebEmpathAPIが使用されている[2]。これは、音声に含まれている物理的な特徴量から感情の情報を独自のアルゴリズムで判定するプログラムを用いて、数万人の音声データベースをもとに喜怒哀楽を判定するものである。

ここで、WebEmpathAPIの主な問題点として、企業による商用のAPIであるため、無料で利用できるのは1か月に300回の呼び出しまでという制限があることが挙げられる。これによって、WebEmpathAPIを利用している本システム全体の利用も月300回に制限されてしまう。

2.2. GAN

GAN[3]とはGenerative Adversarial Network（敵対的生成ネットワーク）を省略したもので、生成モデルの1つとして知られている。GANの概要図を図1に示す。GANはGeneraterとDiscriminaterの2つから構成される。まず、入力したノイズが、生成器（Generater）によって本物データに似せた偽物のデータとして出力される。これを判別器（Discriminater）によって本物のデータと比較・識別させる。以上を繰り返すことで2つのネットワークを競わせ、学習させることで存在しないデータや存在するデータによく似たデータを生成することができる。

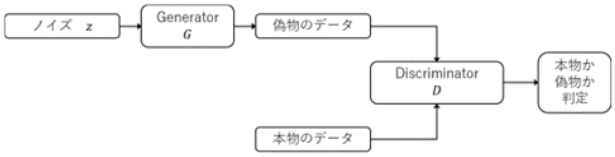


図1　GANの概要

2.3. zi2zi

本研究のシステムの生成モデルとして、zi2zi [4]を使用している。zi2ziはYuchen Tianらが提案したGANの仕組みを利用したpix2pixを漢字に応用させたモデルであり、漢字を異なるスタイルに変化させることができる。

2.4. 提案システム

本研究のシステムの構成を図2に示す。システム全体は学習部分と生成部分から構成されている。学習部分では、zi2ziを利用して感情情報を反映したフォントを生成できるよう学習させる。生成部分では音声を入力し、WebEmpathAPIによる感情の検出と、Google Cloud Speech to Text APIによる音声認識・テキスト化を行い、それらのデータを学習部分で学習済みのGeneraterに渡す。渡されたデータからGeneraterが感情に即して文字を変化させ、フォントを生成する。

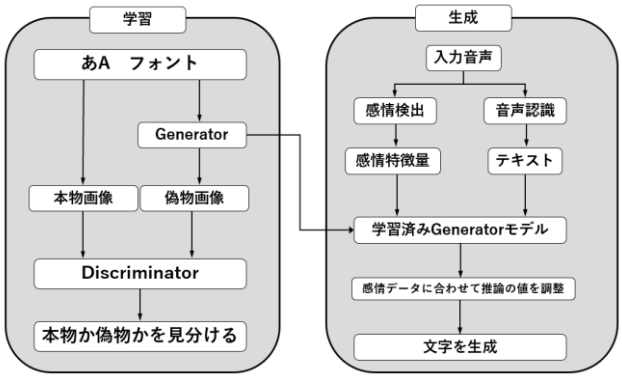


図2　システム全体の構成図

Improved Automatic Font Generation System Reflecting Voice Emotions via Generative Adversarial Networks

†Yu Yamamoto, Qiu Chen, Department of Information and Communications Engineering, Faculty of Informatics, Kogakuin University

‡Yuta Suzuki, Keita Tsuchiya, Graduate School of Engineering, Kogakuin University

3. 実験

3.1. インターフェースの改良

利便性の向上、表現の視認性の向上を図った。改良前のインターフェースを図 3、改良後のインターフェースを図 4 に示す。感情をグラフで可視化し、生成結果を直感的に理解できるようにした。

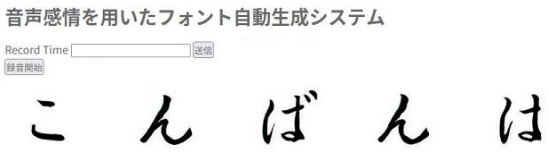
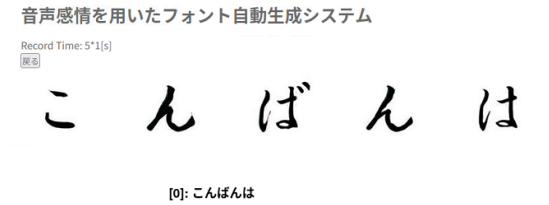


図 3　改良前インターフェース



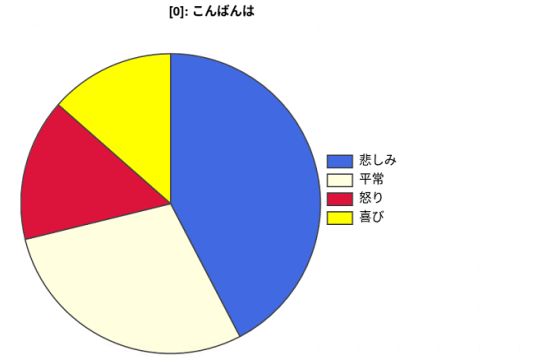


図 4　改良後インターフェース

3.2. WebEmpathAPI の代替

提案システムの問題点となっている WebEmpathAPI を無料で使い続けられるシステムで代替することでシステムの改良を図った。

ここで、人間は文化や言語に依らず幾つかの基本的な感情を有するという仮説[5, 6]に基づいた、言語が異なる感情音声に対し単一の感情認識モデルで感情認識を行う研究（クロスコーパス感情認識）が存在する。

また、公開されている日本語の感情ラベル付きの音声コーパスは、「感情評定値付きオンラインゲーム音声チャットコーパス（OGVC）[7]」のみであり、日本語のみのデータセットで学習を行うのは難しいと考えられる。

以上より、WebEmpathAPI の代替としてクロスコーパス感情認識の手法を用いることが適していると考えられる。本研究ではその一つとして、オープンソースのプログラムである SPEAKER-VGG-CCT[8]を用いる。このシステムでは、入力された音声から物理的な特徴量を抽出し、感情を推測することができる。これを用いて音声から感情分析を行い、WebEmpathAPI との比較を行った。

WebEmpathAPI と SPEAKER-VGG-CCT を用いて、イタリア語の EMOVO データセット[9]と日本語の OGVC データセットで感情推定の検証を行い、精度を比較した結果を表 1 に示す。表 1 に示す通り、どちらの言語でも精度の向上が見られた。

表 1　感情推定の検証結果

データセット	モデル	精度
OGVC(日)	Speaker-VGG-CCT	48.11%
	Empath	28.00%
EMOVO(伊)	Speaker-VGG-CCT	48.50%
	Empath	29.90%

4. まとめ

本研究では、インターフェースの改良、WebEmpathAPI の代替により、実用性を向上させ、使用制限を緩和することで、システムを改良することができた。また、音声感情抽出の精度向上も達成し、より正確に感情をフォントで表現することが可能になった。今後はさらなる精度の向上、システムの音声認識部分のオープンソース化などを検討する。

参考文献

[1] 中村充志，“画像の感性を反映させたフォントの自動生成に関する研究”，平成 29 年度工学院大学卒業論文，2017.

[2] “声ダケノ感情認識テスト Empath”, https://webempath.net/lp-jpn/

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Advances in Neural Information Processing Systems, 27, 2014.

[4] Y. Tian, “zi2zi: Learning Chinese Character style with conditional GAN”, https://github.com/kaonashi-tyc/zi2zi, 2017

[5] C. E. Izard, “Human emotions,” Plenum Press, New York, 1977.

[6] P. Ekman, “An argument for basic emotions,” Cognit. Emot., 6, pp. 169-200, 1992.

[7] Y. Arimoto, H. Kawatsu, S. Ohno and H. Iida: “Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment”, Acoustical Science and Technology, 33, 6, pp. 359-369, 2012.

[8] A. Arezzo, S. Berretti, “SPEAKER VGG CCT: Cross-corpus speech emotion recognition with speaker embedding and vision Transformers”, In Proc. the 4th ACM Int. Conf. on Multimedia in Asia, pp. 1-7, 2022.

[9] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, “EMOVO corpus: An Italian emotional speech database,” In Proc. Int. Conf. on Language Resources and Evaluation, 2014.