Vol.2022-CDS-34 No.7
2022/6/10

GAN を用いた音声感情を反映できる フォントの自動生成システムの検討

土屋奎太 [1]　野中琢登 [2]　陳キュウ [1]

概要：テレビ番組や動画配信サービスにおいて，メディアが伝えたい感情を表すためテロップなどで様々な文字が工夫されて使用されている．しかし，そのような感情を含む文字を作成するには文字のフォントや色などの複雑な情報を決定する必要があるため，高度な専門性が求められている．そこで本研究では，特に専門性が要求されると考えられるフォントの設計を支援するため，敵対的生成ネットワーク(GAN)を用いて音声に含まれる感情とフォントの関係を学習することで，音声感情情報を反映したフォントの自動生成システムの構築を試みた．

キーワード：フォント自動生成，音声感情，GAN

GAN-based Automatic Font Generation System Reflecting Voice Emotions

KEITA TSUCHIYA[†1]　TAKUTO NONAKA[†2]
QIU CHEN[†1]

Abstract: In TV shows and video distribution services, various texts are devised and used in subtitles to express the emotions the media wants to convey. However, in order to generate such emotional text, complex information such as the font and color of the texts needs to be determined, thus requiring a high degree of professionalism. Therefore, in this research, in order to support the design of fonts that are considered to require special expertise, we utilize an adversarial generative network (GAN) to learn the relationship between fonts and emotions contained in voice, and construct an automatic font generation system that reflects voice emotion.

Keywords: Automatic font generation, Voice emotion, GAN

1. はじめに

近年，幅広く普及しているテレビ番組や動画配信サービスにおいて，メディアが伝えたい感情を表す為にテロップなどで様々な文字が工夫されたフォント，デザインで装飾されて使用されることは当然のこととなっている．しかしそのような感情を表現する文字を作成するには文字のフォント，デザインなどの複雑な情報を決める必要がある為，デザインの知識を持ち合わせていない一般のユーザがフォントを作成することは非常に困難であるといえる．

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 [1]がある．本研究では [1]の手法に基づき，リアルタイムで入力された音声に含まれる感情を定量的に解析し，それを反映したフォントの自動生成システムの構築を目的とする．これにより一般ユーザでも容易に高度なフォントのデザインをすることができるようになる．

本研究では，音声感情を反映できるフォントの自動生成システムの構築する．本システムの全体を表す概略を図 1 に示す．本システムの構成は音声入力，音声のテキスト化・感情解析，フォント推論・生成の 3 つの段階に分けられる．第 1 段階では音声入力では入力された音声を音声ファイルに変換する．第 2 段階ではその音声ファイルを入力として Web Empath API [2]による感情解析と Google Cloud Speech to Text API [3]による音声認識・音声のテキスト化を行う．第 3 段階では第 2 段階で得たデータを入力として，フォントの推論・生成を行う．フォントの推論・生成には zi2zi [4]と呼ばれる生成モデルを使用する．

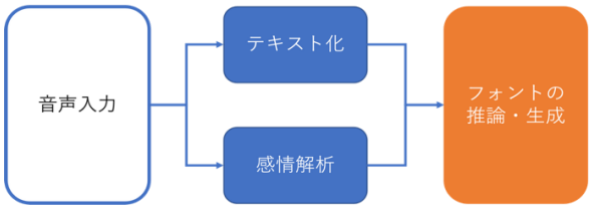

図 1　システム全体の概略図

1 工学院大学大学院
Kogakuin University
2 (株)ヤマトシステム開発
Yamato System Development Co., Ltd.

2. 理論

2.1 Web Empath API

入力音声から感情情報を取得するのに Empath 社の Web Empath API [2]を使用している．Empath と呼ばれる音声等の物理的な特徴量から気分の状態を独自のアルゴリズムで判定するプログラムを用いて，数万人のデータベースを元に喜怒哀楽や気分の浮き沈みを判定する API が Web API 化されたものである．感情の種類を喜び，怒り，悲しみ，平常の 4 つとするとそれぞれの感情の値を 0~50 の値で出力する．

表 1　Web Empath API 対応フォーマット

形式	PCM WAVE 16bit
データサイズ	1.9MB 以下
フォーマット	PCM/FLOAT/PCM_SIGNED/PCM_UNSIGNED いずれか
時間	5.0 秒未満
サンプリング周波数	11025Hz
チャンネル数	1(モノラル)

2.2 Google Cloud Speech to Text API

入力音声を音声認識してテキスト化する機能として Google Cloud Speech to Text API [3]を使用する．

Speech to Text API には同期認識，非同期認識，ストリーミング認識の 3 つの主要な音声認識方法がある．同期認識は音声データを Speech to Text API に送信してデータの認識を行い, 全ての音声が処理されたら結果を返すものである．同期認識リクエストは 1 分以内の音声データに制限される．非同期認識は音声データを Speech to Text API に送信し，長時間オペレーションを開始する．このオペレーションを使用することで，認識結果を定期的にポーリングすることができる．また非同期認識リクエストは長さ 480 分までのデータに使用される．ストリーミング認識は双方向ストリームで提供された音声データの認識を行う．マイクからのライブ音声のキャプチャなどのリアルタイムの認識を目的として設計されているため，音声をキャプチャしながら暫定的な結果を生成して，結果を表示することができる．

2.3 CNN

CNN（畳み込みニューラルネットワーク）とはいくつもの深い層を持ったニューラルネットワークであり，画像認識の分野で主に活躍している．CNN の初期モデルである LeNet [5]の構造を図 2 に示す．

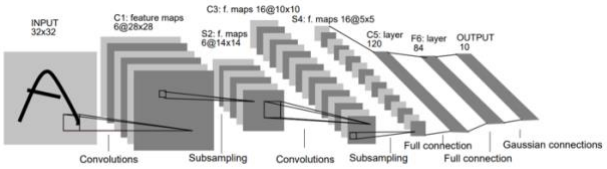

図 2　LeNet の構造 [5]

CNN は畳み込み（Convolution），サブサンプリング（Subsampling）[プーリングに相当] を繰り返した後，全結合を行うことが多い．畳み込み層では画像のより手がかりになる特徴の部分を抽出している．プーリング層では畳み込み層から得た特徴から特定の演算を行い, 縮小している．図 2 では畳み込みを 2 回，プーリングを 2 回行った後，全結合されている．

2.4 U-Net

U-Net [6] は 2015 年に発表された画像セグメンテーション推定するための畳み込みネットワークである．図 3 に U-Net の構造を示す．U-Net は Encoder と Decoder から成るモデルである．後述する zi2zi のモデル構造の一部である Encoder、Decoder は U-Net を使用している．

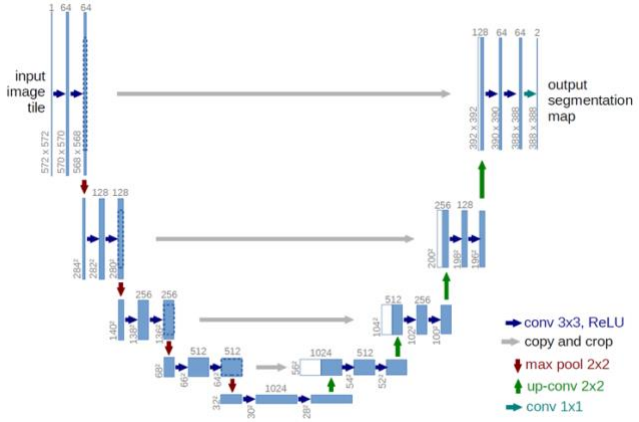

図 3　U-Net の構造 [6]

U-Net の Encoder は入力された画像を何度か畳み込み，その画像の特徴を抽出する．Decoder では Encoder で抽出した特徴を畳み込みとは逆の処理で入力画像と同じサイズにする．また Encoder の情報を各階層で Decoder と連結することで物体の位置情報を捉え続けることができる．

2.5 GAN

GAN [7]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 はデータから特徴を学習することによって存在しないデータを生成することや存在するデータの特徴に近いデータを生成することができる. GAN の概要図を図 4 に示す.

GAN は Generator（生成ネットワーク）と Discriminator（識別ネットワーク）の 2 つから構成されていて，これらを互いに競わせることで学習し，精度を高めることができる．Generator ではノイズが入力されると，そのノイズを本物データに似せた偽物データにして出力を行う．それを Discriminator によって本物データと比較し識別させることを繰り返すことで学習する．

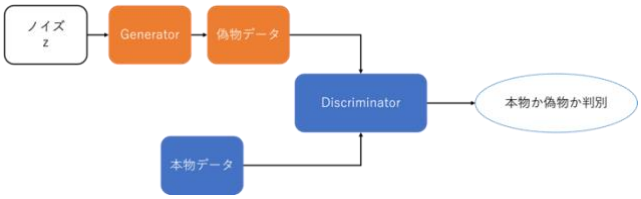

図 4　GAN の概要図

GAN の価値関数の数式を以下に示す．

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}\left[\log\left(1 - D\left(G(z)\right)\right)\right] \quad (1)$$

xは訓練データ，zはランダムノイズ，Dは Discriminator，Gは Generator を示している．Generator は Discriminator が生成画像は偽物であると判定する確率を最小化し，反対に Discriminator は正しく判定する確率を最大化するようにして，学習を行う．

2.6 pix2pix

pix2pix [8]とは GAN を利用した画像生成モデルの一つである．pix2pix では 2 つの画像のペアから画像間の関係を学習することにより，1 枚の画像からその関係を考慮した補間を行なって画像を生成する．pix2pix の概要図を図 5 に示す．なお図 5 では入力画像とそのエッジ画像が例として使用されている．

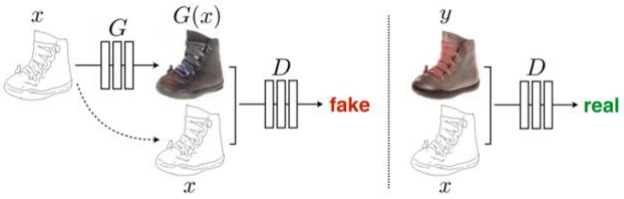

図 5　pix2pi2 の概要図 [8]

図 5 ではyが入力画像，xが入力画像のエッジ画像を示していて，これらの画像を Generator の入力として扱う．その後は GAN のように Discriminator で本物と偽物の判別を行って学習する．

pix2pix はラベルから道路や建物，航空写真から地図，昼から夜など様々な場面で使用可能である．以下に pix2pix の実験結果例を図 6 として示す．

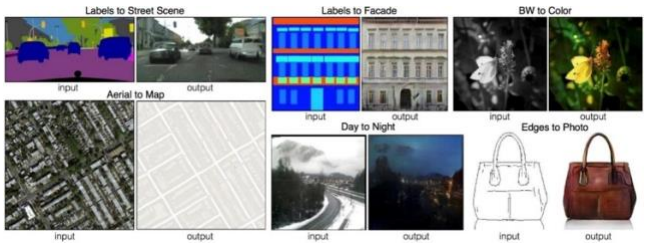

図 6　pix2pix の実験結果例 [8]

2.7 zi2zi

本研究のシステムの生成モデルとしては 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る．漢字を異なるスタイルに変化させることができる．漢字以外にもひらがな，アルファベットなどにも対応可能である．zi2zi の概要図は図 7 に示す．

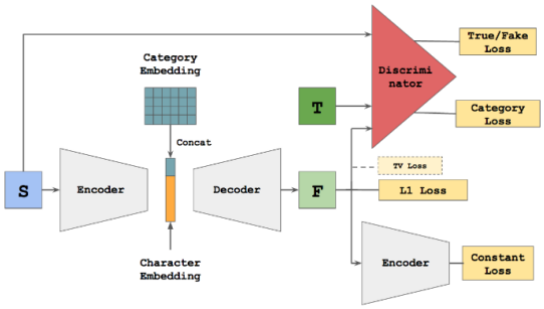

図 7　zi2zi の概要図 [4]

2.8 提案手法

本研究の詳細なシステム構成を図 8 に示す．システム全体は学習部分と生成部分に分けられている．学習部分では zi2zi を利用してフォントの画像を学習させることによって，感情情報を反映した文字フォントを推論可能な Generator を作成する．生成部分では入力音声を Web Empath API による感情検出・特徴抽出と 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 Generator に渡す．学習済みの Generator はそれらのデータから感情データに合わせて文字を変化させ生成する．

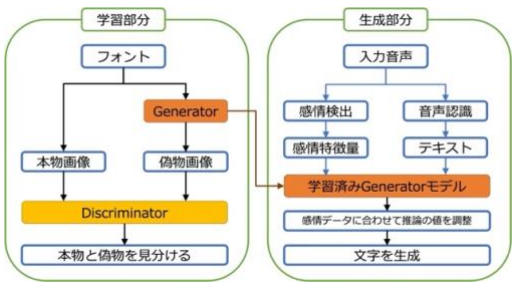

図 8　システム全体の構成図

3. 実験

3.1 実行環境

本システムの実行環境を以下の表 2 に示す.

表 2　実行環境

CPU	Intel Core i3-9100
メモリ	15.6GiB
ストレージ	4TB
GPU	NVIDIA GeForce GTX 1660 SUPER
OS	Ubuntu20.04 LTS
開発言語	python 3.6.13
ライブラリ	imageio 2.9.0
	librosa 0.7.9
	numpy 1.19.2
	pillow 8.2.0
	pyaudio 0.2.11
	requests 2.26.0
	scipy 1.1.0
	tk 8.6.11
	tensorflow 1.9.0

3.2 感情に対応させたフォント

本研究の学習部分で扱う zi2zi では画像のペアを入力として学習させる．それに伴い，プレーンなフォントの画像と感情に対応したフォントの画像を用意する必要がある．

[1]の研究で行われているアンケートから，それぞれの感情に対応したフォントを設定した．それぞれの感情とそれに対応したフォントを以下の表 3 に示す．またそれらのフォントを「世界」という文字で表した例を図 9 に示す.

表 3　感情とフォントの対応表

感情	フォント
喜び	HG 創角英ポップ体
怒り	HOT-大髭 115StdH
悲しみ	HG 行書体
平常	HG 教科書体

図 9　フォントと感情の対応画像

3.3 インターフェース

本システムのインターフェースを図 10，11 に示す．図 10 は生成前の状態，図 11 は生成後を状態である．5 秒の録音を何回ループするのかを 5s×n の欄に書き加え，録音時間決定を押下した後，録音開始ボタンを押下すると録音が開始される.

図 10　インターフェース画像

図 11　インターフェース画像（生成時）

3.4 生成結果

生成結果として得られた画像の例を以下の図 12，13 に示す．また，それぞれの図のタイトルは感情の値がいくつと検出されたのかを示している.

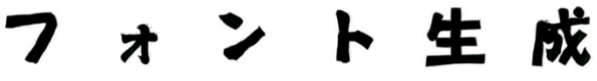

図 12　喜び：30　怒り：15

図 13　平常：38　悲しみ：7

4. まとめ

4.1 結論

本研究において生成システム全体の開発環境の統一とともに，音声の入力から音声認識，感情解析を行い感情情報からフォントスタイルを変化させた文字を自動で生成するシステムを正常に動作させることができた．

またインターフェースに関してはこれから機能を新たに加える上でより汎用的なインターフェースにすることができると考えられる．

4.2 今後の課題

今後の課題としてはシステムの学習部分について評価実験を行うことや，インタフェースの更なる改良，またより滑らかな音声認識の実現が挙げられる．

インタフェースの更なる改良については文章などの大量の文字の表示に対して対応することや，感情リストの明示機能，音声の確認再生機能，録音停止機能などが挙げられる．

滑らかな音声認識については本システムでは 5 秒間を n 回ループさせる形で録音時間を決めているが，5 秒間で区切られていることが原因で文字が不自然に区切られることがある．より滑らかな音声認識を実現する必要がある．

参考文献

[1] 中村充志, 瀧澤生, 星泰成, 綱島秀樹, 陳キュウ. 画像の感情を反映させたフォントの自動生成手法. 日本感性工学会論文誌. 2018, vol. 17, no. 5, p. 523-529.
[2] “Web Empath API について”. https://webempath.net/lp-jpn/, (参照 2022-05-18).
[3] “Speech-to-Text”. https://cloud.google.com/speech-to-text/, (参照 2022-05-18).
[4] “zi2zi: Learning Chinese Character Style with Conditional GAN”. https://github.com/kaonashi-tyc/zi2zi, (参照 2022-05-18).
[5] Haffner, L. Léon, B. Yoshua, B. Patrick, Y.. Gradient-Based Learning Applied to Document Recognition. Proc. the IEEE. 1998, vol. 86, no. 11, p. 2278-2324.
[6] Olaf, R. Philipp, F. Thomas, B.. U-net: Convolutional Networks for Biomedical Image Segmentation. Proc. Int. Conf. on Medical Image Computing and Computer-assisted Intervention. 2015, p. 234-241.
[7] Ian, G. J. et al.. Generative Adversarial Nets. NIPS. 2014, vol. 3, p. 2672-2680.
[8] Phillip, I. et al.. Image-to-image Translation with Conditional Adversarial Networks. Proc. the IEEE Conf. on Computer Vision and Pattern Recognition. 2017, p. 1125-1134.