発話音声の感情を反映したテロップ画像の自動生成手法の検討*

☆中村史也 (神戸大), 相原龍 (三菱電機), 高島遼一, 滝口哲也 (神戸大), △今井良枝 (三菱電機)

1 はじめに

近年，深層学習技術の発展に伴い，ニューラルネットワークによる音声認識 (automatic speech recognition; ASR) の精度が飛躍的に向上している。その結果，スマートフォンの音声アシスタント機能や Web ブラウザの音声検索機能，動画サイトの自動字幕生成機能など，音声認識技術は現在幅広い用途で実用化されている。しかし，例えば皮肉表現や多義語などが発話に含まれる場合，音声を聞かずに音声認識結果のテキストを見るだけでは意味を理解することは困難であると予想される。

一方で，人間は発話内容を理解する際に，言語情報以外にも様々な情報を統合的に利用している。特に，話者の表情や声の調子などから推定される感情情報は，人間のコミュニケーションを円滑にする上で重要な役割を果たしていることが知られている [1, 8]。したがって，音声から発話内容を理解するには，ASR によって得られる言語情報に加えて，音声感情認識 (speech emotion recognition; SER) によって得られる感情情報を利用することが有効であると考えられる。

テレビや YouTube などの動画コンテンツでは，様々なフォントの字幕（テロップ）を用いることで，その場の状況や出演者の感情を表現することがしばしばある。このように，ASR 結果の文字列をもとに，SER で推定した話者の感情にしたがってフォントが変化するようなテロップ画像を生成できれば，発話内容に加えて感情情報も視覚化できるため，例えば難聴者や聴覚障害者の支援システムなどへの応用が期待できる。

ASR と顔画像の感情認識を用いた字幕生成システムとして，NHK テクノロジーズと DNP 社が共同開発した「感情表現字幕システム」[2] がある。従来システムでは，話者の表情を画像認識によってあらかじめ定めた感情クラスに分類し，感情クラスごとに定義したフォントを用いてテロップを生成する，いわばルールベースの手法が用いられている。これに対して本研究は，複数の感情が入り混じったような複雑な感情をフォントによって表現することを目的とする。

画像処理分野では，顔画像から推定した感情を反映したフォント文字画像を GAN によって生成する手法が提案されている [3]。そこで本研究では GAN を用いて発話音声から発話内容と感情を反映したテロップ画像の生成を検討する。

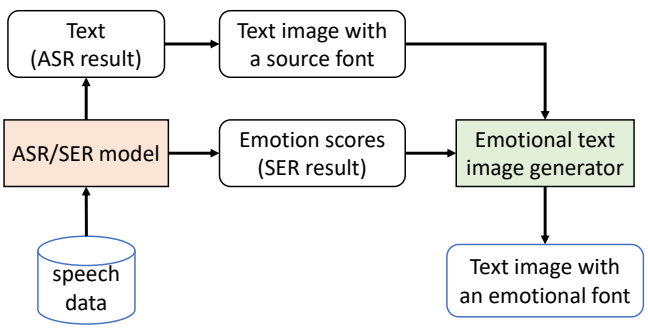


Fig. 1: Procedure for generating subtitle images from speech data

2 感情を反映したテロップの生成手法

2.1 手法の概要

提案するテロップ画像生成手法の流れを Fig. 1 に示す。まず，ASR と SER によって発話音声から発話内容および感情スコアの推定を行う。次に，発話内容のテキストをもとに，ソースとなるフォントで字幕画像を生成する。その後，ソースフォントの字幕画像と感情スコアを事前に学習した感情テロップ画像生成器へ入力し，感情スコアを反映したフォントのテロップ画像を生成する。

2.2 ASR と SER のマルチタスク学習

近年，音声認識の研究では wav2vec 2.0 [4] や HuBERT [5] など，自己教師あり学習モデルが顕著な成果を示している。さらに，これらの自己教師あり学習モデルによる特徴量抽出は，音声感情認識タスクにおいても有効に働くことが示されている [6, 7]。そこで本研究では，字幕文字列および感情スコアを推定するため，wav2vec 2.0 を用いた ASR と SER のマルチタスク学習モデル [8] を使用した。

マルチタスク学習における全体の損失 L_{MTL} は，ASR による CTC 損失 L_{CTC} と SER によるクロスエ

*Automatic generation of subtitle text images reflecting emotions in speech utterances. by Fumiya Nakamura (Kobe University), Ryo Aihara (Mitsubishi Electric Corporation), Ryoichi Takashima, Tetsuya Takiguchi (Kobe University), Yoshie Imai (Mitsubishi Electric Corporation)

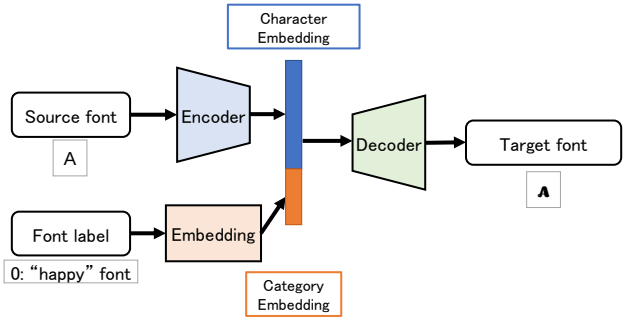


Fig. 2: Model structure of zi2zi generator

ントロピー損失 L_{CE} を用いて以下のように表せる。

$$L_{MTL} = \alpha L_{CTC} + L_{CE} \tag{1}$$

α はマルチタスク学習における CTC 損失の重みで，本研究では $\alpha = 0.1$ としている。

2.3 zi2zi を用いた文字画像生成

テロップ画像の生成モデルには，zi2zi [9] を使用する。zi2zi は，GAN に基づく Image-to-Image モデルの一種である pix2pix [10] を拡張し，フォント文字画像の生成に適用したモデルである。zi2zi の生成器 (generator) 部分を Fig. 2 に示す。pix2pix とは異なり，1 つのソースフォントから複数のターゲットフォントを生成する一対多タスクを取り扱うため，pix2pix の U-Net 構造 [11] の Encoder-Decoder モデルに加えて，フォントラベルを Embedding 層に通すことで得られるカテゴリ埋め込みを Encoder 出力と結合して Decoder に入力している。ここで，生成器において，Encoder がフォントの異なる同じ文字を近いベクトルにマッピングするように学習するため，コンスタント損失 [13] L_{const} が追加されている。さらに，zi2zi の識別器 (discriminator) では，画像がどのフォントであるかクラス分類を行うための全結合層が追加され，カテゴリ損失 [12] $L_{category}$ が導入されている。これは，学習に使用したどのフォントとも異なる画像が生成されるのを防ぐためである。

本研究では，SER タスクでよく使用される「怒り」「悲しみ」「喜び」「平静」の 4 感情にフォントを対応づけているが，後述のフォント補間の精度を上げるため，この 4 フォントを含む 21 種類のターゲットフォントを用意し，多様なフォントを生成できるように zi2zi の学習を行っている。

2.4 zi2zi によるフォントの補間

zi2zi では，入力文字画像に対する Encoder の出力は文字埋め込みに，Embedding 層の出力はフォントのカテゴリ埋め込みにそれぞれ相当する。そこでオリジナルの zi2zi モデル [9] では，異なる 2 つのフォントのカテゴリ埋め込みを線形補間することで，2 フォント間の補間 (interpolation) を行っている。これに対して

Table 1: List of emotional fonts selected by the survey

Emotion label	Font
happy	Birdmathon
angry	Reggae One
sad	しょかきさらり（行体）
neutral	源真ゴシック Light

本研究では，「怒り」「悲しみ」「喜び」「平静」の 4 感情フォントの補間によりテロップ画像を生成する。4 感情のフォントのカテゴリ埋め込みを z_i $(i = 1, 2, 3, 4)$，SER モデルによる感情スコアを w_i としたとき，生成時のカテゴリ埋め込み $\hat{z}$ はそれらの線形結合によって作成する。

$$\hat{z} = \sum_{i=1}^{4} w_i z_i \tag{2}$$

$$\sum_{i=1}^{4} w_i = 1 \tag{3}$$

3 評価実験

3.1 フォントの選定

感情を反映したテロップ画像を生成する zi2zi モデルの学習に当たり，SER の各感情ラベルに対応するフォントを決定する必要がある。視覚情報からどのような感情を表しているのかを判別できるテロップを生成するには，感情ラベルに対してフォントを適切に設定することが重要である。そこで，テロップに使用する感情フォントを選定するため，主観評価によるアンケートを行った。アンケートでは，各感情について 4 種類〜6 種類の感情フォントの候補から，どのフォントが最もその感情の字幕フォントとして適していると感じるかを回答者が選択する。そして，各候補の中から最も得票数の多いフォントを感情フォントとして採用した。14 名のアンケート回答者によって選ばれた感情フォントを Table 1 に示す。

3.2 ASR および SER の実験設定

ASR および SER の学習は，文献 [8] の設定に従って行った。データセットには IEMOCAP [14] を使用した。IEMOCAP は話者 10 人による約 12 時間の英語の対話音声で構成されており，各発話には感情ラベルが付与されている。学習には，happy, angry, neutral, sad, excited の 5 感情のラベルが付与されている計 5,531 発話を使用した。ただし，happy と excited は happy の 1 感情として扱い，計 4 感情のクラス分類を行う。

モデルは，wav2vec2.0 base モデルの最終層に ASR 用の全結合層と CTC，SER 用の pooling 層および全

Table 2: ASR and SER of multi-task learning model

method	ASR WER [%]	SER ACC [%]
[8]	19.29	78.15
ours	19.74	78.03

Table 3: MSE and MAE between generated image and ground truth

Emotion	happy	angry	sad	neutral
MSE	426.38	332.93	429.29	184.43
MAE	2.37	2.18	2.49	1.45

結合層を追加した構造である [8]。wav2vec2.0 base モデルは 7 個の CNN ブロックと 12 個の Transformer ブロックで構成されている。本実験では，LibriSpeech（約 960 時間）による事前学習モデル [15] を用いて fine-tuning を行った。ASR モデルの出力層は，アルファベット 26 文字に句読点などの記号を加えた 32 文字をトークンとして定義し，SER モデルの出力層は，4 感情を出力クラスとして定義した。オプティマイザには AdamW を使用した。ASR および SER において 10-folds で交差検証を行い，Table 2 に示すように文献 [8] と同様の結果が得られている。

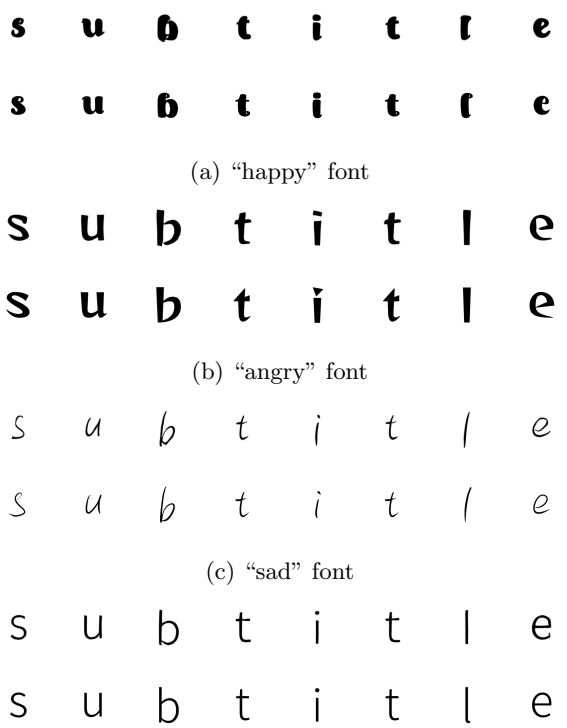
3.3 フォント生成の実験設定

本稿では，IEMOCAP の ASR/SER 結果をテロップ化するため，zi2zi を用いてアルファベット文字の生成を行う。zi2zi の学習では，ソースフォントには Arial を，ターゲットフォントには Table 1 で示した 4 感情のフォントを含む計 21 種類のフォントを設定した。学習には，アルファベット大文字小文字の文字画像計 50 枚を使用した。ただし，小文字の a と g ではフォントごとに字体が異なるため，学習データから除外した。文字画像の大きさは 256×256 で，3 チャネルのカラー画像として生成を行った。データ拡張のため，学習時に一定の範囲内でランダムに文字の回転処理を施した。zi2zi の L1 損失の重みは 100，コンスタント損失 L_{const} の重みは 15 に設定した。オプティマイザは Adam を使用し，2,000 エポック学習時点のモデルでフォント文字画像生成の評価を行った。

3.4 実験結果

zi2zi による感情フォント 4 種の生成結果を Fig. 3 および Table 3 に示す。Table 3 では，生成画像の客観評価として，平均二乗誤差 (Mean Squared Error; MSE) および平均絶対誤差 (Mean Absolute Error; MAE) を示している。Fig. 3 および Table 3 から，zi2zi モデルが各感情フォントの特徴を学習できていることが確認できる。

次に，IEMOCAP の音声データを提案手法でテロップ画像化した結果を Fig. 4 に示す。Fig. 4 (a) のように推定された感情スコアが one-hot に近い場合，Fig. 3 と同様に設定した感情フォントの文字画像を生成した。一方，Fig. 4 (b) のように感情スコアが曖昧な場合，スコアの高い「平静」のフォントと類似するが Fig. 3 (d) と比べると類似性が低い文字が生成された。



(a) “happy” font

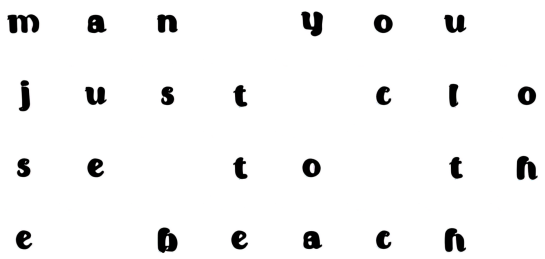
(b) “angry” font

(c) “sad” font

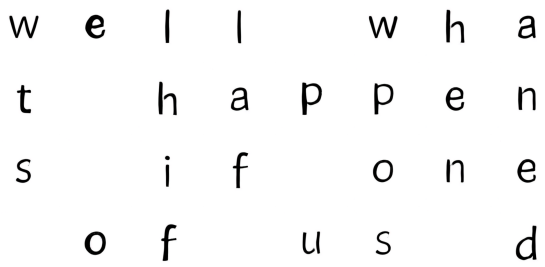
(d) “neutral” font

Fig. 3: Samples of generated fonts (top) and ground truths (bottom).

さらに，フォントの補間に伴う画像の変化の様子を Fig. 5 に示す。ここでは,「怒り」フォントと「悲しみ」フォントの補間を行い，生成された画像と実際のフォント画像との MSE の推移を計測した。Fig. 5 から，カテゴリ埋め込みにおける「悲しみ」フォントの重み係数が大きくなるにつれて,「怒り」フォントとの MSE は大きく,「悲しみ」フォントとの MSE は小さくなり,「怒り」フォントから「悲しみ」フォントへと画像が変化していることが確認できる。しかし，MSE の変化は「悲しみ」フォントの重み係数が 0.3 から 0.7 付近のときに急激に起こっており，重み係数が 0 や 1 に近い場合，生成されるフォントにはほとんど影響していないことが判明した。この傾向は他の感情間の補間でも同様に見られた。以上の結果から，本手法はスコアの最も高い感情に対応するフォントと類似性の高いフォントを生成しているが，Fig. 5 における重み係数 0.3 未満の場合のように，スコアが低い感情をフォントに反映する能力が乏しく，複雑な感情をテロップとして表現するには，より感情



(a) Label: happy, emotion score by SER: [hap 0.997, ang 0.000, sad 0.000, neu 0.003]



(b) Label: neutral, emotion score by SER: [hap 0.379, ang 0.000, sad 0.001, neu 0.619]

Fig. 4: Samples of generated subtitles for IEMOCAP speech data

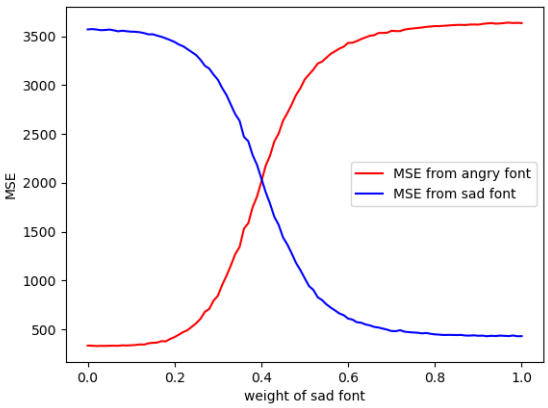


Fig. 5: MSE during interpolation

スコアに沿った視覚表現が必要であると考えられる。

4 おわりに

本研究では，発話に含まれる言語情報と感情情報を視覚的に表現するため，ASR と SER による認識結果を GAN を用いてフォント付きの文字画像として生成する手法を検討した。感情に合わせた文字画像の生成実験を行い，複雑な感情に対してどのようなフォントが生成されるのか調査した。さらに，実際の感情ラベルつき音声データセットに対して，提案手法を用いてテロップ生成を行った。

しかし，フォントの補間によって複雑な感情を表現することは難しいことが明らかになった。したがって，感情に合わせて文字の色も変化するようにするなど，感情の表現方法をさらに検討していくことが今後の課題である。

参考文献

[1] R. Cowie *et al.*, “Emotion recognition in human-computer interaction,” IEEE Signal processing magazine, 18 (1), 32–80, 2001.

[2] https://www.dnp.co.jp/news/detail/10158470_1587.html

[3] 中村 他, “画像の感性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, 17 (5), 523-579, 2017.

[4] A. Baevski *et al.*, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” arXiv:2006.11477, 2020.

[5] W.-N. Hsu *et al.*, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units,” arXiv:2106.07447, 2021.

[6] Y. Wang *et al.*, “A Fine-tuned Wav2vec 2.0/HuBERT Benchmark For Speech Emotion Recognition, Speaker Verification and Spoken Language Understanding,” arXiv:2111.02735, 2021.

[7] L. Pepino *et al.*, “Emotion Recognition from Speech Using wav2vec 2.0 Embeddings,” Proc. Interspeech 2021, 3400-3404, 2021.

[8] X. Cai *et al.*, “Speech Emotion Recognition with Multi-Task Learning,” Proc. Interspeech 2021, 4508-4512, 2021.

[9] https://kaonashi-tyc.github.io/2017/04/06/zi2zi.html

[10] P. Isola *et al.*, “Image-to-Image Translation with Conditional Adversarial Networks,” arXiv:1611.07004, 2017.

[11] O. Ronneberger *et al.*, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” arXiv:1505.04597, 2015.

[12] A. Odena *et al.*, “Conditional Image Synthesis With Auxiliary Classifier GANs,” arXiv:1610.09585, 2016.

[13] Y. Taigman *et al.*, “Unsupervised Cross-Domain Image Generation,” arXiv:1611.02200, 2016.

[14] C. Busso *et al.*, “IEMOCAP: Interactive emotional dyadic motion capture database,” Journal of Language Resources and Evaluation, 42 (4), 335-359, 2008.

[15] https://huggingface.co/facebook/wav2vec2-base