単語埋め込み表現を用いた感情音声からの字幕画像生成手法の検討*

☆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), △伊谷祐介 (三菱電機)

1 はじめに

自動音声認識 (automatic speech recognition; ASR) による自動字幕は，文字という視覚情報で発話内容の理解を補助することができる技術として期待されている。一方で，人間はコミュニケーションにおいて言語情報以外にも様々な情報を統合的に利用しており，中でも話者の表情や声の調子などから推定される感情情報が重要な役割を果たしていることが知られている [1]。例えば皮肉表現や多義語などが発話に含まれる場合には，同じ内容の台詞でも話者がどのような感情で発話しているかによって聞き手の受ける印象は全く違うものとなりうる。したがって，発話内容の理解を補助する字幕生成には，ASR によ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字幕に反映することが有効であると考えられる。

感情情報を字幕で表現する手段の一つとして，文字のフォントが挙げられる。例えば，テレビや YouTube などの動画コンテンツでは，その場の状況や出演者の感情は様々なフォントや色彩の字幕（テロップ）としてしばしば表現される。また，ASR と顔画像の感情認識を用いたフォント字幕生成システムとして,「感情表現字幕システム」（NHK テクノロジーズ，DNP 社）[4] 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感情クラスに分類し，感情クラスごとにあらかじめ定義したフォントを用いてテロップを生成する。

しかし，人間は一つの発話音声から複数の感情を知覚することがある [2, 3] ことから，感情情報の視覚化という観点からは複数の感情を反映した字幕を生成できることが望ましい．そこで，我々の先行研究 [5] では，話者の感情（各感情の確率）に基づき，画像変換モデルを用いて複数の感情フォント間の補間を行うことで複雑な感情を反映したフォントを生成する手法を提案した。しかし，フォントを画像として補間する手法では，複雑な感情を表現することは困難であることが明らかになった。

そこで本研究では，フォントの投稿・検索サイトから大量のフォントとそれに付与されたタグのデータを収集し，単語埋め込み表現を用いて発話の言語情報と感情情報に合った字幕のフォントを決定する手法を提案する。

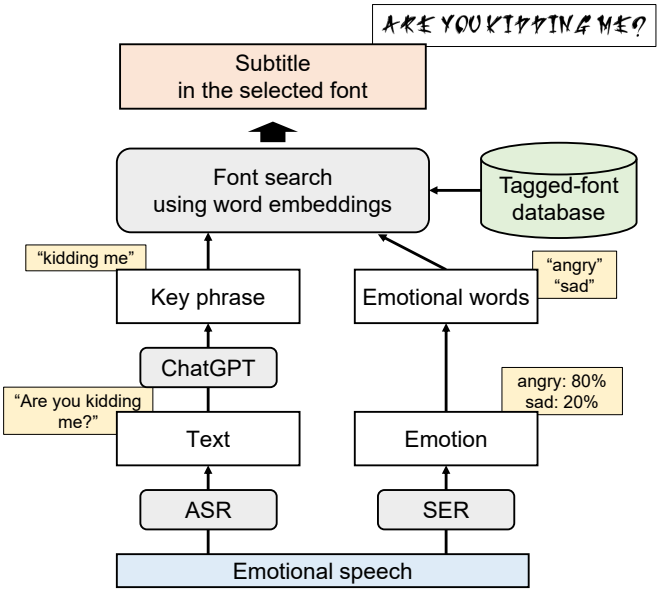


Fig. 1: Procedure for subtitle generation from emotional speech

2 単語埋め込みを用いた感情音声からの字幕生成手法

2.1 手法の概要

提案するフォント字幕の生成手法の流れを Fig. 1 に示す。まず，ASR と SER によって発話音声から発話内容のテキストおよび感情の推定を行う。次に，ChatGPT [15] を用いてテキストから発話において重要と考えられる単語またはフレーズを抽出する。そして，重要フレーズと感情語の単語埋め込みを用いて各フォントとの類似度スコアを計算し，類似度スコアの最も高いフォントで ASR 結果のテキストを字幕化する。ChatGPT による重要フレーズ抽出については 2.2 節で，単語埋め込みを用いたフォントの検索については 2.3 節でそれぞれ詳細を述べる。

2.2 ChatGPT を用いた発話文の重要フレーズ抽出

2.3 節で述べる単語埋め込み表現の獲得の前処理として，ASR によって得られるテキストから ChatGPT を用いて重要フレーズを抽出することで，字幕フォントの検索に不要な情報の除去を行う。

ChatGPT は OpenAI 社による AI チャットボットで，モデルの基本構造は大規模言語モデル GPT-3 [16] の後続である GPT-3.5 系列である。人間のフィードバックを用いた強化学習 [17] によってモデルを更新

*Subtitle generation from emotional speech using word embeddings. by Fumiya Nakamura (Kobe University), Ryo Aihara (Mitsubishi Electric Corporation), Ryoichi Takashima, Tetsuya Takiguchi (Kobe University), Yusuke Itani (Mitsubishi Electric Corporation)

Table 1: Prompt entered into chatGPT. [xxx] is the input sentence that is the target of keyphrase extraction.

Prompt
You will be provided with a line from the conversation, and you are asked to extract one key word or key phrase that is important to the conversation. The following are examples. input: Since fish was cheap today, I decided to make a fish dish. output: fish dish input: Do you think I don't know such a simple thing ...? You should stop making fun of me. output: stop making fun of me input: Thank you for coming this far. output: thank you input: [xxx]

し続ける学習方式を採用しており，ユーザの入力プロンプトに応じて文章要約や文章生成，翻訳，コーディングなどの様々なタスクに柔軟に対応できることが知られている。

この ChatGPT の高い自然言語処理能力を活用するべく，本研究では Table 1 に示す few-shot 形式のプロンプトを与えて発話テキストから重要フレーズの抽出を行った。プロンプトの作成では，OpenAI によるキーワード抽出プロンプトの例 [1]を参考にした。

2.3 単語埋め込みによる字幕フォントの検索

Kulahcioglu ら [6] は，感情語とフォントの属性情報の単語埋め込み表現の距離を用いて単語からフォントを検索するシステムを提案した。これを踏まえて，本稿では感情語に加えて発話内容の重要フレーズも用いてフォントを検索する手法を検討する。

2.3.1 フォント・タグデータの収集

本研究では，重要フレーズと感情語からフォントの検索を行うため，英字フォントの投稿・検索サイトである 1001 Fonts[2]から 10,000 個のフォントおよびユーザによってそれらのフォントに付与されたタグのデータを収集した。収集したデータのうち，以下の条件に当てはまるものは字幕フォントの検索を妨げると考えられるため除外し，結果として 9,221 個のフォントと 1,983 種類のタグで構成されるフォント・タグデータを得た。

- 字幕に不適切であると考えられるタグ (“hard to read”, “dingbat”など) の付与されたフォント
- “serif”, “caps only”など，フォントの字体情報に関するタグ
- 単語埋め込みモデルの学習に含まれないタグ
- （上記の処理を行った状態で）付与されたタグの個数が 3 個未満のフォント

[1]https://platform.openai.com/examples/default-keywords

[2]https://www.1001fonts.com/

2.3.2 単語埋め込み

近年，word2vec [7] や GloVe [8], FastText [9] など，単語を埋め込みベクトル化するさまざまな手法が提案されてきた。しかし，これらの手法は「単語の意味は周囲の単語で形成される」とする分布仮説に基づいているため，似た文脈で使用されることの多い類義語と対義語を区別することができず，対義語同士の分散表現の類似度が高くなるという課題を抱えている。この課題に対して，retrofitting [10] や counterfitting [11] など，類義語や対義語の対で単語ベクトルを fine-tuning する手法が有効であることが示されている。本研究では,「喜び」と「悲しみ」など，別の感情を表す単語同士での類似度計算を行うため，Khosla ら [12] が単語の感情情報データ [13] を用いて counterfitting を行った word2vec[3] (Aff2vec) を利用した。

2.3.3 単語埋め込みによるフォントのスコア計算

感情語 w_e および ChatGPT で抽出した発話の重要フレーズ w_s から，2.3.1 項で収集したタグ情報を用いて各フォントのスコア付けを行う。フォントに付与されたそれぞれのタグ τ と，w_e および w_s の単語埋め込みを用いて，τ に関する類似度スコア s_τ を式 (1) で定義した。

$$s_\tau = \sum_{w_e \in E} p(w_e) sim(v(w_e), v(\tau)) \times sim(v(w_s), v(\tau)) \tag{1}$$

ここで，$sim(x, y)$ はベクトル x と y のコサイン類似度で，式 (2) で示される。

$$sim(x, y) = \frac{x \cdot y}{\|x\| \|y\|} \tag{2}$$

さらに，$p(w_e)$ は SER によって推定される，感情語 w_e に対応する感情の確率であり，E は感情語の集合，$v(a)$ は a の単語埋め込みベクトルである。ただし，a が複数の単語を含む場合，含まれるすべての単語の埋め込み表現の平均を $v(a)$ とする。また，w_s が単語埋め込みモデルにとって未知の単語である場合など，$v(w_s)$ が得られない場合には，

$$s_\tau = \sum_{w_e \in E} p(w_e) sim(v(w_e), v(\tau)) \tag{3}$$

[3]https://bit.ly/2HGohsO

Table 2: Evaluation results of font search by emotional words (top 5 fonts)

Emotion	Accuracy [%]	σ_{acc} [%]
happy	75.6	24.1
angry	75.6	29.8
sad	33.3	36.0
neutral	55.6	30.4
total	60.0	33.1

Fig. 2: Top 2 fonts in search results for “happy”

として計算する。

次に，各フォントには複数個のタグが付与されているため，それぞれのタグに対して式 (1) あるいは式 (3) でスコア付けを行い，フォント f に付与されたタグのスコアを高いものから順に $s_{\tau_1}, s_{\tau_2}, s_{\tau_3}, \ldots$ とする。これらを用いて，フォント f の類似度スコア S_f を式 (4) で定義した。

$$S_f = \sum_{k=1}^{3} \frac{s_{\tau_k}}{k} \tag{4}$$

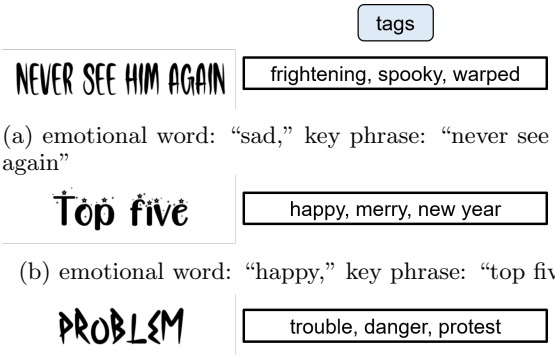
そして，すべてのフォントに対してこれらの計算を行い，最も類似度スコアの高いフォントを字幕フォントとして選択する。

3 評価実験

3.1 感情語によるフォント検索

まず，単語埋め込みを用いたフォント検索の効果を検証するため，感情語のみで類似度スコアの計算を行い，スコア上位のフォントの主観評価実験を行った。実験では，感情語として喜び (happy)，怒り (angry)，悲しみ (sad)，平静 (neutral) の 4 つを対象にフォントの検索を行い，それぞれ類似度スコア上位 5 個，合計 20 フォントのアルファベット文字画像を作成した。評価者はそれらのフォント画像を見て，4 感情のうちどの感情に近い印象を受けるかを回答する。検索に使用した感情語を正解ラベルとして，主観評価結果の感情語ごとの正解率 (Accuracy) およびフォントごとの正解率の標準偏差 (σ_{acc}) を求めた。

日本語母語話者 9 名の評価者による実験結果を Table 2 に示す。感情語ごとの正解率を見ると,「喜び」や「怒り」の場合は比較的高いのに対し,「悲しみ」は正解率が約 33.3%と明らかに低い。これは，収集したデータに含まれる感情語の個数に偏りがあることが原因であると考えられる。全 9,221 個のフォントのうち，“happy”のタグは 139 個に付与されていたのに対し，“angry”や “sad”, “neutral”はそれぞれ 3 個，1 個，0 個であった。“angry”の検索では，“furious”



(a) emotional word: “sad,” key phrase: “never see him again”

(b) emotional word: “happy,” key phrase: “top five”

(c) emotional word: “angry,” key phrase: “problem”

Fig. 3: Samples of selected fonts (with tags) from emotional words and key phrases.

のような類義語のタグが付与されたフォントが得られたが，“sad”の検索では，“bad”や “hateful”などのタグが付与されたフォントが上位であった。このことから,「喜び」や「怒り」に比べて,「悲しみ」を表現することができるフォントが希少である可能性が示唆される。また，フォントごとの正解率の標準偏差が 24.1%〜36.0%と大きいことから推測できるように，同じ感情語の検索結果でも正解率に大きな差が見られた。「喜び」の検索結果上位 2 フォントの例を Fig. 2 に示す。図より，どちらのフォントも “happy” と近い意味のタグが 3 つ付与されており，単語埋め込みの類似度による検索は正しく機能していると考えられる。しかし，上段に示す 1 位のフォントでは評価者による正解率は 44.4%であり，これらのタグはフォントを見た際の印象と必ずしも一致しないことが判明した。

3.2 感情語と重要フレーズによるフォント検索

次に，感情語と重要フレーズを用いたフォントの検索に関する主観評価実験を行った。ASR や SER による認識誤りの影響を無視するため，英語の対話形式の感情音声データセット IEMOCAP [14] から，正解感情ラベルと文字起こしの組をそれぞれ SER 結果，ASR 結果として想定した。感情語と ChatGPT で抽出した重要フレーズを用いて各フォントの類似度スコアの計算を行い，最もスコアが高いフォント画像を正例，スコアが下位 90%に含まれるフォントからランダムに選択したフォントを負例とした。3.1 節で述べた 4 つの感情語を対象に，それぞれ 20 発話分の正例画像と負例画像を作成した。正例画像の例を Fig. 3 に示す。実験において，評価者は感情語と重要フレーズの組から，正例画像と負例画像のうちどちらがその発話の字幕に適しているかを選択する。

9 名の評価者による正例画像の選択率を正解率として求め，Table 3 の結果が得られた。3.1 節の結果と同様に,「怒り」の正解率が高いのに対し,「悲しみ」の正解率は 50%未満と低く，ランダムに選択した負例の方が多く選択されている。また，評価実験の設問群を感情語と重要フレーズの埋め込み表現の類似度

Table 3: Evaluation results of font search by emotional words and key phrases

Emotion label	Accuracy [%]
happy	61.1
angry	86.7
sad	40.0
neutral	65.0
total	63.2

Table 4: Accuracy changes with similarity of emotional words and key phrases

Similarity	Accuracy [%]
low	58.3
moderate	60.0
high	61.1
very high	73.3

によって低い (low)，普通 (moderate)，高い (high)，非常に高い (very high) の 4 つにグループ分けし，グループごとの正解率を算出した結果を Table 4 に示す。感情語と重要フレーズの類似度が「低い」「普通」「高い」の 3 グループでは正解率は約 60%で大きな差は見られないが,「非常に高い」グループでは 70%を超える結果となった。したがって，提案手法は話者の感情と発話内容が一致する場合には適したフォントを検索する能力を有しているが，1 章で述べた皮肉表現のような感情音声の場合には，発話内容の類似度をあまり考慮しないようにするなどの工夫が必要であると考えられる。

4 おわりに

本研究では，発話内容の文字起こしと話者の感情を表す単語から，単語埋め込み表現を用いて発話の言語情報と感情情報を反映したフォントで字幕を生成する手法を提案した。感情音声データセットの文字起こしテキストと感情ラベルを用いて主観評価実験を行い，提案手法は「怒り」の感情音声の場合や話者の感情を表す単語と発話内容の類似度が高い場合には比較的高い性能を示すが,「悲しみ」を表現するフォントの検索に課題があることを明らかにした。今後の課題としては，ASR による認識誤りが重要フレーズの抽出や単語埋め込みの獲得に与える影響を考慮する必要がある点や，提案手法によって複数の感情を表現できるかどうか未検証である点などが挙げられる。

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] 赤木, “音声に含まれる感情情報の認識：感情空間をどのように表現するか,” 音響学会誌, 66 (8), 393-398, 2010.

[3] R. Plutchik, “The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice,” American scientist, 89 (4), 344-350, 2001.

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

[5] 中村ら, “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,” 音講論（春）, 887-890, 2023.

[6] T. Kulahcioglu and G. D. Melo, “Fontlex: A typographical lexicon based on affective associations.” Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[7] T. Mikolov *et al.*, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.

[8] J. Pennington *et al.*, “Glove: Global vectors for word representation,” Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532-1543, 2014.

[9] P. Bojanowski *et al.*, “Enriching word vectors with subword information,” Transactions of the association for computational linguistics 5, 135-146, 2017.

[10] M. Faruqui *et al.*, “Retrofitting word vectors to semantic lexicons,” arXiv preprint arXiv:1411.4166, 2014.

[11] N. Mrkšić *et al.*, “Counter-fitting word vectors to linguistic constraints,” Proceedings of NAACL-HLT, 2016.

[12] S. Khosla *et al.*, “Aff2Vec: Affect–Enriched Distributional Word Representations,” arXiv preprint arXiv:1805.07966, 2018.

[13] A. B. Warriner *et al.*, “Norms of valence, arousal, and dominance for 13,915 English lemmas,” Behavior research methods, 45, 1191-1207, 2013.

[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://chat.openai.com/

[16] T. Brown *et al.*, “Language models are few-shot learners,” Advances in neural information processing systems, 33, 1877-1901, 2020.

[17] L. Ouyang *et al.*, “Training language models to follow instructions with human feedback,” Advances in Neural Information Processing Systems, 35, 27730-27744, 2022.