CF-013

動画メディアコンテンツを対象とした一覧性に着目した音声感情抽出による重要セリフ可視化手法

A Method for Visualizing Important Lines of Dialogue Using Overview-focused Audio Emotion Extraction for Video Media Content

吉井　茜音†　岡田　龍太郎†　峰松　彩子†　中西　崇文†
Akane Yoshii†, Ryotaro Okada†, Ayako Minematsu†, Takafumi Nakanishi†

1. はじめに

近年，YouTube を始めとした動画メディアコンテンツを扱う Web サイトの普及により，我々ユーザが動画メディアコンテンツを楽しむ機会が増大している．それに対し，膨大な量の動画メディアコンテンツの中から視聴者が興味のあるものを効率的に見つけることは難しくなっている．そのため，視聴したい動画の内容を理解し，重要なポイントを把握することが重要となってきている．一般的に，動画メディアコンテンツは時系列の次元を持ったメディアであり，静止画像メディアコンテンツに比べ，ユーザがその動画メディアコンテンツがどのような内容のコンテンツであるかを把握するための一覧性に欠けるという問題がある．しかしながら，現状では動画の探索は静止画像であるサムネイルに依存しており，一覧性が欠落している．この一覧性の欠落は，ユーザが自身の好む動画メディアコンテンツを探索するのに手間を発生させている．つまり，動画メディアコンテンツにおいて，どのような内容であるかを一覧できる可視化手法を実現することにより，ユーザが興味を持つ動画メディアコンテンツを見つける一助になると考える．

本稿では，動画メディアコンテンツを対象とした音声感情抽出による重要セリフ抽出手法について示す．我々はドラマを中心とする動画メディアコンテンツにおいて，感情の起伏が激しいセリフは，その動画メディアコンテンツ内で重要なセリフとなっていることが多いと仮説を立てている．その仮説が正しいなら，動画メディアコンテンツにおける重要セリフの感情の動きを可視化することが可能になれば，動画メディアコンテンツのそのセリフごとの音声感情抽出を行うことによって，重要セリフを抽出することが可能となり，その可視化することによって，ドラマのクライマックスを捉えた新たな要約表現が可能となる．本稿ではヒートマップを用いて，動画メディアコンテンツを対象としたセリフの重要度可視化手法を実現する．本手法により，動画メディアコンテンツの内容の一覧性に優れた可視化が実現される．本手法により，ユーザがより興味に合致した動画像メディアコンテンツを楽しむ機会を増大させることが可能となる．

本稿では，上記を実現する実験システムを構築し，被験者実験を行うことで，本方式の有効性を示す．

2. 関連研究

R. Hirano ら[1]は，YouTube の動画の字幕データとその動画に付いてコメントの内容を分析し，視聴者の反応喚起因子となる重要単語を，抽出する手法を提案している．M. Xu ら[2]は，DVD/DivX ビデオの字幕を用いて映像を台詞で分割し，感情関連単語と音声イベントの検出を利用することで，視聴者の反応や感情を重視した効果的なビデオ解析を行い，感情的なコンテンツを抽出する手法を提案している．L. Chen ら[3]は，音声の音量，調子，そしてリズムを分析することで，会話の盛り上がりやアクションが激しいシーンを効果的に特定する手法を提案している．A. Alatan ら[4]は，音声情報と視覚情報を用いた隠れマルコフモデル (HMM)の状態遷移によってマルチメディアコンテンツ中の対話シーンにインデックスを付ける自動アルゴリズムを提案している．O. Fried ら[5]は，話者のセリフのフィラー語を削除するために，トランスクリプトを編集するだけでシームレスな音声と映像の流れを維持する新しい方法を提案している．B. Schuller ら[6]は，音声チャネルに依存せず，音響特徴と SVM-SFFS ラッパーベースの探索，および ASR 出力のベイジアンネット分析を組み合わせた感情認識システムを提案している．

これらの研究では，いずれも機械学習や統計手法を用いて映像や音声を解析し，視聴者の反応や感情を喚起する重要なコンテンツやシーンを自動的に抽出することが可能となっている．しかしながら，これらの手法は動画全体の視聴者に与える感情を一覧して見せるものにはなっていない．本稿では，入力した動画に対して，感情の起伏が激しい部分のセリフは，その動画内で重要な場面であるという仮説のもと，機械学習を用いた音声感情抽出により，ヒートマップを用いて一覧として可視化することによって，ユーザが重要セリフを発見するのを補助する新たな要約表現を提案している．

3. 提案方式

3.1 システム全体像

本提案方式の全体像を図 1 に示す．本方式は，大きく分けて，動画メディアコンテンツ抽出機能，センテンス別音声抽出機能，音声感情抽出機能，音声感情可視化機能からなる．

ユーザは，動画メディアコンテンツの選択をし，その URL を入力する．入力された URL から動画メディアコンテンツを取得し，字幕情報からタイムスタンプ付きセンテンスを抽出する．抽出したセンテンス別に音声を分割し，音声感情の抽出を行う．音声感情の抽出には，感情ラベル付き動画メディアコンテンツデータベースをラベルごとに学習させた分類モデルを用いる．このプロセスにより，センテンス別の感情確率値を抽出することができるため，最終的にヒートマップ化することで，感情を一覧化した要約表現をユーザに提供する．

3.2 動画メディアコンテンツ選択機能

動画メディアコンテンツ選択機能は，実験対象とする動画を保存する機能である．ユーザは YouTube から動画メディアコンテンツを選択し，それに対応する URL を入力する．pytube ライブラリを使用して，選択した動画メディアコンテンツを mp4 形式で保存する．

3.3 センテンス別音声抽出機能

† 武蔵野大学データサイエンス学部データサイエンス学科　Department of Data Science, Musashino University

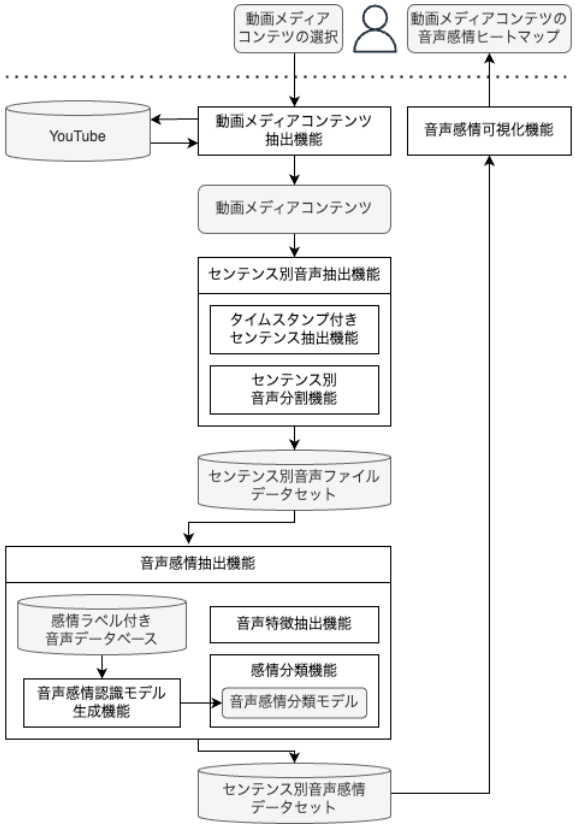

図 1：感情重要キーワード抽出方式の全体像

センテンス別音声抽出機能は，タイムスタンプ付きセンテンス抽出機能とセンテンス別音声分割機能からなる．それぞれ，動画メディアコンテンツから字幕を抽出し，その字幕に対応するタイムスタンプを取得する機能と，このタイムスタンプを用いて，動画メディアコンテンツを各センテンスに分割し，wav 形式の音声ファイルとして出力する機能で構成されている．

3.3.1 タイムスタンプ付きセンテンス抽出機能

タイムスタンプ付きセンテンス抽出機能は，動画メディアコンテンツから字幕を抽出し，その字幕に対応するタイムスタンプを同時に取得する機能である．ここでは mp4 形式の動画メディアコンテンツに対して，Whisper ライブラリの large モデルを適用し，音声認識を行う．出力として，動画メディアコンテンツの各センテンスに区切られた字幕と，そのセンテンスが話されたタイムスタンプを抽出する．

3.3.2 センテンス別音声分割機能

センテンス別音声分割機能では，mp4 形式である動画メディアコンテンツから各センテンスに分割された wav 形式を抽出する機能である．音声の抽出には ffmpeg を使用する．タイムスタンプ付きセンテンス抽出機能で抽出されたセンテンス，及び，そのセンテンスが話されたタイムスタンプを用いて，mp4 形式の音声付き動画を各時間単位に分割し，複数のセンテンス別 wav ファイルを出力する．このようにして生成された wav ファイルは，センテンス別音声ファイルデータセットとして扱う．

3.6 音声感情抽出機能

音声特徴抽出機能は，音声特徴機能，音声感情認識モデル生成機能，感情分類機能からなる．

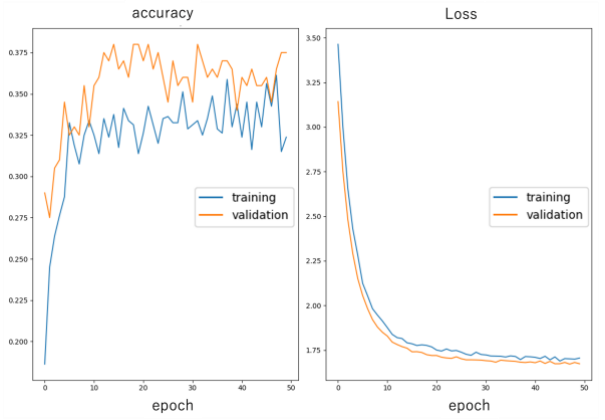

図 2：音声感情認識モデルの精度

音声特徴抽出機能では，センテンス別音声ファイルデータセットを用いて，音声データの特徴抽出を行う．音声感情認識モデル生成機能では，ラベル付けされている感情ラベル付き音声データベースの各感情の音声データを機械学習させて，音声感情分類モデルを生成する．最後に，感情分類機能では，感情ラベル付き音声データベースから，音声感情分類モデルを用いて，感情別に確率値を算出し，センテンス別音声感情データセットを出力する．

3.6.1 音声特徴抽出機能

音声特徴抽出機能は，センテンス別音声ファイルデータセットの各音声データに対して，音声の特徴抽出を行う機能である．Liblosa ライブラリを使用し，センテンス別音声ファイルデータセットの音声データから，MFCC(Mel-frequency cepstral coefficients)特徴量を抽出する．その後，抽出された MFCC の各係数について，時間軸に並んだ全データの平均値を計算する．その後，平均値の標準化を行い，スケーリングされた形式に変換する．これによって，各センテンスの音声データの特徴は 20 次元のベクトルとなり，機械学習モデルへの入力データとして利用可能な形式となる．

3.6.2 音声感情認識モデル生成機能

音声感情認識モデル生成機能では，感情ラベル付き音声データベースを用いて，音声感情認識モデルを生成する．このデータベースには，The RAVDESS データセット[7]を使用する．RAVDESS データセットは，24 人のプロの俳優(女性 12 人，男性 12 人)が，1 人につき 60 個の録音を行った英語の音声データである．これらは 8 つの感情(neutral, calm, happy, sad, angry, fearful, disgust, surprised)でラベル付けされた音声ファイルであり，全部で 1,440 個の音声ファイルで構成されている．本研究では，感情の分類を，表情の感情の分類に用いられる 7 つの感情(neutral, happy, sad, angry, fearful, disgust, surprised)[8]に合わせるため，calm のラベルが付いた音声ファイルは除外した．これによって残るのは 1,260 個のファイルとなった．このデータセットの音声は，すべて英語で話されているが，本研究で入力とする動画メディアコンテンツは日本語である．言語が違うにも関わらず英語のデータを用いているのは，英語と日本語の両方を話す中での感情表現の音声データには差がないという仮説に基づいている．

モデル生成に用いるデータとして，説明変数は音声特徴抽出機能で抽出した 20 次元ベクトルに対応した特徴とし，目的変数は感情のラベルをカテゴリ形式に変換したものとする．機械学習のモデルのアーキテクチャは，中間層 1

表 1：実行環境 – コンピュータスペック

項目	詳細
OS	macOS Sonoma 14.4.1
モデル	MacBook Air (M1, 2020)
チップ	Apple M1
メモリ	16GB

表 2：実行環境 – 主要 Python ライブラリ

ライブラリ名	バージョン
numpy	1.25.2
Tensorflow	2.15.0
Librosa	0.10.2
whisper	20231117
matplotlib	3.7.1
seaborn	0.13.1
ffmpeg	1.4
glob	10.4.1
pytube	15.0.0

つのニューラルネットワークとした．学習の結果を図 2 に示す．Loss が収束していることが図から読み取れる．ここでの精度は約 35%である．

3.6.3 感情分類機能

感情分類機能では，音声感情認識モデル生成機能で生成された音声感情分類モデルを用いて，音声データの感情分類を行う．ここで分類される 7 つの感情は，neutral，happy，sad，angry，fearful，disgust，surprised であり，それぞれの確率値が出力される．ここで，センテンス別音声ファイルデータセットを音声特徴抽出機能にかけて，音声データごとの特徴量を用いて，感情分類機能にかけることによって，センテンス別の音声感情データを抽出することができる．ここでは分類モデルを用いているが，確率値が最大のもののみを採用するのではなく，各感情の確率値をすべて保持する．このデータをセンテンス別音声感情データセットとして扱う．

3.7 音声感情可視化機能

音声感情可視化機能は，音声感情抽出機能で抽出された，センテンス別音声感情データセットを，ヒートマップとして可視化する機能である．縦軸を 7 つの感情ラベルとし，横軸を動画メディアコンテンツのセンテンスでヒートマップを生成する．このヒートマップは動画メディアコンテンツの音声感情を視覚的に表現し，ユーザへ提供する．

4. 評価実験

本節では，本方式の評価実験について述べる．3 節で示した方式を実装し， 本方式の有効性の検証を行うために 2 つ実験を行う．

本方式の有効性の検証にあたって，入力した動画音声を分析し，感情値が高い箇所を確認して，それが重要な箇所を抽出できていれば本方式が有効であると言えると考える．これを踏まえ，実験 1 では，最終的にユーザに提供される出力である，音声感情ヒートマップの有効性の検証を行う．検証は，アンケート調査を行う．実験 2 では，センテンスを可視化するにあたって，文字起こしが正確に出力されているかの精度についてアンケート調査を行った．

表 3：実行環境 – 入力する動画

動画タイトル	URL
テレビ朝日：『くるり～誰が私と恋をした？～』5/21(火)第 7 話 記憶をめぐる大喧嘩で恋の四角関係に大変化が…!?【TBS】	https://www.youtube.com/watch?v=eux_jCiDXFc
テレビ朝日：『9 ボーダー』5/17(金)第 5 話 大事なあの人と急接近！3 姉妹の恋の行方は…【TBS】	https://www.youtube.com/watch?v=810MbeW6SrQ?si
テレビ朝日：『アンチヒーロー』6/2(日)第 8 話 遂に明かされる，明墨の過去…【TBS】	https://www.youtube.com/watch?v=1ZYfEUsRcpM?si

表 4：実験 1 の結果

ドラマ名	とてもマッチしている	ややマッチしている	あまりマッチしていない	全然マッチしていない
くるり	13.3%	33.3%	53.3%	0%
9 ボーダー	6.7%	26.7%	60%	6.7%
アンチヒーロー	13.3%	40%	40%	6.7%

4.1 では実験環境について述べる．4.2 節では実験 1 の類似度計量機能の有効性の検証ついて述べる．4.3 節では実験 2 のシステム全体の有効性の検証ついて述べ，4.3 節では本研究による実験結果について考察を行う．

4.1 実験環境

システムの構築には Python 言語を用いる．Python バージョンは 3.10 を使用する．実行環境におけるコンピュータのスペックを表 1 に，主要 Python ライブラリを表 2 に示す．

また，本実験に用いるデータは，7 つの感情 (neutral, happy, sad, angry, fearful, disgust, surprised) にラベル付けされた 24 人の俳優の音声データである RAVDESS を使用し，感情ごとの確率値のそれぞれを抽出する．

入力に用いる動画を表 3 に示す．『くるり～誰が私と恋をした？～』は論文中では『くるり』と記述する．

4.2 実験 1

実験 1 では，『くるり』，『9 ボーダー』，『アンチヒーロー』の 3 つのドラマ予告動画を用いて，メディアコンテンツの音声感情ヒートマップの有効性の検証を行う．

実験結果として，出力された 3 つのヒートマップを図 3，4，5 に示し，アンケート結果を表 4 に示す．すべてのヒートマップにおいて，「disgust」である確率が高いと判別された．その次に，数値が高くなりやすい感情は「angry」であるとヒートマップを見て取れる．

有効性の検証を行うアンケート調査を行なった．アンケート内容としては，各動画を視聴した後にセンテンスごとの感情とマッチしているかの評価を，「とてもマッチしている」「ややマッチしている」「あまりマッチしていない」「全然マッチしていない」の 4 段階評価で調査を行った．実験結果をまとめて表 4 に示す．

結果として，すべてのドラマ予告において，「あまりマッチしていない」の確率が高いことがわかる．他の項目と比較しても，「ややマッチしている」の確率が次に高く，マッチしている部分もあるが，そうでもない部分もあると言える評価を受けた．

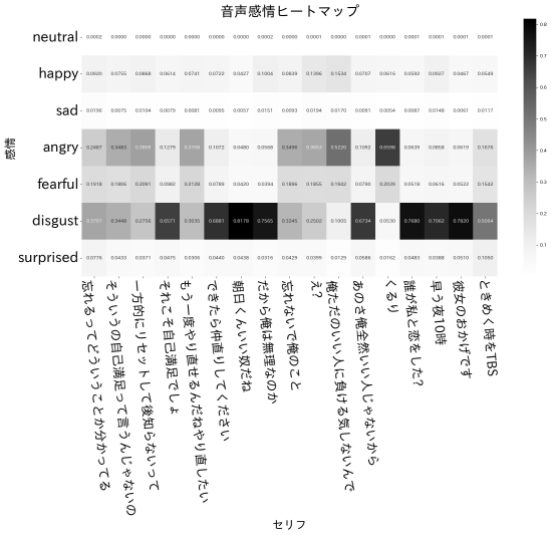

図 3:『くるり』音声感情ヒートマップ

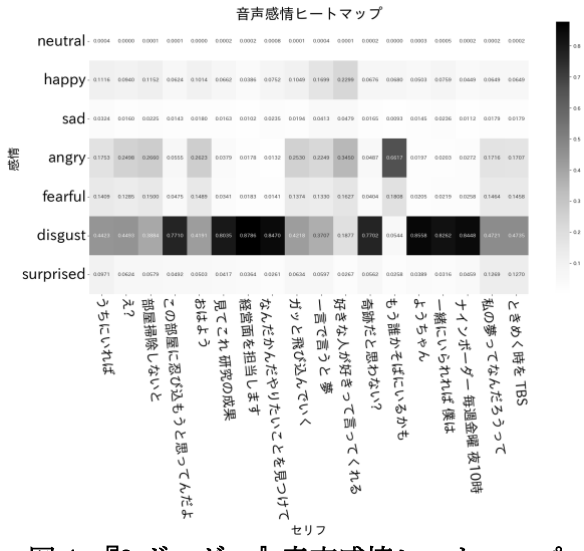

図 4:『9 ボーダー』音声感情ヒートマップ

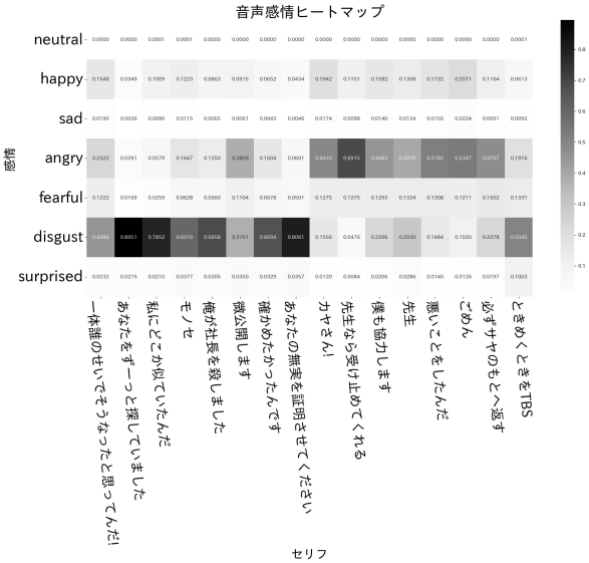

図 5:『アンチヒーロー』音声感情ヒートマップ

表 5：実験 2 の結果

ドラマ名	非常に良い	普通	悪い
くるり	73.3%	26.7%	0%
9 ボーダー	80%	20%	0%
アンチヒーロー	40%	60%	0%

4.3 実験 2

実験 2 では，実験 1 と同様に『くるり』,『9 ボーダー』,『アンチヒーロー』の 3 つのドラマ予告動画を用いて，動画メディアコンテンツのセンテンス別音声抽出機能で抽出した，センテンスを正確に抽出できているかの精度の検証を行う．ここでセンテンスの抽出が上手く出来ているとは，文字が正確に抽出されていることと，動画が適切に分割されていることを指している．

評価方法としてアンケートを用いた．センテンスの抽出の精度に対して「非常に良い」,「普通」,「悪い」の 3 択で回答してもらった．実験結果を表 5 に示す．『くるり』と『9 ボーダー』について，センテンス抽出について，「非常に良い」割合が高いため，可視化するにあたって，精度が高くセンテンス抽出ができていることがわかる．一方で『アンチヒーロー』は，「普通」であると回答した人が 60%いる．理由としては，図 5 の出力で横軸にセンテンスが記載されているが，その部分の一つに「モノセ」というものがある．これを動画メディアコンテンツで確認したところ，人の名前「ももせ」を発していたことがわかった．他にも，「尾行を開始します」という言葉が「微公開します」と認識されていた．これらから，音声認識の精度に改善の余地があると考えられる．

4.4 考察

本節では，実験の結果を踏まえて考察を行う．実験 1 では，選択した 3 つのメディアコンテンツと，それらに対応するメディアコンテンツ音声感情ヒートマップとのマッチ度を評価する．本実験の手順は，動画をセンテンスごとに分割して音声ファイルを生成し，音声感情認識によって得られた感情確率値をヒートマップ化し，選択した動画との適合性を評価するものである．実験 2 では，センテンス別音声抽出機能で抽出した，センテンスの抽出精度を検証するものである．

上記の実験 1 の結果から，3 つのコンテンツを比較したところ，『アンチヒーロー』が最も評価が高かった．この理由として，『アンチヒーロー』の中には怒っているシーンや脅迫しているシーンが多く含まれており，これらのシーンは音声認識において感情が明確に表現されやすいため，感情認識の精度が高くなったと考えられる．

『9 ボーダー』の予告動画が最も精度が低かった原因として，予告動画用に音量調整が行われている可能性が挙げられる．学習データセット内の音声は，それぞれのラベルによって音量の大きさが異なっている．そのため，入力した動画の落ち着いた場面で「disgust」と認識されることがあった．この現象は，データセット内でのラベル付き音声の音量を大きくしているラベルがあるため，ドラマ予告のような一定の音量を保っている音声を認識する際に偏った結果が出やすくなると考えられる．

実験 1 において，全てのヒートマップのセンテンスの最後には「ときめく時を TBS」と発言されている部分がある．このセンテンスは，ドラマの中の登場人物が発言している

のではなく，TBS テレビのブランドメッセージであり，ナレーションが発しているものである．しかし，図 3,4,5 からわかるように，全てのヒートマップでこの部分が「disgust」と認識される確率が高いことが示されている．ドラマの主人公が特定のセンテンスで「disgust」と判別されることは問題ないが，ナレーションが発した言葉に対して「disgust」である確率が高いと判別される原因として，音量の大きさが影響しているのではないかと考えられる．他にも，ヒートマップを見る限り，neutral, sad, surprised の値があまり変動していないため，これも音量の大きさが影響している可能性がある．また，RAVDESS データセットは英語の音声を用いているが，日本語との音声感情表現に差異があった可能性もあると考えられる．

実験 2 について，タイムスタンプ付きセンテンス抽出機能の，センテンス抽出の精度を上げるためには，Whisper の large モデルに加えて，そのドラマの主人公の名前や，あらすじからドラマ全体の単語データベースを作成することが必要である．出力されたセンテンスの中に存在しない単語や文字列が発見された場合，そのドラマごとに作った単語データベースを参照することによって話された言葉を推測し，候補を提示することで，センテンス抽出の精度を高めることができる可能性がある．

今後の課題としては，ドラマの予告動画を入力する際に音量を一定に整えることが重要であると考えられる．これにより，音声感情認識の精度が向上すると期待される．この方法によって，データセット内で感情ラベルに対する音量の不一致を補正し，予告動画のように均一な音量を持つ音声をより正確に分類できるようになると考える．他にも，実験では日本語の動画メディアコンテンツを使用したが，感情認識に用いた RAVDESS のデータセットは英語の音声データである．このため，感情の表現に言語の違いがある場合，最終的な出力結果が異なり，精度に影響を与える可能性がある．より高い精度を出力するためには，日本語の感情ラベル付き音声データセットを使用することが望ましい．あるいは， RAVDESS のデータセットを使用して，精度を確認する場合は，英語で話されている動画メディアコンテンツを入力データとして使用する必要がある．このようにして，言語の違いが感情認識の結果に与える影響を最小限に抑えることができると考えられる．

5. おわりに

本稿では．動画メディアコンテンツを対象とした音声感情抽出による重要セリフ抽出手法について示した．本方式は，動画メディアコンテンツを入力とし，それをセンテンス別音声データに変換し，事前に用意した， 7 感情ラベルごとに学習させた音声感情分類モデルを用いて，センテンス別音声感情データを作成することによって，動画メディアコンテンツの音声感情ヒートマップを生成し出力とする．また，入力した動画メディアコンテンツの音声感情ヒートマップの有効性を検証し音声感情可視化機能の有効性を確認した．さらに，動画メディアコンテンツのセンテンス別音声抽出機能で抽出した，センテンスを正確に抽出できているかの精度の検証をした．

今後の課題として，使用する感情ラベル付き音声データセットの言語と入力言語を一致させる必要がある．また，動画編集による音量調整を考慮した上で，データセット内の音量を調整することが必要となる．これらを行うことによって，より精度の高い音声感情認識の実現が可能であると考えられる．

参考文献

[1] Hirano Ryuichi, Okada Ryotaro, Nakanishi Takafumi, "Extraction Method for Important Words as a Viewer’s Reaction Arousal Factor from YouTube-Transcription," in 2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI), IEEE, pp. 651-652, (2022).
[2] Xu Min, Chia Liang-Tien, Yi Haoran, Rajan Deepu, "Affective content detection in sitcom using subtitle and audio," in 2006 12th International Multi-Media Modelling Conference, IEEE, pp. 6, (2006).
[3] Chen Lei, Rizvi Shariq J., Özsu M. Tamer, "Incorporating audio cues into dialog and action scene extraction," in Storage and Retrieval for Media Databases 2003, SPIE, Vol. 5021, pp. 252-263, (2003).
[4] Alatan A. Aydin, "Automatic multi-modal dialogue scene indexing," in Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205), IEEE, Vol. 3, pp. 374-377, (2001).
[5] Fried Ohad, Tewari Ayush, Zollhöfer Michael, Finkelstein Adam, Shechtman Eli, Goldman Dan B., Genova Kyle, Jin Zeyu, Theobalt Christian, Agrawala Maneesh, "Text-based editing of talking-head video," in ACM Transactions on Graphics (TOG), Vol. 38, No. 4, pp. 1-14, (2019).
[6] Schuller Björn, Reiter Stephan, Muller Ronald, Al-Hames Marc, Lang Manfred, Rigoll Gerhard, "Speaker independent speech emotion recognition by ensemble classification," in 2005 IEEE International Conference on Multimedia and Expo, IEEE, pp. 864-867, (2005).
[7] Livingstone Steven R., Russo Frank A., "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," in PloS One, Vol. 13, No. 5, pp. e0196391, (2018).
[8] 吉井茜音，岡田龍太郎，峰松彩子，中西崇文, "動画メディアコンテンツを対象とした人物表情認識を用いた感情重要キーワード抽出方式," 第 85 回全国大会講演論文集, (2023).