

音声エンタテインメントからの ウェブ音声マイニングの可能性

山西 良典^{1,a)} 田中 一星² 井本 桂右^{3,b)} 山下 洋一^{2,c)}

受付日 2020年1月28日, 採録日 2020年9月10日

概要: ウェブ上には様々なマルチメディアで構成されたユーザ参加型のエンタテインメントコンテンツが存在している。これらのエンタテインメントコンテンツからは、統制された条件に従った映像や音声を取得できる可能性がある。本稿では、音声データの活用に焦点を当て、ウェブ上のエンタテインメントコンテンツからの統制された環境下での音声データの収集をウェブ音声マイニングとして提案する。ウェブ音声マイニングの基本的な手続きのフレームワークを示し、ウェブ上のエンタテインメントコンテンツから音声データセットを取得した。音声コンテキスト認識と t-SNE 法を用いた 2 次元空間上への可視化を通して、取得した音声データセット中の発話に見られる音響特徴の傾向について基礎的な考察を行った。その結果、各発話は課題コンテキストごとに複数の発話者で共通の音響特徴を示す傾向であることを確認し、ユーザ発信型のエンタテインメントコンテンツ中の音声をラベル付き音声データとして研究用途に応用できる可能性を示した。

キーワード: ウェブ音声マイニング, エンタテインメントの活用, 音声情報処理

Application Vision of Web Speech Mining from Vocal Entertainment Contents

RYOSUKE YAMANISHI^{1,a)} ISSEI TANAKA² KEISUKE IMOTO^{3,b)} YOICHI YAMASHITA^{2,c)}

Received: January 28, 2020, Accepted: September 10, 2020

Abstract: There is a lot of participatory entertainment consisting of varied multimedia on Web. From such entertainment contents, we believe that it should be possible to acquire multimedia data such as movie and audio under the fixed condition. This paper focuses on the application of the speech data, and proposes the framework that acquires speech data under the fixed condition from vocal entertainment contents on Web as **Web speech mining**. In this paper, basic procedures of Web speech mining were introduced and the speech dataset was constructed from the entertainment content on Web. The speeches in the constructed dataset were foundationally studied based on their acoustic features through speech context recognition and visualization using t-SNE method. As the result, we confirmed that speeches in the constructed dataset showed a trend that the speakers commonly expressed specific acoustic features for each context. The results also pointed the application vision of Web speech mining, where speeches in user-generated entertainment contents can be applied to labeled speech data for research use.

Keywords: Web speech mining, application of entertainment, speech processing

¹ 関西大学総合情報学部
Faculty of Informatics, Kansai University, Takatsuki, Osaka
569-1095, Japan

² 立命館大学情報理工学部
College of Information Science and Engineering,
Ritsumeikan University, Kusatsu, Shiga 525-8577, Japan

³ 同志社大学理工学部
Faculty of science and engineering, Doshisha University,
Kyotanabe, Kyoto 610-0394, Japan

a) ryama@kansai-u.ac.jp

b) keisuke.imoto@ieee.org

c) yyama@is.ritsumei.ac.jp

1. はじめに

ウェブ上では、日々、ユーザ発信型の新たなエンタテインメントが創造されている。ユーザが共通のコンテンツに対して、それぞれのオリジナリティを適用するエンタテインメント現象は、インターネット黎明期から盛り上がりを見せている。たとえば、「ボカロイドによる楽曲カバー」、「ゲーム実況動画」、「踊ってみた動画」などでは、様々な作品がニコニコ動画、YouTubeなどのウェブサービス上で共有され、楽しまれている。

ユーザ発信型エンタテインメントの1つに、「じゃがりこ^{*1}」という菓子商品名を、様々な感情や状況に適した表現方法で発話する演技力じゃがりこ面接^{*2}がある。演技力じゃがりこ面接では、33種類の異なるコンテキストに応じた表現で声優やYouTuber、V-tuberなどがそれぞれの発話をインターネット動画共有サイトなどへ投稿している。発話文となる「じゃがりこ」自体は、感情や状況といった「コンテキスト」を内包しないため、取得された音声データのラベル間での音声特徴の差異は、課題となったコンテキストそのものの特徴としてとらえることができる。課題には表現が難しいような複雑なコンテキスト（たとえば、「告白しながら」や「必殺技の」など）も用意されているが、演技音声面接と題することで難しいコンテキストの音声発話に対して挑戦することへのモチベーションが創出されている。このしかけによって、一般的な実験環境下ではタスク難易度が高く、発話者の作業負担が大きくなってしまような複雑なコンテキストに対応した音声表現も取得可能である。さらに、課題として「普通に」が用意されていることにより、「普通に」を基準とした話者ごとの正規化も可能である。このようなコンテンツの特性から、「演技力じゃがりこ面接」の音声をラベル付き音声データとして活用する可能性を考えた。

このアイデアの背景には、ウェブ上に投稿された画像から研究用の画像データセットを構築する研究（たとえば、文献[1]や文献[2]など）がある。このようなウェブ上の画像を利用したデータセットの構築は、スマートフォンの普及によって一般ユーザが生活のあらゆる場面で容易に写真を撮影可能な環境が用意され、撮影した写真を共有する楽しみが生まれたことに起因すると考えられる。現在、スマートフォンの高性能化とインターネット通信回線の高速化によって、写真だけでなく動画を撮影して共有することがエンタテインメントとして普及しつつある。このことから、ウェブ上で共有されて蓄積されていく動画を構成する音声やモーションを統計的機械学習用のラベル付きデータとして活用する発想を得た。音声やモーションは画像とは

異なる“時系列データ”であるため、ラベル付与とセグメンテーションに課題が残る。この課題解決において、ユーザ参加型エンタテインメントコンテンツの特徴が有効に働くと考えた。ユーザ参加型のエンタテインメントコンテンツは、実験統制環境ではなく、各ユーザがそれぞれの環境でコンテンツを制作しているにもかかわらず、条件や文脈を共有している（たとえば、同一楽曲に対するダンスなど）ことが多い。そのため、ユーザが投稿した動画には一定のラベル付与とセグメンテーション共有が期待される。

本稿では、エンタテインメントコンテンツによって制御された発話音声の音響特徴量に着目する。我々は、「BGMが用いられ、発話とリズム音楽の混合音であったとしても、エンタテインメントコンテンツによって制御された発話者の音声の特徴量は、感情やコンテキストといった課題の特性に応じて分布する」という仮説を立てた。この仮説を検証するためにユーザ発信型エンタテインメントの1つである演技力じゃがりこ面接から、様々なコンテキストについての発話を収録した音声データセットを構築する。構築した音声データセット中の各コンテキストの認識実験と音声の多次元空間上への可視化から、構築した音声データセット内での演技音声の音響特徴量の傾向を考察する。考察結果をもとに、音声の音響特徴量の観点から、ウェブ上のエンタテインメントコンテンツから取得した音声データセットの統計的機械学習での利用、つまりウェブ音声マイニングの可能性について基礎検討を行う。

2. 関連研究

音声情報処理分野では、様々なラベルつき音声データセット（たとえば、文献[3]や文献[4]）が公開されている[5]が、A) 特定のプロフィール（年代、性別など）を持つ少数話者が同一文を発話したもの、B) 話者は多くても異なる文を発話したもの、C) 多くの話者が同一文を発話していても音声表現の感情やコンテキストの種類数が少ないもの、といったデータセットであることが多い。本稿では、「演技力じゃがりこ面接」からラベル付き音声データを取得することで、 α) 様々なプロフィールの多数の発話者、 β) 同一文（「じゃがりこ」に統一）に対する多様な感情やコンテキストで表現された音声、 γ) 複雑なコンテキストを意図した発話音声、といった既存の音声データセットにはない性質を持ったデータセットの構築を目指す。

これまでにも、ウェブ上でユーザが発信したコンテンツを研究用データとして応用することを試みた数多くの報告がある。このうち、画像情報処理研究に多大な影響を与えた研究として、Imagenet[1]がある。Imagenetでは、ウェブ上に投稿された無数の写真（画像）を収集し、ラベルを付与することによって、それまでの画像処理研究で用いられてきたCaltech-256[6]やPASCAL[7]といった人手で用意されたデータセットに比べて大量で多様な画像を収集し

^{*1} <https://www.calbee.co.jp/jagarico/> (accessed 2019-12-23)

^{*2} <https://nana-music.com/sounds/008054f0> (accessed 2019-05-15)

ている。機械学習、特に深層学習のアプローチを用いる研究においては、大量のラベルつき学習データの準備はそれぞれの研究のタスクにおける性能に大きく影響を与える一方で、多大なコストを要する問題となっている。この問題を、「インターネット上で画像を共有する」といった社会的な動向に着目することで解決した点において、Imagenetはその功績を特徴づけられる。ウェブ上の画像を収集し、データセットとして応用するための研究は「ウェブ画像マイニング」と呼ばれ、Imagenetの他にも様々なアプローチが報告されている [2], [8].

エンタテインメントの活用の観点からは、本稿はエンタテインメントコンテンツの Human Computation [9] への応用として位置づけられる。本稿で目指すウェブ音声マイニングと同様のコンセプトでウェブ上のエンタテインメントコンテンツからモーション情報を抽出した研究として、LiらはYouTube上にアップロードされた動画エンタテインメントである *MannequinChallenge* を基にデータセットを構築し^{*3}、動画中の深度情報を推定している [10]. 従来のCrowdsourcing [11] の多くでは、賃金 [12] や情報 [13] を対価としてユーザにマイクロタスクを実施させている。一方で、ユーザ発信型のエンタテインメントコンテンツの場合には、ユーザが主体的にクオリティの高いコンテンツをウェブ上へ発信する傾向にある。そのため、エンタテインメントコンテンツからは、“結果的にマイクロタスクを実施することになるユーザにとっての心的負担も少なく、低コストでありながら高クオリティなラベル付きデータの取得が期待される。つまり、“楽しんだこと”を利用して研究課題に必要なデータセットを獲得する試みといえる。

3. データの準備

「演技力じゃがりこ面接」の音源を使用した動画をWebから取得した。動画の収集源とするWebサイトはYouTubeを使用し、検索クエリを「じゃがりこ面接」として得られた動画に対して、再生回数が多い順にソートしてから研究対象とする動画を取得した。「演技力じゃがりこ面接」の音源では好きにアドリブを入れて良いという指示がされており、コンテキストによっては「じゃがりこ」と発話されていなかったり、「じゃがりこ」以外にも発話内容があるものも多い。そこで、検索結果として得られた180件の動画から、タイトルに異なる発話内容や「動物の声真似」などの条件が含まれており、課題音声が発話されている可能性が低くアドリブが多いと推察されるものを第2著者の主観で判断し視聴対象から除外し、120動画を取得した。その後、同じく第2著者が動画の音声を聴取し、“録音を1人でやっていること”、“【part1】演技力じゃがりこ面接”を利用して”を基準として主観で判断した。また、

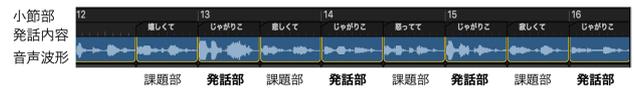


図1 ジャがりこ演技音声面接から取得した音声データの構造。コンテキストの提示と演技音声の発話が一定間隔で交互に繰り返される

Fig. 1 The structure of the speech data obtained from Jagarico performed speech interview. The task and the performed speech are alternately repeated in a certain period.

本稿では、同一発話文での音声を収集することを目的としているため、著者のうち1名が主観で“過大なアドリブが含まれているもの（たとえば、まったく異なる単語を叫んでいるものなど）”を判断して除外した。これらのデータクレンジングによって、検索結果として得られた180動画(5,760発話)から、最終的には全15名(男性5名、女性10名、480発話)の動画を選出した。このとき、データクレンジングの過程におけるデータの除去理由とその分布によるデータ収集の容易性や収録条件などによる音声品質についての議論は、本稿の主旨とは異なるため今後の課題とする。

取得した動画の音声は、各コンテキストに分割される必要がある。取得した動画から得られる音声データは、図1に示すように、リズムに合わせて各コンテキストが課題として提示され、続いてその回答として課題コンテキストの「じゃがりこ」が発話される。楽曲投稿サイトnanaでは、コンテキスト（つまり、課題部の音声）とリズム音楽のみで構成された8ビートのBGM^{*4}（図1の発話部の音声が含まれていない音源）がテンプレートとして用意されている。そこで、音楽編集ソフトLogicのトラック1に動画から取得したBGMを含む音声データ、トラック2にテンプレートのBGM音源をそれぞれ入力した。まず、著者のうち1名が、音楽編集ソフト上で目視と聴取によってによりBGMのバスドラムの位置を基準として動画から取得した音声データとテンプレートのBGM音源を同期させた。トランジェント検出によって得られたバスドラムを基準としてBGMのテンポを計算したところ、BPMは135であった。その後、1/2小節ごとにデータを分割することで、音声データ中のコンテキスト提示部分とそれに対応づく回答としての発話を切り分けた。音源によっては回答が切り分けた範囲に収まっていないものもあり、これらについては分割後に著者が聴取してデータの抽出範囲を手作業で調整した。ここで、BGMが大きく異なる音楽特徴を示すコンテキストについては、音声表現による差異よりもBGMによる差異による影響が大きくなることが考えられるため、対象データから取り除くものとした。表2に、除外したコンテキストと除外理由となったBGMにおけるそれぞれの

^{*3} <https://google.github.io/mannequinchallenge/www/index.html> (accessed 2019-12-23)

^{*4} <https://nana-music.com/sounds/008054f0> (accessed 2019-05-15)

表 1 音声データセットとして収録した 22 コンテキストを、4 属性 (感情, 状況, 話相手, 特殊) に分類

Table 1 The 22 contexts in the speech dataset are classified into four types of category: **emotion**, **situation**, **speech partner**, and **special**.

感情	状況	話相手	特殊
嬉しくて	失恋して	隣の人に	必殺技の
悲しくて	喧嘩して	遠くの人に	関西風に
怒ってて	疲れ果てて	知り合いに	関東風に
寂しくて	寒すぎて	友達に	
嫉妬して	暑すぎて	恋人に	
驚いて	眠たくて		
	食べながら		
	告白しながら		

表 2 音声データセットから除外したコンテキストとその除外理由となった音楽特徴

Table 2 The contexts eliminated from the speech dataset and each corresponding music features to eliminate the context.

コンテキスト	除外理由となった音楽特徴	コンテキスト	除外理由となった音楽特徴
がっかりして	ドラムフィル	テンション上がって	バスドラムがない
感謝して	ドラムフィル	テンション下がって	バスドラムがない
赤の他人に	異なるリズムパターン	英語風に	ドラムロール
恋のビームの	ドラムロール	中国風に	異なるリズムパターン
2次元を見て	バスドラムがない	普通に	BGM なし
3次元を見て	バスドラムがない		

音楽特徴を示す。

以上のプロセスによって得られた演技音声を音声データセットとして用いる。本稿では、構築された音声データセットを **JSD** (**J**agarico **S**peech **D**ataset) と呼ぶ。なお、JSD では 22 コンテキストすべての発話をあらかじめ表 1 のように 4 つの属性「感情」、「状況」、「話し相手」、「特殊」に分類した。

4. 音響特徴量の抽出

3 章で得られた JSD 内の各発話から音響特徴量を抽出する。ここで、音響特徴量は BGM も含めた各発話部から抽出するが、表 2 に示した除外されたコンテキスト以外では課題間で BGM の音楽特徴量は共通しているため、音響特徴量の差異は発話音声の差異として仮定する。音響特徴量の抽出には、OpenSMILE [14] にコンフィグファイル IS09_emotion [15] を適用した。ここで、分析フレーム長は 25 ms、フレームシフト幅は 10 ms であった。表 3 に示し

表 3 OpenSMILE で IS09_emotion.conf を用いて得られる音響特徴。感情の推定に関わる合計 32 種類の音響特徴を取得する

Table 3 Acoustic features obtained by using OpenSMILE with IS09_emotion.conf. Those are 32 acoustic features concerning emotion recognition.

特徴 ID p	特徴	説明
1	RMS energy	エネルギーの 2 乗平均平方根
2	RMS energy differential	エネルギーの 2 乗平均平方根の微分
3	F0	基本周波数
4	F0 differential	基本周波数の微分
5-16	MFCC 1-12	1~12 次のメル周波数ケプストラム係数
17-28	MFCC 1-12 differential	1~12 次のメル周波数ケプストラム係数の微分
29	ZCR	ゼロ交差率
30	ZCR differential	ゼロ交差率微分
31	voiceProb	その時点での音が声である確率
32	voiceProb differential	その時点での音が声である確率微分

表 4 表 3 に示した音響特徴それぞれについての統計特徴量。表 3 に示した特徴量について、本表中の各素性値を取得する。

Table 4 Stastical features for each acoustic feature shown in Table 3. The values in this table are calculated for each feature shown in Table 3.

統計量 ID m	素性値	説明
1	max	最大値
2	min	最小値
3	range	最大値と最小値の差
4	maxPos	最大値の絶対位置
5	mixPos	最小値の絶対位置
6	amean	算術平均
7	linregc1	線形近似の勾配
8	linregc2	線形近似のオフセット
9	linregerrQ	線形近似と 2 乗誤差
10	stddev	値の標準偏差
11	skewness	歪度
12	kurtosis	尖度

た計 16 種類の音響特徴それぞれに対して、表 4 に示す統計特徴量を算出し、1 発話から全 384 個の特徴量 (32 特徴 \times 12 統計量) を抽出した。ここで、特徴 ID p と統計量 ID m を用いて、音響特徴量を $f_{p,m}$ として表す。たとえば、MFCC ($p = 7$) についての算術平均 ($m = 6$) は $f_{7,6}$ として示される。

話者 j のコンテキスト i についての発話を s_i^j としたとき、この発話の $f_{p,m}$ は $f_{p,m}(s_i^j)$ として表される。それぞれの発話を「普通に」からの変化量、平均・分散を用いて標準化する。まず、式 (1) に従って、 $f_{p,m}(s_i^j)$ の「普通に」発話からの変化量 $v_{f_{p,m}}(s_i^j)$ を算出する。ただし、「普通

に] について $i = \phi$ とする.

$$v_{-f_{p,m}}(s_i^j) = \begin{cases} f_{p,m}(s_i^j)/f_{p,m}(s_\phi^j) & (p = 1, \dots, 4), \\ f_{p,m}(s_i^j) - f_{p,m}(s_\phi^j) & (\text{others}), \end{cases} \quad (1)$$

ここで、音響特徴の性質を考慮し、RMS energy と F0 に関わる特徴量については除算式、その他の特徴量については減算式をそれぞれ用いて「普通に」発話と比較した変化量を算出する. 次に、式 (2) に従って、 $v_{-f_{p,m}}(s_i^j)$ を特徴 p と統計量 m の組合せごとの平均と分散を用いて、「普通に」を除くすべてのコンテキスト i で標準化した $sv_{-f_{p,m}}(s_i^j)$ を算出する.

$$sv_{-f_{p,m}}(s_i^j) = \frac{v_{-f_{p,m}}(s_i^j) - \overline{v_{-f_{p,m}}(s^j)}}{\sigma_{p,m}^j}, \quad (2)$$

ここで、 $\sigma_{p,m}^j$ と $\overline{v_{-f_{p,m}}(s^j)}$ は、それぞれ、話者 j のすべての発話から得られる $v_{-f_{p,m}}(s_i^j)$ の分散と平均を示す. これらの処理によってすべての話者の、すべてのコンテキストについての発話を比較可能にする.

5. 演技音声のコンテキスト認識と可視化

JSD 内での演技音声はコンテキストごとに特徴づけられて発話されているかを確認する. 4 章で得られた音響特徴量を用いて、話者オープン・発話クローズ (異なる話者の「じゃがりこ」音声に対する認識) での音声コンテキスト認識実験と t-SNE 法による可視化を行った.

音声コンテキスト認識実験では、4 章で抽出した音響特徴量を用いて SVM によるコンテキスト認識を行った. SVM の認識には、機械学習ライブラリ scikit-learn を利用し、判別関数には linear kernel を用いた. 本稿で用意した JSD 内の音声データは、各コンテキストに対する認識モデルを構築するための学習用データとしてはデータ量の観点で不安が残る. そこで、各音声データから音響特徴量を抽出する際に分析開始フレームを複数用意することで、データオーグメンテーションする. 既存研究 [16] では、音声データの分析では分析フレームの開始時刻が数 ms ずれることによって取得される特徴量に変化し、音声認識の性能にも影響を与えることが示されている. 異なるフレーム開始地点で得られた音響特徴量を用いることで、一種のデータオーグメンテーションの効果が得られる. 本稿では、音響特徴量の抽出で採用した分析フレーム長とフレームシフト幅を考慮し、7 種類の分析開始フレーム (0 ms, 2 ms, 4 ms, 5 ms, 6 ms, 8 ms, 10 ms: ただし、ファイルの先頭位置は 0 ms とする) を用意し、それぞれの分析開始フレームを採用して抽出された音響特徴量は同一のラベルを持つ別発話から抽出された音響特徴量として扱うことでデータ量を 7 倍に増加させた^{*5}. テスト話者を除いた話者の発話

^{*5} 分析開始フレーム 0 ms と 10 ms では表 4 に示す統計量の算出範囲が異なることになる.

に対してデータオーグメンテーションした音声特徴量を学習データ、テスト話者の分析開始フレーム = 0 ms の発話をテストデータとした話者オープンな音声コンテキスト認識実験を行った.

音声データセットの可視化には、高次元データの可視化に用いられる非線形次元削減の手法の 1 つである t-SNE 法 [17] を用いた. t-SNE 法では、高次元のデータ集合を 2, 3 次元に配置する際に高い確率で類似した集合が近傍に、異なる集合が遠方になるように対応づける. 上述の処理によって得られた $sv_{-f_{p,m}}(s_i^j)$ に対して t-SNE 法を適用することで、特徴量の距離に基づいて発話の集合が形成され、コンテキスト内での発話のまとめりやコンテキスト間の距離が可視化される. t-SNE に必要となる各パラメータについては複数の値を用いて可視化結果を出力し、著者らの主観により 2 次元空間内で課題ごとの発話が最もまとまって見られるものを採用した. ただし、これらの t-SNE のパラメータは、多次元空間内のデータを 2 次元空間上で可視化するために用いるものであり、音響特徴量に基づくデータ間の相対関係を恣意的に操作するものではない.

5.1 全コンテキストに対するコンテキスト認識と可視化

「普通に」を除いた全 22 種類のコンテキストの認識について、正解率: 40.91%, 平均適合率: 41.60%, 平均再現率: 40.91%, 平均 F 値: 40.59% であった. 表 5 に、コンテキストごとの認識結果の詳細を示す. 22 種類と音声コンテキストの認識としては多クラスで複雑な課題を含むコンテキスト認識であるにもかかわらず、比較的高い精度での認識結果が得られた. このことから、22 種類のコンテキストは話者によらず共通した音響特徴で演じ分けられていることが示唆された. 特に、「嬉しくて」、「寂しくて」、「喧嘩して」、「告白しながら」、「遠くの人に」、「恋人に」、「関東風」については高い F 値が示されており、これらのコンテキストに対しては話者に共通して特徴的な演技音声が発話されていたことが推察される. 一方で、「疲れ果てて」、「寒すぎて」、「知り合いに」、「友達に」、「関西に」については、チャンスレベルに対しては大幅に高い精度であったが、他のコンテキストに比べると比較的低い認識結果を示した. 低い認識結果を示したコンテキストについては、誤認識の傾向を分析したが誤認識先に大きな偏りは見られなかった.

図 2 に、t-SNE 法によって出力された可視化結果を示す. このとき t-SNE のパラメータは、 $Perplexity = 30$, $Learning\ rate = 700$ とした. 同図から、認識精度が高かった「嬉しくて」や「寂しくて」などは空間上で発話が比較的まとまって分布していることが見て取れる. また、「寂しくて」、「怒ってて」、「嫉妬して」などのネガティブな感情に近い空間に分布していることから、これらの音声表現に近い傾向が見られることが推察される. 一方で、コン

表 5 22 種類の音声コンテキスト認識についての適合率, 再現率, F 値. 表中の値は%を示す. 太字は 50.00 以上, 下線は 30.00 未満の F 値をそれぞれ示す

Table 5 The recall, precision and F -value of recognition results for 22 kinds of context. The values in the table indicate %. Each bolded and underlined value indicates over than 49.99 and less than 30.00 F -value, respectively.

コンテキスト	適合率	再現率	F 値	コンテキスト	適合率	再現率	F 値
嬉しくて	54.55	80.00	64.87	眠たくて	37.50	40.00	38.71
悲しくて	41.18	46.67	43.75	食べながら	62.50	33.33	43.48
怒ってて	37.50	40.00	38.71	告白しながら	66.67	66.67	66.67
寂しくて	58.82	66.67	62.50	隣の人に	29.41	33.33	31.25
嫉妬して	33.33	40.00	36.36	遠くの人に	70.00	46.67	56.00
失恋して	35.00	46.67	40.00	知り合いに	26.32	33.33	<u>29.41</u>
喧嘩して	53.85	46.67	50.00	友達に	22.22	13.33	<u>16.66</u>
驚いて	40.00	40.00	40.00	恋人に	50.00	53.33	51.61
疲れ果てて	27.27	20.00	<u>23.08</u>	必殺技の	31.25	33.33	32.26
寒すぎて	14.29	13.33	<u>13.79</u>	関西風に	26.67	26.67	<u>26.67</u>
暑すぎて	33.33	33.33	33.33	関東風に	63.64	46.67	53.85

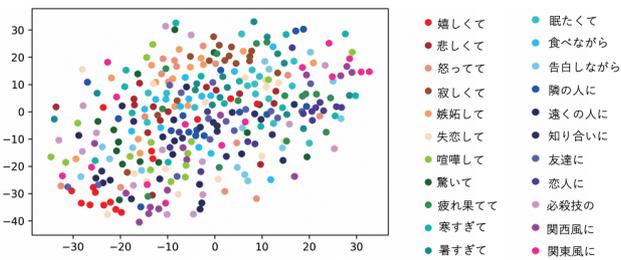


図 2 全 22 種類のコンテキストについての全発話を可視化した結果
Fig. 2 The visualization of speech for all of the 22 kinds of contexts.

表 6 各コンテキスト属性の可視化に採用した t-SNE のパラメータ値

Table 6 Parameters used in t-SNE for visualizing speeches in each context category.

	感情	状況	話し相手	特殊
<i>Perplexity</i>	20	40	20	30
<i>Learning rate</i>	500	500	500	500

テキストの属性によるまとまりは見られず, 22 種類のコンテキストが全体的に分散していることが分かる.

5.2 コンテキスト属性ごとのコンテキスト認識と可視化

表 1 に示したコンテキスト属性ごとに SVM を用いて音声コンテキスト認識モデルを構築し, 同一属性内でのコンテキスト認識を行った. また, 各コンテキスト属性に該当するコンテキストの発話のみを対象として t-SNE 法を適用して 2 次元空間を構築し, それぞれの属性ごとに各コンテキストについての発話の可視化を行った. このとき, コンテキスト属性ごとに, 15 人全員分の発話の可視化空間についての了解性が高いパラメータ (表 6 参照) を採用した. これらの分析により, 類似した概念内でのコンテキスト間の演技音声の傾向を考察した.

表 7 コンテキスト属性「感情」中での 6 種類の音声コンテキスト認識についての適合率, 再現率, F 値. 表中の値は%を示す. 太字は 70.00 以上の F 値を示す

Table 7 The recall, precision and F -value of recognition results for six kinds of context in “emotion.” The values in the table indicate %. The bolded value indicates over than 70.00 F -value.

コンテキスト	適合率	再現率	F 値
嬉しくて	66.67	80.00	72.73
悲しくて	70.59	80.00	75.00
怒ってて	62.50	33.33	43.48
寂しくて	68.75	73.33	70.97
嫉妬して	44.44	53.33	48.48
驚いて	69.23	60.00	64.29
平均	63.70	66.33	62.49

5.2.1 コンテキスト属性「感情」についての考察

表 7 に, コンテキスト属性「感情」内での 6 コンテキストに対する認識結果を示す. 「怒ってて」と「嫉妬して」については比較的低い認識率であったものの, その他のコンテキストにおいておおむね高い認識率が示され, 話者によらず各コンテキストに対して特徴的に発話されていたことが示唆された. 特に, 「嬉しくて」, 「悲しくて」に対しては適合率, 再現率ともに高い値が示された. 認識率が比較的低かった「怒ってて」については, 「寂しくて」と「嫉妬して」に多く誤認識されていることが分かった.

図 3 に示した可視化結果では, 「嬉しくて」と「悲しくて」は空間上で対極の位置に離れて分布している. また, 「悲しくて」が分布している図中左下付近には, 「寂しくて」「嫉妬して」といったネガティブなコンテキストが集まっていることが見て取れる. これらのことから, ポジティブとネガティブなコンテキスト群どうしは離れて位置しており, Russell の円環モデル [18] における Arousal と Valence

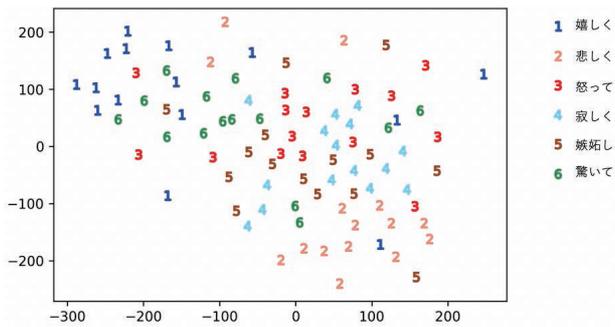


図 3 コンテキスト属性「感情」についての発話を可視化した結果
Fig. 3 The visualization of speech for category “emotion.”

表 8 コンテキスト属性「状況」中での 8 種類の音声コンテキスト認識についての適合率, 再現率, F 値. 表中の値は%を示す. 太字は 70.00 以上の F 値を示す

Table 8 The recall, precision and F -value of recognition results for eight kinds of context in “situation.” The values in the table indicate %. The bolded value indicates over than 70.00 F -value.

コンテキスト	適合率	再現率	F 値
失恋して	47.62	66.67	55.56
喧嘩して	76.92	66.67	71.43
疲れ果てて	55.56	33.33	41.67
寒すぎて	47.06	53.33	50.00
暑すぎて	40.00	40.00	40.00
眠たくて	36.84	46.67	41.18
食べながら	66.67	53.33	59.26
告白しながら	85.71	80.00	82.76
平均	57.05	55.00	55.23

で構成される空間と類似した t-SNE 空間が確認された。

5.3 コンテキスト属性「状況」についての考察

表 8 に, コンテキスト属性「状況」内での 8 コンテキストに対する認識結果を示す. 全体的な推定性能は, チャンスレベル (12.50%) を考慮すると高い性能が認められた. 一般的な音声コーパスで指示されることが多い「感情」に比べて, 複雑な「状況」のコンテキストであることを考えると十分に高い性能でコンテキスト認識が実現されているといえる.

特に, 「喧嘩して」や「告白しながら」は非常に高い精度で認識されており, これらの状況での発話については話者間で共通した音響特徴で演じられていると見られる. 属性「状況」についての発話を可視化した図 4 から, コンテキストごとに発話がまとまって分布している様子が見て取れる. これらの「状況」を指示した複雑なコンテキストに対しても話者は特徴的に発話を演じ分けていることが示唆された.

5.3.1 コンテキスト属性「話相手」についての考察

表 9 に, コンテキスト属性「話相手」中での 5 種類の音声コンテキスト認識結果を示す. こちらも, 「状況」の認識

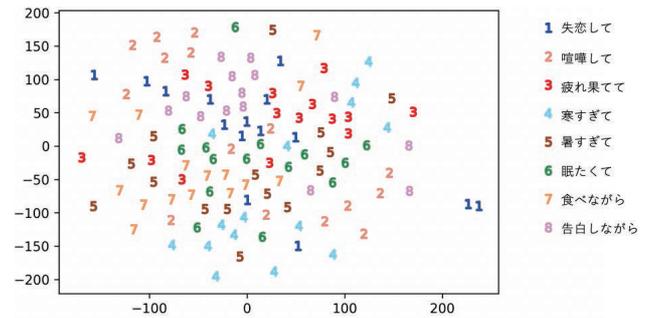


図 4 コンテキスト属性「状況」についての発話を可視化した結果
Fig. 4 The visualization of speech for category “situation.” The legend indicates the context index i .

表 9 コンテキスト属性「話相手」中での 5 種類の音声コンテキスト認識についての適合率, 再現率, F 値. 表中の値は%を示す. 太字は 65.00 以上, 下線は 40.00 未満の F 値をそれぞれ示す

Table 9 The recall, precision and F -value of recognition results for five kinds of context in “speech partner.” The values in the table indicate %. Each bolded and underlined value indicates over than 65.00 and less than 40.00 F -value, respectively.

コンテキスト	適合率	再現率	F 値
隣の人に	33.33	40.00	<u>36.36</u>
遠くの人に	64.71	73.33	68.75
知り合いに	47.06	53.33	50.00
友達に	50.00	26.67	<u>34.79</u>
恋人に	66.67	66.67	66.67
平均	52.35	52.00	51.31

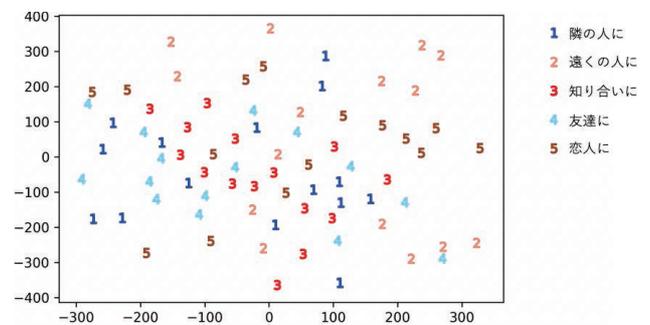


図 5 コンテキスト属性「話相手」についての発話を可視化した結果
Fig. 5 The visualization of speech for category “speech partner.”

と同様に一般的な音声コーパスでは指示されることがない「友達に」「恋人に」のような複雑なコンテキストを含んでいたが, チャンスレベル (20.00%) を超える精度でコンテキストが認識された. 「遠くの人に」については, 全 22 課題の認識においても高い精度を示したが, 「話し相手」のコンテキスト属性においても比較的高い精度が確認された. このことから, 「遠くの人に」向かっては話者間で共通の音響特徴で発話されていることが示唆される.

一方で全体的な傾向としては, 図 5 に示した可視化結果からも, 各コンテキストの発話が空間上に分散して位置し

表 10 コンテキスト属性「特殊」中での 3 種類の音声コンテキスト認識についての再現率, 適合率, F 値. 表中の値は%を示す. 太字は 70.00 以上の F 値を示す

Table 10 The recall, precision and F -value of recognition results for three kinds of context in “special.” The values in the table indicate %. The bolded value indicates over than 70.00 F -value.

	適合率	再現率	F 値
必殺技の	53.85	46.67	50.00
関西風に	52.63	66.67	58.82
関東風に	76.92	66.67	71.43
平均	61.13	60.00	60.08

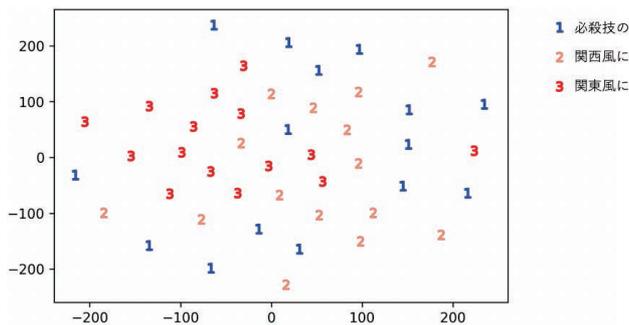


図 6 コンテキスト属性「特殊」についての発話を可視化した結果
Fig. 6 The visualization of speech for category “special.”

ていることが分かる。「隣の人に」「遠くの人に」以外については、3 種類とも親しい人間関係への発話をコンテキストとして想定しており、これらの演じ分けは難しい課題であることが推察される。認識対象であるコンテキスト数が少ないにもかかわらず、「感情」や「状況」のように非常に高い精度でのコンテキスト認識精度は得られなかった要因として、これらのコンテキストの演じ分けの難易度が考えられる。

5.3.2 コンテキスト属性「特殊」についての考察

表 10 に示したコンテキスト認識結果から、「関東風に」については高い認識精度が得られたことが分かる。全体的に、チャンスレベル (33.33%) より高い精度での認識結果が得られたものの、3 クラスのコンテキスト認識としては十分に高い結果とはいえない。誤認識結果を確認すると、多くの発話が「必殺技の」は「関西風に」に、「関西風に」は「必殺技の」にそれぞれ互いに誤って認識されており、「関西風に」の発話と「必殺技の」の発話が互いに類似した音響特徴量であったと考えられる。

図 6 の可視化結果でも、「関東風に」についてはまとまって分布しているのに対して、「必殺技の」と「関西風に」は空間上に混合して分散していることが示された。他のコンテキスト属性に比べてチャンスレベルが高く分類問題としての難易度は低いにもかかわらず、各コンテキストでまとまりを持った分布とはなっておらずこれらのコンテキストについては発話者間で音響特徴量に対して共通認識が得ら

れていなかったことが示唆された。

6. おわりに

本稿では、ウェブ上で共有されているユーザ発信型エンタテインメント中の音声の研究用途へと応用するウェブ音声マイニングのフレームワークを提案した。ウェブ音声マイニング実現の基礎検討として、「演技力じゃがりこ面接」から抽出した音声によって構成される複雑なコンテキストに対する演技音声データセット JSD を構築した。JSD における各コンテキストに対する複数話者の発話から抽出した音響特徴を用いて、音声コンテキスト認識実験と t-SNE 手法によって構成された可視化を行った。全コンテキスト 22 種類に対するコンテキスト認識結果は、40.91%の正解率と 40.59%の F 値が確認された。また、属性ごとに行ったコンテキスト認識では、「感情」や「状況」の属性ではチャンスレベルを大幅に上回る平均精度でコンテキストを認識可能であることが確認され、類似した概念内においても、コンテキストごとに異なる音響特徴で発話が特徴的に演じ分けられていることが示唆された。特に、従来の音声コーパスにはないような複雑なコンテキストに対しても一定の認識性能が認められ、JSD 内では統計的機械学習による認識が可能なレベルで演技音声の特徴づけられて発話されていることが明らかになった。また、可視化結果の考察からも、それぞれのコンテキスト内では複数の話者の発話間距離が近くまとまって分布していることが確認された。同一のコンテキストを課題として音声表現された発話や類似したコンテキストは音響特徴量空間で相対的に近い位置に配置されることが確認された。たとえば、図 4 における「暑すぎて」と「寒すぎて」のコンテキストでは、気温に関する音声表現として相対的に近い位置に配置されつつも、これら 2 種類のコンテキストが分離して配置されていることを確認した。これらの考察から、エンタテインメントコンテンツによって制御された発話者の音声の特徴量は、発話者によらず課題コンテキストの特性に応じて分布することが確認された。以上から、「演技力じゃがりこ面接」を通してウェブ上に公開された音声は、コンテキストラベルが付与された音声データとして応用できる可能性が示唆された。

本稿で検討対象とした「演技力じゃがりこ面接」の特性を考察すると以下のような条件を有しているコンテンツは、ウェブ音声マイニングの対象として利用できる可能性が高まると示唆された。

- 複数の話者が共通のコンテンツを発話している。
- 音声表現 (発話内容とコンテキスト) が課題あるいはコンテンツによって制御されている。
- 発話が重なっておらず、発話部は 1 発話のみで構成されている。

これに加え、「演技力じゃがりこ面接」では、テンプレートとして利用可能な音声ファイルが提供されていることに

よって、発話部のセグメンテーションにかかる作業コストが大幅に削減された。したがって、以下の条件を満たすコンテンツは、ウェブ音声マイニングの対象としてより有用性が高いと考えられる。

- 時間的なセグメンテーションや課題に対するセグメンテーションを同定可能にするテンプレートとして利用可能な音声ファイルが提供されている。

テンプレートとして利用可能な音声ファイルが用意されていないコンテンツを対象とする場合には、音声認識や発話区間検出技術を用いてセグメンテーションする必要がある。ウェブ上には、演技力じゃがりこ面接の他にも同一の童話に対する複数話者の読み聞かせなど^{*6}が公開されている。音声認識や音声区間の自動検出技術を用いて発話内容に対するセグメンテーションを行えば、これらの音声データも共通した発話コンテキストにおける多様な音声表現のデータとして扱える可能性が考えられる。

今後は、データ収集の容易性の観点から、データクレンジングの自動化についても、音声認識技術による発話内容の認識や音声区間の自動検出技術、クラウドソーシングの利用を含めて検討していく。また、エンターテインメントコンテンツとして収録された演技音声は、実環境下で音声のコンテキスト認識においてどの程度実用可能であるかについても検証していきたい。本稿で構築したデータセットでは、BGMを含んだ音声ファイルが収録されている。本データセットをそのままコンテキスト認識に利用する際には、入力音声にテンプレートとなる課題音声ファイルを重ねたうえで音響特徴量を抽出する必要がある。構築したデータセットを一般発話音声のコンテキスト認識に用いるための検討（データセット中の各発話とBGMを分離することによる認識性能への影響の分析や、発話内容に対してオープンなコンテキスト認識）は今後の課題とする。一般発話音声に適用できれば、たとえばスマートスピーカーにおいて「ただいま」の一言からユーザのコンテキスト（疲れ果てている、あるいは、眠たくて仕方ないなど）を認識してシステム側から能動的に情報提示を行うことが可能になる。また、「明日の天気は？」というユーザからの同一の問合せに対しても、「暑い」と感じているユーザには「明日も暑そうです」、「寒い」と感じているユーザには「明日は暖かくなりそうです」のような対応が可能になるであろう。このような情報呈示が可能になれば、人間とシステムのインタラクションデザインは、既存の“ユーザのコマンドに制御された受動的なインタラクション”から“ユーザのコンテキストに応じたシステムからの能動的なインタラクション”へ変化することが期待される。これらの項目に取り組み、楽しまれたコンテンツを統計的機械学習の学習データとしての利用することを想定したエンターテインメントコンテン

ツのデザイン（たとえば、ゲームデザイン [19] など）についても、エンターテインメントコンピューティングと Human Computation の共通トピックとして議論していきたい。

謝辞 本研究は、一部、立命館大学アトリサーチセンター「日本文化資源デジタル・アーカイブ国際共同拠点国際共同研究」の支援のもと実施した。記して謝意を表す。

参考文献

- [1] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L.: Imagenet: A large-scale hierarchical image database, *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp.248–255 (2009).
- [2] Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z. and Zheng, Y.: NUS-WIDE: A Real-world Web Image Database from National University of Singapore, *Proc. ACM Int'l Conf. Image and Video Retrieval*, pp.48:1–48:9 (2009).
- [3] Arimoto, Y., Kawatsu, H., Ohno, S. and Iida, H.: Emotion recognition in spontaneous emotional speech for anonymity-protected voice chat systems, *Proc. Annu Conf. Int'l Speech Communication Association*, pp.322–325 (2008).
- [4] 森山 剛, 森 真也, 小沢慎治: 韻律の部分空間を用いた感情音声合成, *情報処理学会論文誌*, Vol.50, No.3, pp.1181–1191 (2009).
- [5] NII: 音声資源コンソーシアム, 入手先 (<http://research.nii.ac.jp/src/>).
- [6] Griffin, G., Holub, A. and Perona, P.: Caltech-256 Object Category Dataset, Technical Report 7694, California Institute of Technology (2007).
- [7] Everingham, M., Van Gool, L., Williams, C.K., Winn, J. and Zisserman, A.: The PASCAL visual object classes (VOC) challenge, *Int'l Journal of Computer Vision*, Vol.88, No.2, pp.303–338 (2010).
- [8] Fergus, R., Perona, P. and Zisserman, A.: A Visual Category Filter for Google Images, *Proc. European Conf. Computer Vision*, pp.242–256 (2004).
- [9] Quinn, A.J. and Bederson, B.B.: Human computation: A survey and taxonomy of a growing field, *Proc. CHI Conference on Human Factors 2011*, pp.1403–1412 (2011).
- [10] Li, Z., Dekel, T., Cole, F., Tucker, R., Snively, N., Liu, C. and Freeman, W.T.: Learning the Depths of Moving People by Watching Frozen People, *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2019).
- [11] Howe, J.: The rise of crowdsourcing, *Wired Magazine*, Vol.14, No.6, pp.1–4 (2006).
- [12] amazon: amazon mechanical turk, available from (<https://www.mturk.com/>).
- [13] Google: reCAPTCHA, available from (<https://www.google.com/recaptcha/intro/v3.html>).
- [14] Eyben, F., Wöllmer, M. and Schuller, B.: openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor, *Proc. ACM Multimedia*, pp.1459–1462 (2010).
- [15] Schuller, B., Steidl, S. and Batliner, A.: Emotion recognition in spontaneous emotional speech for anonymity-protected voice chat systems, *Proc. Annu Conf. Int'l Speech Communication Association*, pp.312–315 (2009).
- [16] 伊藤彰則: 音声認識におけるフレームシフト再考, *情報処理学会研究報告*, Vol.2016-SLP-112, No.10 (2016).
- [17] van der Maaten, L. and Hinton, G.: Visualizing Data

*6 たとえば, <http://aozoraroudoku.jp/> など (accessed 2020-05-14)

using t-SNE, *Journal of Machine Learning Research*, Vol.9, pp.2579–2605 (2008).

- [18] Russell, J.A.: A Circumplex Model of Affect, *Journal of Personality and Social Psychology*, Vol.39, No.6, pp.1161–1178 (1980).
- [19] 三輪聡哉, 中村聡史: マイクロタスク埋め込み型音楽ゲームの提案, 情報処理学会研究報告エンタテインメントコンピューティング (EC), Vol.2014-EC-34, No.2 (2014).



山西 良典 (正会員)

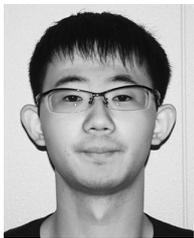
2007年名古屋工業大学工学部知能情報システム学科卒業。2009年同大学大学院工学研究科情報工学専攻博士前期課程修了。2012年同博士後期課程修了。博士(工学)。同年立命館大学情報理工学部助手, 2013年同特任

助教。2014年同助教, この間, UBC(カナダ)客員助教。2018年同講師。2020年関西大学総合情報学部准教授。現在に至る。感性情報処理, Webインテリジェンス, マルチメディア情報処理, 音楽情報処理等のコンテンツ処理研究に従事。電子情報通信学会, 人工知能学会, 日本感性工学学会, 日本音響学会, 芸術科学会, ACM, ACL各会員。



山下 洋一 (正会員)

1984年大阪大学工学研究科電子工学専攻前期課程修了。同年4月大阪大学産業科学研究所文部技官, 1993年1月同助手, 1994年8月同講師, 1997年4月立命館大学理工学部助教授, 2001年4月同教授, 2004年4月同大学情報理工学部教授, 現在に至る。博士(工学)。音声情報処理に関する研究に従事。電子情報通信学会, 日本音響学会, 人工知能学会, ISCA, IEEE各会員。



田中 一星

2020年立命館大学情報理工学部卒業。在学中は, 音声コンテンツからのWebマイニングに関する研究に従事。



井本 桂右

2010年京都大学大学院修士課程修了。同年4月日本電信電話(株)入社。2017年3月総合研究大学院大学博士課程修了。博士(情報学)。2017年4月立命館大学情報理工学部助教。2020年4月同志社大学理工学部准教授, 現在に至る。音響イベント検出, アレイ信号処理に関する研究に従事。電子情報通信学会, 日本音響学会, IEEE各会員。