エンタテインメントコンテンツの ラベル付きデータとしての活用可能性

田中 一星[1,a)]　井本 桂右[1,b)]　山西 良典[1,c)]　山下 洋一[1,d)]

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．本稿では，このうちエンタテインメントメントコンテンツから取得可能な音声データの活用に焦点をあてた．音声の感情認識などではラベル付き音声データを学習することで統計的機械学習モデルを構築する必要があるが，多数話者による大量の音声データを準備することは難しい．本稿では，「じゃがりこ演技力面接」の音声から抽出した特徴量を基に t-SNE 法によって各ラベル間の関係を 2 次元空間に可視化し，エンタテインメントコンテンツ内の音声を研究用のラベル付き音声データとして応用可能であるか考察した．

キーワード：ウェブ音声マイニング，音声情報処理，エンタテインメントの活用

Application Vision of Entertainment Contents for Labeled-data

Abstract: There is a lot of participatory entertainment consisting of varied multimedia on Web. From such entertainment contents, we can acquire visual and audio data under the fixed condition. This paper focuses on the application vision of the speech data from such entertainment contents. Labeled-speech should be prepared to construct the estimation model for emotion and context of the speech, however, it is a hard task to prepare a large amount of speech data spoken by multiple speakers. In this paper, we acquire the amount of labelled speech data from the entertainment content on Web named "Jagarico performed speech interview." The acquired speech data is visualized by using the t-SNE method with acoustic features. Through the discussion of the visualization, we study whether the speeches in the entertainment contents could be effective as the labelled-data for research-use.

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

1. はじめに

「じゃがりこ [*1]」．ウェブ上ではこの菓子商品名を，様々な感情や状況に適した表現方法で発話するエンタテイメントがある．これは，「じゃがりこ演技音声面接[*2]」と題され，表 1 に示した 33 種類の課題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等へ投稿している．このような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．この他にも，「ボー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，ユーザ発信型のエンタテイメントは数多くあり，ニコニコ動画，YouTube，TikTok など様々なウェブサービス上で共有され，楽しまれている．

これらのユーザ参加型のエンタテイメントコンテンツの特徴として，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関わらず，条件や文脈を共有していることが多い．本研究では，これらのエ

[1] 立命館大学
Ritsumeikan University
[a)] is0361hf@ed.ritsumei.ac.jp
[b)] k-imoto@fc.ritsumei.ac.jp
[c)] ryama@media.ritsumei.ac.jp
[d)] yyama@is.ritsumei.ac.jp

[*1] https://www.calbee.co.jp/jagarico/（retrieved on July 25, 2019）
[*2] https://nana-music.com/sounds/008054f0（retrieved on July 25, 2019）

ンタテイメントコンテンツを，統計的機械学習アプローチに用いるラベル付きデータとして利用する可能性を考えた．

1.1 背景

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試みは数多く報告されている．画像情報処理研究に多大な影響を与えた研究としては，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 [2] や PASCAL [3] といった人手で用意されたデータセットに比べて大量の画像をインターネット画像共有サービスである Flickr から収集している．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課題となっている．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いてる [4], [5]．ウェブ画像マイニングの盛り上がりは，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．一方，現在は，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，写真のみならず動画を撮影して共有することが一般ユーザにとってのエンタテイメントとして普及しつつある．以上の背景から，我々は今後インターネット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ラベル付きデータとして活用する可能性，つまり，**ウェブ音声マイニング**および**ウェブモーションマイング**の実現可能性をにらんだ．しかしながら，音声やモーションは画像とは異なって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メントコンテツの特徴が有効に働くと考える．上述のように，ユーザ参加型のエンタテイメントコンテンツでは，複数ユーザが条件や文脈を共有してコンテンツを制作している．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実現されていると考えた．

1.2 本稿の主旨

本稿では，ウェブ音声マイニング実現の端緒として，「じゃがりこ演技音声面接」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検討する．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（例えば，文献 [6] や文献 [7]）が公開されている [8] が，

(1) 特定のプロファイル（年代，性別など）をもつ少数話者が同一文を発話したもの
(2) 話者は多くても異なる文を発話したもの
(3) 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，「じゃがりこ演技音声面接」からラベル付き音声データを取得し，

(1) 様々なプロファイルの多数の発話者
(2) 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
(3) 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もったデータセットの構築を目指す．

発話文となる「じゃがりこ」は，文自身が「感情」や「コンテキスト」を内包しない内容であるため，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，課題となったラベルそのものの特徴として捉えることができる．課題には表現が難しいような複雑なコンテキスト（例えば，「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課題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う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，課題に「普通に」が用意されていることにより，課題「普通に」を基準とした話者毎の正規化も可能であると考えられる．

以下では，音声データの取得方法，音声特徴量の抽出と正規化の方法を述べた後，取得したラベル付き音声データの特徴量空間へのマッピングを考察する．そして，じゃがりこ演技音声面接の動画から取得した音声データを音声の感情やコンテキストの自動認識のために用いるラベル付きデータとしての活用することの有用性について検討する．

2. データの準備

「じゃかりこ面接」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．このとき，音源として楽曲投稿サイト nana に投稿されている “「【part1】演技力じゃがりこ面接」を利用していること”，“録音を 1 人で行なっていること”，“アドリブが少ないこと”，の 3 点を基準として選別した．「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，課題によって「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かった．本研究では，同一発話

表 1　じゃがりこ演技音声面接の課題 33 種類．ID は，「じゃがりこ演技音声面接」での出題順序に従う．

課題 ID i	課題内容	課題 ID i	課題内容	課題 ID i	課題内容
1	嬉しくて	12	暑すぎて	23	必殺技の
2	悲しくて	13	眠たくて	24	恋のビームの
3	怒ってて	14	食べながら	25	2 次元を見て
4	寂しくて	15	告白しながら	26	3 次元を見て
5	嫉妬して	16	感謝して	27	テンション上がって
6	失恋して	17	隣の人に	28	テンション下がって
7	喧嘩して	18	遠くの人に	29	関西風に
8	がっかりして	19	知り合いに	30	関東風に
9	驚いて	20	赤の他人に	31	英語風に
10	疲れ果てて	21	友達に	32	中国風に
11	寒すぎて	22	恋人に	33	普通に

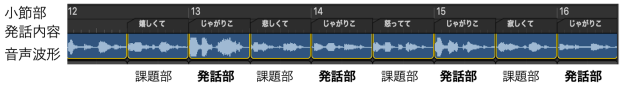


図 1　じゃがりこ演技音声面接から取得した音声データの構造．課題の提示と演技音声の発話が一定間隔で交互に繰り返される．

文での感情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を判断して，除外した．具体的には，検索結果として得られた約 180 動画のうち，動画タイトルからアドリブが多いと推察されるものを除外した約 120 動画を第一著者が視聴した．このうち，上述の 3 点の基準に従って動画を選出したところ，全 15 名（男性 5 名，女性 10 名）の動画が得られた．

取得した動画の音声を各課題ごとに分割する必要がある．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課題とその回答が発話される．楽曲投稿サイト nana では，課題とリズム音楽のみで構成された課題 BGM が用意されている．そこで，音楽編集ソフトのトラック 1 に音声データ，トラック 2 に課題 BGM をそれぞれ入力した．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課題 BGM のテンポを計算したところ，BPM は 135 であった．まず，第一著者が，音楽編集ソフト上で課題 BGM のバスドラムの位置を基準として，目視で音声データのリズムを適合させた．その後，1/2 小節ごとにデータを分割することで，音声データ中の課題部分とその回答を切り分けた．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については，分割後に第一著者が聴取してデータの抽出範囲を手作業で調整した．

3. 分析手法

2 章で得られた各動画の演技音声の発話部分から音響特徴量を抽出する．音響特徴量の抽出には．OpenSMILE [9] にコンフィグファイル IS09_emotion [10] を適用した．表 2 に示した計 16 種類の音響特徴それぞれに対して，表 3 に示

表 2　IS09_emotion.conf で得られる音響特徴．感情の推定に関わる合計 32 種類の音響特徴を取得する．

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 didifferential	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 x つ微分
29	ZRC	ゼロ交差率
30	ZRC 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 3　IS09_emotion.conf で得られる各音響特徴についての統計特徴量．表 1 に示した特徴量について，本表中の各素性値を取得する．

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

表 4 「普通に」を除いた 32 項目を，4 属性（感情，状況，話相手，特殊）に分類.

属性 1	感情	属性 2	状況	属性 3	話相手	属性 4	特殊
課題 ID i	課題内容	課題 ID i	課題内容	課題 ID i	課題内容	課題 ID i	課題内容
1	嬉しくて	6	失恋して	17	隣の人に	23	必殺技の
2	悲しくて	7	喧嘩して	18	遠くの人に	24	恋のビームの
3	怒ってて	10	疲れ果てて	19	知り合いに	29	関西風に
4	寂しくて	11	寒すぎて	20	赤の他人に	30	関東風に
5	嫉妬して	13	眠たくて	21	友達に	31	英語風に
8	がっかりして	14	食べながら	22	恋人に	32	中国風に
9	驚いて	15	告白しながら	25	2 次元を見て		
27	テンション上がって	16	感謝して	26	3 次元を見て		
28	テンション下がって						

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 × 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 $k = p \times m$ とし，音響特徴量を f_k として表す．例えば，3 次元目の MFCC（$p = 7$）についての算術平均（m=6）は f_{42} として示される．

話者 j の課題 i についての発話を s_i^j としたとき，この発話についての f_k は $f_k(s_i^j)$ として表される．式 (1) に従って，$f_k(s_i^j)$ の「普通に」発話からの変化量 $v\_f_k(s_i^j)$ を算出する．

$$v\_f_k(s_i^j) = \begin{cases} f_k(s_i^j)/f_k(s_{33}^j) & (k = 1, \ldots, 48), \\ f_k(s_i^j) - f_k(s_{33}^j) & (others), \end{cases} \quad (1)$$

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．つぎに，式 (2) に従って，$v\_f_k(s_i^j)$ を特徴量 k ごとにすべての課題 i を参照して標準化を行うことで，標準化された「普通に」の発話からの変化量 $sv\_f_k(s_i^j)$ を算出する．

$$sv\_f_k(s_i^j) = \frac{v\_f_k(s_i^j) - \overline{v\_f_k(s^j)}}{\sigma_{j,k}}, \quad (2)$$

ここで，$\sigma_{j,k}$ は話者 j のすべての発話から得られる特徴量 k についての $v\_f_k(s_i^j)$ の分散を示す．これらの処理によってすべての話者の，すべての課題についての発話を比較可能にし，同一空間内での各発話の分布を考察可能にする．

分析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一つである t-SNE 法 [11] を用いる．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 $sv\_f_k(s_i^j)$ に対して t-SNE 法を適用することで，課題毎に発話の集合が形成され，課題間の距離を可視化する．実装には Python の機械学習ライブラリである scikit-learn[*3] を

[*3] https://github.com/scikit-learn/scikit-learn（retrieved on July 25, 2019）

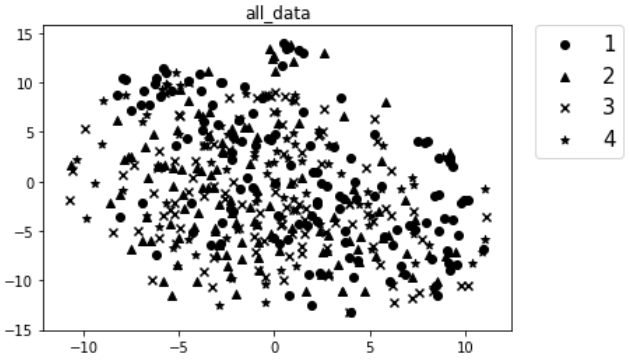


図 2 全課題についての全 15 人分の発話を可視化した結果．図中の凡例は，表 4 に示した属性の番号を示す.

表 5 各属性の可視化に採用した t-SNE のパラメータ値.

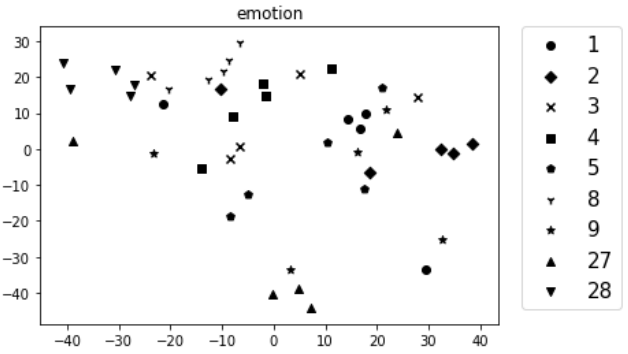
	感情	状況	話し相手	特殊
Perplexity	20	20	30	30
Learning rate	200	200	500	900

使用する．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，空間に対する了解性が最も高いものを第一著者が採用して考察する．なお，「普通に」を除いた 32 課題すべての発話をあらかじめ表 4 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. 分析結果

まず，予備的な分析として，4 属性の 32 課題の発話を全発話データを用いて構築した 1 つの 2 次元空間内に可視化した．このとき，発話のラベルについては属性（感情，状況，話相手，特殊）の 4 種類を用いて，課題の種類ごとの分布に特徴が見られるのかを検討した．つまり，それぞれの属性ごとの発話の集合が確認できるのかを検証した．図 2 に，3 章に示した分析手法に従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，$Perplexity = 50$, $Learning\ rate = 200$ とした．同図から，課題の属性によらず空間内に各発話が分散しており，属性ごとの発話の特徴は確認されなかった．演技発話においては，一般的な発

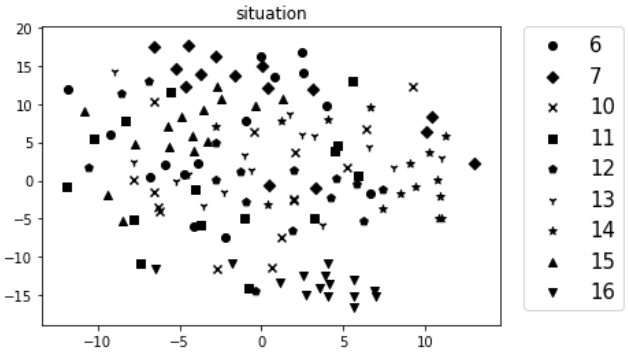
(a) 15 名全員の発話を可視化した結果.



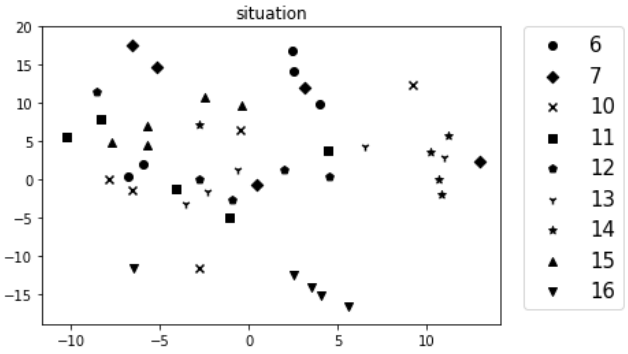
(b) 男性の発話のみを抽出して可視化した結果.

(c) 女性の発話のみを抽出して可視化した結果.

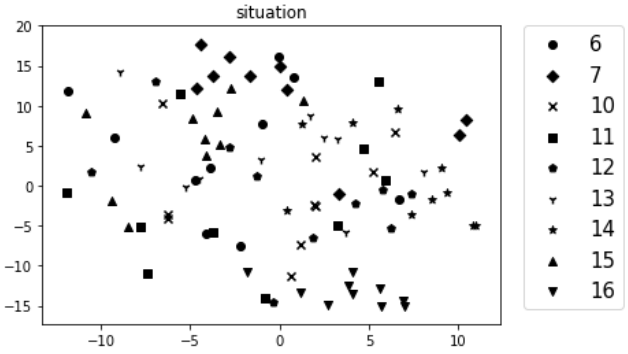
図 3 「感情」の属性についての発話を可視化した結果．図中の凡例は，課題 ID の i の値を示す.



(a) 15 名全員の発話を可視化した結果.



(b) 男性の発話のみを抽出して可視化した結果.



(c) 女性の発話のみを抽出して可視化した結果.

図 4 「状況」の属性についての発話を可視化した結果．図中の凡例は，課題 ID の i の値を示す.

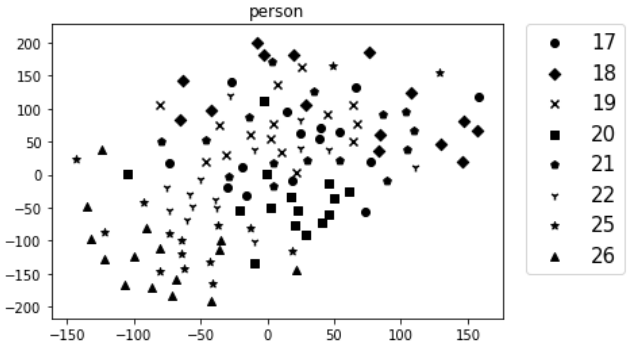
話ラベルに用いられる感情表現と複雑なコンテキストである状況や話相手，特殊といった課題の間には音声特徴量として差異がないことが示唆された.

つぎに，各属性に該当する課題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課題の可視化を行った．図 3 から図 6 に，表 4 に示した属性ごとに出力された可視化結果を示す．このとき，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を採用し，図 3(a)，図 4(a)，図 5(a)，図 6(a) とした．表 5 に，各属性の可視化に採用した t-SNE のパラメータを示す．そして，それらの空間内の男性発話と女声発話をそれぞれ抽出して，「男性のみ」についての可視化空間（図 3(b)，図 4(b)，図 5(b)，図 6(b)）と「女性のみ」についての可視化空間（図 3(c)，図 4(c)，図 5(c)，図 6(c)）をそれぞれ示した.

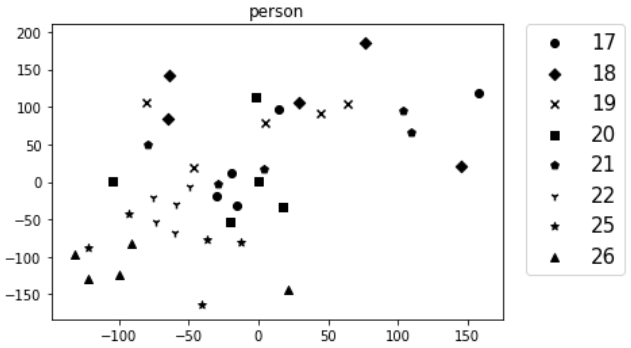
4.1 「感情」についての発話に関する考察

図 3(a) において，それぞれの課題ごとに見てみると，「嬉しくて（$i = 1$）」「寂しくて（$i = 4$）」「嫉妬して（$i = 5$）」「テンション上がって（$i = 27$）」「テンション下がって（$i = 28$）」の発話は密集して分布していることがわかる．課題間で見てみると，「嬉しくて（$i = 1$）」と「テンション上がって（$i = 27$）」についての発話が近い位置に分布しており，その対局位置に「寂しくて（$i = 4$）」「嫉妬して（$i = 5$）」「がっかりして（$i = 8$）」「テンション下がって（$i = 28$）」の発話が集まっていることがわかる.

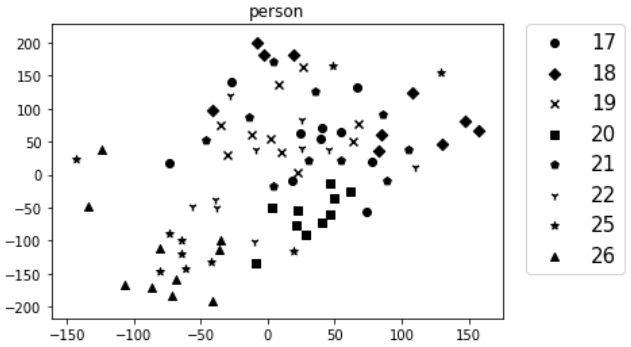
図 3(b) と図 3(c) からは，男性よりも女性の方が課題毎に発話が特徴量空間内で近い領域内に集まっていることが



(a) 15 名全員の発話を可視化した結果.

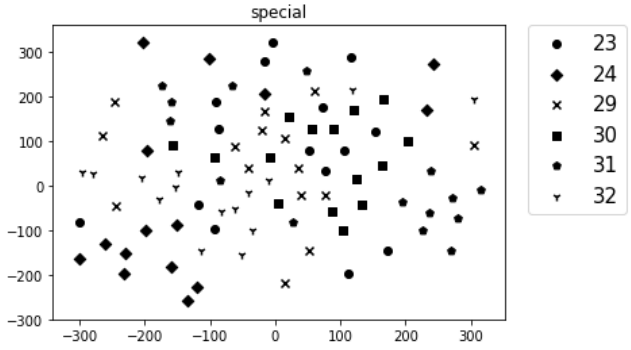


(b) 男性の発話のみを抽出して可視化した結果.

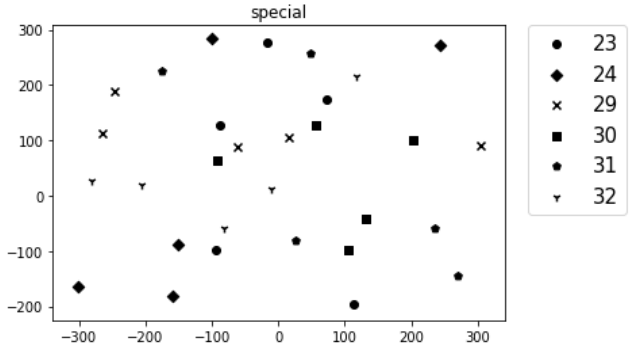


(c) 女性の発話のみを抽出して可視化した結果.

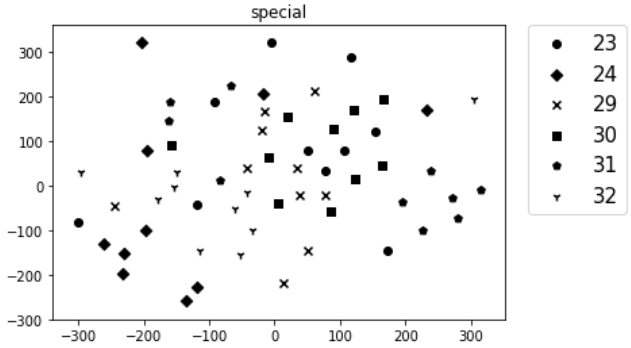
図 5 「話相手」の属性についての発話を可視化した結果．図中の凡例は，課題 ID の i の値を示す.



(a) 15 名全員の発話を可視化した結果.



(b) 男性の発話のみを抽出して可視化した結果.



(c) 女性の発話のみを抽出して可視化した結果.

図 6 「特殊」の属性についての発話を可視化した結果．図中の凡例は，課題 ID の i の値を示す.

わかる．課題間の位置関係を見てみても，女性のみの発話についての可視化結果（図 3(c)）では，意味的に類似した感情である「嬉しくて（$i = 1$)」と「テンション上がって（$i = 27$)」が近い位置に分布しているのに対し，男性のみの発話についての可視化結果では「嬉しくて（$i = 1$)」の分布は「テンション上がって（$i = 27$)」とは離れた位置に分布していることがわかる.

4.2 「状況」についての発話に関する考察

属性「状況」についての発話を可視化した図 4(a) では，「喧嘩して（$i = 7$)」「食べながら（$i = 14$)」「告白しながら（$i = 15$)」「感謝して（$i = 16$)」それぞれについての発話が比較的近い位置に集まっていた．課題間の位置関係に着目すると，「喧嘩して（$i = 7$)」「告白しながら（$i = 15$)」は近い空間に分布しており，その反対側に「感謝して（$i = 16$)」の発話が分布するという結果になった.

図 4(b) と図 4(c) を見ると，「感謝して（$i = 16$)」「食べながら（$i = 14$)」の発話は男女ともに近い位置に分布していた．また，「喧嘩して（$i = 7$)」と「告白しながら（$i = 15$)」は男女で発話の分散に差異がみられるものの，同様の範囲に位置していた．「眠たくて（$i = 13$)」の発話については，男性のみの発話では中央に分布しているのに対して，女性のみの発話では空間内に分散していることが見てとれる.

4.3 「話相手」についての発話に関する考察

図 5(a) に示した「話相手」についての発話では，「赤の他人に（$i = 20$)」「恋人に（$i = 22$)」「2 次元を見て（$i = 25$)」

「3 次元を見て（$i = 26$）」の発話が近い位置に分布していた．課題間には，極端な分布の特徴は見当たらなかったが，「2 次元を見て（$i = 25$）」「3 次元を見て（$i = 26$）」「恋人に（$i = 22$）」「知り合いに（$i = 19$）」の順に左下から中央上にかけて分布していた．

図 5(b) と図 5(c) の男女別の可視化結果を比較すると，「赤の他人に（$i = 20$）」の発話が男性のみの分析結果では分散が大きかったのに対して，女性のみの分析結果方では空間内で比較的近い領域に集中していることがわかる．また，「恋人に（$i = 22$）」については，男性のみの発話では中央左下のみに分布しているが，女性のみの発話では中央左下と中央付近と 2 種類のクラスタが形成されていることが伺える．

4.4 「特殊」についての発話に関する考察

図 6(a) に示した属性「特殊」に関する発話では，「恋のビームの（$i = 24$）」「英語風に（$i = 31$）」「中国風に（$i = 32$）」の発話が，上述の他属性の可視化結果に比べれば広範囲への分散となるが，比較的集中して分布していた．課題間では，「関西風に（$i = 29$）」「関東風に（$i = 30$）」が近い位置に分布しており，「英語風に（$i = 31$）」と「中国風に中国風に（$i = 32$）」は離れた位置に分布していた．このことから，日本語の方言についての演技は類似した傾向があるのに対して，言語の違いについては異なる音声特徴をもつ演技が発話されていることが示唆される．

男女の発話について，図 6(b) と図 6(c) を比較してみると，男性に比べて女性の方が課題ごとに比較的近い位置に発話が集中していると見られる．「特殊」に分類した課題は，曖昧で演技難易度が高い課題が多く含まれているが，女性の発話は一定の領域に発話が集中しており，発話者同士で課題に対して共通了解性を有していたと推察される．

4.5 全体の考察

可視化結果において，発話同士の距離が近いということは，課題に対する演技の表現方法が類似していると考えることができる．また，同一の課題についての発話が空間内で集中しているほど話者間の表現方法に差異が小さく，散らばりが大きいほど話者間で表現方法の差異が大きいと捉えられる．発話が集合している課題としては「嬉しくて（$i = 1$）」「寂しくて（$i = 4$）」「嫉妬して（$i = 5$）」「テンション上がって（$i = 27$）」「テンション下がって（$i = 28$）」「喧嘩して（$i = 7$）」「食べながら（$i = 14$）」「告白しながら（$i = 15$）」「感謝して（$i = 16$）」「赤の他人に（$i = 20$）」「恋人に（$i = 22$）」「2 次元を見て（$i = 25$）」「3 次元を見て（$i = 26$）」が該当する．特筆すべきは，一般的な音声コーパスに収録されている感情表現のみならず，状況や話相手といった複雑なコンテキストを課題とした発話に対しても，話者間で類似した音声表現がされているということが音響特徴量空間の考察から明らかになったことである．

課題間の距離は，課題ごとの表現方法の違いと考えられる．特に感情については「嬉しくて（$i = 1$）」や「テンション上がって（$i = 27$）」といったどちらもポジティブな感情についての発話は近い位置に分布しており，「寂しくて（$i = 4$）」「嫉妬して（$i = 5$）」「テンション下がって（$i = 28$）」といったネガティブな感情も近い位置に分布している．また，これらのポジティブとネガティブな課題群同士は離れて位置しており，Russell の円環モデル [12] における Arousal と Valence で構成される空間と類似した t-SNE 空間が得られたと考えられる．

演技音声のコンテキスト認識への応用を考えた場合，発話の課題内分散は認識率，課題間距離は分類タスクの難易度に影響すると考えられる．今後は，本稿で得られたデータを拡充するとともに，発話のコンテキスト分類モデルを構築し，その認識精度を参照とした客観的な考察を行っていく．このとき，「じゃがりこ」以外の音声を入力とした発話オープンな認識についても実験し，その性能を考察する．また，既存の感情音声コーパスを用いて構築された音声の感情分類モデルと「じゃがりこ演技音声面接」の発話を用いて構築された感情認識モデルとの性能比較も行っていく．

5. おわりに

本稿では，ウェブ音声マイニング実現のための基礎検討として，「じゃがりこ演技音声面接」における各課題に対する複数話者の発話の音響特徴を分析し，t-SNE 手法によって構成された可視化結果を考察した．考察の結果，課題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．特徴量空間に分布した課題ごとに考察すると，感情モデルで定義される Arousal や Valence に相当する平面も確認された．以上から，「じゃがりこ演技音声面接」でウェブ上に公開された音声を，感情音声ラベルが付与された音声データとしての応用できる可能性が示唆された．また，特徴量空間を話者の性別ごとに考察したところ，同一の課題に対して女性話者の発話は男性話者の発話に対して，比較的狭い領域に分布していることが確認され，男性と女性の感情音声の演技の傾向についても示唆的な結果が得られた．今後は，本稿で得られた音声データを学習データとした発話の感情認識モデルを構築し，感情音声の認識精度などの客観的な指標に基づいた考察を行う．このとき，「じゃがりこ」以外の発話を認識モデルへの入力とした，発話オープンな感情認識精度についても検証していく．

エンタテイメントの活用の観点からは，エンタテイメントコンテンツを統計的機械学習アプローチに用いるラベル付きデータとして応用可能であることが示唆された．つまり，エンタテイメントコンテンツの Human Compu-

tation [13] への応用可能性を示したこととなる．従来の Crowdsourcing [14] の多くでは，賃金 [15] や情報 [16] 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，“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のラベル付きデータが生成されることになると期待される．このようなユーザ参加型のエンタテイメントコンテンツのデザインについては，今後，エンタテイ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扱っていくことを提起したい．

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. on Computer Vision and Pattern Recognition, CVPR*, pp. 248–255 (2009).

[2] Griffin, G., Holub, A. and Perona, P.: Caltech-256 Object Category Dataset, Technical Report 7694, California Institute of Technology (2007).

[3] 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).

[4] 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. of the ACM Int'l Conf. on Image and Video Retrieval*, pp. 48:1–48:9 (2009).

[5] Fergus, R., Perona, P. and Zisserman, A.: A Visual Category Filter for Google Images, *Proc. of European Conf. on Computer Vision, ECCV*, pp. 242–256 (2004).

[6] Arimoto, Y., Kawatsu, H., Ohno, S. and Iida, H.: Emotion recognition in spontaneous emotional speech for anonymity-protected voice chat systems, *Proc. of the Annu Conf. of the Int'l Speech Communication Association, INTERSPEECH*, pp. 322–325 (2008).

[7] 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).

[8] NII: 音声資源コンソーシアム，`http://research.nii.ac.jp/src/`. (retrieved on July 25, 2019).

[9] Eyben, F., Wöllmer, M. and Schuller, B.: openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor, *Proc. of ACM Multimedia*, pp. 1459–1462 (2010).

[10] Schuller, B., Steidl, S. and Batliner, A.: Emotion recognition in spontaneous emotional speech for anonymity-protected voice chat systems, *Proc. of the Annu Conf. of the Int'l Speech Communication Association, INTERSPEECH*, pp. 312–315 (2009).

[11] van der Maaten, L. and Hinton, G.: Visualizing Data using t-SNE, *Journal of Machine Learning Research*, Vol. 9, pp. 2579–2605 (2008).

[12] Russell, J. A.: A Circumplex Model of Affect, *Journal of Personality and Social Psychology*, Vol. 39, No. 6, pp. 1161–1178 (1980).

[13] Quinn, A. J. and Bederson, B. B.: Human computation: a survey and taxonomy of a growing field, *Proc. of CHI 2011*, pp. 1403–1412 (2011).

[14] Howe, J.: The rise of crowdsourcing, *Wired magazine*, Vol. 14, No. 6, pp. 1–4 (2006).

[15] amazon: amazon mechanical turk, `https://www.mturk.com/`. (retrieved on July 25, 2019).

[16] Google: reCAPTCHA, `https://www.google.com/recaptcha/intro/v3.html`. (retrieved on July 25, 2019).