パラ言語情報によるラベリングに基づく音声合成用の演技音声の検索システム

吉田 聡平
Sohei Yoshida
法政大学情報科学部ディジタルメディア学科
sohei.yoshida.6e@stu.hosei.ac.jp

Abstract

When using synthesized speech in media content, the element based on the user's impression is important. For example, from comic's automatic dubbing, it is necessary to select the voice that fits the impression of the character. Elements that determine the impression of the character include ones that isn't directly related to the voice. Therefore, in this research, the applicability of multi-labeling by paralinguistic information for automatic selection of acted voice is verified. For that purpose, experiment assuming to automate retrieval of voice which becomes material for comic's automatic dubbing database of multi-labeling acted voices. The content of the experiment is a listening experiment in which they search acted voice imagined for specific lines of comic by inputting multi-labels. In this time, learning by GMM and performance evaluation before the experiment were conducted. As a result, the acted voice used for training data was highly likelihood for GMM.

1 音声合成による漫画の自動アフレコ

近年、音声合成による自動音声読み上げが様々な場面で使われるようになっている。現在は、スマートフォンの対話アプリや、公共施設の案内音声など、文字情報以外の情報を必要としない場面での利用が多いが、今後メディアコンテンツなどでの利用などを考えてゆく上では、ユーザーが聞いた印象に関する要素が重要となってくる。例えば、音声合成を用いてまだセリフがあてられていない漫画に自動でセリフをあてることを考えるとき、キャラクターのイメージに合う音声が必要となる。キャラクターのイメージを決定する要素には、性格やポーズ、髪色や身長などの身体的特徴、背景や状況など様々なものがあり、その中には声が直接関係の無いものもある。例えば、治部らの研究ではキャラクターのジェスチャーが与える印象について述べられている [1]。また、松井らの研究では、キャラクターの顔によって与えられるイメージについて述べられている [2]。

そこで本研究では、Similarity Search of Acted Voices for Automatic Voice Casting[3] という論文で提案された、声の類似性の知覚におけるパラ言語内容の重要性に基づき、パラ言語情報による演技音声へのマルチラベルスコアリングの、演技音声の自動選択への応用可能性を検証する。今回は漫画を音声合成によってフルボイス化する際の元となる音声の選択を自動化することを想定し実験を行う。パラ言語情報とは、リズム、ポーズ、声質といった発声と同時に伝わる文字情報以外の情報のことを指す。なお、実際に音声の選択を行う際には、漫画に登場する同一登場人物は別のコマであっても同じ声でなければならないが、今回は前段階として、ひとつのコマそれぞれに対しての適切な演技音声の検索を行う。本研究で演技音声は、アニメやゲームにおいて登場するキャラクターの発言として発せられる音声と定義する。

2 演技音声の知覚におけるパラ言語内容

本研究の先行研究として、Nicolas Obin, Axel Roebel による Similarity Search of Acted Voices for Automatic Voice Casting[3] が挙げられる。音声波に含まれる声道特性を抽出して認識する話者認識技術を適用すると、音響空間における声の類似度を直接的に測定できるが、この類似性尺度は音響空間全体で測定してしまうため、声の類似性の認識を反映しているかは分からない。この研究では、各国で利用な映画やビデオゲームなどのマルチメディアコンテンツの吹き替えの際のキャスティングを自動化するという観点から、音響による類似度とパラ言語による類似度どちらが優れているかという検証を行っている。ビデオゲーム「MassEffect3」のアメリカ英語版をもとの言語、フランス語版を翻訳先の言語とし、もとの言語の音声と翻訳先の言語の音声から、混合ガウスモデルベースの音響モデルによるスコアリングと、パラ言語情報をラベルとして複数付与したマルチラベルによるスコアリングによって類似した音声を探した。評価ではアンケートを用いてどちらがより人間の類似性の知覚に近い結果を得られるかを検証した。アンケートは 30 人のフランス人によって行われ、とても似ている (-2)、似ている (+1)、どちらとも言えない (0)、似ていない (-1)、非常に似ていない (-2) の 5 段階評価の平均をとった結果は表 1 のようになった。

表 1. 先行研究結果

手法	平均値の 95 %信頼区間
マルチラベル	0.75 ∼0.88
音響モデル	-0.03 ∼0.11

この結果から、マルチラベルスコアリングは音響スコアリングより優れており、声のカテゴリ (話者の特性および状態) への抽象化が、演技音声の類似性の人間の知覚において重要な役割を果たす、という結果が得られた。

3 パラ言語内容によるラベリングに基づく演技音声検索

本研究では、先行研究にて音声間の認識の類似度を求めるのに用いられたマルチラベル分類を、元の声が存在しないような状況でも応用できるかを検証する。アニメやゲームに実際に使われている演技音声をデータベースとし、声の当てられていな

Supervisor: Prof. Katunobu Itou

いキャラクターやセリフのイメージからマルチラベルを選択して似たマルチラベルが付与されている演技音声を検索、その結果が入力したユーザーのイメージと合っていれば応用可能であると判断できる。

3.1 演技音声の収集とマルチラベリング

実験には、実際にアニメやゲームで使われている演技音声を利用する。同一キャラクターのセリフでもシチュエーションによってさまざまなラベルが付与できるため、1 キャラクターあたり複数の種類のセリフを収集する。なお、今回は実験を簡単にするため、キャラクターによって一意に定まる性別は片側 (今回は男性キャラクター) のみに絞った。主に収集した作品はスマートフォンアプリゲーム 3 作品、ブラウザゲーム 1 作品、ドラマ CD1 作品である。内訳を表 2 に示す。

表 2. 収集作品例

作品名	キャラクター数	音声数
アイ★チュウ	32 人	812 個
あんさんぶるスターズ	20 人	2950 個
刀剣乱舞	42 人	574 個
ヒプノシスマイク	12 人	456 個
Fate Grand Order	12 人	403 個
合計	118 人	4487 個

なお、音声の収集方法に関しては、スマートフォンアプリゲームは iPhone の画面録画機能を使い録音したのちパソコンに取り込んで wav ファイルへ変換する方法で、ブラウザゲームはプレイをしながらシステム音を直接 wav ファイルとして録音する方法で、ドラマ CD は wma でパソコンに取り込んだのち wav ファイルへ変換する方法で行った。録音された音声は無音区間の検出によって自動で切り分けた。切り分けた音声から再生時間が 2 秒に満たない音声と、効果音や BGM が被っていてデータベースとして適さない音源は手動で取り除いた。

3.2 ラベルの種類

パラ言語情報のラベルはカテゴリごとにより当てはまるものを選択する形で付与する。先行研究は演技音声を聞いた上で用いられる形容表現となっているが、今回は漫画のセリフを基準とした形容表現を扱う必要がある。そこで声の無い漫画のセリフを基準としたラベルを作成する。使用するラベルは熊本らの印象に基づく検索のための印象語選定法 [4] の提案に従う。また今回、カテゴリごとのラベルは反対の意味を持つ言葉によって構成される。具体的な手順を以下に示す。

1. 男性キャラクターが喋っている漫画のコマを用意する。今回はスマートフォンの無料漫画アプリ GANMA で読むことができる作品を対象とした。人気ランキングの上位 20 位のうち、過度に性的な作品と男性キャラクターの登場しない作品を除いた 15 作品から様々な状況のコマを選択し、それぞれ 1〜15 までの番号を振った。使用した画像を図 1 に示す。
2. 研究内容を知らない人間に対し聴取実験を行う。今回は漫画のコマとともに「以下の 1〜15 の 15 個の漫画コマそれぞれに声優によって声をあてることを考えるとき、どのような演技音声を想定するかを、1〜15 それぞれについて自由に記述してください。」という文章を送付した。熊本らの研究では回答者数が多かったため質問は各々に対し 2 つずつであったが、今回は聴取実験の対象者が限られるため 15 個のコマそれぞれに対し回答をさせることで母数を確保した。対象者は 20 代の男性 4 人、20 代の女性が 4 人の計 8 人だった。
3. 自由記述の集計結果から印象語を抽出する。例えば、回答が「悲しんで声を震わせながら大声で怒りをぶつけるように」だった場合は「悲しい」「声を震わせながら」「大声」「怒りをぶつけるような」という印象語が抽出される。今回は全部で 156 種類の印象語を得た。
4. 熊本らの研究に従い同義語、反義語、類義語を定義し、言語データから各印象語の出現頻度を求める。各印象語は自分より出現頻度が低く、かつ自分との共起頻度が tv_1 未満である同義語および反義語とグループを構成する。各グループは出現頻度が低く、かつ自分どの共起頻度が tv_2 未満である類義語を自グループに加える。ここで共起頻度は、ある特定の形容表現の組について一つのコマに対する回答の中で同時に出現した回数とする。グループの数が規定数に達するまでこれを繰り返し、出現頻度の一番高いグループを選定する。形容表現の組み合わせは 229 通りあり、各表記頻度の該当組は表 3 のようになった。よって今回は実験的に、tv_1=2,tv_2=2, グループの数を 10 とした。

図 1. アンケートに使用した画像

表 3. 共起頻度

共起頻度	該当数	組み合わせ例
頻度 3	3	「焦り」と「早口」など
頻度 2	8	「呆れた」と「テンションが低い」など
頻度 1	218	「明るい」と「優しい」など

今回は表 4 のようなラベルが選定された。括弧内は出現頻度を表す。なお、熊本らの研究では SD 法を想定しているため対義語を同じグループとして扱っているが、今回は検索に用いるラベルのため、対義語としてグループ化されたものも形容表現として選出した。

3.3 ラベルの学習

演技音声へのラベルの付与は、一部を手動にて行い、それをもとに学習をすることによって残りの演技音声にも付与を行う。ラベルの学習は Lie Lu らによる Automatic mood detection and tracking of music audio signals[5] で用いられた手法に従い、混合ガウスモデル (Gaussian Mixture Model) を用いて行う。以下、混合ガウスモデルを GMM と呼ぶ。GMM の確率密度関数は式 1 で表される。

$$p(x|\lambda) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \sum k) \quad (1)$$

表 4. 選定されたラベル

ラベル	形容表現例 (類義語 及び 対義語)
声が高い (23) 声が低い	声が高い (23) 声が低い (9)
声が大きい (19) 声が小さい	小さい (4) 小さい (4)
優しい (18)	優しい (10), 柔らかい (2), 冷淡な (1)
焦り (18)	焦り (8), 落ち着いた (6)
怒り (16)	怒り (10), 苛立っている (1)
喜び (16) 悲しみ	喜び (6), 嬉しそうな (3) 悲しい (3)
早口な (12) ゆっくりと	早口な (6), 早い (2)) ゆっくりと (4)
強い (11) 弱い	強い (9), 力強い (1)
テンションが低い (8) テンションが高い	テンションが低い (5) 軽快な (1)
呆れた (7)	呆れた (7)

これは、多次元正規分布の確率密度関数に混合比をかけ、和をとったものである。ここで、p 次元の多次元正規分布の確率密度関数は式 2 で表される。

$$N(x|\mu_k, \sum k)$$

$$=\frac{1}{(2\pi)^p/2|\sum k|^1/2}exp-\frac{1}{2}(x-\mu_k)^T\sum_k^{-1}(x-\mu_k) \tag{2}$$

また、ここで π_k は混合比を表す。

各ラベルに該当する演技音声をデータベースの中から 30 個〜50 個抽出し、GMM を用いてモデルを形成する。演技音声は通し番号で管理され、GMM によって形成されたモデルから算出される、各ラベルに対する尤度を保持する。GMM に適用する音響特徴量には、平賀らによる主観評価に基づく楽曲間類似度算出モデル [6] や、松澤らによる自然対話中の発話対における音響特徴量に基づく感情の程度推定 [7] などを参考に、メル周波数ケプストラム係数 (MFCC)、F0 平均、F0 範囲、パワー平均、ラウドネス、スペクトル重心、話速を使用した。混合数は実験的に 6 とした。

3.3.1 メル周波数ケプストラム (MFCC)

人間の音声知覚の特徴を考慮した声道特性を表す特徴量である。

3.3.2 F0 平均・F0 範囲

音声の音程の高さを表す特徴量である。今回は自己相関を用いて測定を行っている。

3.3.3 パワー平均

音声の音量に関する特徴量である。以下の式によって表される。

$$p(k) = log(\sum_{t=kT}^{(k+1)T-1} x(t)^2) \tag{3}$$

3.3.4 ラウドネス

音声の大きさを表す特徴量である。

3.3.5 スペクトル重心

音声のスペクトルの形状を表す特徴量のひとつ。以下の式によって表される。

$$f_c(t) = \frac{\sum_{f=0}^{f_0/2} X(t,f)\cdot f}{\sum_{f=0}^{f_0/2} X(t,f)} \tag{4}$$

3.3.6 話速

発話の速さを表す特徴量である。単位時間当たりのピークの数によって表される。

4 システムの実装と評価

検索システムの評価は、再現率を求める実験と適合率を求める実験の 2 つの実験によって行う。以下に評価に使った画像 1 と画像 2、および画像 3 と画像 4 を示す。

図 2. 評価に使用した画像

4.1 再現率

再現率を求めるには、全データの中にどれだけ正解が含まれているかという情報が必要となる。しかし、4487 個の全ての演技音声から正解数を数えて算出するのは困難である。そこで Introduction to Information Retrieval[8] の pooling を参考に、演技音声全体の 5 %をチェックして正解の割合を推定し、全体の正解数だけシステムで演技音声を検索して正解率を求める、という手法を取った。以下がその結果である。画像 1,3,4 が再現率 50 %以上となっている中で、画像 2 のみ再現率が 0 %となってしまっている。これは画像 2 から得られる主な印象が「焦り」であるにも関わらず、形容表現全てに対して尤度の高い音声を検索しているため、「焦り」の尤度が高い音声が検索されなかったためと考えられる。

表 5. 再現率

画像番号	推定個数	正解音声数	再現率 (%)
1	80	50	62.5
2	40	0	0.00
3	60	42	70.0
4	80	45	56.3

4.2 適合率

適合率は主観評価実験によって評価する。ユーザーは各ラベルについて当てはまると思ったら Yes、当てはまらないと思ったら No を選択する。システムは各ラベルに重み付けを行って検索をし、予め用意しておいたマルチラベル付きの演技音声をを尤度が高い順に表示する。重み付けに関しては、コマごとの印象に大きな差が出るものを重視することを想定する。形容表現の選定と同様にコマの画像を用意し、SD 法に基づき各ラベ

ルに対して当てはまらない、あまり当てはまらない、どちらでもない、少し当てはまる、当てはまるの 5 段階でアンケートをとり、その分散を用いて行った。表 4.2 にそれぞれの重みを示す。

表 6. ラベルごとの重み

ラベル	重み
声が高い	1.50
声が低い	1.50
声が大きい	1.79
声が小さい	1.66
優しい	1.60
焦り	1.55
怒り	1.62
喜び	1.70
悲しみ	1.60
早口な	1.45
ゆっくりと	1.47
強い	1.42
弱い	1.55
テンションが低い	1.66
テンションが高い	1.70
呆れた	1.79

評価者はメディアコンテンツの利用者にあたる人物 8 人で行う。以下に手順を示す。

1. まず評価者にまだ声が当てられていない漫画のコマを見せる。これはは元の声が存在しない状況を想定するためである。
2. その中に登場する登場人物のセリフを指定し、選定したラベルに基づき特徴を Yes か No で入力させる。
3. 評価者の入力に基づき、データベースの中からシステムが最もキャラクターのセリフとあっていると判断した演技音声を出力する。
4. 出力された音声がイメージした演技音声と合っていると思うかを、評価者にアンケート形式で解答させる。なお、先行研究では-2〜+2 の五段階評価であったが、今回は検索性能の検証であるため、「あっている」もしくは「あっていない」の二択とした。このとき対照実験として、データベースの中から無作為に選んだ演技音声と同時に評価させる。

その結果、システムによって尤度が高いと判断された上位 5 位の演技音声に適するものを含むのは 16 件中 11 件で適合率は 32.5 %、無作為に選出された演技音声に適するものを含むのは 16 件中 4 件で適合率は 5.0 %であった。以下に詳細を示す。適合率は 32.5 %と低いものの、無作為な選出よりははるかに良い結果であり、また上位 5 件の中に適切な音声が含まれている割合も高いため、検索システムとしてはある程度よい結果といえる。

4.3 考察

検索システムとして有用である一方で、再現率・適合率どちらも極端に低くなってしまう場合がある。これは「怒り」や「焦り」や「悲しみ」など、 必須であるラベルが反映されていないため起こると考えられる。その根拠として、主観評価実験においてシステムによって選出した音声をすべて不適としたユーザーから、画像 1 に対し「“声が高い”を選択したのに全て低く

表 7. 適合率

被験者番号	画像番号	システム適合率 (%)	無作為適合率 (%)
1	1	60	0
1	2	0	0
2	1	60	0
2	2	0	0
3	1	0	0
3	2	20	0
4	1	40	0
4	2	0	0
5	3	60	0
5	4	80	20
6	3	40	0
6	4	0	20
7	3	40	0
7	4	20	20
8	3	40	0
8	4	60	20

感じた」という意見や、画像 2 に対して「焦っている音声がなかった」という意見が寄せられた。よって、精度を向上させるためには入力するユーザーに合った重み付けが必要といえる。

5 結論

パラ言語情報によるラベリングに基づく音声合成用の演技音声の検索システムを作成した。精度は無作為選出より著しく高く、検索システムとしての機能は果たすと言える。更なる精度向上のためには、ユーザーによるラベルの重み付けを検討する必要がある。

参考文献

[1] 治部 晶子 ; Dilokwattanakoon kamolphan ; , 木谷 庸二, “パッケージ上のキャラクターのジェスチャーが製品及びキャラクターの印象に及ぼす影響の研究”, 日本デザイン学会研究発表大会概要集 64(0), 182, 2017

[2] 松井 哲也“キャラクターの顔イメージの形成”, JSAI 大会論文集 JSAI2014(0), 1E4OS23a3-1E4OS23a3, 2014

[3] Nicolas Obin ; Axel Roebel“Similarity Search of Acted Voices for Automatic Voice Casting”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, Volume: 24, Issue: 9, pp1642-1651.

[4] 熊本 忠彦 ; 太田 公子 , “印象に基づく検索のための印象語選定法の提案”, 情報処理学会論文誌 44(7), 1808-1811, 2003-07-15

[5] Lie Lu ; D. Liu ; Hong-Jiang Zhang “Automatic mood detection and tracking of music audio signals”, IEEE Transactions on Audio, Speech, and Language Processing Volume: 14 , Issue: 1 , Jan. 2006

[6] 平賀 悠介 ; 大石 康智 ; 武田 一哉 , “主観評価に基づく楽曲間類似度算出モデル”, 研究報告音楽情報科学（MUS）2009-MUS-81(2), 1-6, 2009-07-22

[7] 松澤直之 ; 政倉祐子 ; 大野澄雄 , “自然対話中の発話対における音響特徴量に基づく感情の程度推定”, 第 75 回全国大会講演論文集 2013(1), 503-504, 2013-03-06