# 英語話者の感情を保持した日本語吹替音声合成の検討

Japanese-dubbed speech synthesis that preserves the emotions of English language speakers

田中 貴裕
Takahiro Tanaka

岡山大学 阿部研究室
Abe Laboratory, Okayama University

**概要** 本研究では，映画・動画等の映像のうち英語音声で収録されたものに対し，日本語吹替音声を合成する際に，英語話者の感情を自然に反映させることを目指す．本報告では，英語と日本語における感情を x-vector として抽出し，t-SNE により可視化した結果，日本語と英語で明確な違いがあることが分かった．また，ランダムフォレストによる英語感情音声認識器は学習データを発話テキスト単位で切り出した場合と，話者単位で切り出した場合では，後者で学習させた場合の精度の方が男声:0.527，女声:0.320 高いことが分かった．

## 1 はじめに

既に一部の動画サイト等では，外国語の音声を認識し自動翻訳を行ったものを字幕として表示する機能がある．これを使用して吹替音声を作る方法の一つとして，合成音声によって吹替を作成することが考えられる．このとき，字幕の文字情報と組み合わせてより自然な音声を合成する方法として，元言語の発話音声に含まれる感情を認識し，吹替後にも反映させるということを考えた．具体的な手順としては，(1) 英語音声の感情認識，(2) 英語音声の感情から日本語の感情への変換 を検討している．感情の変換は，感情発話の埋め込みベクトルの変換を想定しており，日本語 TTS モデルを埋め込みベクトルにより条件付けする方法 [1] で音声合成を行うことを検討している．

本報告では，(1) に関連して，英語と日本語の音声感情データセットを用いて感情認識を行った結果を述べ，(2) に関連して，同じ感情での日本語・英語間の差について調べた結果を述べる．

## 2 感情音声データセット

実験に用いる感情音声データセットとして，日本語は JTES[2]，英語は ESD[3] を用いた．JTES は男女各 50 名の話者の喜び・怒り・悲しみ・平常の各 50 文を計 100 名が発話した 20,000 発話（23.5 時間）の日本語データセットであり，収録は 16kHz，16bit で行われている．ESD は英語話者男女各 5 名と中国語話者男女各 5 名の喜び・怒り・悲しみ・驚き・平常の各感情 350 文の音声で構成されたデータセットであり，収録は 16kHz，16bit で行われている．なお，今回は ESD データセットのうち英語話者の発話のみを用いた．

実験では 2 つのデータセット間で共通する感情：喜び（joy）・怒り（angry）・悲しみ（sad）・平常（neutral）で分類を行う．まず話者の中から男女各 5 名を選び，各話者で 4 感情の発話の中から感情ごとに 50 発話を取り出した $5 \times 4 \times 50 = 1000$ 発話を男女それぞれ選択する．ただし，全話者で取り出す発話テキストのバリエーションは共通とする．その中で発話内容による影響を評価するために，(1) 5 名の話者の発話の各感情の 50 発話を学習データとテストデータで 40:10 の比率で分割する．次に，話者個人の影響を評価するために，(2) 5 名の話者に割り当てられた発話のうち，4 名分を学習データに，1 名分をテストデータとして分割を行う．

## 3 感情認識器

### 3.1 学習方法

本実験では感情認識器として，機械学習のアルゴリズムの一つであるランダムフォレストを用いた．これはランダムにサンプリングしたデータで学習した複数の決定木の出力から多数決で最終的な出力を決めるアンサンブル学習のアルゴリズムである．決定木の最大の深さは 4 とした．

音声データから抽出した特徴量を用いて学習を行うが，これには音声ツールキット SpeechBrain[4] を用いて，VoxCeleb[5] を ECAPA-TDNN[6] で学習したモデルで抽出した x-vector を用いる．x-vector は 192 次元のベクトルで，認識器は x-vector を入力として受け取り，出力として 4 感情に対応した {0, 1, 2, 3} のいずれかの感情 ID を得る．最終的には学習モデルが正解感情 ID を出力できるように学習を行い，2 章で示した (1)，(2) の異なる指標で 5 分割したデータに対し，Cross Validation により認識精度の平均値を求め，言語ごとにどのような違いが生じるかを見る．

### 3.2 実験結果

表 1: 感情認識精度

| | ESD | | JTES | |
|---|---|---|---|---|
| 分割方法 | male | female | male | female |
| (1) 発話テキスト | 0.331 | 0.549 | 0.768 | 0.729 |
| (2) 話者 | 0.819 | 0.869 | 0.422 | 0.484 |

各データセットを 2 つの方法で分割したデータ学習

したモデルの認識精度を表 1 に示す．ESD は，話者でデータを分割した方が精度が高い．つまり，言語的特徴が同じ発話であれば，未知の話者に対しても認識がしやすく，同じ話者であっても，未知のテキストの感情を識別するのは難しいと考えられる．一方で，JTES は発話テキストで分割した方が精度が高く，同じ話者であれば未知のテキストであっても高い精度で識別が可能だが，未知の話者に対しては既に学習済みのテキストの発話であっても認識精度が低くなると言える．

## 4　言語間の感情の比較

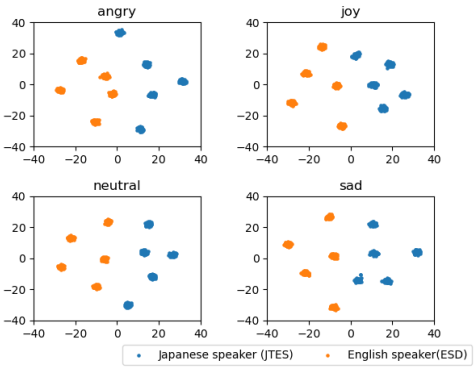


図 1: t-SNE による感情ごとの比較（男声）

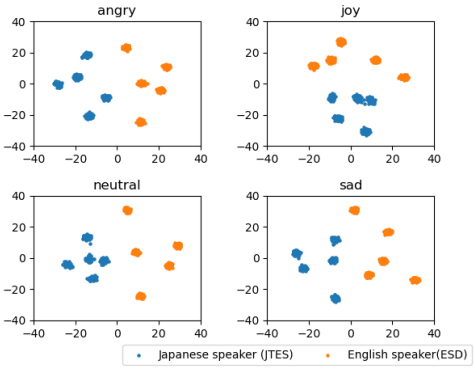


図 2: t-SNE による感情ごとの比較（女声）

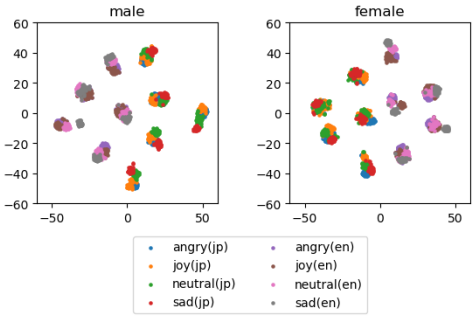


図 3: t-SNE による全感情の比較

次に，言語間の同一感情間での違いを調べる．実験に用いた音声を男声，女声に分け，x-vector に変換し，感情ごとに t-SNE により 2 次元平面上に点としてプロットしたものを図 1，図 2 に示す．例として男声の angry を見てみると，2 言語の話者 5 名の計 10 名がクラスタとして表れていることがわかり，ほかの感情についても同様の状態である．そして日本語と英語で明確な境界線を引くことができ，これが言語による感情の違いとなる．図 3 に日英 4 感情の x-vector を全て t-SNE によりプロットした結果を示す．この結果から，x-vector による発話分類に影響を与える要素は話者性が最も強く，次に，大まかに話者が言語で分かれていることから，言語性が強いといえる．感情は言語，話者ごとに 4 感情がセットでクラスタが表れているので，最も弱い．

## 5　まとめ

x-vector を特徴量として使用し英語音声の感情認識を行った．その結果として，未知の話者であっても既存テキストの発話であれば，感情認識が比較的正確に行われることが分かったが，未知テキストの発話については更に検証が必要である．

また，x-vector を用いる手法は，発話データ数の少ない話者に対する感情認識が困難であると考えられ，今後はそういった場合でも，認識精度を上げる方法を検討する必要がある．

## 参考文献

[1] 小原他, “音声対話システムのための入力音声の感情に同調する声質変換と x-vector 埋め込みを用いたテキストからの音声合成方式の検討,” 電子情報通信学会技術研究報告, vol. 122, no. 389, SP2022-73, pp. 203-208, Mar. 2023.

[2] E. Takeishi, *et al.*, “Construction and analysis of phonetically and prosodically balanced emotional speech database,” 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), pp. 16–21, 2016.

[3] K. Zhou, *et al.*, “Seen and Unseen Emotional Style Transfer for Voice Conversion with A New Emotional Speech Dataset,” ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 920–924, 2021

[4] M, Ravanelli, *et al.*, “SpeechBrain: A General-Purpose Speech Toolkit,” 2021. `https://arxiv.org/abs/2106.04624`

[5] A, Nagrani, *et al.*, “ VoxCeleb: A Large-Scale Speaker Identification Dataset,” Proc. Interspeech 2017, pp. 2616–2620, 2017

[6] B, Desplanques, *et al.*, “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification,” Proc. Interspeech 2020, pp. 3830–3834, 2020