演技発話における x-vector 埋め込みの感情と話者性の表現の検討

Investigation of emotion and speaker's representation in visualization of x-vector embedding in acted speech.

天田 侑輝

Yuki Amada

岡山大学 原研究室

Hara Laboratory, Okayama University

概要 本研究では，感情と話者性の表現について検討した．データセットとして，広島市立大学の感情音声コーパスを用いた．3 種類の x-vector 埋め込みのモデル，t-SNE による次元削減による可視化により，感情，話者性，演技技法，セリフそれぞれで分離できるか実験を行った．結果として，話者性の分離は見られた．感情，セリフで部分的な分離が見られた．演技技法では分離は見られなかった．また，3 種類の x-vector の各モデルによる違いも確認できた．

1 はじめに

感情は，コミュニケーションに影響を与える．話し手は自身の感情に影響されて話し方や話す内容を変化させており，聞き手も話し相手の感情を認識し相手に合わせた応答を行うことで円滑な対話を実現している．さらに，感情は人のパフォーマンスに影響を与える．感情がポジティブであれば，創造性や生産性、意欲などが高まり，感情がネガティブであれば，健康被害を及ぼす．今後，人の機械との自然な音声コミュニケーションやパフォーマンスの最大化には，感情を認識する技術や，認識した感情に応じた行動制御を行う技術が不可欠であろう．

しかし，現在において音声認識や音声合成などの音声処理技術と比べると，音声感情認識は一般に普及しているとは言えず，技術の到達点や課題が広く知られているわけではない．

本研究では x-vector 埋め込みの感情と話者性の表現に焦点を当て，その有効性を評価することを目的とする．x-vector は主に話者の識別を目的として用いられてきたが，その適用範囲を感情認識に拡張する試みもなされている [7][8].

2 実験方法

2.1 データセット

広島市立大学の目良和也先生らによって構築された「広島市立大学 感情音声コーパス（HCUDB）」[1] を用いている．声優・ナレータ等のプロの演者が同一のセリフを複数の感情で演じた音声が収録されており，HCUDB1 と HCUDB2 の 2 種類で構成されている．

HCUDB1 は 14 名の演者（男性 6 名，女性 8 名，いずれも 20〜60 代）が，「そうなんですか」「どうなってるの」等の 10 種類のセリフについて，それぞれ Russell の感情円環モデルに基づく「驚き」「怒り」「軽蔑」「眠い・疲れた」等の 11 種類の感情で，それぞれ 3 テイクずつ発声している（計 4,620 発話）．またこの全発話に対して，16 名の評価者により，3 種類の話者感情評価を行っている（快不快および覚醒の度合いの 7 段階評価，11 感情の該否判定，自然性の 3 段階評価）．

HCUDB2 は 20 名の演者（男性 7 名，女性 13 名，いずれも 20〜60 代）が，HCUDB1 の 10 種類のセリフのうち 5 種類について，HCUDB1 と同様の 11 種類の感情を，2 種類の演技手法で演じ分けて，それぞれ 3 テイクずつ発声している（計 6,600 発話）．他者による感情評価は付与されていない．演技手法の一つは，技術的演技である．これは，感情を伝達するために意図的に口調を変えるような技術を使った演技をする手法である．もう一つの演技手法は，感情移入演技である．これは，自分の中でイメージを作り感情移入をしてから演技をする手法である．

2.2 x-vector 埋め込みとその可視化

VoxCeleb データセット [4] の学習済みモデルである，TDNN アーキテクチャ [2]，ECPA-TDNN アーキテクチャ [3] を用いる．本報告では，それぞれのモデルを TDNN，ECPA-TDNN と呼ぶ．また，JTubeSpeech コーパス [5] と呼ばれる，YouTube から収集した日本語音声から学習されたモデルの x-vector[6] を用いている．以降，3 つ目のモデルを `x-vector_jtubespeech` と呼ぶ．

また，可視化するために非線形次元削減手法である t-SNE を用いている．t-SNE を用いて 2 次元に削減している．t-SNE のパラメータは `perplexity` を 5 としている．また，`learning_rate` は 200 としている．他のパラメータはデフォルトである．

3 実験結果

実験結果を図 1，図 2，図 3，図 4 に示す．

3.1 感情

図 1 に示すような，感情により色分けした t-SNE による可視化の結果を確認する．図より，`x-vector_jtubespeech` において，感情の分離がわずかに見られた．「狂気・楽しい」，「憂鬱・悲しい」が分離していることが図から読み取れる．また，Russell の

感情円環モデルにおいて，対をなす「狂気・楽しい」「憂鬱・悲しい」は，対をなして離れているように見られた．ただし，他の感情においてはそのような対となる分離は見られなかった．ECPA-TDNN と TDNN においては明確な感情の分離が見られなかった．

3.2 話者性

図 2 に示すような，話者により色分けした t-SNE による可視化の結果を確認する．図より，ECPA-TDNN において，最も話者の分離が見られた．TDNN においても，十分な分離が見られた．`x-vector_jtubespeech` においては，あまり見られなかった．

3.3 演技技法

図 3 に示すような，演技技法により色分けした t-SNE による可視化の結果を確認する．凡例の G は技術的演技であり，K は感情移入演技である．どのモデルにおいても分離は見られなかった．

3.4 セリフ

図 4 に示すような，セリフにより色分けした t-SNE による可視化の結果を確認する．図より，`x-vector_jtubespeech` においては，話者性よりもセリフによって分離していることが読み取れる．ECPA-TDNN と TDNN においては，それぞれの話者の中で分離が見られた．

4 考察

感情において十分な分離は見られなかったが，話者性の分離においても顕著な分離は見られなかったことから，音声の処理単位が文章ではなく，1 発話，1 フレーズ程度であることが一つの原因として考えられる．さらに，今回用いた音声コーパスが，自然発話に近いことも原因として考えられる．また，t-SNE のパラメータの設定を今回はあまり変えて試せておらず，変更していけばよりよい分離がみられると考えられる．

演技技法による分離は見られなかったことから，感情と話者性には演技技法は関係がないと考えられる．

話者とセリフおいて，ECPA-TDNN と TDNN では前者でより分離が見られた．ECPA-TDNN は TDNN を改良したアーキテクチャである [3] ことが原因として考えられる．

また, ECPA-TDNN, TDNN と `x-vector_jtubespeech` による違いは，学習に用いたデータセットの影響を受けていることが考えられる．`x-vector_jtubespeech` だけが音声コーパスとして用いた言語とおなじ日本語での学習を行っているが，モデルの違いが顕著に出たのはセリフであり，話者性には他のモデルに比べてより良い分離が見られなかったことから，学習に用いる言語による影響は話者性は少なく，セリフの特徴を抽出しやすいと考えられる．

さらに，今回の実験では感情を最も分離できたのは `x-vector_jtubespeech` であるが，感情音声認識の側面から考えるとセリフの特徴を抽出してしまうといえるため，モデルの学習にはむしろ認識したい言語とは違う言語を使ったほうが良いとも考えられる．

5 まとめ

本報告では，演技発話を用いた x-vector 埋め込みの可視化による感情と話者性の表現について述べた．感情と話者性の表現における有効性が示唆された．

考察でも述べたように感情音声認識の側面から考えると，感情以外の特徴を抽出してしまうともいえる．このため，対策として感情の成分を強め，感情以外の成分を弱める必要が考えられる．論文 [10] のように，損失関数とニューラルネットワーク構造を工夫する方法や分類器を用いる方法が考えられる．

また，感情の分離がわずかに見られた．話者と話者の間の音声合成において，シームレスな話者性の制御ができないという論文 [9] がある．感情においても同様のことが起こるとすると，感情と感情の間の音声合成において役立つことが考えられる．

今回の実験では 3 種類の x-vector のモデルを用いたが，他のモデルにおける感情と話者性の表現についても検討したい．

参考文献

[1] Kazuya Mera, *et al.*, Hiroshima City University (2024): Hiroshima City University Japanese Emotional Speech Corpus (HCUDB). Speech Resources Consortium, National Institute of Informatics. (dataset). `https://doi.org/10.32130/src.HCUDB`

[2] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey,and S. Khudanpu,*et al.*, “X-Vectors: Robust DNNEm- beddings for Speaker Recognition “, Proceedings of 2018 IEEE International Conferenceon Acous- tics Speech and Signal Processing (ICASSP), pp. 5329-5333, 2018

[3] B. Desplanques, J. Thienpondt, and K. Demuynck,*et al.*, “ECAPA-TDNN: Emphasized Chan- nel Attention, Propagation and Aggregation in TDNN Based Speaker Verification “, Proceeding of Interspeech 2020, pp.3830-3834, 2020

[4] A. Nagrani, J. S. Chung, and A. Zisserman,*et al.*, “VoxCeleb: a large-scale speaker identification dataset “, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-9, 2017

[5] Shinnosuke Takamichi, Ludwig Kurzinger, Takaaki Saeki, Sayaka Shiota, Shinji Watanabe,*et al.*, ”JTubeSpeech: corpus of Japanese speech collected from YouTube for speech recognition and speaker verification”

[6] `https://github.com/sarulab-speech/xvector_jtubespeech`

[7] J. Williams, and S. King, *et al.*, Disentangling Style Factors from Speaker Representations,” Pro- ceedings of Interspeech 2019, pp. 3945–3949,

2019.
[8] R. Pappagari, T. Wang, J. Villalba, N. Chen, andN. Dehak, *et al.*, “X-vectors meet emotions: A study on dependencies between emotion and speaker recognition,” Proceedings of 2020 IEEE Interna- tional Conference on Acoustics, Speech and Sig- nal Processing (ICASSP), pp. 7169-7173, 2020.
[9] 青谷他，”話者特徴量の操作によりシームレスに話者性を制御できる End-to-End 音声合成方式の検討”
[10] Jing-Xuan Zhang *et al.*, ”Non-Parallel Sequence-to-Sequence Voice Conversion With Disentangled Linguistic and Speaker Representations”

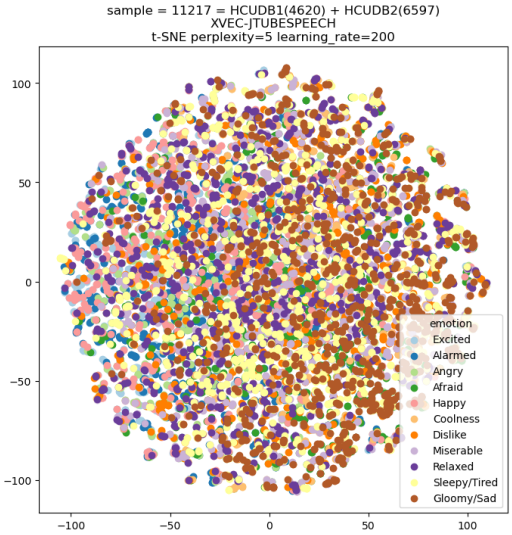

図 1: 感情による色分け
（x-vector_jtubespeech）

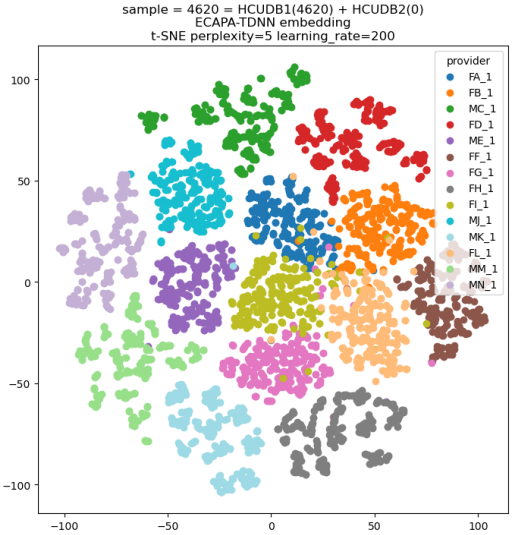

図 2: 話者による色分け
（ECPA-TDNN）

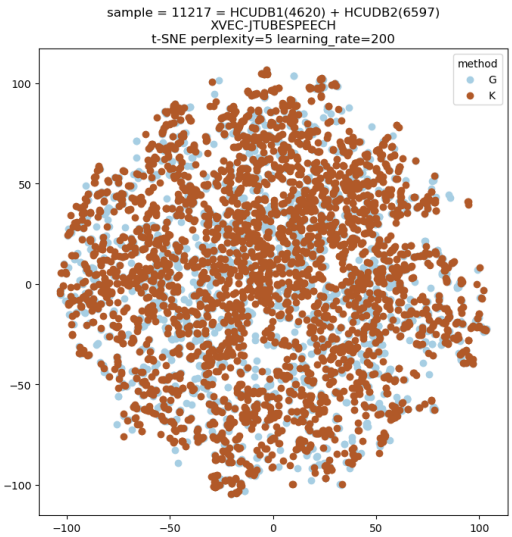

図 3: 演技技法による色分け
（x-vector_jtubespeech）

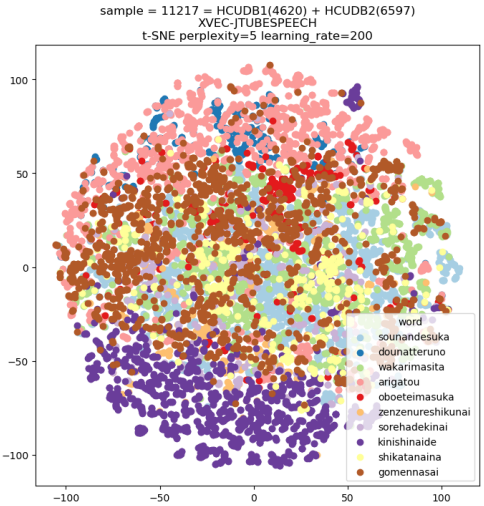

図 4: セリフによる色分け
（x-vector_jtubespeech）