演技音声に対する話者感情推定におけるセリフの違いの影響の分析

An analysis of the effect of differences in dialogue on the estimation of speaker emotion for acting speech

宮本　侑季

Yuki Miyamoto

広島市立大学　言語音声メディア工学研究室

Language and Speech Research Laboratory, Graduate School of Information Sciences, Hiroshima City University

概要 本研究では，発話音声から話者感情を推定する機械学習器の学習データとしての利用を想定し，演技音声データベースを構築する際の「セリフ」の影響について分析を行った．まず，10 種類のセリフと 11 種類の感情の組み合わせについて 14 名に演じてもらうことで演技音声データベース（HCUDB1）を独自に構築した．そしてこのデータベースを用いて，発話音声に対して他者が感情評定を行う際にセリフの違いが与える影響について検証した．

1　はじめに

近年，接触型デバイスを用いない「発話音声からの話者感情推定技術」が注目されている．音声対話システムにこのような技術を組み込む際は自然な会話音声（自発音声）から感情推定を行うことになるため，感情推定用機械学習器に用いる学習データも解析対象と同じ自発音声が最も適している．

しかし多様な感情ラベルが付与された大規模な自発音声データベースを構築するにはいくつかの問題がある．まず，何も統制していない会話から発話音声を収集する場合，収集できる音声の感情クラスに偏りがでたり，感情クラスごとに話者が異なったりするという問題が起こる．とはいえ特定の感情，特にネガティブな感情状態における自発音声を収集するために実験参加者の感情をネガティブに誘導することは倫理的な観点から慎重な対応が求められる．さらに，収集した自発音声に対して話者感情を評定する作業については，複数の評価者による大量のデータへのアノテーション作業のコストの問題や，他者による評定結果が話者自身が感じた感情と完全に一致するわけではない[1]という問題がある．

一方，感情が生起しているかのように演技して発話した音声（演技音声）であれば，感情の種類に偏りが出ず，また発話音声への感情アノテーションも容易であることから，自発音声からの感情推定においても演技音声を学習データに用いることも多い．このような演技音声による感情音声データベースを構築する際は，演者の年齢や性別，発話音声のセリフや演技する感情の種類，さらには演技する際の技法など様々な条件を統制することができる．しかし，条件の数を増やしすぎると収集データの質は良くなるが量を収集することは難しくなるため，データベースを構築する際にはどのような条件が収集データの質の向上に有効かについて見極める必要がある．

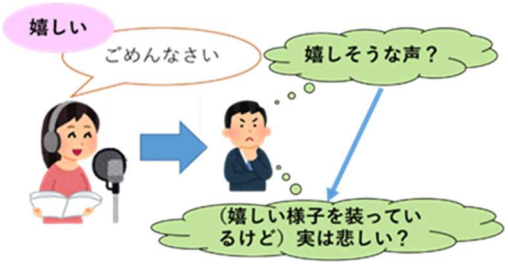


図 1. 発話内容が感情推定に与える影響

そこで本研究では感情を込めた音声を演じる際の「セリフ」の影響に注目する．セリフを自由にすることで起こりうる問題として感情アノテーションへの影響が考えられる．他者評定によって収録音声に感情ラベルを付与する際，セリフの印象が感情評定に影響するおそれがある．図. 1 に示すように，「ごめんなさい」というセリフを嬉しい感情で発話した際，口調的には嬉しいと認識しているにも関わらずセリフの影響で別の感情と認識されてしまうことがある．

そこで本研究では，発話音声に対して他者が感情評定を行う際にセリフの違いが与える影響について，独自に構築したデータベースを使って分析を行う．さらにセリフ別に機械学習分類器を構築し，感情推定実験を行う．

2　感情音声データベースの構築

2.1　感情音声データの収録

本研究では独自に構築した広島市立大学感情音声データベース（HCUDB1）[2]を用いる．HCUDB1 ではプロ声優，俳優，ナレーターなど男女 14 名（男性 6 名，女性 8 名）に 10 種類の台詞を 11 種類の感情で 3 パターンずつ発話してもらった音声を収録している．演じた感情は，狂喜・嬉しい，驚き，怒り，恐れ，余裕・楽しい，冷静，嫌い，軽蔑，リラックス・気楽，眠い・疲れた，憂鬱・悲しいの 11 種類とした．セリフは「そうなんですか」「どうなってるの」「分かりました」「ありがとう」「覚えていますか」「全然嬉しくない」「それはできない」「気にしないで」「仕方ないな」「ごめんなさい」の 10 種類である．収録した音声データ数は合計 4,620 発話である．

2.2　感情ラベルの付与

HCUDB1 では，各演技音声に対して話者が演じようとした感情である「演技感情ラベル」と，音声を聞いた他者が推定した話者感情である「他者評価感情ラベル」の 2 種類のラベルを付与している．

演技感情ラベルは，収録の際に演者に指示した感情クラスと同じものを付与している．

他者評価感情ラベルは，各演技音声に対して評価者（21～24 歳の男子大学生 10 名）が 2.1 節の 11 種類の感情クラスから話者感情に該当するものを複数選択可能で回答した結果を集計したものを付与している．そのため，HCUDB1 の他者評価感情ラベルは，一つの演技音声に対して 11 種類の感情の評価値が付与される形式となっている．

3　セリフによる感情評価傾向の違い

本章では，各セリフの演技音声において演技感情と他者評価感情がどれだけ一致しているかについて HCUDB1 を用いて分析する．分析の指標として，感情認識率（ある演技感情で発した演技音声に対してある他者評価感情と評価された割合）を定義する．計算式を(1) に示す．

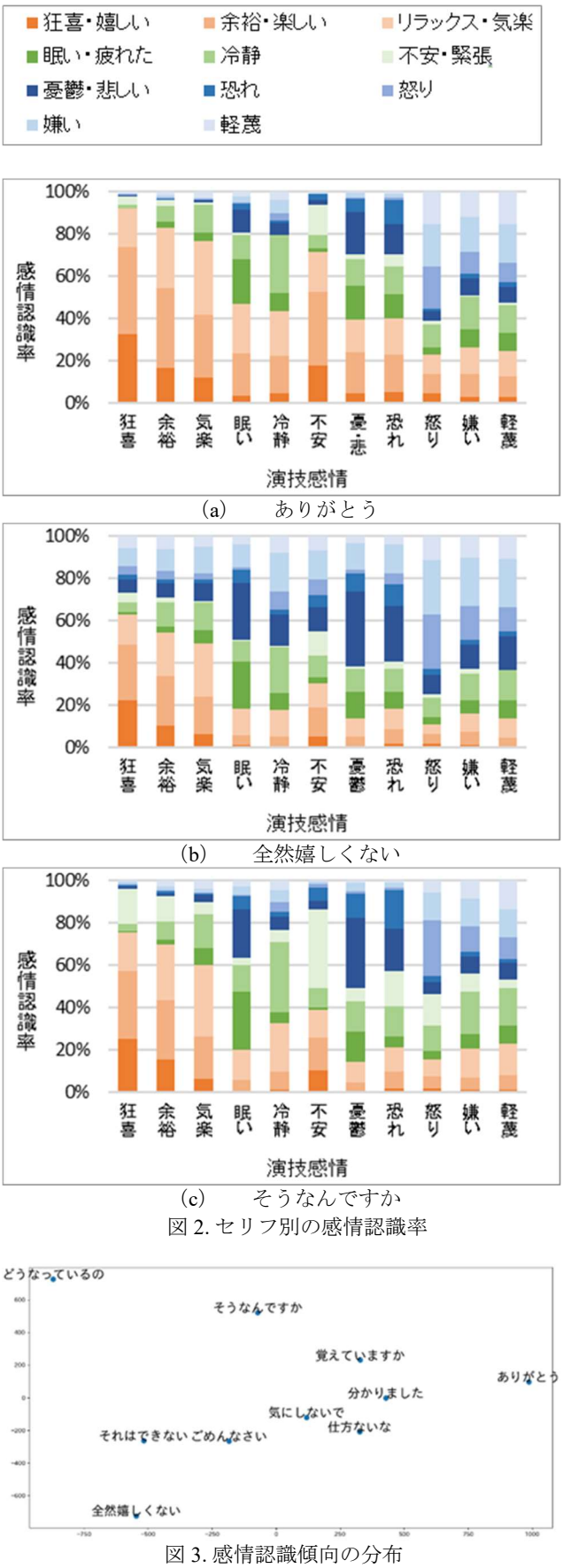
$$感情認識率_{act}^{eva} = \sum_{v \in ACV} \frac{tpa_v^{eva}}{|TPA_v|} \qquad (1)$$

$$TPA_v = \{tpa_v^{狂喜}, tpa_v^{余裕}, \cdots, tpa_v^{軽蔑}\}$$

式中のactは演技感情，evaは他者評価感情，vはある演技音声，ACVは演技感情actで演じられた演技音声の集合，TPA_vは演技音声vに対する他者評価感情ラベル，tpa_v^eは演技音声vが感情eに該当すると評価した評価者の人数を表す．

「ありがとう」，「全然嬉しくない」，「そうなんですか」の 3 つのセリフによる演技音声に対して算出した感情認識率を図. 2 に示す．図.2 (a)より，「ありがとう」というセリフの演技音声では，ニュートラルな感情やネガティブな感情を演じた場合でも全体的にポジティブな話者感情と認識されやすい結果となった．逆に，「全然嬉しくない」というセリフの演技音声では，ポジティブな感情やニュートラルな感情を演じた場合でも全体的にネガティブな話者感情と認識されやすい傾向がみられた（図.2 (b)参照）．同様の傾向がみられたセリフとしては，「ごめんなさい」，「それはできない」，「どうなっているの」があった．そして，「そうなんですか」というセリフは，「ありがとう」や「全然嬉しくない」のようにポジティブやネガティブな印象に偏ることがなく話者感情が推定される傾向がみられた（図.2 (c)参照）．同様の傾向がみられたセリフとしては，「分かりました」，「仕方ないな」，「覚えていますか」，「そうなんですか」，「気にしないで」があった．

さらに図.2 の各感情認識率に対して主成分分析を行った．第一主成分を横軸，第二主成分を縦軸としてプロットした散布図を図.3 に示す．その結果，左側に「ごめんなさい」，「全然嬉しくない」といったネガティブな印象のセリフ，中央に「そうなんですか」，「仕方ないな」といったニュートラルなセリフ，右側に「ありがとう」のようなポジティブな印象のセリフが集まっており，セリフの印象の感情極性ごとに感情評価結果が似ていることが確認された．



(a)　ありがとう

(b)　全然嬉しくない

(c)　そうなんですか

図 2. セリフ別の感情認識率

図 3. 感情認識傾向の分布

4　セリフによる機械学習性能の違い

本章では，学習データに用いる発話音声のセリフの違いが機械学習器の性能に及ぼす影響について調査するため，同一セリフの音声のみを学習データとする多クラス分類器を構築し，感情推定性能について比較考察する．

4.1　実験条件

機械学習器には Support Vector Machine を使用した．カーネルは RBF カーネル，ハイパーパラメータは C=1.0，gamma=1 に設定した．説明変数に用いた音響特徴量は音響分析ツール OpenSMILE[3]の eGeMAPSv02 特徴量セットである．本特徴量セットでは，基本周波数，ラウドネス，メル周波数ケプストラム係数（MFCC），ジッター，シマー，Harmonic-to-Noise Ratio（HNR）などから算出される静的特徴量 88 種類が算出される．そして各特徴量に対して平均=0，分散=1 になるよう正規化を行ってから，機械学習実験に用いた．

学習データに使用した HCUDB1 のセリフ別演技音声の個数は 252 個である．その際，狂喜・嬉しい，余裕・楽しいはともに「喜び」として扱うため，両感情から 126 個ずつ無作為に抽出したものを喜びの学習データとした．なお本実験では「他者評価の代用として演技感情ラベルをそのまま使えないか」という観点から，学習データに付与する感情ラベルとして他者評定感情ではなく演技感情クラスを用いる．

テストデータには感情評定値付きオンラインゲーム音声チャットコーパス（OGVC）[4]の自発音声データを使用した．本実験では，OGVC の喜び，怒り，嫌悪，悲しみ，恐れ，平静の 6 感情に HCUDB1 の感情ラベルから近いものを対応付けて実験を行った．感情ラベルの対応を Table 1 に示す．なお OGVC の収録音声には 3 人の評定者による感情評定結果が付与されているため，本実験では 3 名中 2 名以上が当該感情と評定した音声データをテストデータとして使用する．その際，6 感情の中で最も少ない“恐れ”にテストデータ数を合わせるため，“恐れ”以外の感情についても 3 名全員が当該感情と評定したデータの数を恐れと同じ 108 個，3 名中 2 名が当該感情と評定したデータの数を同 33 個となるよう無作為にデータを抽出した．

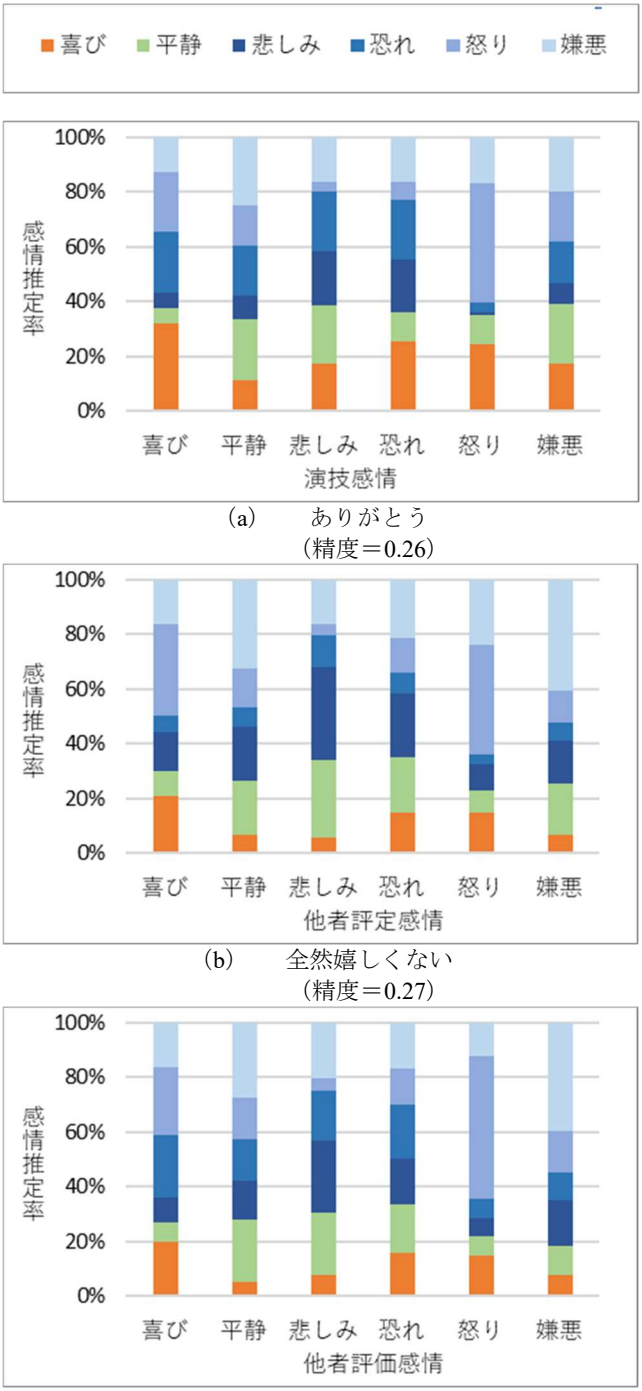
4.2　実験結果

「ありがとう」，「全然嬉しくない」，「そうなんですか」の 3 セリフを使った機械学習実験結果を図.4 に示す．分析の指標として，感情推定率（ある他者評価感情に対してある感情と推定された割合）を式(1)と同様の手順で算出している．

図 4 より「ありがとう」というセリフのみを使って学習した機械学習器では，怒りや悲しみといったネガティブ感情の自発音声でも全体的に喜びと推定されやすいという結果となった．「分かりました」についても同様の感情推定傾向がみられた．一方「全然嬉しくない」というセリフのみを使って学習した機械学習器では，喜びや恐れといった感情の自発音声でも全体的に悲しみや嫌悪と推定されやすいという結果となった．同様の傾向のセリフとして「ごめんなさい」があった．「そうなんですか」というセリフで演じた演技音声では，他のセリフで学習させた機械学習器

表 1 データベース間の感情ラベルの対応

OGVC	HCUDB1
喜び	狂喜・楽しい，余裕・嬉しい
怒り	怒り
嫌悪	嫌い
悲しみ	憂鬱・悲しい
恐れ	恐れ
平静	冷静



(a)　ありがとう
（精度＝0.26）

(b)　全然嬉しくない
（精度＝0.27）

(c)「そうなんですか」
（正解率＝0.30）

図 4. セリフ別の 6 感情推定実験結果

ほど大きく偏って推定される感情はなく，また正解率も他のセリフと比べて高かったことから平均的に推定されている結果ことが分かった．

4.3　考察

実験の結果，セリフによって同じ感情で演じても推定される感情の傾向に違いがあることが分かった．特にポジティブな印象がある「ありがとう」や「分かりました」といったセリフで学習した機械学習器はポジティブでない感情の自発音声に対してもポジティブな感情と誤推定しやすい傾向がみられた．一方，ネガティブな印象がある「ごめんなさい」や「全然嬉しくない」といったセリフで学習した機械学習器はネガティブでない感情の自発音声に対してもネガティブな感情と誤推定しやすい傾向がみられた．このような傾向の原因として，感情移入による演技手法を用いている演者にとって日常の中で想像しにくいセリフと感情の組み合わせの演技が難しかったのではないかと推察される．これに対してニュートラルなセリフであれば感情推定傾向に偏りが出ておらず，また正解率が高かったことから演技感情ラベルがそのまま正解ラベルとして使えるのではないかと考えられる．

5　おわりに

本実験ではセリフによって他者に与える印象の違いについて独自に構築したデータベースを使って分析を行った．さらにセリフ別に機械学習器を構築し，感情推定実験を行った．その結果，ポジティブやネガティブな印象があるセリフは他者評価や機械学習器の感情推定結果に偏りが出るが，ニュートラルなセリフでは偏りが出なかった．このことから演技音声を用いて感情推定を行う際にはポジティブでもネガティブでもないセリフを指定したほうがよいということが分かった．

今後は，自発音声を学習データとした場合との比較実験と，本実験で構築した HCUDB1 を公開するための準備を行う予定である．

謝辞

本研究は JST，COI，JPMJCE1311 の助成を受けたものです．また，演技音声収録においてはヒューマンアカデミー広島校パフォーミングアーツカレッジ様およびサウンドオフィスクロスロード様にご協力いただきました．

6　参考文献

[1] Saitou Shota, Mera Kazuya, Kurosawa Yoshiaki, Takezawa Toshiyuki. “Correlation Analysis between Subjectively Annotated Emotions and Objectively Annotated Emotions”, IMECS 2019, pp.141-146(2019)

[2] 目良和也，谷有希，村田唯，黒澤義明，竹澤寿幸“演技感情と推定感情のタグを付与した感情音声コーパスの構築”，日本音響学会 2017 年春季研究発表会講演論文集，pp.1471-1474 （2017）

[3] Eyben, F., Wllmer, M. and Schuller, B. (2010) Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. Proceedings of the 18th ACM International Conference on Multimedia, 1459-1462.

[4] Y. Arimoto, H. Kawatsu, S. Ohno, and H. Iida, “Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment,” Acoustical Science and Technology, vol. 33, no. 6, pp. 359-369, 2012.