

日本語音声のパラ言語的特徴をもとにした話者の印象推定手法の提案

中川撫々子[†] 木村昌臣[‡]

[†] [‡] 芝浦工業大学工学部 〒135-8548 東京都江東区豊洲 3-7-5

[‡] Universiti Malaysia Kelantan Pengkalan Chepa, 16100 Kota Bharu, Kelantan, Malaysia

E-mail: [†] al22033@shibaura-it.ac.jp, [‡] masaomi@shibaura-it.ac.jp

あらまし

近年、オンライン会議やライブ配信コンテンツの音声中心のコミュニケーションが普及している。このようなコミュニケーションでは、話し手の印象が声の抑揚、話速、声質などのパラ言語的特徴に強く左右される。しかし、こうした印象は必ずしも話者の意図通りに伝わるとは限らず、話の内容が同じでも話し方によって誤解や印象の齟齬が生じるという問題がある。本研究は、話者の印象をパラ言語的の手がかりに着目して定量的に評価し、機械学習によって話者の印象を推定する手法を提案するものである。本手法により、話し方による印象形成への理解を深め、より円滑なコミュニケーションへの支援を目指す。

キーワード 音声印象推定, 感情, 感性, 認知, 深層学習

1. 研究背景と目的

近年、オンライン会議やライブ配信コンテンツの普及により、音声中心のコミュニケーションが増加している。このような状況では話し手に関する視覚的な手がかりが乏しいため、聞き手が声から感じ取る印象が一層重要になる。声には、抑揚、話速、声質などのパラ言語的特徴が含まれ、話者の印象形成に大きな影響を与えることが、明らかになっている[1]。パラ言語的特徴とは発話内容（言語情報）そのものではなく、声の出し方・話し方に関する非言語的の手がかりである。しかし、こうした印象は必ずしも話者の意図通りに伝わるとは限らず、話の内容が同じでも話し方によって誤解や印象の齟齬が生じることがある。特に女性は男性よりも他者からの評価や印象を意識する傾向が強いことが明らかになっている[2]。このため、女性話者の印象をパラ言語的の手がかりに着目して定量的に評価し、声の印象を理解・調整できれば、より円滑なコミュニケーションにつながると考えられる。

そこで本研究では、パラ言語的特徴を扱う機械学習モデルによって、複雑な話者の印象形成を捉え、女性話者の印象を推定する手法の提案を行う。

2. 先行研究

橋本ら[1]は、音声に対する印象や話者の発言中の状態に関する印象を「音声状態印象」と呼び、性格特性の推論がパラ言語的特徴から直接行われるのか、あるいは音声状態印象の形成を経て行われるのかを検証した。その結果、パラ言語的特徴は性格特性に対して直接的に影響を与えるだけでなく、音声状態印象を介して間接的に影響を与えることが示された。このことから、話者の印象形成においてパラ言語的特徴そのものだけではなく、それを聴者がどのように知覚し音声状態印象を形成するかが重要であると考えられる。しか

しながら、先行研究では話者の印象推定方法は検討されていない。

Mohammadi[3]らは、抑揚や声の大きさ、話速といった韻律的特徴量を音声から抽出し、それらに基づき、聞き手が音声から受けた「印象」として回答した Big Five を推定するモデルを構築している。Big Five とは以下の5つの因子をもとに行う性格特性の分類法である。

- 外向性： 社交的で活発かどうか。
- 協調性： 思いやりがあり、友好的かどうか。
- 勤勉性： 責任感があり、計画的かどうか。
- 神経症傾向： 不安やストレスを感じやすいかどうか。
- 開放性： 新しい経験やアイデアに対して柔軟かどうか。

推定は Big Five の各次元について High/Low の二値分類として行われ、SVM を用いた場合の精度は特性により 60～72%程度であった。一方で、パラ言語的特徴やそのときの音声状態印象が複雑に関与する性格特性を、韻律統計量と SVM のような比較的単純な枠組みだけで十分に捉えきれない可能性が示唆され、精度面でも改善の余地が残る。

3. 提案手法

3.1 話者印象評価データの収集

本研究では、女性話者の音声に対して聞き手が抱く印象を、アンケート形式により定量的に評価、収集する。対象音声には JVS (Japanese Versatile Speech) コーパスから選定した女性 40 話者の読み上げ音声を用いる。また、読み上げる内容による印象差を無くすため、全話者で同一文章の音声を用いる (4 発話, 約 30 秒)。

評価項目は、先行研究[1]を拡張させ、①聞き手が知

覚するパラ言語的特徴、②話者の状態に関する印象、③性格特性（Big Five）の3観点から構成し、合計19項目をVAS形式のSD法で回答させる。①②では表1に示す項目を使用し、③性格特性評価ではBig Five評価尺度であるTIPI-J[4]を用いる。ここで得られる評価データを、後段の印象軸抽出および推定モデルで使用する。

音声 状態 印象	①知覚する パラ言語的特徴	早口なーゆっくり話す
		声の高いー声の低い
		声の大きいー声の小さい
		固いー柔らかい
		澄んだー濁った
	②話者の状態印象	心地よいー不快な
		明るいー暗い
		緊張しているー落ち着いている
		威圧的ー優しい
	③性格特性 (BigFive)	外向性
		協調性
		勤勉性
		神経症傾向
		開放性

表 1：評価項目の構成

3.2 主成分分析による印象軸の抽出

音声に対する多次元な印象評価を主成分分析によって要約することを目的とし、評価データに主成分分析を適用する。話者印象は主成分軸（印象軸）上の値として表現し、この印象軸の値を推定モデルの目的変数として用いる。

3.3 推定モデル

本研究では、発話音声から得られる音響統計特徴（発話時間、ポーズ率、F0 平均・標準偏差・レンジ、音圧(dB)平均・標準偏差）に加え、メルスペクトログラム画像を入力として用い、話者の印象軸の値を推定する。モデル全体の構成を図1に示す。

まず、メルスペクトログラム画像に対して CNN（事前学習済み ResNet18）を適用し、画像特徴量を抽出する。抽出された特徴量は小規模なフィードフォワードネットワーク（FFNN）で変換した後、話者に対応する複数画像（複数区間）から得られた特徴を平均し（mean pooling）、話者単位で特徴量を得る。音響統計特徴は、別の FFNN に入力し、統計特徴に基づく埋め込み表現へ変換する。最後に、メル画像特徴と統計特徴を結合し、統合した特徴に対して FFNN を適用することで、印象軸の推定値を出力する。

なお、目的変数には各話者について得られた複数の印象評価（5～6 名の評価者による印象軸値）の平均値を用いる。これは、印象評価には評価者間のばらつき

や外れ値が含まれやすく、その影響をおさえるためである。

学習時の損失関数には平均二乗誤差（MSE）を採用し、推定値と目的変数の誤差を最小化するように学習する。

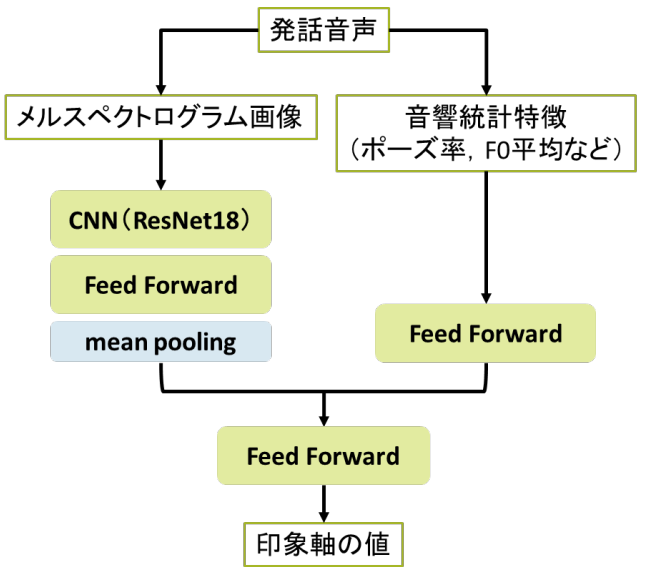


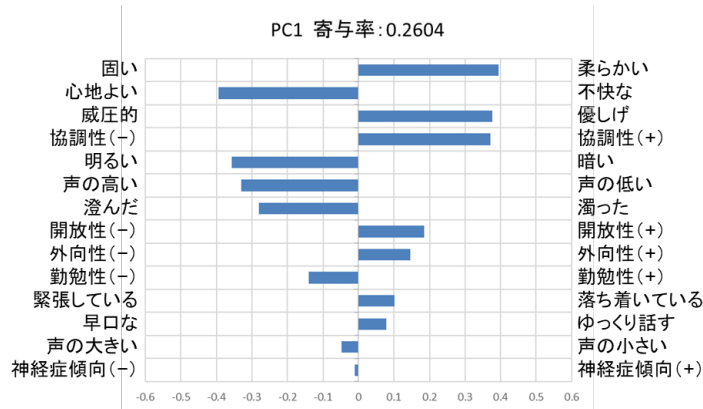
図 1：推定モデルの全体像

4. 実験

4.1 実験 1：印象軸の抽出

実験 1 では、評価データに対し主成分分析を用いて印象軸を抽出した。

評価実験は 18～22 歳の男性 28 人、女性 19 人、合計 47 人を対象として実施し、各被験者は 4 話者分の評価を行った。その結果、1 セット 14 項目からなる合計 188 セットのデータを得た。このデータに対し、主成分分析を行い 6 つの主成分を抽出した。このうち寄与率が高い主成分 PC1～PC3 における因子負荷量を図 2 に示す。



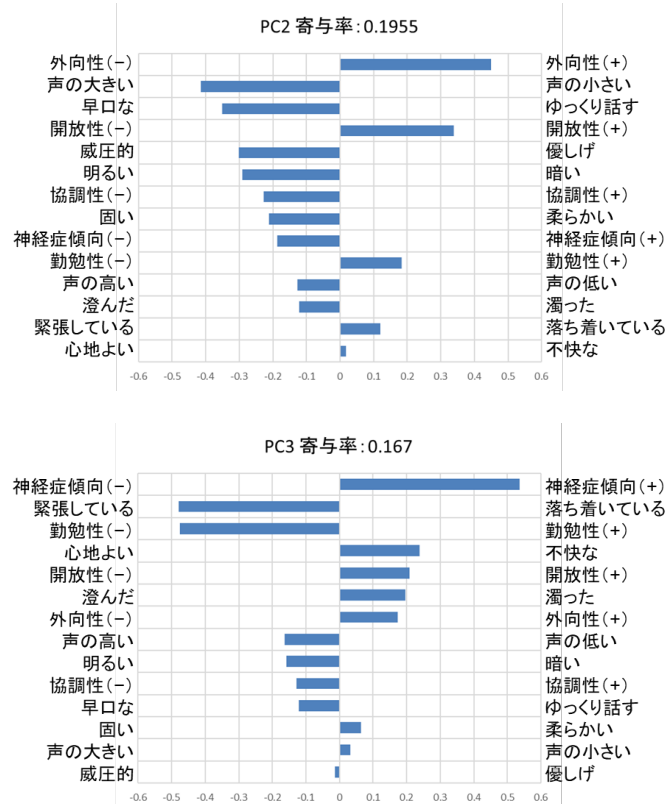


図 2：PC1~PC3 の因子負荷量

各主成分軸が表す印象の解釈を行った。

PC1（寄与率 0.2604）は「協調性」「心地よい-不快な」「固い-柔らかい」「威圧的-優しい」などの項目と強く対応しており、対人印象としての親しみやすさ・柔らかさに関わる軸と考えられる。したがって PC1 は、温かく親しみやすい声の印象を表す軸と解釈できる。

PC2（寄与率 0.1955）は「外向性」に加え、「声の大きい-声の小さい」「早口な-ゆっくり話す」など発話の活性度を表す項目の寄与が大きく、エネルギッシュさ・社交性と関係すると考えられる。よって PC2 は、元気で社交的な声の印象を表す軸と解釈できる。

PC3（寄与率 0.167）は「神経症傾向」「緊張している-落ち着いている」の寄与が顕著であり、神経質で不安定な印象を表す軸と解釈できる。

また、既存研究[5]は、本研究とは異なる SD 項目を用いて、挨拶程度の短い音声から話者の性格印象を評価した研究である。既存研究[5]における主成分分析では「温かみ」「支配性」の 2 軸が抽出され、それぞれ本研究の PC1 と PC2 に概ね対応し、整合性が確認された。

実験 1 では、主成分分析により解釈可能な複数の印象軸を抽出した。これらの軸は、音声特有の印象要素と性格特性の評価を統合した指標と考えられる。さらに、各軸は互いに異なる印象を持っていた。

4.2 実験 2：推定モデルの検証

実験 2 では、3.3 で示した推定モデルの性能評価を

行った。評価では $k=5$ の k -fold 交差検証を行い、対象である 40 話者を 5 分割（5-fold）して、1 分割を検証用データ、残り 4 分割を学習用データとして学習・検証を 5 回繰り返した。各 fold において学習は最大 30epoch で実施し、指標には平均二乗誤差（MSE）を用い、各 fold における学習曲線を算出した。fold 平均の学習曲線を図 3 に示す。

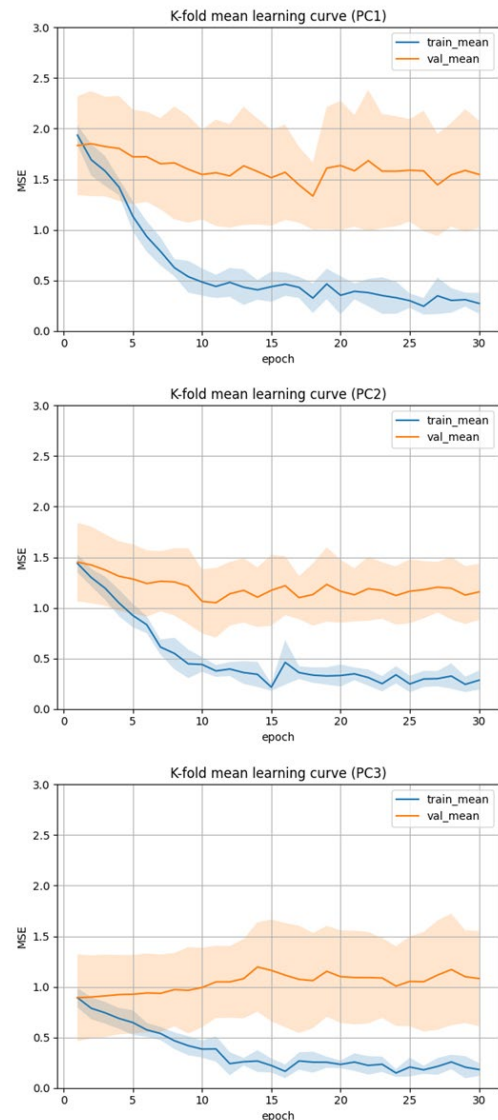


図 3：PC1~3 の学習曲線

PC1~PC3 のいずれにおいても、汎化性能が低い傾向が確認された。学習誤差は epoch の進行に伴い大きく低下した一方で、検証誤差は低下後ほぼ横ばい、または微増で推移した。また、fold 間での検証誤差のばらつきも比較的大きく、推定精度が安定しないことが示された。

これらの原因として目的変数の使用データの偏りがあげられる。目的変数としている話者平均得点の分布は図 4 から分かるように中心付近に偏り、両端の極端な値の話者が少ない。その結果 fold によっては極端

値の話者が学習用データに含まれず、検証で大きな誤差が発生し、汎化性能の低下、精度の不安定につながった可能性が考えられる。

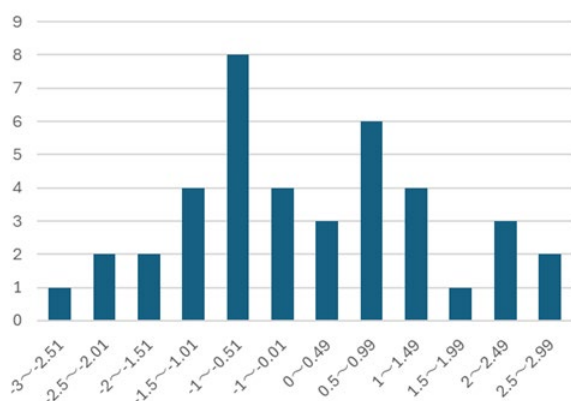


図4：目的変数（話者平均得点）のヒストグラム

実際の推定でも、図5のようにデータ数の少ない極端な値を持つ話者の推定では誤差が大きくなっている傾向がみられた。また、他のフォールドでも同様の傾向が確認されている。ただし、0付近の値を無差別に出力しているわけではなく、負に比べてデータの多い正の極端な値を持つ話者では、より正解値に近づいている場合が数件みられた。

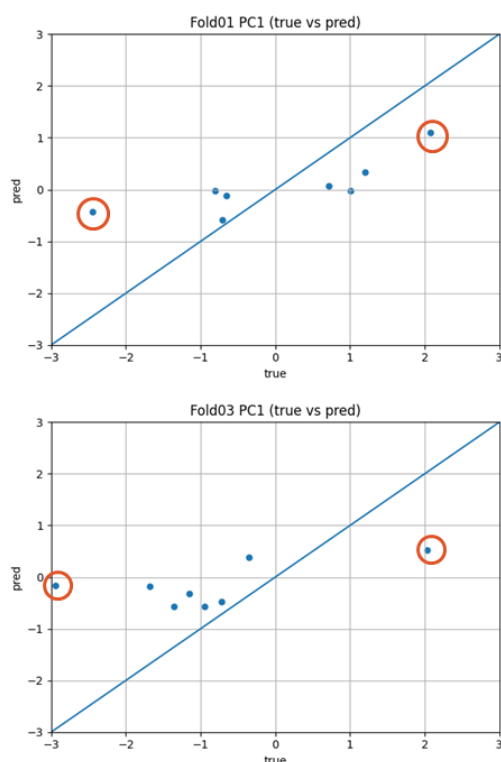


図5：正解値と推定値

また、40話者に対してパラメータ自由度が大きく学習用データに適しやすく、検証用データでは誤差が大きくなった可能性がある。

加えて、本モデルでは話速・ポーズ・リズムといっ

た時間的変化の特徴を十分に捉えられておらず、性能の伸び悩みにつながった可能性が考えられる。

5. まとめと展望

本研究では、日本語音声のパラ言語的特徴に着目し、女性話者の印象を定量化し推定する手法を提案した。印象評価データを収集し、主成分分析により異なる特徴を持つ解釈可能な印象軸を得た。さらに、音響統計特徴とメルスペクトログラム画像を入力とし、CNNとFFNNを用いて印象軸の値を回帰推定するモデルを構築した。推定モデルは汎化性能が低い傾向と検証話者によって精度が大きく変動する傾向が見られた。

今後は、話者得点分布の偏りや少数データを考慮したデータ拡張やモデル設計を検討する。また、話速・ポーズ・リズムなどの時間的変化を考慮した特徴量の導入や、主観評価のばらつきを考慮した目的変数設計や学習方法の工夫を行い、推定精度と安定性の向上を目指す。

参考文献

- [1] 橋本和奈実, 古屋健. 発話速度と声の高さが特性推論に及ぼす影響-二段階推論仮説に基づいて-. Japanese Journal of Applied Psychology, Vol. 45, No. 1, p15-25, 2019
- [2] Rosanna E. Guadagno, Robert B. Cialdini. Gender Differences in Impression Management in Organizations: A Qualitative Review. Springer Science + Business Media, LLC, 2007.
- [3] Mohammadi, Vinciarelli. Automatic Personality Perception: Prediction of Trait Attribution Based on Prosodic Features, IEEE Transactions on Affective Computing, July 2012
- [4] 小塩真司, 阿部晋吾, カトローニ ピノ (2012). 日本語版 Ten Item Personality Inventory (TIPI-J) 作成の試み. パーソナリティ研究, 21(1), 40-52. doi:10.2132/personality.21.40
- [5] Phil McAleer, Alexander Todorov, Pascal Belin. How Do You Say 'Hello'? Personality Impressions from Brief Novel Voices. PLoS ONE, 9(3), e90779. <https://doi.org/10.1371/journal.pone.0090779>, 2014