高齢者音声のエンゲージメント推定

Engagement Assessment using Prosodic Information of the Elderly

高橋 光一[1] 黄 宏軒[1]*

[1] 福知山公立大学情報学部
[1] Faculty of Informatics, The University of Fukuchiyama

Abstract: 増加傾向にある独居高齢者のコミュニケーション不足による健康被害のリスク低減に対話エージェントが有効であると考えられる．本研究は，それを実現するために高齢者の音声から対話への積極性の度合いを意味する「エンゲージメント」の推定を試みる．無音区間で区切った高齢者の音声データにエンゲージメントを 5 段階でアノテーションをし，韻律的特徴を用いて深層学習をおこなった．

1 はじめに

日本は，2007 年から高齢化率が 21 %を超えて超高齢社会に突入し，2021 年では世界で最も高い高齢化率をキープしている [1]．高齢者増加に伴って家族構成の核家族化も進み，一人暮らしの高齢者の世帯数が増加している [2]．それにより高齢者は若年層と比較して所属するコミュニティが少なくなり，持病等も影響してコミュニケーションをする機会が減少傾向にある．こうしたコミュニケーション不足は，鬱病や認知症といった健康被害のリスクの要因となっている [3]．高齢者のコミュニティの主となる介護施設では，介護業界における人材不足により入居できない高齢者も増加しているため，コミュニケーション不足解消が可能となる機会が必要とされている [4]．現在では，このような問題を解決するシステムとして対話エージェントが期待されている．

対話することのできるシステムとして，Apple 社の Siri などタスク指向型対話システムは多くの機能を実現し，利用することである程度の対話が可能となっている．このようなシステムは，対話からタスクの進行が目的で自然な対話ではなく，対話体験としては楽しむことはできずコミュニケーション不足解消の手段としては十分ではない．従って，実際に人間と対話しているような機能が求められ，人間にあるような機能が対話エージェントには不可欠である．そこで本研究は，高齢者の音声からエンゲージメントを推定をすることで，対話相手の対話への積極性を自動で判断することを目的とする．

*連絡先：福知山公立大学情報学部情報学科
〒 620-0886 京都府福知山市堀 3370
E-mail: hhhuang@acm.org

2 関連研究

音声から対話者の状態の推定として，感情の推定が挙げられる．話者の感情状態は声の大きさや高さなどといった声の調子に表れるとして，音響的特徴を音声データから算出し，機械学習機に学習させて分類をおこなった [5]．音響的特徴量の抽出にはオープンソースソフトウェア openSMILE を用いている．これによって音声から相手の状態を推定することが可能であるとしている．

エンゲージメントを推定する研究では，マルチモーダル学習がおこなわれている [6]．マルチモーダルは，音声を含む韻律的，視覚的情報など複数のデータ情報からエンゲージメントを推定することで学習の精度を高めることができ，精度が最大 96.93 %まで特徴量の組み合わせによっては達成している．特徴量として，音声の韻律的特徴や拍手や笑い，顔の視覚的特徴量などを用いて分析をしている．

本研究では，openSMILE を用いて高齢者の音声から韻律的特徴量を抽出して深層学習をおこない，エンゲージメントを推定する．感情推定のように精度が高くなるか検証する．エンゲージメントを推定したいと考えているため，マルチモーダルのような複数の特徴量から推定する方が精度は高くなると考えられるが，音声のみからエンゲージメントを推定が可能であれば，対話エージェントに搭載する際にコストや処理速度など有効となるため実現する価値がある．

3 対話コーパス

深層学習に用いる高齢者の音声データとして対話コーパス [7] のうち 1 つを用いる．話し相手役の健常な高齢者（69-73 歳）を 4 名 (男女 2 名ずつ) と聴き手役の

若者（22 歳）を 4 名（男女 2 名ずつ）の 15-30 分の 16 セッションのデータである．インターネットを通じて高齢者の自宅からビデオカメラで撮影し Skype で接続されているため，音声にノイズや音割れ，音ズレ等問題があったのでできうる限り除去している．その音声データを音声分析ソフトフェアの Praat を用いて自動で無音区間 200ms で区切った後，高齢者の音声のみ抽出するため手動で調整した．アノテーションは音声および動画の注釈を作成が可能なソフトフェアの ELAN を用いた．音声のエンゲージメントは主観で評価し，その他の音声との相対的判断で数値が高いほどエンゲージメントも高いとして 5 段階でアノテーションをおこなった．処理した音声データは，300 件程度で 1〜3 秒である．

4　エンゲージメント推定

深層学習には，機械学習ライブラリの TensorFlow を用いた．処理した音声データは 300 件程度，openSMILE の GeMAPSv01a で特徴量を抽出して特徴量は 62 件のデータを多クラス分類をおこなった．図 1 では学習モデルを 10 分割交差検証をおこなった．結果として精度が高くなっているが，エポック数が増加するたびに精度も増加していることから過学習が起きている可能性がある．音声データは 300 件程度と少なく，かつ特徴量も 62 件と少ないため過学習で精度が高くなっていると考えられる．

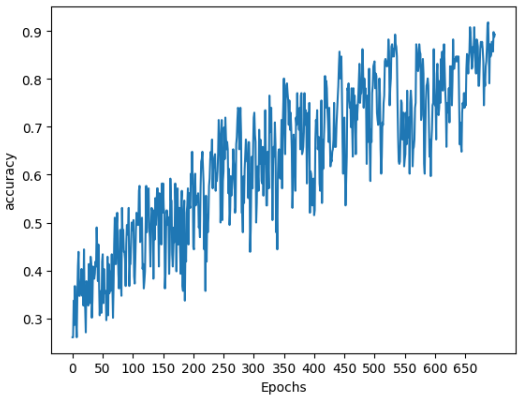

図 1: 学習の過程.

5　おわりに

本研究は，対話エージェントが高齢者のコミュニケーション不足の解消に有効であると考え，それを実現するために高齢者の音声から表面化するエンゲージメントを推定することを可能としたい研究である．音声の情報を抽出する上で，感情を推定することが可能な韻律的特徴を用いて感情の推定同様にエンゲージメントを推定した．高齢者の音声からエンゲージメントを推定する手法として，音声の韻律的特徴を用いて深層学習をおこなった．精度は高いという結果になったが，音声データが足りておらず過学習になっていると考えられる．openSMILE の特徴量を抽出する機能として，本研究で用いた特徴量より多くの特徴量が抽出することも可能であるため，これを用いてより正確な推定を可能としたい．また，エンゲージメントをアノテーションする際に主観でおこなったが，エンゲージメントの意味を明確に定義する必要があったと考えている．

参考文献

[1] 総務省統計局, 統計からみた我が国の高齢者, 2021.

[2] 内閣府, 平成 29 年版高齢社会白書, 2017.

[3] L. Fratiglioni, H.X. Wang, K. Ericsson, M. Maytan, and B. Winblad: Infuence of social network on occurrence of de-mentia: a community-based longitudinal study, 2000.

[4] 厚生労働省, 介護人材確保に向けた取り組み, 2021.

[5] Tang Ba Nhat, 目良和也, 黒沢義明, 竹沢寿幸. 音声に含まれる感情を考慮した自然言語対話システム, HAI 2014.

[6] Fahim A. Salim, Fasih Haider, Owen Conlan, Saturnino Luz, Nick Campbell, Analyzing Multimodality of Video for User Engagement Assessment, 2015.

[7] Huang, H.H., Masato Fukuda, Toyoaki Nishida, An Investigation on the Effectiveness of Multimodal Fusion and Temporal Feature Extraction in Reactive and Spontaneous Behavior Generation RNN Models for Listener Agents, HAI 2019.