論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	A Multimodal Model for Personality Recognition through Speech
著者(和文)	Nah Nathania, 土屋 ゆり, 越仲 孝文, 篠田 浩一
Authors(English)	Nathania Nah, Yuri Tsuchiya, Takafumi Koshinaka, Koichi Shinoda
出典(和文)	日本音響学会講演論文集, vol. 150, no. , pp. 1323-1324
Citation(English)	, vol. 150, no. , pp. 1323-1324
発行日 / Pub. date	2023, 9

A Multimodal Model for Personality Recognition through Speech

○ Nathania Nah (Tokyo Tech), △ Yuri Tsuchiya (Tokyo Tech),
Takafumi Koshinaka (YCU), Koichi Shinoda (Tokyo Tech)

Abstract

Exploring the field of affective computing is important for understanding how humans think and interact with each other. Personality computing focuses on methods of performing the automatic detection of human traits which compose their personality. Using the Five Factor Model of Personality as a measure to describe a subject's personality, temperament, and psyche, this work employs a multimodal model to perform automatic personality recognition on speech. We employ the use of speaker and phone disentanglement in speech representation learning, a technique known to be effective in emotion recognition, to predict scores for personality traits trained on the UDIVA dataset and outperform current methods that use visual features.

1 Introduction

Understanding another's thought patterns and motivations is often essential for effective communication. This type of information can often be captured through understanding their personality. One's personality consists of stable characteristics that drive their own motivations, behaviors, emotions, etc. and contributes to the individuality [3].

Our work explores how to infer personality from audio and transcribed text, building upon a previous multimodal audio-textual method for emotion recognition [5] that disentangles personality features from phone and speaker characteristics from speech.

2 Related work

One of the most common models used in personality psychology is the Big Five Model of Personality, which is also often referred to as the Five-Factor or OCEAN Model [3]. The traits described by the Big Five are: Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism. This model assigns a numerical value to each of these traits, representing personality in a five-dimensional vector.

Several methods have been introduced to predict self-reported personality scores, but most of them heavily rely on visual information [6]. In this work, we aim to perform personality recognition without the use of visual features, as previous works have also found that speech features can be informative for the personality recognition task [1].

3 Methodology

Our model follows an audio and text bimodal structure with late fusion. We train each modality separately to predict personality scores and combine those results for our final output.

3.1 Audio Modality

For our audio speech model, we use an encoder-decoder model to generate speech representations that contain personality information. We perform feature extraction on our input to generate speech spectrograms, phone sequences, speaker identity embeddings, and wav2vec2.0 features [2].

The representational encoder takes the wav2vec2.0 features and speaker identity embeddings as input to generate a speech representation. The decoder uses this representation with the phone sequence embeddings and speaker identity embeddings to reconstruct the mel-frequency spectrograms. These speech representations are also used in our regressor to predict personality scores. We minimize the reconstruction loss from the decoder and MSE loss from the regressor.

3.2 Text Modality

For the text-based personality recognition task, we implement an ASR component to generate text transcriptions for each speech sample. From these transcriptions, we extract features using a pre-trained BERT model [4] with which to train our text-based personality recognition model using a convolutional neural network [5].

3.3 Multimodal Fusion

Each modality produces personality score predictions for each input. To combine the results, we assign a weight vector to each modality to apply different weights for each individual trait.

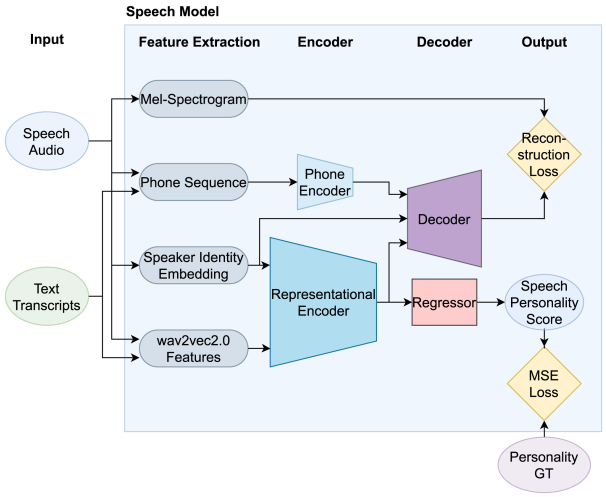

Fig. 1 Architecture for the speech modality, as described in Section 3.1.

Method	O	C	E	A	N	Avg ↓
UDIVA	0.74	0.79	0.89	0.65	1.01	0.82
SMART-SAIR	0.71	0.72	0.87	**0.55**	1.00	0.77
FGM Utrecht	0.75	0.69	0.92	0.67	1.10	0.82
Speech* (ours)	0.27	0.65	0.29	0.61	0.72	0.51
Text* (ours)	0.21	0.42	0.19	0.69	1.25	0.55
Fusion* (ours)	**0.17**	**0.41**	**0.17**	0.59	**0.68**	**0.41**

*results from the English subset

Table 1 Comparison of the results of our proposed method with challenge results from [6]. The values on the table represent MSE. OCEAN initials refer to the Big Five traits.

4 Experiments

For our personality recognition task, we use the UDIVA dataset, which contains video recordings of dyadic sessions between participants and their normalized self-reported personality scores [6]. In this work, we only use the English sessions of the dataset. During our training and evaluation, we perform five fold cross-validation to separate our data.

Comparing our results to the existing methods in Table 1, we observe that our model outperforms existing methods on nearly all personality traits. On average, our mean squared error for the fusion model is 0.41, which is much smaller than that of the average performance of the winning solution of the ChaLearn challenge at 0.77 [6].

5 Conclusion

The multimodal personality recognition system presented here outperforms existing methods evaluated on the UDIVA dataset. Disentangling speaker and phone information from speech representations allows our system to better identify personality cues from speech audio signals. This allows us to outperform other methods which use other modalities such as vision within the English subset of the dataset.

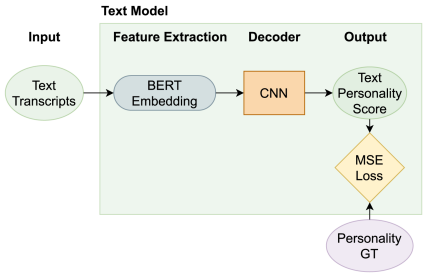

Fig. 2 Architecture for the text modality, as described in Section 3.2.

References

[1] Guozhen An and Rivka Levitan. Lexical and Acoustic Deep Learning Model for Personality Recognition. In *Interspeech*, pages 1761–1765, 2018.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020.

[3] Boele de Raad. *The Big Five Personality Factors: The psycholexical approach to personality.* Hogrefe & Huber Publishers, 01 2000.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL: Human Language Technologies*, volume 1. ACL, June 2019.

[5] Mariana Rodrigues Makiuchi, Kuniaki Uto, and Koichi Shinoda. Multimodal emotion recognition with high-level speech and text features. In *2021 IEEE ASRU Workshop*, pages 350–357. IEEE, 2021.

[6] Cristina Palmero, German Barquero, et al. Chalearn LAP challenges on self-reported personality recognition and non-verbal behavior forecasting during social dyadic interactions: Dataset, design, and results. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, volume 173 of *PMLR*, pages 4–52, 16 Oct 2022.