Personality Trait Perception from Speech Signals Using Multiresolution Analysis and Convolutional Neural Networks

Ming-Hsiang Su[*], Chung-Hsien Wu[*,†], Kun-Yi Huang[*], Qian-Bei Hong[†] and Hsin-Min Wang[†]
[*] Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan
E-mail: {huntfox.su, chunghsienwu, iamkyh77}@gmail.com Tel: +86-06-2089349
[†] PhD Program for Multimedia Systems and Intelligent Computing,
National Cheng Kung University and Academia Sinica, Taiwan
E-mail: qbhong75@gmail.com, whm@iis.sinica.edu.tw Tel: +86-2-2788-2399#1714, 1507

***Abstract*— This study presents an approach to personality trait (PT) perception from speech signals using wavelet-based multiresolution analysis and convolutional neural networks (CNNs). In this study, first, wavelet transform is employed to decompose the speech signals into the signals at different levels of resolution. Then, the acoustic features of the speech signals at each resolution are extracted. Given the acoustic features, the CNN is adopted to generate the profiles of the Big Five Inventory-10 (BFI-10), which provide a quantitative measure for expressing the degree of the presence or absence of a set of 10 basic BFI items. The BFI-10 profiles are further fed into five artificial neural networks (ANN), each for one of the five personality dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism for PT perception. To evaluate the performance of the proposed method, experiments were conducted over the SSPNet Speaker Personality Corpus (SPC), including 640 clips randomly extracted from the French news bulletins in the INTERSPEECH 2012 speaker trait sub-challenge. From the experimental results, an average PT perception accuracy of 71.97% was obtained, outperforming the ANN-based method and the Baseline method in the INTERSPEECH 2012 speaker trait sub-challenge.**

I. INTRODUCTION

Personality traits (PT) can be defined as the characteristics of a person that uniquely influence their cognitions, motivations, and behaviors in different situations [1]. Understanding the personality of other people is extremely useful in establishing effective relationships with others. In recent years, automatic personality recognition (APR) and automatic personality perception (APP) systems in speech and essays have received considerable attention [2-8]. In psychology, the Big Five model is the most common representation of personality [9], and the Big Five Inventory (BFI) [10] has been developed for measuring the five dimensions: Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N). Table I shows the BFI-10 questionnaire [10] and instructions used to evaluate a speaker's PT. The manually evaluated PT scores can be obtained using the item scores rated by the judges through a simple calculation. For example, the score of extraversion can be obtained by subtracting the evaluation score of *Q1* from that of *Q6*.

TABLE I
THE BFI-10 QUESTIONNAIRE USED IN THE EXPERIMENTS AND THE INSTRUCTION TO EVALUATE THE SPEAKER'S PT (AS PROPOSED IN [10]).

Instruction: How well do the following statements describe the speaker's personality?

I see the speaker as someone who…	**D_S**	**D_L**	**N**	**A_L**	**A_S**
… is reserved	(1)	(2)	(3)	(4)	(5)
… is generally trusting	(1)	(2)	(3)	(4)	(5)
… tends to be lazy	(1)	(2)	(3)	(4)	(5)
… is relaxed, handles stress well	(1)	(2)	(3)	(4)	(5)
… has few artistic interests	(1)	(2)	(3)	(4)	(5)
… is outgoing, sociable	(1)	(2)	(3)	(4)	(5)
… tends to find fault with others	(1)	(2)	(3)	(4)	(5)
… does a thorough job	(1)	(2)	(3)	(4)	(5)
… gets nervous easily	(1)	(2)	(3)	(4)	(5)
… has an active imagination	(1)	(2)	(3)	(4)	(5)

D_S: Disagree strongly; **D_L**: Disagree a little; **N**: Neither agree nor disagree; **A_L**: Agree a little; **A_S**: Agree strongly.

Although various studies on PT topic in speech and essays have demonstrated the benefits using different features and classifiers [5], most PT studies in speech only considered the feature set in the full band of speech segments. However, microscopic characteristics of speech signals may convey pertinent information for characterizing a speaker's PT. As feature sets in different frequency bands of speech segments may provide different contributions to PT, multiresolution analysis is thus beneficial to PT perception.

In this study, the Big Five model is used to represent the five dimensions of personality, namely OCEAN. The proposed method is divided into three phases. In the first phase, the initial speech signals are decomposed, using wavelet transform, into various resolution levels, and a set of features are extracted using the openSMILE feature extraction toolkit [11] for the signals at each resolution level. The second phase includes the

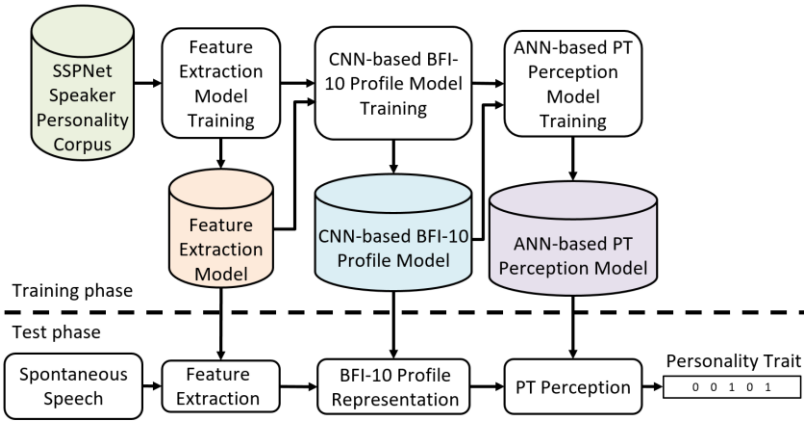

Fig. 1 Schematic diagram of PT perception system.

convolutional neural network (CNN) [12-13] which is adopted to generate the BFI-10 profiles for each speech segment. The multi-dimension neuron property in CNN is used to characterize the feature set in different band of speech segments to generate the BFI-10 profile of a speaker. The third phase is PT perception using an ANN model to determine the PT of a speaker from the BFI-10 profiles.

Fig. 1 shows the block diagram of the proposed PT perception system. In the training phase, the SSPNet Speaker Personality Corpus [5] is used for training and evaluation. The corpus includes 640 speech clips (10 seconds each) for a total of 322 subjects. Each clip was assessed by 11 judges in terms of the Big Five Personality traits. First, the speech and PT evaluation scores in SSPNet SPC were used for feature extraction and the CNN-based BFI-10 profile model training. The CNN-extracted BFI-10 profiles of the speech signals in the corpus were then used to construct an ANN-based PT perception model. In the test phase, for a spontaneous speech, multi-resolution frequency bands were decomposed by using the wavelet transform, and the acoustic features were extracted using openSMILE toolkit. The BFI-10 profiles were then obtained by feeding the acoustic features into the CNN-based BFI-10 profile model. Finally, given the generated BFI-10 profiles, the PT of an individual was perceived using the ANN-based PT perception model.

The rest of this paper is organized as follows: Section II presents the state-of-the-art approaches. Next, the system framework is described in Section III. Section IV presents experimental setup and results. Finally, conclusions and future works are described in Section V.

II. STATE-OF-THE-ART APPROACHES

In recent studies, textual and speech features extracted from speech, social media content, and essays have been adopted to analyze PTs (Table II). Mohammadi and Vinciarelli [5] presented the experimental results by performing over the largest database used so far in terms of the numbers of samples and individuals on prosody-based APP. The experimental results show an accuracy ranging between 60 and 72 percent (depending on the trait) in predicting whether a speaker is perceived to be high or low with respect to a given trait. Mairesse et al. [14] used linguistic and prosodic features for recognizing PTs in both speech and essays and reported the experimental results. Mohammadi and Vinciarelli [15] proposed an approach to automatically predicting traits that listeners attribute to a speaker they have never heard before. They employed a logistic regression model and support vector machine (SVM) to classify each PT dimension by using prosodic features, and they reported an APP accuracy of 60%-72% for various PTs. The study in [15] showed that the perception accuracy is higher for extroversion and conscientiousness. These two traits tend to be perceived with higher consensus in zero acquaintance scenarios. Salamin et al. presented experiments on APR and conflict handling style based on nonverbal communication. The experimental results show that the performance higher than chance to a statistically significant extent can be achieved for one personality trait (Neuroticism) and two conflict handling styles (Dominating and Obliging). Su et al. [2] proposed an approach to PT perception of two speakers over dyadic conversations using RNNs and C-HMMs, and the overall accuracy achieved 85.61%. Mairesse et al. [14] reported experimental results for APR, in both conversation and text, utilizing both self and observer ratings of personality. Their results show that for some traits, any type of statistical model performed

TABLE II
PERSONALITY-RELATED RESEARCH LITERATURE.

Data type	Method	Assessment	Questionnaire
	SVM [5]	Perceived and Self-assessed	NEO-FFI and BFI-10
	SMO [14]	Perceived	BFI-44
	Logistic Regression and SVM [16]	Perceived	BFI-10
	SVM [7]	Perceived	BFI-10
	C-HMM [2]	Perceived	BFI-10
	SMO [14]	Self-assessed	BFI-44
	weighted ML-kNN model [8]	Self-assessed	BFI-44

significantly better than the baseline, but ranking models performed the best overall. Zuo et al. [8] introduced a weighted k-nearest neighbor model to predict a user's PT based on linguistic and emotional features.

III. System Framework

A. Wavelet Transform

The wavelet transformation [16] is one of the most popular time-frequency transformations, and is also a form of data compression well suited for image compression and audio compression. Fig. 2 illustrates the decomposition procedure. Original signal *S* is filtered by high and low pass filters respectively, and the corresponding coefficients are extracted. Several identical wavelet filters with different scales and shift factors are performed in the decomposition processes, and the length of coefficients is reduced to half of original signal each time during wavelet filtering. Related equation is referred to Eq. 1, and $s(t)$ represents the original signal. $\Psi((t\text{-}b)/a)$ is the wavelet filter of different parameters where a is scale parameter and b is shift parameter. $S(a,b)$ is then the coefficients after decomposition. Since Daubechies filters are the most widespread and effective mother wavelet, this study adopts the Daubechies filters for wavelet decomposition in our experiments, and outputs the decomposed signals Lo_3, Hi_3, Hi_2, and Hi_1.

$$S(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} \overline{\psi(\frac{t-b}{a})} s(t) d(t) \tag{1}$$

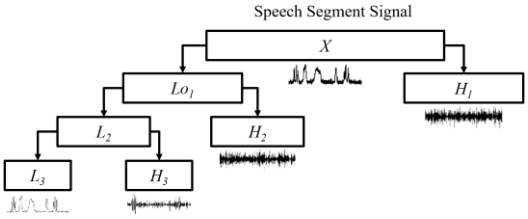

Fig. 2 Process of wavelet decomposition.

B. CNN for BFI-10 Profile Representation

Recently, many researchers have kept their eyes on CNNs [17] which have been shown to be effective for sentence classification tasks [18-19] and other related tasks [20-21]. In this study, the BFI-10 profiles are extracted by using the CNN-based models, as shown in Fig. 3.

Let $x_i^B \in \mathbb{R}^k$ be the k-dimensional acoustic feature corresponding to the i-th segment in the B filter output, where each segment is the k-dimensional acoustic features which are extracted by using a fixed window size with fifty percent overlap. The acoustic features of the speech signal with segment length N are represented as

$$x_{1:N}^B = x_1^B \oplus x_2^B \oplus \ldots \oplus x_N^B \tag{2}$$

where $\oplus$ is the concatenation operator, $B \in \{L_3, H_3, H_2, H_1\}$. A convolution operation involves D filters $W_B^d \in \mathbb{R}^{hk}$, which is applied to a window of h segments to produce the new features. For example, the features are generated from a window of segment $x_{i:i+h-1}^B$ by

$$c_{segment,i}^d = f(W_B^d \cdot x_{i:i+h-1}^B + b) \tag{3}$$

where $b \in \mathbb{R}$ is a bias term and f is a non-linear function such as the sigmoid function. Finally, we can produce k feature maps

$$c_{frame} = [c_{frame,1}, c_{frame,2}, \ldots, c_{frame,n-h+1}] \tag{4}$$

We then apply a max-overtime pooling operation over the feature maps and take the maximum value as the feature corresponding to this particular filter. These features are passed to a fully connected layer to output the BFI-10 profiles.

C. ANN for PT Perception

An ANN model (or backpropagation network) consists of an input layer, hidden layers and a final layer of output neurons [22]. Each connection between different layer neurons is associated with a numeric number called weight. The slot output of node i in the output layer is computed as

$$sl_i(t) = a(\sum_{j=1}^{H} W_{ij}(a(\sum_{k=1}^{N} W_{jk} F^t + T_j^{hid})) + T_i^{out}) \tag{5}$$

where $a(\cdot)$ is the activation function, N is the number of input nodes, H is the number of hidden nodes, W_{ij} and W_{jk} are the interconnection weights, F^t is the BFI-10 profile, T_j^{hid} is the threshold of the j-th hidden node and T_i^{out} is the threshold of the i-th output node. The BFI-10 profile is fed to the ANN-based PT perception model to determine the final PT.

IV. Experiment Setup and Experimental Results

A. Experimental setup

This study focused on three different methods for PT perception. The first method was to consider the entire contents of a speech segment and evaluate the accuracy of PT perception using ANN, which has been used in several approaches [16, 23]. The second and third methods were to consider multilevel wavelet decomposition with reconstructed single branch from wavelet coefficients and approximation coefficients. The second method, multilevel wavelet decomposition with reconstructed single branch from wavelet coefficients, computed the vector of the reconstructed coefficients, and reconstructed the single branch. The third method, multilevel wavelet decomposition with approximation and detail coefficients, computed the approximation and detail coefficients of a one-dimensional signal. The "wavedec" and "wrcoef" functions in Matlab were used to implement the second method, and we used "wavedec", "detcoef" and "appcoef" functions to implement the third method.

The experimental setup was based on the K-fold (K = 5) cross-validation method, and totally 5757-dimensional acoustic features were extracted by using the openSMILE feature extraction toolkit. For implementing the proposed

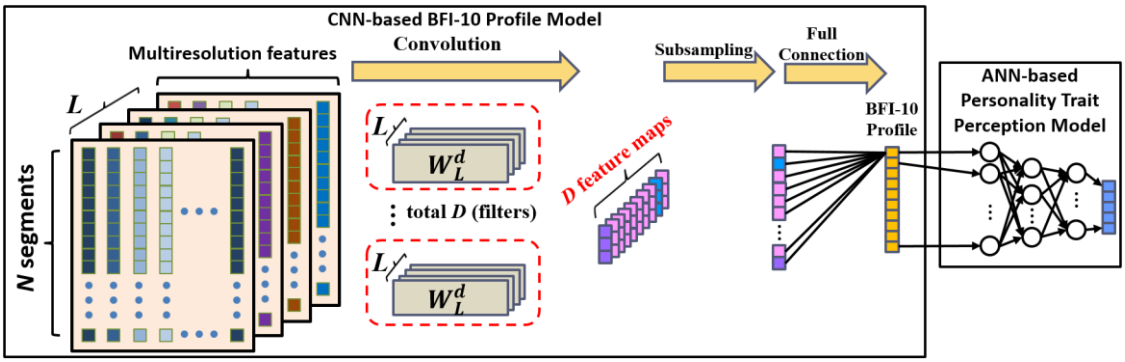

Fig. 3 Schematic diagram of the CNN for PT profile generation.

method, the DeepLearnToolbox was used to construct the CNN and ANN models [24]. For the ANN-based classifiers, we trained five ANN models corresponding to five PTs; The ANN input was 5757-dimensional acoustic features, and the ANN output was the high/low value for each of the five PTs.

B. Evaluation of PT Perception

Table 3 shows the accuracy of the Baseline system in the INTERSPEECH 2012 speaker trait sub-challenge [25], ANN-based and the proposed methods with three different experimental settings. In Table 3, **Best*** is the best accuracy for each of OCEAN in the INTERSPEECH 2012 Speaker Trait Sub-Challenge; **WT-CNN-L2** is the proposed wavelet-based multiresolution analysis using multilevel wavelet level-2 decomposition with approximation coefficients followed by CNN; **WT-CNN-L3** is the proposed method using multilevel wavelet level-3 decomposition with approximation coefficients and CNN; **WT-CNN-L3-Recon** is the proposed method using multilevel wavelet level-3 decomposition with reconstructed single branch from wavelet coefficients and CNN.

In the experimental results, PT perception achieving the best accuracy was the Conscientiousness trait, and the worst was Openness trait in all methods. For Openness trait, the proposed method using multilevel wavelet level-3 decomposition with approximation coefficients achieved the best, but the obtained accuracy was lower than the best accuracy of the PT challenge by 6%. For Conscientiousness trait, the proposed method with multilevel wavelet level-2 decomposition with approximation coefficients outperformed other methods, and the accuracy was higher than the best accuracy of the PT challenge by 10.45%. For Extraversion and Agreeableness traits, the method using multilevel wavelet level-3 decomposition with reconstructed single branch from wavelet coefficients obtained the best accuracy. The accuracy of Extraversion was lower than the best accuracy of the PT challenge by 1.25%, and the accuracy of Extraversion was higher than the best accuracy of the PT challenge by 3.28%. For Neuroticism trait, the method using multilevel wavelet level-3 decomposition with approximation coefficients performed the best, and the accuracy was higher than the best accuracy of the PT challenge by 1.77%.

Overall, the proposed methods with different wavelet settings are better than the Baseline and the ANN-based methods, and the proposed method using the multilevel wavelet level-2 decomposition with approximation coefficients obtained the best accuracy of **71.97**%. The experimental results show that the accuracy of PT perception can be improved by considering the contributions from different resolutions of the speech signals.

TABLE III
EXPERIMENTAL RESULTS OF THE METHODS FOR COMPARISON

	Best* [25]	**Baseline** [25]	**ANN**	**WT-CNN-L2**	**WT-CNN-L3**	**WT-CNN-L3-Recon**
O	**62.50%**	59.00%	48.13%	56.69%	**57.26%**	54.33%
C	80.10%	79.10%	89.85%	**90.55%**	90.32%	85.04%
E	**79.20%**	75.30%	77.66%	73.23%	72.58%	**77.95%**
A	66.80%	64.20%	66.72%	69.29%	68.55%	**70.08%**
N	69.20%	64.00%	67.81%	70.08%	**70.97%**	67.72%
Average Accuracy	-	68.32%	70.03%	**71.97%**	71.94%	71.02%

V. CONCLUSIONS

This work presents an approach to PT perception of speech signals using multiresolution analysis and CNN, followed by the ANN-based perception model. First, the wavelet transform is employed for high and low frequency separation for the speech signal of a speaker. Then, the acoustic features are extracted by using openSMILE toolkit. Next, the BFI-10 score of each signal is evaluated by using the CNN-based BFI-10 profile model. The BFI-10 profiles of the speech signals are fed to the ANN-based PT perception model to determine the final PT. From the experimental results, an average PT perception accuracy of 71.97% was obtained, outperforming the Baseline method in the INTERSPEECH 2012 speaker trait sub-challenge and the traditional ANN-based method.

There are several issues needed to be further explored in the future. First, the relationship between emotion and personality is worth exploring. The user's personality and emotional information can provide the human computer interaction system with flexible and versatile reaction. Second, facial expression is a complex signal, diversity of observations is useful for PT perception. Therefore, the PT perception based on speech-based emotional expression [26-27] and facial expression can be employed in our future work.

REFERENCES

[1] M. McTear, Z. Callejas, and D. Griol, "Emotion, Affect, and Personality," in *The Conversational Interface*, Springer International Publishing, May 2016, pp. 309–327.

[2] M.-H. Su, C.-H. Wu, and Y.-T. Zheng, "Exploiting turn-taking temporal evolution for personality trait perception in dyadic conversations," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no 4, pp. 733-744, February 2016.

[3] R. Gao, B. Hao, S. Bai, L. Li, A. Li, and T. Zhu, "Improving user profile with personality traits predicted from social media content," in *the 7th ACM Conf. Recommender Syst.*, October 2013, pp. 355–358.

[4] D. Markovikj, S. Gievska, M. Kosinski, and D. Stillwell, "Mining Facebook data for predictive personality modeling," in *the 7th Int. AAAI Conf. Weblogs Social Media (ICWSM)*, June 2013, pp. 1–4.

[5] G. Mohammadi and A. Vinciarelli, "Automatic personality perception: Prediction of trait attribution based on prosodic features," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 273–284, April 2012.

[6] T. Polzehl, K. Schoenenberg, S. Möller, F. Metze, G. Mohammadi, and A. Vinciarelli, "On speaker-independent personality perception and prediction from speech," in *Interspeech*, September 2012, pp. 258–261.

[7] H. Salamin, A. Polychroniou, and A. Yinciarelli, "Automatic recognition of personality and conflict handling style in mobile phone conversations," in *14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, October 2013, pp. 1–4.

[8] X. Zuo et al., "A weighted ML-KNN model for predicting users' personality traits," in *International Conference on Information Science and Computer Applications (ISCA)*, September 2013, pp. 345–350.

[9] G. Saucier and L. Goldberg, "The language of personality: Lexical perspectives on the five-factor model," in *The Five-Factor Model of Personality: Theoretical Perspectives*, J. Wiggins, Eds. New York, USA: The Guilford Press, 1996, pp. 21–50.

[10] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the big five inventory in English and German," *Journal of research in Personality*, vol. 41, no. 1, pp. 203–212, February 2007.

[11] F. Eyben, and B. Schuller, "openSMILE:) The Munich open-source large-scale multimedia feature extractor," *ACM SIGMM Records*, vol. 6, no. 4, pp. 4-13, December 2014.

[12] Y. Zhang, and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," *arXiv preprint arXiv:1510.03820*, 2015.

[13] Y. Zhang, S. Roller, and B. Wallace, "MGNC-CNN: A Simple Approach to Exploiting Multiple Word Embeddings for Sentence Classification," in *the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, June 2016, pp. 1522-1527.

[14] F. Mairesse, M. A. Walker, M. R. Mehl and R. K. Moore, "Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text," *Artificial Intelligence Research*, vol. 30, pp. 457-500, September 2007.

[15] T. Polzehl, K. Schoenenberg, S. Möller, F. Metze, G. Mohammadi, and A. Vinciarelli, "On Speaker-Independent Personality Perception and Prediction from Speech," in *Interspeech*, September 2012, pp. 258-261.

[16] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Transactions on Information Theory*, vol. 36, no 5, pp. 961-1005, September 1990.

[17] M. Francis-Landau, G. Durrett, and D. Klein, "Capturing semantic similarity for entity linking with convolutional neural networks," in *the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, June 2016, pp. 1256-1261.

[18] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *the 52nd Annual Meeting of the Association for Computational Linguistics*, June 2014, pp. 655–665.

[19] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in *the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, October 2014, pp. 1746-1751.

[20] L. Dong, F. Wei, M. Zhou, and K. Xu, "Question Answering over Freebase with Multi-Column Convolutional Neural Networks," in *the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, July 2015, pp. 260-269.

[21] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "Learning semantic representations using convolutional neural networks for web search," in *the 23rd International Conference on World Wide Web*, April 2014, pp. 373-374.

[22] S.-C. Wang, "Artificial neural network," in *Interdisciplinary computing in java programming*, Springer US, 2003, pp. 81-100.

[23] R. Gao, B. Hao, S. Bai, L. Li, A. Li and T. Zhu, "Improving user profile with personality traits predicted from social media content," in *the 7th ACM conference on Recommender systems*, October 2013, pp. 355-358.

[24] R. B. Palm, *Prediction as a candidate for learning deep hierarchical models of data*, Technical University of Denmark, 2012.

[25] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, and G. Mohammadi, "A survey on perceived speaker traits: personality, likability, pathology, and the first challenge," *Computer Speech & Language*, vol. 29, no. 1, pp. 100-131, January 2015.

[26] C.-H. Wu and W.-B. Liang, "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels," *IEEE Trans. Affective Computing*, Vol. 2, No. 1, January-March 2011, pp. 10~21.

[27] C.-H. Wu, J.-C. Lin, W.-L. Wei, "A Survey on Audiovisual Emotion Recognition: Databases, Features, and Data Fusion Strategies," *APSIPA Transactions on Signal and Information Processing*, Vol. 3, e12, DOI: 10.1017/ATSIP.2014.11, Published online: 11 November 2014.