This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2531286, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Exploiting Turn-Taking Temporal Evolution for Personality Trait Perception in Dyadic Conversations

Ming-Hsiang Su, Chung-Hsien Wu, *Senior Member, IEEE*, and Yu-Ting Zheng

***Abstract*—In dyadic conversations, turn-taking is a dynamically evolving behavior strongly linked to paralinguistic communication. Turn-taking temporal evolution in a dyadic conversation is inevitable and can be incorporated into a modeling framework for characterizing and recognizing the personality traits (PTs) of two speakers. This study presents an approach to automatically predicting PTs in a dyadic conversation. First, a recurrent neural network (RNN) was used to model the relationship between Big Five Inventory 10 (BFI-10) items and linguistic features of spoken text in each turn of a speaker (speaker turn) to output a BFI-10 profile. The RNN applies a recurrent property to characterize the short-term temporal evolution of a dialog. Second, the coupled hidden Markov model (C-HMM) was employed to model the long-term turn-taking temporal evolution and cross-speaker contextual information for detecting the PTs of two individuals for the entire dialog represented by the BFI-10 profile sequence. The Mandarin Conversational Dialogue Corpus was used for evaluation. The evaluation result shows that an average perception accuracy of 79.66% for the big five traits was achieved using five-fold cross validation. Compared with conventional HMM and support vector machine-based methods, the proposed approach achieved a more favorable performance according to a statistical significance test. The encouraging results confirm the usability of this system for future applications.**

***Index Terms*—Personality Trait Perception, Big-Five Personality Trait, Dyadic Conversation, Coupled Hidden Markov Model.**

I. Introduction

DYADIC communication refers to an interaction between two individuals and "*each participant affects and is affected by the other* [1]". Thus, in daily conversation, two individuals can acquire useful information regarding each other's roles and needs through dyadic communication [2]. Spoken expression is a dynamically evolving interaction in dyadic conversations. In a dyadic conversation, temporal evolution information is a prominent cue for characterizing a speaker's characteristics. In addition, interactions in dyadic conversations have various degrees of mutual influence caused by turn-taking dialogs between two individuals [3]. Turn-taking temporal evolution has proven to be crucial for automatically recognizing the participants' state, such as interaction style, attitude, emotion, and personality [4]. Understanding the participants' states is beneficial for providing harmonious communication between humans and computers [5-7].

As described in [8], personality is the collection of all attributes, such as emotional, mental, temperamental and behavioral states, associated with an individual. In recent years, automatic personality recognition (APR) and automatic personality perception (APP) systems have received considerable attention [9-17]. In psychology, the *Big Five model*, also called the *Five-Factor model*, is the most common representation of personality and has five major dimensions of individual differences that "*appear to provide a set of highly replicable dimensions that parsimoniously and comprehensively describe most phenotypic individual differences*" [18]. Several inventories, such as the Eysenck Personality Questionnaire (EPQ) [19], Revised NEO Personality Inventory (NEO-PI-R) [20], and the Big Five Inventory (BFI) [21], have been developed for measuring the five dimensions.

Among these inventories, BFI-10 is an abridged version of the well-established BFI, and this questionnaire can be completed in less than one minute [22]. Table I shows the BFI-10 questionnaire [21] and instructions used to evaluate a speaker's PT. For manual PT evaluation based on the BFI-10 questionnaire for the input spoken text, five score levels, ranging from 1 to 5, were used to represent the degree of each BFI-10 item for the input spoken text perceived by the judges.

The manually evaluated PT scores can be obtained using the item scores rated by the judges through a simple calculation (*Qi* is the evaluation score of item *i*). For example, the score of extraversion can be obtained by subtracting the evaluation score of *Q1* from that of *Q6*. Each PT score is calculated as follows:

1) **Extraversion**: *Q6 - Q1*
2) **Agreeableness**: *Q2 - Q7*
3) **Conscientiousness**: *Q8 - Q3*
4) **Neuroticism**: *Q9 - Q4*
5) **Openness**: *Q10 - Q5*

This work was supported in part by the Ministry of Science and Technology under Contract MOST102-2221-E-006-094-MY3 and the Headquarters of University Advancement at the National Cheng Kung University, which is sponsored by the Ministry of Education, Taiwan. The authors are with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan (e-mail: huntfox.su@gmail.com; chunghsienwu@gmail.com; starbuckbox@gmail.com).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2531286, IEEE/ACM Transactions on Audio, Speech, and Language Processing

TABLE I
THE BFI-10 QUESTIONNAIRE USED IN THE EXPERIMENTS AND THE INSTRUCTION TO EVALUATE THE SPEAKER'S PT (AS PROPOSED IN [21])

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 perception in speech and essays have demonstrated the benefits using different features and classifiers [13], such studies have rarely evaluated the dynamic aspects in PT perception, particularly dyadic conversations. In general, temporal evolution in a PT expression contains a lot of detailed information. Most PT perception studies only considered a speech segment/sentence or an individual essay. In a dyadic conversation, perception of another person's personality may change over time. In the first time people may perceive others belonging to extraversion, but later found that this is not the case. Only considering a speech segment in a dyadic dialog may lead to inaccurate modeling of a dyadic dialog and achieve unsatisfactory accuracy for PT perception. In addition, cross-speaker contextual information in a conversation between two individuals could show different personalities and should be assessed in PT modeling. In the current study, to precisely perceive the personalities of two individuals, we propose an approach to modeling the evolution of expressions over time and cross-speaker contextual information in a dyadic conversation for PT perception.

In this study, the *Big Five model* was used to represent the five dimensions of personality, namely openness (O), conscientiousness (C), extraversion (E), agreeableness (A), and neuroticism (N). The proposed method is divided into two phases. The first phase includes the recurrent neural networks (RNNs) [23-25] and it entails generating the BFI item scores for each turn of a speaker (speaker turn). During a conversation, an individual spoken text might be projected onto a 10-dimensional point in the BFI space (Fig. 1). In this study, the RNNs were used to describe a short-term point movement by assessing the temporal evolution in the BFI space. The RNNs can output ten BFI-10 scores of spoken text for a single speaker turn, one for each BFI-10 item, to form a BFI-10 score vector, denoted as BFI-10 profile. The BFI-10 profile provides

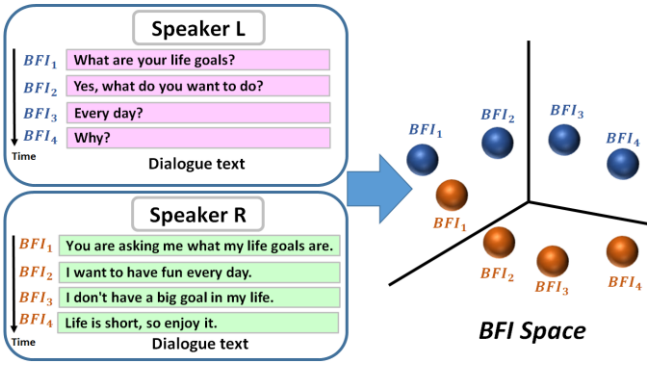

Fig. 1. Illustration of the temporal evolution during a dyadic conversation.

a quantitative measure for expressing the degree of presence or absence of the ten BFI items for each speaker turn. The recurrent property in RNN is used to characterize the short-term temporal evolution of a speaker's state in a dialog. The predicted BFI-10 profile for each speaker turn is used as an input feature vector of the second phase. The second phase includes the coupled hidden Markov models (C-HMMs) [26] and it entails constructing a multiple speaker turn personality perception model by evaluating the long-term turn-taking temporal evolution and cross-speaker contextual information. Each C-HMM models one combination of high or low level for each PT of the two speakers. Finally, the five PTs with high or low level for the two speakers can be determined on the basis of the scores obtained from the C-HMMs.

Fig. 2 shows a block diagram of the proposed PT perception system. In the training phase, the Mandarin Conversational Dialog Corpus (MCDC) [27] was used as the dyadic conversation corpus for training and evaluation. First, the spoken texts of each speaker turn and the corresponding turn-based PT evaluation scores in MCDC were used to train the RNN-based BFI-profile generation model. The linguistic features of the spoken texts in MCDC were then fed into the trained RNN-based BFI-profile generation model to generate the BFI-10 profiles. The generated BFI-10 profiles were finally used to train the C-HMM-based PT perception model.

In the test phase, for a new dyadic conversation, linguistic features from the spoken text of each speaker turn were extracted. The BFI-10 profiles were then obtained by feeding the linguistic features into the RNN-based BFI-profile generation model. Finally, given the generated BFI-10 profiles, the PTs of the two individuals were detected using the C-HMM-based PT perception model.

The rest of this paper is organized as follows. Section II introduces state-of-the-art approaches. Section III describes the procedures involved in data collection and annotation. Section IV presents PT perception based on RNNs and C-HMMs. Section V presents experimental setup and results. Finally, Section VI presents the conclusion and future work.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2531286, IEEE/ACM Transactions on Audio, Speech, and Language Processing

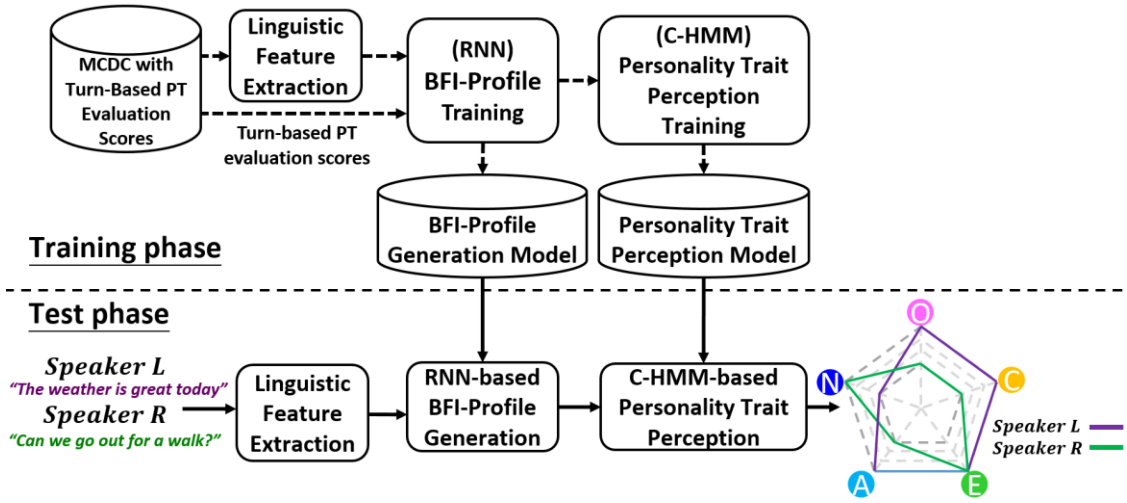

Fig. 2. The basic concept of the proposed method for the perception of Extroversion traits of two individuals in a dyadic conversation.

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). Mairesse et al. [8], [11] have used linguistic and prosodic features for recognizing PTs in both speech and essays and reported the experimental results. Mohammadi and Vinciarelli [13] proposed an approach for 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) classifier to classify each PT dimension by using prosodic features, and they reported an APP accuracy of 60%–72% for various PTs. The study [13] 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. Zuo et al. [17] introduced a weighted k-nearest neighbor model to predict a user's PT based on linguistic and emotional features. Iacobelli et al. [28] attempted to use data-driven approaches for extracting linguistic characteristics directly from collected blog posts, and they adopted word *n*-grams as the linguistic features. In their study, the personality perception model that was based on word *n*-grams achieved a higher prediction performance compared with other models. The experimental results were consistent with their expected positive results. Mohammad and Kiritchenko [29] developed state-of-the-art SVM classifiers and indicated that lexical categories corresponding to fine-grained emotions were considerable indicators of personality. The experimental results revealed that the advantages associated with fine affect categories were not gained when coarse affect categories or specificity features were used alone. Argamon et al. [30] proposed a method for classifying texts according to the personality of the author from casual written text by using functional lexical features and machine learning. The experimental results confirmed the utility of functional lexical features for psychological profiling and indicated the necessity for further refinement in feature sets and learning algorithms.

TABLE II
PERSONALITY-RELATED RESEARCH LITERATURE

Data type	Ref.	Method	Assessment	Questionnaire
Speech-based	[8]	SMO	Perceived and self-assessed	BFI-44
	[13]	Logistic regression and SVM	Perceived	BFI-10
	[14]	SVM	Perceived and self-assessed	NEO-FFI and BFI-10
	[15]	SVM	Perceived	NEO-FFI
	[16]	SVM	Self-assessed	BFI-10 and ROCI-II
Text-based	[8]	SMO	Self-assessed	BFI-44
	[17]	Weighted ML-kNN model	Self-assessed	BFI-44
	[28]	SVM	Self-assessed	IPIP -41
	[29]	Category-based and word-based analyses	Self-assessed	NEO-FFI and NEO-PI-R
	[30]	SVM, KNN and NB	Self-assessed	NEO-PI-R and IPIP
	[31]	SVM	Self-assessed	BFI-44
	[33]	TiMBL	Self-assessed	MBTI
	[34]	Regression model	Self-assessed	BFI
	[35] [36]	Egogram Model	Self-assessed	TEG2
	[37]	SMO	Self-assessed	NEO-FFI
	[39]	NB and SVM	Self-assessed	IPIP

Because of the increasing prevalence of social media websites recently, research related to PT perception has gradually focused on online platforms. Yarkoni [31] conducted a large-scale analysis of personality and words used in blogs. The results showed high-level associations between PTs and aggregate word categories at the linguistic inquiry and word count (LIWC) categorical and single-word levels and emphasized the relevance of complementary approaches for the study of personality. Farnadi et al. [32] collected user activities, including posted articles and user activity time, from Facebook and mainly focused on the relationship between

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2531286, IEEE/ACM Transactions on Audio, Speech, and Language Processing

users' activities and their PTs. Luyckx and Daelemans [33] analyzed the user's personality and achieved satisfactory results for six of eight binary classification tasks, including Introverted, Extraverted, Intuitive, Sensing, Feeling, Thinking, Judging and Perceiving. Gill et al. [34] used a large blog corpus annotated using the LIWC text analysis program and examined the blog content to gain insight into the role of personality in motivation. The experimental results revealed that bloggers tended to adapt to the possibilities of the medium rather than attempting to present themselves differently. Minamikawa and Yokoyama [35-36] have conducted a personality estimation by using Japanese weblog text. They executed the estimation by using the multinomial naïve Bayes classifier with feature words, and the proposed method achieved significant improvements in personality estimation. Nowson and Oberlander [37] used an *n*-gram-based approach to identifying phrases associated with differences in the Big Five dimensions and suggested that opinion mining could benefit from personality information. This result suggests that incorporating personality models into other tasks may improve the accuracy of the models. Tausczik and Pennebaker [38] explored the links between word usage and basic social and personality processes and determined the correlations of word categories through numerous studies. Oberlander and Nowson [39] classified weblog authors according to four crucial PTs by using a corpus of personal weblogs and reported the results by employing a relatively novel approach involving automatic classification of author personality. The results were highly promising and comparable across all four PTs.

III. DATA COLLECTION AND ANNOTATION

A. Corpus Collection

The MCDC [27] comprises dialogs with turn-taking conversations between two speakers. This corpus includes 16 subjects that participate in multiple conversations. During the corpus collection process, two conversing speakers met each other for the first time, and no specific topics were provided to the speakers. They had complete freedom to choose and change topics. The original content of the MCDC contains text and speech files. In this study, we used only manually transcribed texts of the dialogs in MCDC.

For the corpus processing, we first removed the backchannel in MCDC, such as "yes" and "uh-huh," without relinquishing the turn to ensure that no interposition existed in each dialog when one of the two speakers is speaking. For convenience, we denoted the two speakers as "*Speaker L*" and "*Speaker R*." Fig. 3 illustrates an example of a dialog fragment in the corpus. The time variable t denotes the speaker turn. For example, "$L(t)$" represents the t-th speaker turn of Speaker L.

Finally, a total of 310 dialogs were collected, comprising 6510 single speaker turns. In this corpus, the number of speaker turns ranges from 2 to 29, but the data with turn number smaller than 6 or larger than 20 are quite few. Table III lists the detailed information of the corpus.

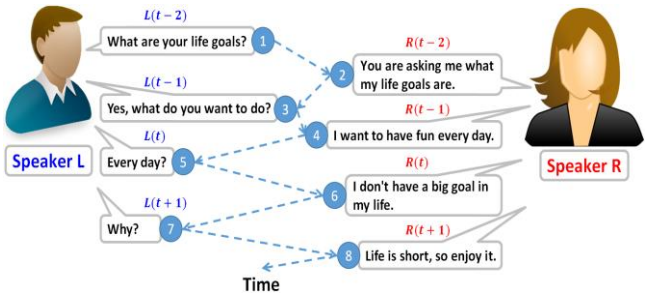

Fig. 3. Example of a dialog in the corpus (fragment only).

TABLE III
CORPUS INFORMATION (AFTER ARRANGEMENT)

Total	
Number of subjects	16
Number of dialogs	310
Number of speaker turns for two speakers	6510
Word count	85951

For each dialog of a speaker	
Avg. number of speaker turns in each dialog of a speaker	10.5
Avg. number of words in each speaker turn	13.08
Avg. time (sec) of each speaker turn	4.875

B. Corpus Annotation

In this study, to obtain the PT scores of a dialog by using the Big Five model, we adopted the BFI-10 questionnaire for manual evaluation [22]. We recruited four judges for evaluating PTs. The personality evaluation process for each dialog is described as follows.

1) For each single speaker turn, each judge used the BFI-10 questionnaire to evaluate each item in BFI-10 and assign a score ranging from 1 to 5. During the evaluation for a specific speaker turn, the contents of its previous two turns produced by the same speaker and the other speaker were concurrently assessed. That is, the judges were requested to listen to the contexts of the current turn $L(t)$ and its preceding turns, $L(t - 1)$ and $R(t - 1)$, to assess the PT. For example, to analyze the PT for the speaker turn $L(t)$, we evaluated the contents of speaker turns $L(t - 1)$ and $R(t - 1)$ simultaneously. This was mainly conducted to ensure that cross-speaker contextual information and turn-taking temporal evolution were assessed.

2) For each dialog, each judge used the BFI questionnaire for evaluating the dialog-based PTs of two speakers, and each speaker obtained a PT evaluation score for the entire dialog. Fig. 4 depicts an example of PT evaluation of a dialog. For example, the speaker turn-based BFI-10 evaluation scores of one sentence in speaker turn $L(t - 2)$ are "1354532211" for the 10 BFI-10 items, respectively, and the dialog-based BFI-10 evaluation scores according to the entire dialog for *Speaker L* are "2341234153." For each speaker turn and dialog, the average item scores over judges were rounded to the nearest integer ranging from 1 to 5 and were used to represent the score for each BFI-10 item.

For manual PT evaluation on the MCDC corpus, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2531286, IEEE/ACM Transactions on Audio, Speech, and Language Processing

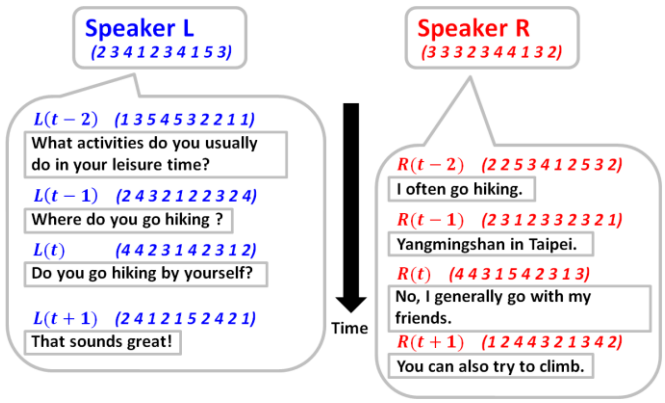

Fig. 4. Example of PT evaluation for a dialog in the corpus (fragment only).

Kendall's W coefficient [40-42] was employed to evaluate the agreement of manual PT evaluation scores among assessors. In agreement evaluation for the corpus with 310 dialogs, the proportion of the 310 dialogs for the consistency higher than "Normal" achieved 98.36%. For the corpus from 6510 single speaker-turns, the proportion with consistency higher than "Normal" was 92.28%. On the whole, the high consistency confirms the reliability of manual PT evaluations.

IV. PT PERCEPTION BASED ON RNN AND C-HMM

A. Feature Extraction

Although the original content of the MCDC contains text and speech files, only the manually transcribed texts of the dialogs were used in this study. In feature extraction, the Chinese Linguistic Inquiry and Word Count (CLIWC) [43-44] and simplified tag set of parts-of-speech (POSs), also called the Chinese Knowledge Information Processing (CKIP) POS tag [45] developed by the Chinese Knowledge Information Processing group of Academia Sinica, were adopted to extract linguistic features; these features comprised a 71-dimensional CLIWC feature vector and a 14-dimensional CKIP POS feature vector for each speaker turn. CLIWC categorizes daily used words to 71 vocabulary categories, including 29 categories of linguistic characteristics and 42 categories of psychological characteristics. Each category includes tens to thousands of words in Chinese. Each word in Chinese may be classified into several categories. There are some examples of CLIWC vocabulary categories in Table IV, and some examples of CKIP POS categories in Table V. These two feature vectors were concatenated to form an 85-dimensional feature vector. Finally, for evaluating the turn-taking temporal evolution, feature vectors from three consecutive speaker turns produced by the two speakers in the dialog were concatenated as the input linguistic feature vector.

The input linguistic feature vector in one dialog with data sequence length T is expressed as follows.

$$X_1^T = x_1, x_2, ..., x_T \tag{1}$$

where x_t represents the linguistic feature vector of the t-th speaker turn.

TABLE IV
EXAMPLES OF CLIWC CATEGORIES

Category name	Abbreviation	Words in category	Examples
Anxiety	anx	129	不安(uneasy)、掙扎(struggled)、緊繃(tight)
Cognitive process	cogmech	1255	理解(comprehend)、選擇(select)、質疑(doubt)
Insight	insight	328	了解(understand)、恍然大悟(take a tumble)、體會(experience)
Causation	cause	128	引起(arouse)、使得(make)、變成(become)
Friend	friend	44	同伴(companion)、朋友(friend)、麻吉(machi)

TABLE V
EXAMPLES OF CKIP POS CATEGORIES

Category	Frequency	Meaning of Category
A	1453	non-predictive adjective
Na	34372	common noun
Nb	14813	proper noun
Nc	9688	location noun
Nd	2264	time noun
VA	6466	active intransitive verb

B. BFI Profile Generation Based on RNN

An RNN [23-25] is an extended version of the conventional artificial neuron network (ANN). The main difference between an RNN and ANN is that the network architecture in an RNN exhibits a recurrent property. In each learning iteration, a feedback value is returned from a hidden layer of an RNN. Therefore, the recurrent property was adopted in the current study to characterize short-term temporal evolution in a dialog.

The activation function of an RNN for the hidden layer is a hyperbolic tangent, and the activation function for the output layer is linear. For an input data I_i at time t, the k-th hidden unit output $V_k(t)$ is computed as follows [46]:

$$V_k(t) = \tanh\left(\sum_{i=1}^{N_1} w_{ki} I_i(t) + \sum_{i=N_1+1}^{N_1+N_2} w_{ki} V_{i-N_1}(t-1)\right) \tag{2}$$

and the output values in the output layer are derived as follows:

$$Y_j(t) = \sum_{k=1}^{N_2} w_{jk} V_k(t), \qquad j = 1, 2, ..., N_3 \tag{3}$$

In (2) and (3), N_1, N_2, and N_3 represent the numbers of units in the input layer, hidden layer, and output layer, respectively. $I_i(t)$ is the i-th input unit at time t, $V_k(t)$ is the output of the k-th hidden unit at time t, $Y_j(t)$ is the output of the j-th output unit at time t, w_{ki} is the weight between the i-th input unit and the k-th hidden unit, w_{jk} is the weight between the k-th hidden unit and the j-th output unit. According to (2), the input value of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2531286, IEEE/ACM Transactions on Audio, Speech, and Language Processing

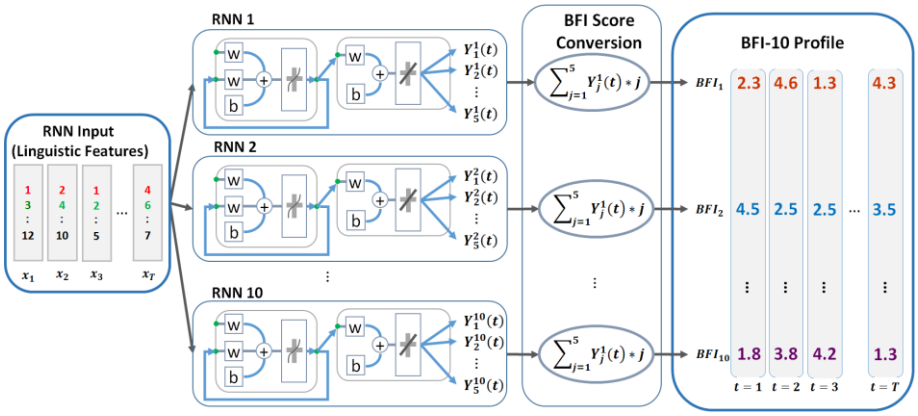

Fig. 5. The architecture for BFI-10 score generation in the BFI-10 profile generation model, in which 10 RNNs, one for each item, were used.

hidden layer in each iteration is referred to the output value from the previous iteration in the hidden layer.

During learning, the weights are updated according to the following expression:

$$\Delta w_{p,q}^{m}(r) = (1-\alpha)\eta\sum_{t=1}^{T}\delta_{p}^{m}(t)V_{q}^{m-1}(t) + \alpha\Delta w_{p,q}^{m}(r-1) \quad (4)$$

where p and q represent the unit number, r is the iteration step, η is the learning rate, α is the momentum parameter, m is the layer number of the RNN. Here, m = 0, 1, and 2 represent the input layer, hidden layer, and output layer, respectively. δ^{m} is obtained as

$$\delta_{j}^{m=2}(t) = D_{j}(t) - Y_{j}(t),\ j = 1,2,...,N_{3} \quad (5)$$

$$\delta_{k}^{m=1}(t) = (1-V_{k}^{2}(t))\sum_{j}W_{jk}^{m=2}\delta_{j}^{m=2}(t),\ k = 1,2,...,N_{2} \quad (6)$$

where $D_{j}(t)$ is the desired output at the j-th output unit at time t. The parameters η and α can be automatically adjusted during training, by using an adaptive learning mechanism [47]:

$$\Delta\eta = \begin{cases} +a\eta, & if\ \Delta E < 0 \\ -b\eta, & if\ \Delta E > 0 \\ 0, & otherwise \end{cases} \quad (7)$$

where ΔE is the cost function change and a and b are appropriate positive constants.

Regarding the generation of the BFI-10 profile, 10 RNNs were independently constructed to model the 10 items in BFI-10. Because each BFI-10 item was evaluated using a score ranging from 1 to 5, an RNN with a coding schema (e.g., 00001 for 1, 00010 for 2, etc.) was employed to model each BFI-10 item. Based on this coding scheme, the desired output of the j-th bit in each RNN is coded with 1 for the score of value j and 0 for the other bits as shown in the first column of Table VI. According to the definition of the correspondence between the codes and RNN output values, the likelihood value for each BFI-10 item can be obtained from the five RNN output nodes. The use of the 5-way bit coding schema is to obtain a more discriminative output value of RNN, preventing the inputs labeled with different classes from receiving the same output value, a condition known as data squashing. Besides, this schema is advantageous for speeding up the convergence of RNN. Fig. 5 illustrates the architecture of the BFI-10 profile generation model. As shown in Fig. 5, the linguistic feature vector in one dialog with a data sequence of length T is used as the input data and the BFI-10 profile vector in one dialog with a data sequence of length T is generated as the output data. The BFI-10 scores for the corresponding spoken text in a speaker turn should be encoded to conform to the desired outputs of the nodes, each with a value of either 0 or 1, in the RNN output layer.

Because each RNN for one BFI-10 item has five coded desired outputs (in which 00001 is assigned to a value of 1 and 00010 assigned to 2), the obtained likelihood values and their corresponding codes were used to obtain the final scores for the linguistic feature vector of the input spoken text. According to the definition of the correspondence between the codes and RNN output values, the likelihood value for each BFI-10 item can be obtained by summing the scores from the five RNN output nodes. Table VI shows that the sixth BFI score is 3.81.

TABLE VI
AN EXAMPLE FOR CONVERTING THE RNN OUTPUTS TO THE BFI-10 SCORE OF THE SIXTH ITEM

Coded Desired RNN Outputs	Normalized RNN Output of Node j	Likelihood Value Conversion	BFI-10 Item Score
00001 (value = 1)	0.06 (j =1)	0.06*1=0.06	Summation of the five values = 3.81
00010 (value = 2)	0.10 (j =2)	0.10*2=0.20	
00100 (value = 3)	0.12 (j =3)	0.12*3=0.36	
01000 (value = 4)	0.41 (j =4)	0.41*4=1.64	
10000 (value = 5)	0.31 (j =5)	0.31*5=1.55	

For a dyadic dialog, the contents of the conversation between two speakers are decomposed into various speaker turns. Considering an individual PT, the linguistic features of the t-th speaker turn x_t are fed into 10 RNNs. For the estimation of the score, the BFI-10 item score of the l-th item BFI_t^l is obtained from the five output values, each calculated by multiplying each normalized RNN output value $Y_j^l(t)$ by its

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2531286, IEEE/ACM Transactions on Audio, Speech, and Language Processing

corresponding code value *j*, as shown in (8).

$$BFI_t^l \equiv P(item^l | x_t) = \sum_{j=1}^{5} \frac{Y_j^l(t)}{\sum_{i=1}^{5} Y_i^l(t)} \times j \quad (8)$$

where $item^l$ represents the *l*-th item in the BFI-10 questionnaire. As the calculated score is a weighted sum of the five RNN outputs, each corresponding to one of the five defined BFI item scores (from 1 to 5), the calculated score could be regarded as the likelihood of the *l*-th item BFI_t^l. Finally, after the scores for the 10 BFI-10 items are calculated, the BFI-10 item profile vector for the *t*-th speaker turn, represented by $BFI_t = [BFI_t^1, BFI_t^2, ..., BFI_t^{10}]$, is generated.

C. PT Perception Using C-HMM

To perceive the PTs of two individuals through temporal evolution evaluation, the C-HMM was adopted in the present study to construct the PT perception model. The C-HMM-based PT perception model accepts the BFI-10 profiles of two speakers in a dialog as the input and returns a high or low level for each of the five PTs for the speakers as the output.

The C-HMM is an extended version of the HMM. The main difference between the C-HMM and HMM is that the C-HMM comprises two HMMs that describe the state transitions of two objects, respectively, and the state transitions in each HMM mutually influence each other [4], [48]. This C-HMM property conforms to our aforementioned research aim. In this study, we employed the C-HMM for modeling turn-taking temporal evolution during the entire dialog and assessed cross-speaker contextual information simultaneously.

Fig. 6 shows the C-HMM architecture, indicating that two state transition sequences correspond to two HMMs. The main feature of the C-HMM is that it assesses cross-speaker contextual information between current state (s_t^L, s_t^R) and its previous state (s_{t-1}^L, s_{t-1}^R).

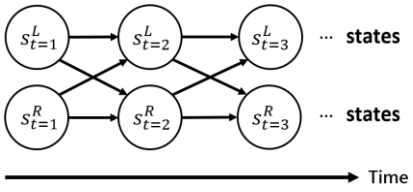

Fig. 6. The architecture of a C-HMM.

Given a sequence of BFI-10 profile $BFI_1^T = [BFI_1, BFI_t, ..., BFI_T]$ with T speaker turns, the likelihood of the state transition sequence with respect to the C-HMM is estimated as follows:

$$P(S_{(L_{pl},R_{pl})} | BFI_{t=1}^T) = P_{s_{t=1}^L} b_{s_{t=1}^L}(BFI_{s_{t=1}^L}) P_{s_{t=1}^R} b_{s_{t=1}^R}(BFI_{s_{t=1}^R}) \times \prod_{t=2}^{T} P_{s_t^L|s_{t-1}^L} P_{s_t^R|s_{t-1}^R} P_{s_t^R|s_{t-1}^L} P_{s_t^L|s_{t-1}^R} b_{s_t^L}(BFI_{s_t^L}) b_{s_t^R}(BFI_{s_t^R}) \quad (9)$$

In (9), $S_{(L_{pl},R_{pl})}$ denotes the state sequence in the C-HMM with PT-level (high or low) pair (L_{pl}, R_{pl}), $b_{s_t^L}(BFI_{s_t^L})$ is the probability of the output of a given state in chain *L*, $BFI_{s_t^L}$ is the BFI-10 profile sequence in chain *L*, $BFI_{s_t^R}$ is the BFI-10 profile sequence in chain *R*, $P_{s_t^L|s_{t-1}^R}$ is the probability of a state in chain *L* given the previous state in chain *R*, and s_t^L and s_t^R denote the states of the opposing HMMs at time *t*.

In this study, 20 C-HMMs are defined and trained to predict the high or low levels of the five personality traits (**O**, **C**, **E**, **A**, **N**) for each speaker in a dialog (Fig. 7). For each PT, the PT-level pair (L_{pl}, R_{pl}) for the two speakers are defined as (0,0), (0,1), (1,0), and (1,1), where 0 represents *low* level, 1 represents *high* level (Table VII). Each C-HMM is used to predict a combination of high or low levels for a specific PT. For example, for evaluating the extraversion PT of two speakers denoted as (***E***, ***E***) in TABLE VII, the PT-level pair (0,1) indicates that the **extraversion** PT level for *Speaker L* is *low* and that for *Speaker R* is *high*.

Finally, considering the input sequence of BFI-10 profile $BFI_1^T = [BFI_1, BFI_t, ..., BFI_T]$, for a specific personality trait $pt \in \{O,C,E,A,N\}$, the PT-level pair (L_{pl}, R_{pl}) of two speakers

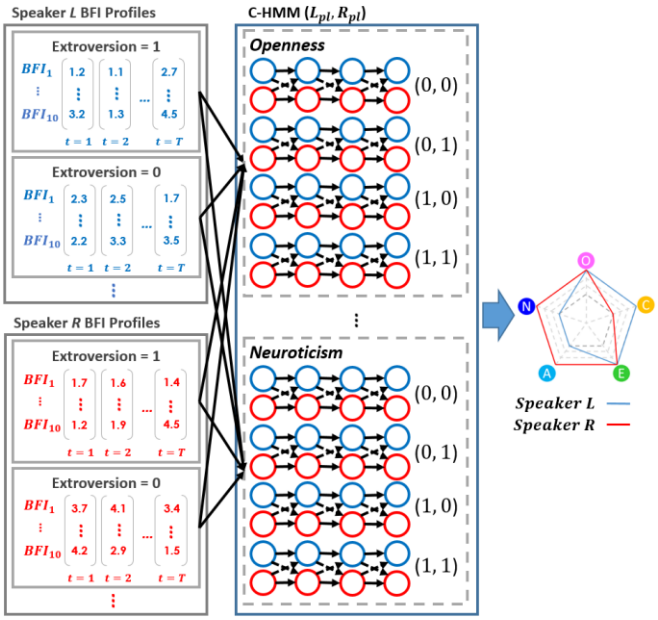

Fig. 7. The architecture of PT perception based on C-HMM.

TABLE VII
20 C-HMMS, ONE FOR EACH PT LEVEL PAIR (L_{PL}, R_{PL})

(O,O)	(C, C)	(E, E)	(A, A)	(N, N)
0:0	0:0	0:0	0:0	0:0
0:1	0:1	0:1	0:1	0:1
1:0	1:0	1:0	1:0	1:0
1:1	1:1	1:1	1:1	1:1

0 represents the *low* extent and 1 represents the *high* extent.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2531286, IEEE/ACM Transactions on Audio, Speech, and Language Processing

(*L* and *R*) in the dialog can be obtained by selecting the PT-level paired C-HMM with maximum likelihood among the four PT-level paired C-HMMs: (0,0), (0, 1), (1, 0) and (1, 1).

$$(L_{pl}, R_{pl})^{pt} = \underset{L_{pl}\in\{0,1\}, R_{pl}\in\{0,1\}}{\operatorname{argmax}} P_{pt}(S_{(L_{pl},R_{pl})} \mid BFI_1^T) \qquad (10)$$

where L_{pl} and R_{pl} are the PT-level values (0 or 1) of speaker *L* and *R*, respectively.

The PT-level paired C-HMM with the highest likelihood is determined, and the corresponding PT pair is considered the PTs of the two speakers. For a detailed demonstration of the perceived PTs, the five PT values of each speaker are individually calculated by evaluating one chain in the determined PT-level pair C-HMM and further normalized to a range of 0–1.

V. EXPERIMENTAL SETUP AND RESULTS

A. Experimental Setup

This study focused on three methods for evaluating the accuracy of PT perception. The first method entailed assessing a speaker's entire contents in a conversation, and the accuracy of PT perception was evaluated using an SVM classifier, which has been successfully used in several research areas [8], [9], [12], [17]. The libSVM was used to construct the SVM-based classifier for PT perception [49]. The second technique involved assessing the BFI space and the evolution of expressions over time in a dyadic conversation, and the accuracy of PT perception was evaluated using RNNs and HMMs. The third method is the approach proposed in this study.

For implementing the BFI-10 profile generation model, the NN Toolbox [50] was used to construct Elman-type RNNs. CHMMBOX v1.2 [51], a third party toolkit developed by the University of Oxford, was used for implementing the C-HMM for PT perception.

B. Parameter Tuning

Parameter tuning was conducted to ensure a fair comparison and evaluation of the SVM- and HMM-based methods. For the SVMs, several kernel functions and parameters were evaluated and fine tuned. According to the results, we set svm_type = C-SVC, kernel_type = radial basis, cost = 0.125 and gamma = 0.0078 which achieved the most favorable results, and these parameter settings were used to construct the SVM-based system for comparison. Regarding the HMMs, the state number from two to six were evaluated. According to the results, the five-state HMM achieved the best performance and therefore was selected as the HMM-based system for the following comparisons. Finally regarding the proposed method, the optimal number of hidden nodes in RNNs, the optimal state number in the C-HMMs and the concatenation type of the linguistic features were evaluated.

In this study, we determined the optimal number of hidden nodes in the RNNs for the proposed method. A 255-dimensional linguistic feature vector was used as the input for the RNNs, and the BFI-10 evaluation scores corresponding to this linguistic feature vector were the desired outputs. We adopted the mean squared error as the evaluation criteria. Table VIII shows the evaluation results, indicating that the mean squared errors tend to converge when the hidden node number was set to 40. Therefore, we set the hidden node number to 40 for the subsequent experiments.

TABLE VIII
EVALUATION RESULTS FOR DIFFERENT HIDDEN NODE NUMBER IN RNNS BASED ON MEAN SQUARED ERROR

HNN	Mean Squared Error									
	BFI-1	**BFI-2**	**BFI-3**	**BFI-4**	**BFI-5**	**BFI-6**	**BFI-7**	**BFI-8**	**BFI-9**	**BFI-10**
5	0.137	0.152	0.154	0.159	0.099	0.139	0.159	0.140	0.151	0.125
20	0.135	0.142	0.146	0.151	0.105	0.140	0.160	0.140	0.147	**0.123**
40	**0.133**	0.144	**0.145**	**0.148**	**0.097**	**0.138**	0.152	**0.139**	**0.145**	0.126
90	0.136	**0.140**	0.155	0.150	0.098	0.143	**0.151**	0.141	0.149	0.129
130	0.135	0.145	**0.145**	0.148	0.101	0.142	0.154	0.142	0.147	0.125

HNN: Hidden Node Number

We also determined the optimal state number *k* in the C-HMMs and the concatenation type of the linguistic features for the proposed method. In this experiment, we tested four concatenation types of linguistic features in the RNN training process to verify the effect of using cross-speaker contextual information in a dialog.

In the first type, we concatenated 85-dimensional linguistic features extracted from a single speaker turn, denoted as C-HMM (*L*, *R*). In the second type, we concatenated two 85-dimensional linguistic feature vectors extracted from two consecutive single speaker turns of the same speaker, denoted as C-HMM (*LL*, *RR*). In the third type, we concatenated three 85-dimensional linguistic feature vectors extracted from three consecutive single speaker turns of the same speaker, denoted as C-HMM (*LLL*, *RRR*). In the fourth type, we concatenated three 85-dimensional linguistic feature vectors extracted from three consecutive single speaker turns between two speakers alternately, denoted as C-HMM (*LRL*, *RLR*).

Table IX shows the experimental results. In Table IX, we tried to find the best performance in considering the optimal state number *k* in the C-HMMs and the concatenation type of the linguistic features for the proposed method. The manually evaluated PT scores could be obtained using the item scores rated by the judges through a simple calculation (i.e., *Q6-Q1*, *Q2-Q7*, etc.), and the calculated score range was between -4 and +4. To obtain the high or low binary values, the score range between -4 to 0 was regarded as low and the score range between 0 to +4 was defined as high. We evaluated the performance by using Eq. (10) to calculate PT scores for determining if the score was high or low, and we compared the binary results with the ground truth PT binary values (high or low) which was converted from the manually evaluated scores. According to the experimental results, the C-HMM (*LRL*, *RLR*) achieved optimal recognition accuracy when the state number *k* was set to 4. Therefore, we chose *k* = 4 for training C-HMMs

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2531286, IEEE/ACM Transactions on Audio, Speech, and Language Processing

and used the type of C-HMM (*LRL*, *RLR*) in the subsequent experiments.

In order to evaluate the effectiveness of the RNN followed by C-HMM, the PT predictor using RNN only without C-HMM was constructed for comparison. Table X shows the comparison between RNN only and C-HMM using 255-dimensional linguistic feature vectors (*LRL*, *RLR*) to show whether modelling temporal sequence/variability (via an HMM-like model) is important. In the experiment, we compared the proposed C-HMM with the method using only RNNs (without HMM or C-HMM) by averaging over the RNN outputs and directly converting the BFI-10 values into 5-way PT values. The input of the RNN was 255-dimensional linguistic feature vectors (LRL, RLR), and the output of RNN was the BFI-10 values. The final PT scores were calculated by using the item scores through a simple calculation (i.e., *Q6-Q1*, *Q2-Q7*, etc.). The results for C-HMM were obtained from the best results in Table IX. The comparison results reveal that the proposed C-HMM (using the BFI-10 values from RNN as the inputs) outperformed the RNN only for PT perception.

TABLE IX
EVALUATION RESULTS USING DIFFERENT CONCATENATION TYPE AND STATE NUMBER IN C-HMM

	(*L,R*)	(*LL,RR*)	(*LLL,RRR*)	(*LRL,RLR*)
Openness				
k=2	58.74%	58.08%	58.41%	69.44%
k=3	60.37%	63.98%	66.84%	73.93%
k=4	60.17%	64.16%	65.75%	**76.80%**
k=5	62.61%	64.13%	64.88%	75.93%
Conscientiousness				
k=2	55.96%	59.45%	59.10%	71.27%
k=3	60.57%	66.66%	63.68%	73.82%
k=4	59.64%	66.77%	66.48%	**77.10%**
k=5	61.46%	66.13%	67.53%	75.39%
Extroversion				
k=2	69.45%	74.90%	77.28%	80.48%
k=3	68.00%	75.83%	78.18%	82.98%
k=4	67.29%	76.40%	77.77%	**85.65%**
k=5	62.72%	76.83%	77.93%	83.91%
Agreeableness				
k=2	57.54%	59.49%	65.06%	64.43%
k=3	61.72%	63.68%	67.33%	68.07%
k=4	63.17%	65.79%	71.35%	**69.90%**
k=5	63.55%	67.81%	69.69%	67.27%
Neuroticism				
k=2	63.41%	66.45%	72.77%	83.16%
k=3	68.74%	72.55%	74.61%	84.50%
k=4	73.17%	74.81%	72.27%	**88.83%**
k=5	68.25%	71.81%	72.96%	87.66%

TABLE X
COMPARISON BETWEEN RNN ONLY AND C-HMM USING 255-DIMENSIONAL LINGUISTIC FEATURE VECTORS (*LRL*, *RLR*)

PT	RNN	Proposed C-HMM
Openness	43.07%	76.80%
Conscientiousness	16.57%	77.10%
Extroversion	47.63%	85.65%
Agreeableness	52.37%	69.90%
Neuroticism	24.85%	88.83%

TABLE XI
DIFFERENT EXPERIMENTAL SETTINGS IN HMM AND C-HMM FOR PT PERCEPTION

PT	HMM-G	C-HMM-G	Proposed C-HMM
Openness	53.55%	40.24%	76.80%
Conscientiousness	35.80%	16.57%	77.10%
Extroversion	13.91%	47.63	85.65%
Agreeableness	17.16%	52.37%	69.90%
Neuroticism	92.31%	24.85%	88.83%

In the next experiment, the gold-standard human-annotated BFI-10 values were used as the input of the HMM and C-HMM (denoted as HMM-G and C-HMM-G, respectively) to evaluate the effect of using RNN for generating BFI-10 values. It is interesting to note that the HMM-G and C-HMM-G methods underperformed the proposed C-HMM using the RNN-generated BFI-10 values as the results in Table XI. A possible reason may result from the recurrent properties of the RNN. The RNN applies a recurrent property to characterize the short-term temporal evolution of a dialog. Because of the recurrent properties, the temporal information could be propagated forward to endow the RNN outputs with historical information.

Furthermore, this study used 10-dimensional BFI value vectors in the RNN classifiers to obtain BFI profiles, rather than using these to derive 5-dimensional Big Five vectors. The intuition for using 10-dimensional BFI value vectors is that it learns intermediate representations between BFI value inputs and Big Five outputs and these intermediate representations could generate better transfer across feature spaces. This concept is quite similar to the hidden layers in deep neural network. This study also conducted the experiments for comparing the effect of using these two feature representations. The results show that the average PT perception accuracy of the RNN classifiers using 10-dimensional BFI value vectors achieved 56.88%, while using 5-dimensional Big Five vectors was only 35.20%. The results are consistent with the intuition for favoring 10-dimensional BFI value vectors.

C. Evaluation of PT Perception of Two Speakers in a Dialog

We evaluated the performance of PT perception on the SVM- and HMM-based methods. The experimental setup is based on the K-fold (K = 5) cross-validation method [52]. In 5-fold cross-validation, the 310 dialogs were randomly partitioned into 5 subsets of equal size, in which the dialog from one subject were grouped together and organized into one subset for subject-independent evaluation. Of the 5 subsets, one subset is retained as the validation data for testing, and the remaining 4 subsets were used for training. For the SVM-based classifiers, we trained five SVMs corresponding to five PTs; each SVM-based classifier was used to predict the extent of a specific trait as either high or low. For the evaluation of the SVM-based classifiers, two types of input data were

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2531286, IEEE/ACM Transactions on Audio, Speech, and Language Processing

considered as the input vector of the SVM-based classifiers. One is the vector sum of all 255-dimensional linguistic feature vectors (*LRL*, *RLR*) of a specific speaker in a dialog, denoted as SVM-255, and the other is the vector sum of 85-dimensional linguistic feature vectors (*L*, *R*) of a specific speaker in a dialog (SVM-85). The first input type considers the conversations between two speakers, while the second input contains only one speaker's conversational content. Similar to previous definition, the manually evaluated PT scores were converted into binary variables and used as the desired outputs of SVM. In the test phase, the SVM-based classifiers were used to detect the PT, and the results were compared with the corresponding ground truth PT binary values (high or low) which were converted from the manually evaluated scores.

For the HMM-based classifiers, we trained 10 HMMs corresponding to high or low extent with respect to the five PTs. In the training phase of the HMM-based classifiers, we used the RNNs output as the input of the HMM-based classifiers. The corresponding dialog-based BFI-10 evaluation scores of this speaker were used to define the HMM-based classifiers. In the test phase, the HMM-based classifiers were used to estimate the PT score by assessing the cross-speaker contextual information between speaker turns. The HMM with the highest score was regarded as the perceived PT. The results were compared with the corresponding manually evaluated PT binary values to estimate the perception accuracy.

Each PT was individually evaluated using the mentioned methods to describe its distinct properties. Fig. 8 illustrates the evaluation results of the five PTs, indicating that the proposed method is superior to the SVM- and HMM-based approaches. As the data with speaker turn number smaller than 6 or larger than 20 are quite few, in this figure, only the evaluation results for the speaker turn number ranging from 6 to 20 are illustrated for detailed comparison. For overall evaluation using five-fold cross validation, the average PT perception accuracy achieved 79.66%. In addition, the accuracy of the proposed method increased slightly with the increase of the number of speaker turns. From the evaluation results, the HMM-based method did not perform well for PT perception. One possible reason is that the HMM-based method is not suitable to model the dialog data, in which cross-speaker contextual information is not carefully considered. Similarly, SVM-255 and SVM-85 also obtained unsatisfactory results. For SVM-255, the linguistic features of the conversational contents of two speakers were merged together making it difficult to distinguish between the PTs of two speakers. This is also a potential reason why SVM-255 was inferior to SVM-85 which considered the linguistic features from only one speaker.

We conducted a *t*-test to analyze the differences in performance among these three methods (Table XII). As

TABLE XII
T-TEST RESULTS FOR PT PERCEPTION USING C-HMM, HMM, AND SVM

PT	C-HMM and HMM	C-HMM and SVM-85
Openness	$t = 7.70$, $p = 0.002$**	$t = 3.28$, $p = 0.023$*
Conscientiousness	**$t = 3.55$, $p = 0.006$****	**$t = 4.78$, $p = 0.004$****
Extroversion	$t = 2.11$, $p = 0.044$*	$t = 2.04$, $p = 0.044$*
Agreeableness	$t = 4.42$, $p = 0.002$**	**$t = 1.77$, $p = 0.063$**
Neuroticism	**$t = 4.94$, $p = 0.008$****	**$t = 5.51$, $p = 0.003$****

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

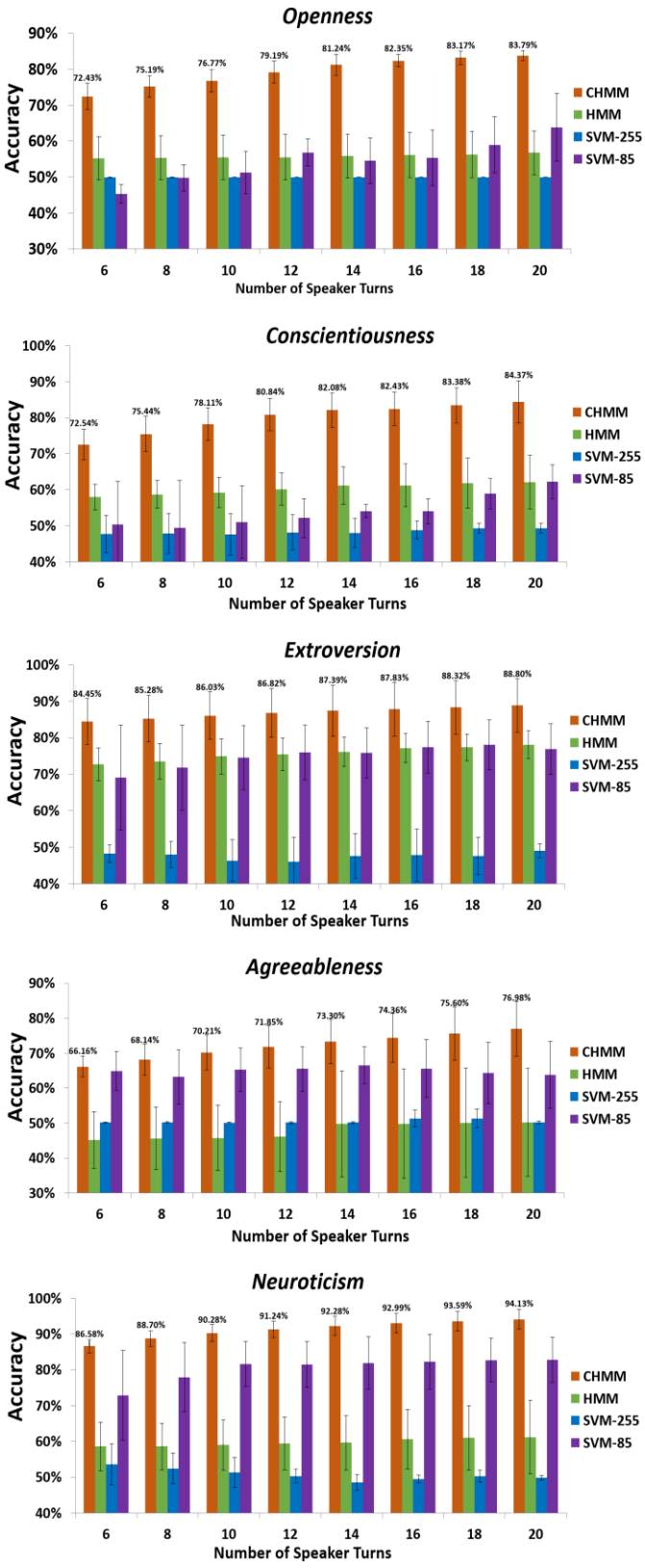

Fig. 8. Accuracy in different speaker turns for the perception of PT.

expected, significant differences existed between the proposed and the SVM- and HMM-based approaches for conscientiousness and neuroticism. However, for agreeableness, the proposed method had no significant difference compared with the SVM-based approach ($t = 1.77$, $p = 0.063 > 0.05$). The evaluation results for the other traits (Fig. 8) revealed that the proposed method was superior to the SVM- and HMM-based approaches. The accuracy of the proposed method increased with an increase of the number of speaker turns. However, the accuracy of the SVM-based method increased only for extraversion and neuroticism. In addition to extraversion, the HMM-based method showed low accuracy for other traits. The results were the same as those obtained in previous experiments.

Several findings were derived from the aforementioned experiments. First, PTs were perceived through linguistic features, such as LIWC and Sogou Industry lexicon features [9], in dyadic conversations. We observed that a favorable PT perception result was obtained when we concatenated linguistic feature vectors extracted from three consecutive single speaker turns between two speakers alternatively. Second, considering only the content of one speaker in a conversation was not sufficient, and the experimental results from the SVM- and HMM-based approaches confirmed this premise. In addition, we found that the contents of two speakers in a conversation influenced each other, and an unsatisfactory result of PT perception was obtained when the content of only one speaker was evaluated. Third, a person perceives other people's personalities by assessing their cross-speaker contextual information in a dyadic conversation. For example, PT perception might be helped from long conversation and interpersonal interaction. Finally, the experimental results showed that extraversion and neuroticism had more satisfactory perception results of 88.8% and 94.13%, respectively, compared with the other dimensions. The results conform to the fact that people with high extraversion are easily perceived as expressive, outgoing, and easily captivate attention in social situations. By contrast, people with high neuroticism are unattractive because attraction generally involves a positive affective evaluation of a person and they are offered few social rewards [53]. In summary, for all PTs, the proposed method outperformed the SVM- and HMM-based classifiers in a dyadic conversation with various speaker turns.

VI. CONCLUSION AND FUTURE WORK

This study presents an approach for evaluating PT perception of two speakers in a dyadic conversation by using RNNs and C-HMMs. The linguistic features of the transcribed spoken texts were extracted to generate BFI-10 profiles by using RNNs, which were used to characterize short-term temporal evolution in the dialog. The C-HMMs were employed to detect the personalities of two speakers across speaker turns in each dialog by using long-term turn-taking temporal evolution and cross-speaker contextual information. To evaluate the proposed method, an automatic PT perception system was constructed. For evaluation, an average perception accuracy of 79.66% for the big five traits was achieved using five-fold cross validation. Compared with the HMM- and SVM-based methods, the proposed approach exhibited higher performance. These encouraging results confirm the practicality of this system for future applications. We can improve the performance of PT perception to a greater extent by assessing temporal evolution and cross-speaker contextual information.

Several concerns require further investigation. First, the ability to perceive a speaker's personality provides a human–computer interaction (HCI) with a flexible and versatile reaction. Personality complementarity and similarity are crucial factors influencing a user's acceptance of an interface [54]. Overall, understanding a person's personality should be of great help for improving the experience of HCI. Second, the relationship between emotion and personality is worth exploring. Four traits (neuroticism, extraversion, agreeableness, and openness) are related to emotional dispositions; their relation to emotional dispositions is attributable to their correlations with explicit emotional disposition measures, such as the trait form of the positive and negative affect schedule [55]. The user's personality and emotional information can help the dialog system to produce more appropriate response sentences. Therefore, establishing a complementary personality module and a mechanism for detecting personalized auxiliary emotions is beneficial for HCI and can be investigated in the future.

REFERENCES

[1] W. W. Wilmot, Dyadic communication, New York: Random House, 1987.

[2] M. Allen, R. W. Preiss, B. M. Gayle and N. Burrell, Interpersonal communication research: Advances through meta-analysis, New Jersey: L. Erlbaum, 2002.

[3] J. K. Burgoon, L. A. Stern and L. Dillman, Interpersonal adaptation: Dyadic interaction patterns, Cambridge University, 2007.

[4] C. C. Lee, C. Busso, S. Lee and S. S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions," in *Proc. INTERSPEECH*, 2009.

[5] C.-H. Wu, Z.-J. Chuang and Y.-C. Lin, "Emotion Recognition from Text using Semantic Label and Separable Mixture Model," *ACM Trans. on Asian Language Information Processing*, vol. 5, no. 2, pp. 165-182, 2006.

[6] C.-H. Wu and W.-B. Liang, "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels," *IEEE Transactions on Affective Computing (TAC)*, vol. 2, no. 1, pp. 10-21, 2011.

[7] W.-L. Wei, C.-H. Wu, J.-C. Lin and H. Li, "Exploiting Psychological Factors for Interaction Style Recognition in Spoken Conversation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 659-671, 2014.

[8] 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, no. 1, pp. 457-500, 2007.

[9] 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 *Proc. the 7th ACM conference on Recommender systems*, 2013.

[10] J. Golbeck, C. Robles, M. Edmondson and K. Turner, "Predicting Personality from Twitter," in *Proc. 2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing*, 2011.

[11] F. Mairesse and M. Walker, "Words mark the nerds: Computational models of personality recognition through language," in *Proc. the 28th Annual Conference of the Cognitive Science Society*, 2006.

[12] D. Markovikj, S. Gievska, M. Kosinski and D. Stillwell, "Mining Facebook Data for Predictive Personality Modeling," in *Proc. the 7th international AAAI conference on Weblogs and Social Media (ICWSM 2013)*, 2013.

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

[14] T. Polzehl, K. Schoenenberg, S. Möller, F. Metze, G. Mohammadi, and A. Vinciarelli, "On Speaker-Independent Personality Perception and Prediction from Speech," in *Proc. INTERSPEECH 2012*, 2012.

[15] T. Polzehl, S. Moller and F. Metze, "Automatically assessing personality from speech," in *Proc. 2010 IEEE Fourth International Conference on Semantic Computing (ICSC)*, 2010.

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

[17] X. Zuo, B. Feng, Y. Yao, T. Zhang, Q. Zhang, M. Wang and W. Zuo, "A Weighted ML-KNN Model for Predicting Users' Personality Traits," in *Proc. 2013 International Conference on Information Science and Computer Applications (ISCA 2013)*, 2013.

[18] 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, Ed., New York, The Guilford Press, 1996, pp. 21-50.

[19] H. J. Eysenck and S. B. G. Eysenck, Manual of the Eysenck Personality Questionnaire (adult and junior), London: Hodder & Stoughton, 1975.

[20] P. T. Costa and R. R. McCrae, The NEO Personality Inventory manual, Odessa, FL: Psychological Assessment Resources, 1985.

[21] O. P. John, L. P. Naumann and C. J. Soto, "Paradigm Shift to the Integrative Big-Five Trait Taxonomy: History, Measurement, and Conceptual Issues," in Handbook of personality: Theory and research, O. P. John, R. W. Robins and L. A. Pervin, Eds., New York, Guilford Press, 2008, pp. 114-158.

[22] 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, 2007.

[23] J. L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179-211, 1990.

[24] M. Henderson, B. Thomson, and S. Young, "Deep neural network approach for the dialog state tracking challenge," in *Proc. the SIGDIAL 2013 Conference*, 2013.

[25] G. Mesnil, Y. Dauphin, K. Yao, Y. Bengio, L. Deng, D. Hakkani-Tur, X. He, L. Heck, G. Tur, D. Yu and G. Zweig, "Using Recurrent Neural Networks for Slot Filling in Spoken Language Understanding", in *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 3, pp. 530-539, 2015.

[26] M. Brand, N. Oliver and A. Pentland, "Coupled hidden Markov models for complex action recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Juan, 1997.

[27] S. C. Tseng, "Processing spoken Mandarin corpora," *Traitement Automatique des Langes . Special Issue: Spoken Corpus Processing*, vol. 45, no. 2, pp. 89-108, 2004.

[28] F. Iacobelli, A. J. Gill, S. Nowson and J. Oberlander, "Large scale personality classification of bloggers," in *Proc. the 4-th international conference on Affective computing and intelligent interaction*, Memphis, TN, USA, 2011.

[29] S. M. Mohammad and S. Kiritchenko, "Using Nuances of Emotion to Identify Personnality," in *Proc. the ICWSM Workshop on Computational Personality Recognition*, Melon Park, 2013.

[30] S. Argamon, S. Dhawle, M. Koppel and J. Pennebaker, "Lexical predictors of personality type," in *Proc. Interface and the Classification Society of North America*, St. Louis, MO, USA, 2005.

[31] T. Yarkoni, "Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers," *Journal of research in Personality*, vol. 44, no. 3, pp. 363-373, April 2010.

[32] G. Farnadi, S. Zoghbi, M. F. Moens and M. De Cock, "Recognising personality traits using facebook status updates," in *Proc. the 7-th international AAAI conference on weblogs and social media*, Cambridge, Massachusetts, USA, 2013.

[33] K. Luyckx and W. Daelemans, "Using syntactic features to predict author personality from text," in *Proc. Digital Humanities*, Oulu, Finland, 2008.

[34] A. J. Gill, S. Nowson and J. Oberlander, "What are they blogging about? personality, topic and motivation in blogs," in *Proc. the The International AAAI Conference on Weblogs and Social Media*, San Jose, California, 2009.

[35] A. Minamikawa and H. Yokoyama, "Personality estimation based on weblog text classification," in *Proc. Modern Approaches in Applied Intelligence*, Syracuse, NY, USA, Springer, 2011, pp. 89-97.

[36] A. Minamikawa and H. Yokoyama, "Blog tells what kind of personality you have: Egogram estimation from Japanese weblog weblog," in *Proc. the ACM 2011 conference on Computer supported cooperative work*, Hangzhou, China, 2011.

[37] S. Nowson and J. Oberlander, "Identifying more bloggers: Towards large-scale personality classification of personal weblogs," in *Proc. the International Conference on Weblogs and Social Media*, Boulder, Colorado, U.S.A., 2007.

[38] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *Journal of language and social psychology*, vol. 1, pp. 24-54, 2010.

[39] J. Oberlander and S. Nowson, "Whose thumb is it anyway? Classifying author personality from weblog text," in *Proc. the Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, 2006.

[40] S. Siegel and J. N. J. Castellan, Nonparametric statistics for the behavioral sciences, New York: Mcgraw-Hill, 1988.

[41] B. A. Jones, C. L. Dozier and P. L. Neidert, "An evaluation of the effects of access duration on preference assessment outcomes," *Journal of Applied Behavior Analysis*, vol. 47, no. 1, pp. 209-213, 2014.

[42] P. Legendre, "Species associations: the Kendall coefficient of concordance revisited," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 10, no. 2, pp. 226-245, 2005.

[43] J. L. Huang, C. k. Chung, N. Hui, Y. C. Lin, S. Yi-Tai, B. C. P. Lam, W. C. Chen, M. H. Bond and J. W. Pennebaker, "The development of the Chinese Linguistic Inquiry and Word Count dictionary," *Chinese Journal of Psychology*, vol. 55, no. 2, pp. 185-201, 2012.

[44] R. Gao, B. Hao, H. Li, Y. Gao and T. Zhu, "Developing Simplified Chinese Psychological Linguistic Analysis Dictionary for Microblog," in *Brain and Health Informatics*, Maebashi, Gunma, Japan, 2013.

[45] CKIP, "The content and illustration of Sinica corpus of Academia Sinica," Technical Report no. 95-02, Institute of Information Science, Academia Sinica, Taiwan, Taipei, 1995.

[46] J. G. Wu and H. Lundstedt, "Prediction of geomagnetic storms from solar wind data using Elman recurrent neural networks," *Geophysical research letters*, vol. 23, no. 4, pp. 319-322, 1996.

[47] J. Hertz, A. Krough, and R. G. Palmer. "Introduction to the theory of neural computation", volume 1 of Santa Fe Institute, Studies in the sciences of complexity, lecture notes. Addison-Wesley, 1991.

[48] N. Brewer, N. Liu, O. De Vel and T. Caelli, "Using coupled hidden Markov models to model suspect interactions in digital forensic analysis," in *Proc. International Workshop on Integrating AI and Data Mining*, 2006. AIDM'06, 2006.

[49] C. C. Chang, and C. J. Lin, "LIBSVM: a library for support vector machines", *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, article 27, 2011.

[50] R. Lina, L. Yanxin, R. Zhiyuan, L. Haiyan and F. Ruicheng, "Application of Elman Neural Network and MATLAB to Load Forecasting," in *Proc. International Conference on Information Technology and Computer Science*, Kiev, 2009.

[51] I. Rezek, M. Gibbs and S. J. Roberts, "Maximum a posteriori estimation of coupled hidden Markov models," *Journal of VLSI signal processing systems for signal*, vol. 32, no. 1-2, pp. 55-66, August 2002.

[52] P. A. Devijver and J. Kittler, Pattern Recognition: A Statistical Approach, Prentice Hall, 1982.

[53] H. R. Riggio, P. P. Lui, A. L. Garcia, B. K. Matthies, G. Culbert, and J. Bailey, "Initial validation of a self-report measure of perceptions of interpersonal attraction", *Personality and Individual Differences*, vol. 74, pp. 292-296, 2015.

[54] C. Nass and K. M. Lee, "Does computer-generated speech manifest personality? An experimental test of similarity-attraction," in *Proc. the SIGCHI conference on Human Factors in Computing Systems*, The Hague, The Netherlands, 2000.

[55] R. Reisenzein and H. Weber, "Personality and emotion," in the Cambridge handbook of personality psychology, Cambridge, Cambridge University, 2009, pp. 54-71.

Ming-Hsiang Su received the B.S. degree in computer science and information engineering from the Tunghai University, Taichung, Taiwan, in 2001, the M.S. degree in management information systems from the National Pingtung University of Science and Technology, Pingtung, Taiwan, in 2003, and the Ph.D. degrees in computer science and information engineering from Chung Cheng University (NCKU), Tainan, Taiwan, in 2013.

He is currently a postdoctoral fellow in the Department of Computer Science and Information Engineering, NCKU, Taiwan. His research interests include e-learning, artificial intelligence, machine learning, multimedia signal processing, and personality detection.

Chung-Hsien Wu received the B.S. degree in electronics engineering from National Chiao Tung University in 1981, and the M.S. and Ph.D. degrees in electrical engineering from National Cheng Kung University (NCKU), Taiwan, in 1987 and 1991, respectively. Since 1991, he has been with the Department of Computer Science and Information Engineering, NCKU. He became the distinguished professor in 2004. From 1999 to 2002, he served as the Chairman of the Department. He served as the deputy dean of the College of Electrical Engineering and Computer Science, NCKU, in 2009~2015. He also worked at Computer Science and Artificial Intelligence Laboratory of Massachusetts Institute of Technology (MIT), Cambridge, MA, in summer 2003 as a visiting scientist. He received the Outstanding Research Award of National Science Council in 2010 and the Distinguished Electrical Engineering Professor Award of the Chinese Institute of Electrical Engineering, Taiwan, in 2011. He was the associate editor of IEEE Trans. Audio, Speech and Language Processing (2010~2014) and IEEE Trans. Affective Computing (2010~2014). He is currently the associate editor of ACM Trans. Asian and Low-Resource Language Information Processing, and APSIPA Transactions on Signal and Information Processing. Dr. Wu served as the Asia Pacific Signal and Information Processing Association (APSIPA) Distinguished Lecturer and Speech, Language and Audio (SLA) Technical Committee Chair in 2013~2014. His research interests include affective computing, speech recognition/synthesis, and spoken language processing.

Yu-Ting Zheng received the B.S. degree in computer science and information engineering from Chang Gung University, Taipei, Taiwan, in 2012, and the M.S. degree in computer science and information engineering from National Cheng Kung University, Tainan, Taiwan, in 2014. His research interests include multimedia signal processing, and personality detection.