NNIME: The NTHU-NTUA Chinese Interactive Multimodal Emotion Corpus

Huang-Cheng Chou
Department of Electrical Engineering
National Tsing Hua University
Email: hc.chou@gapp.nthu.edu.tw

Wei-Cheng Lin
Department of Electrical Engineering
National Tsing Hua University
Email: winston810719@gmail.com

Lien-Chiang Chang
Department of Drama
National Taiwan University of Arts
Email: len6184@ntua.edu.tw

Chyi-Chang Li
Graduate School of Arts and Humanities
Instruction National Taiwan University of Arts
Email: lialan@ntua.edu.tw

Hsi-Pin Ma
Department of Electrical Engineering
National Tsing Hua University
Email: hp@ee.nthu.edu.tw

Chi-Chun Lee
Department of Electrical Engineering
National Tsing Hua University
Email: cclee@ee.nthu.edu.tw

***Abstract*—The increasing availability of large-scale emotion corpus along with the advancement in emotion recognition algorithms have enabled the emergence of next-generation human-machine interfaces. This paper describes a newly-collected multimodal corpus, i.e., the NTHU-NTUA Chinese Interactive Emotion Corpus (NNIME). The database is a result of the collaborative work between engineers and drama experts. This database includes recordings of 44 subjects engaged in spontaneous dyadic spoken interactions. The multimodal data includes approximately 11-hour worth of audio, video, and electrocardiogram data recorded continuously and synchronously. The database is also completed with a rich set of emotion annotations on discrete and continuous-in-time annotation by a total of 49 annotators. Thees emotion annotations include a diverse perspectives: peer-report, director-report, self-report, and observer-report. This carefully-engineered data collection and annotation processes provide an additional valuable resource to quantify and investigate various aspects of affective phenomenon and human communication. To our best knowledge, the NNIME is one of the few large-scale Chinese affective dyadic interaction database that have been systematically collected, organized, and to be publicly-released to the research community.**

1. Introduction

Emotion is a core fundamental internal attribute of humans that governs our behaviors and decision-makings. There has already been a tremendous research effort in modeling humans using a variety of measurable signals, which aims at enabling machines to sense emotional states automatically [1], [2]. One key components in advancing such research is the availability of databases for researchers to develop robust recognition algorithms and carry out meaningful analyses. This paper presents an additional novel resource, the NTHU-NTUA Chinese Interactive Multimodal Emotion Corpus (NNIME), for the affective computing community. The NNIME is a result of the collaborative works between engineers and drama experts. The collection of the database is designed with three major elements: 1) adopt the use of dyadic interactions for natural elicitation of affective behaviors, 2) collect both external behaviors and internal physiology of the dyad simultaneously, and 3) annotate rich emotion attributes of the interacting dyads from different perspectives. This publicly-available large-scale Chinese affective database opens up potential opportunities for researchers to advance research in directions, such as cross-cultural (language) emotion recognition, interaction dynamics modeling, joint external-internal behavior-physiology modeling, and perceptual analyses on diverse emotion evaluations.

Human interaction often involves complex processes of communicative goals and emotional behaviors, which not only are expressed verbally and nonverbally but also are reflected in the inner responses of humans [3]. Dyadic interaction is a basic unit of face-to-face interaction providing an important gateway for humans to convey emotion, communicate information, and foster mutual understanding [4], [5], [6], [7]. Hence, the use of dyadic improvisation, i.e., a creative performance where actors collaborate and coordinate in real-time to create a coherent viewing experiences [8], within a pre-designed scenario setup offers a unique window into understanding the dyadic affective interaction processes. Improvised acting has been considered as a viable research methodology for studying human emotions and communication [9], [10]. We adopt a similar approach in the creation of the NNIME. The collection protocol is designed based on situating the dyadic interaction in daily-life scenario settings, where each interaction lasts approximately 3-minute long with an overall pre-defined target affective atmosphere (i.e., angry, happy, sad, neutral, surprise, and frustration). Professional director and trained actors are involved to ensure the quality and the naturalness of the performances. Except for the requirement that the scene needs to resemble real-life scenarios with an target overall affective tone, there is no further restriction in order for the recorded interactions to resemble natural behaviors in daily life as close as possible.

978-1-5386-0563-9/17/$31.00 ©2017 IEEE

TABLE 1. OVERVIEW OF EXISTING AVAILABLE DATABASES: **SUBJS** DENOTES THE NUMBER OF SUBJECTS COLLECTED IN THE DATABASE; **RATERS** SHOWS THE NUMBER OF RATERS USED FOR EMOTION ANNOTATION; **DATA** INDICATES THE TYPES OF MULTIMODAL DATA RECORDED; **SETTINGS** INDICATES WHETHER THE DATABASE INCLUDES ONLY *single* SPEAKER SPEAKING, *dyadic* INTERACTIONS, OR A MIX OF INTERACTION TYPES FROM *TV shows*; **LANGUAGE** INDICATES THE LANGUAGE USED IN THE DATABASE; **METHODS** DENOTES WHETHER THE SUBJECTS ARE GIVEN FIXED *scripted* TEXT OR ARE FREE TO *spontaneously* USE THEIR OWN WORDINGS; **LABELS** INDICATES THE TYPES OF EMOTION ATTRIBUTES ANNOTATED: *Discrete* (A SINGLE CATEGORICAL AND/OR DIMENSIONAL ATTRIBUTES ANNOTATED FOR A SPECIFIED DURATIONAL SEGMENT) OR *Continous* (REAL-VALUED DIMENSIONAL ATTRIBUTES ANNOTATED CONTINUOUSLY IN TIME)

Databases	Subjs	Raters	Data	Settings	Language	Methods	Labels
DES [11]	4	20	Audio	Single	Danish	Script	Discr.
eNTERFACE [12]	2	42	Audio, Video	Single	English	Script	Discr.
HUMAINE [13]	4	4	Audio, Video	TV show	English	Script & Spont.	Conti.
IEMOCAP [14]	10	6	Audio, Video, MOCAP	Dyadic	English	Script & Spont.	Discr.
CreativeIT [15]	16	3	Audio, Video, MOCAP	Dyadic	English	Script & Spont.	Discr. & Conti.
SEMAINE [16]	20	8	Audio, Video	Dyadic	English	Script	Conti.
RECOLA [17]	46	6	Audio, Video, ECG, EDA	Dyadic	French	Spont.	Conti.
CHEAVD [18]	238	4	Audio, Video	TV show	Chinese	Spont.	Discr.
NNIME	44	49	Audio, Video, ECG	Dyadic	Chinese	Spont.	Discr. & Conti.

Comparing to existing emotional corpora, the novel contributions of the NNIME are the following: first, it is one of the few large-scale publicly-available affective databases in Mandarin Chinese and is additionally valuable due to its collection setting of spontaneous dyadic interactions; second, it includes multimodal data streams of both speakers' audio-video-text and physiology information; lastly, it is completed with rich emotion annotations, e.g., continuous-in-time annotation, session-level discrete attributes, and a combination of peer-report, director-report, self-report, and observer-report assessments of emotion attributes. The rest of this paper is organized as follows. Section 2 describes existing public multimodal databases. Section 3 introduces the corpus construction and design, including collection protocol, multimodal recording setup, and emotion annotations. Section 4 provides preliminary baseline emotion recognition results. Section 5 concludes with future potential of this newly-collected database.

2. Existing Multimodal Emotional Databases

In the last two decades, there has been a number of multimodal emotion databases collected for algorithmic development of automatic emotion recognition through modeling of human behaviors. Ringeval et al. provided a latest and extensive overview of existing emotional corpora when introducing their newly-collected corpus, i.e., the RECOLA database [17]. Table 1 shows a selected list of existing emotion databases with their key design attributes, including number of subjects and annotators, recorded behavior modalities, settings of data collection (single speaker, dyadic interactions, or TV shows), language used, methods (scripted or spontaneous speaking), and types of emotion labels (categorical, dimensional, or time-continuous).

Some notable exemplary databases include: Engberg et al. recorded 30 minutes of scripted Danish speech to analyze the five major emotion categories, i.e., happy, sad, angry, neutral, and surprise (DES) [11]. Douglas-Cowie et al. recorded audio-video data by asking subjects to utter a fixed number of sentences each with six different manifestations of emotion expressions (eNTERFACE) [12]. In 2007, Douglas-Cowie et al. summarized an extensive analyses, e.g., on the perspective of scope, naturalness, and context, in the collection of the well-known HUMAINE audio-video emotion database (HUMAINE) [13]. Around the same time, Busso et al. presented a novel database (IEMOCAP), which used the setting of dyadic interactions as the emotional behavior elicitation technique [14]; this database includes speech and unique detailed facial expressions that were collected using motion capture technology. An extension of IEMOCAP resulted in the release of another dyadic affective interactions database (CreativeIT) [15]. Metallinou et al. adopted the use of theatrical dyadic improvisation technique to elicit expressive speech and body language behaviors, where both actors' full body movements were captured using motion capture technology. Mckeown et al. collected a multimodal database consists of emotional conversations between human subjects and a computer conversational agent (SEMAINE) [16]. Lastly, Ringeval et al. recently presented a database consists of spontaneous collaborative and affective interactions in French (RECOLA) [17]. The database included 46 participants with synchronized mulit-modal data, i.e., audio, video, electrocardiogram (ECG) and electrodermal activity (EDA).

Most of the systematic collection of large-scale emotion corpora are in Western languages. The ones in Mandarin Chinese are often much smaller in scale and usually only include single modality. For example, Morrison et al. collected spontaneous spoken data from a call center including 388 utterances from 11 speakers for classifying between anger and neutral emotion [19]. Nogueiras et al. used a database containing 720 utterances from 12 speakers in six different emotions for the task of developing speech-based emotion recognizer [20]. Fu et al. recorded 7 actors, and each with 20 spoken utterances of fixed emotion categories [21]. Zhou et al. collected 1200 utterances from 4 nonprofessional actors for six emotions and named the database as the CLDC [22]. All of the databases mentioned above include only single, i.e., speech, modality. Recently, Bao and

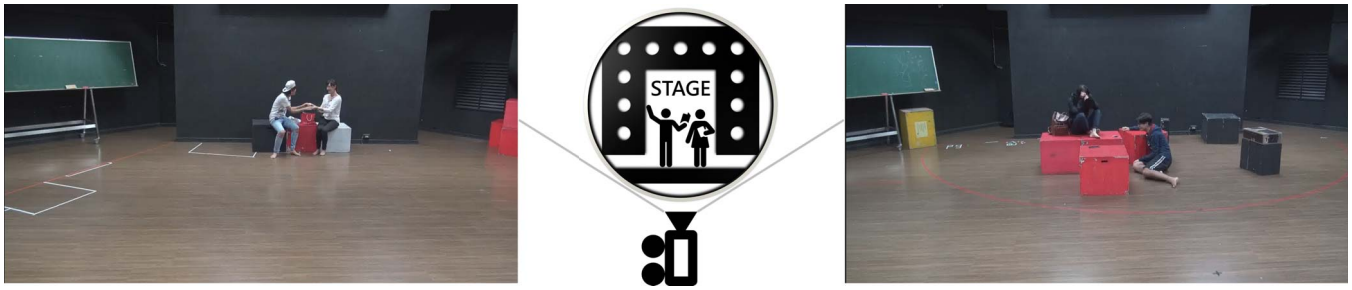

Figure 1. It depicts an actual snapshot of two different recording sessions extracted from the stage front-facing video camcorder (left and right). The middle depicts the camera setup in relation to the stage

Li et al. constructed a CASIA Natural Emotional Audio-Visual Database (CHEAVD), i.e., similar in manner with the HUMAINE. It contained 140 minutes of spontaneous emotional audio-video segments of data extracted from 238 speakers from films, TV plays, and talk shows with 26 non-prototypical emotional states labeled by four Chinese native speakers [18], [23].

The emotion elicitation design of the NNIME database is similar in nature with the IEMOCAP and the CreativeIT databases. In the NNIME, we elect to use carefully-designed spontaneous dyadic interactions that would be ideal in resembling real-life human communications and be suitable for multimodal affective analyses. The description on the scope, protocol, annotation, and multimodal data will be detailed in Section 3.

3. NNIME Database Construction and Design

3.1. Collection Protocol and Procedure

We first describe our subjects pool, the recording environment, and the design of collection procedure in this section. The participants were recruited from the Department of Drama at the National Taiwan University of Arts (NTUA). There were a total of 44 subjects (22 females, 20 males) with age ranged from 19 to 30. All of the subjects had prior real-life experiences in professional acting performances and were all enrolled in the dramatic arts degree program at the NTUA at the time of recordings. All of the subjects were native Mandarin Chinese speakers. The acting was carried out in one designated performance room in the NTUA with a central stage.

In order to induce natural affective interactions, the scenarios were set up as spontaneous dyadic interaction settings. There was a professional director involved in ensuring the affective quality and the naturalness of the performances at every recording session. The 44 subjects were paired into dyadic groups (seven female-female, ten female-male, and five male-male pairs). Each pair were instructed to interact freely on the stage with each other in order to spontaneously act out a short scene, i.e., approximately 3 minutes long, that targeted the overall performance tone to be one of the six pre-specified affects (anger, sadness, happiness, frustration, neutral, and surprise). The hypothesized stage was assumed to be in a real-life home setting, such as living room, dormitory, or bedroom. Every pair of actors had the scene rehearsal presented to the director prior to the actual onset of recordings. Some examples of the scenes included situations such as one of the two was being late to an important meeting, one of them accidentally figured out he had won a lottery, or a married couple found that the wife was pregnant. The data collection lasted for two months with most of the pairs completing all six affective atmospheres; this resulted in a total of 102 dyadic interaction sessions with roughly 11 hours worth of audio-video data (the duration of each session is $\mu = 195.35s, \sigma = 73.26s$).

3.2. Multimodal Recording Setup

We describe our multimodal recording setup in this section. One thing to note is that since these are professional actors, which are trained to work under different conditions as well as to be minimally affected by the particularities of the performance environment, we could ensure that wearing sensors would not interfere with their natural performances. For each dyadic interactions, we record the following three raw signal modalities:

Audio Recording: Each actor in the dyadic scene wore a Bluetooth wireless closed-up microphone (CAROL BTM-210C). We chose wireless microphones to avoid interfering with actors natural behaviors. The two wireless audio signal streams were transmitted to a multi-channel digital recorder (ROLAND BR-800) for dual-channel time synchronization. The setting was set to sample at 44.1 kHz with 24-bit AD conversion.

Video Recording: Each interaction session was recorded using a high definition camcorder (SONY HDR-P J790V); the specification of the raw video data was 1920 x 1080, 60P, and 28Mbps. The video camera was placed in front of the acting area at a fixed position that enabled capturing of both actors' movements on scene (see Figure 1). The actors were free to move as naturally as possible and to stay within the field view of the HD camera. These videos were also used for annotation of emotion attributes.

Electrocardiogram (ECG) Recording: In every session, each actor wore a multichannel physiological integrated circuit from Texas Instrument (ADS1292R) as the analog front-end circuit [24]. This device was capable of detecting human ECG signals. ADS1292R is a low-power integrated circuit with a 24-bit analog-to-digital converter (Figure 2 shows an image of the actual device used). The sampling rate was 250 Hz.

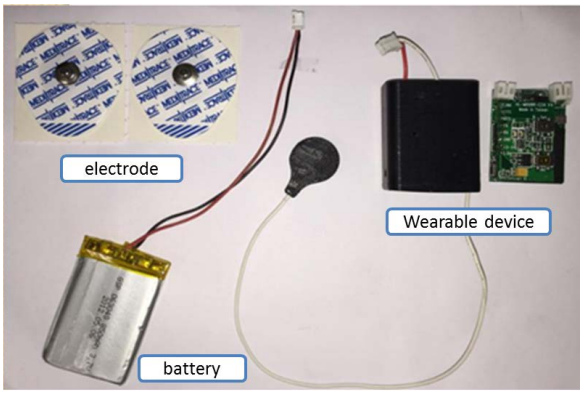

Figure 2. A picture of ECG collection device: a multichannel physiological integrated circuit from Texas Instrument (ADS1292R device).

Except that the audio data from the dyads were already synchronized with the audio mixer, we had to manually synchronize other modalities by using a clap board. The resulting signal spikes in the audio-recording of the SONY camcorder were used to synchronize between audio and video data. The ECG signal were synchronized also manually directly during the recording when the operator heard the clap sound at the scene.

3.3. Post-Processing

We further carried out the following post-processing procedures for the NNIME corpus: manual utterance segmentations, non-verbal vocalization markings, and ECG signal de-noising with R-R intervals extractions.

Each audio file corresponds to data collected from one of the microphones of each session. We manually segmented all audio files (two in every session with each lasted approximately 3 minutes long) into spoken utterances. This resulted in a total of 6701 utterances. Further, we marked each utterances as speech, laugh, sigh, sobbing, or audience background noise in order to enable further studies in understanding the role of non-verbal vocalizations in affective interactions. We also manually completed the transcripts for all of the sessions. All of these post-processing for audio-video data were properly time-stamped to ensure its future applicability in cross-modality and cross-speaker analyses.

Furthermore, since subjects had a lot of movements during interactions on stage, the ECG raw recordings may include unwanted motion artifacts. We first utilized discrete wavelet transform (DWT) to remove baseline variability, motion artifact, and other high frequency noise such as electromyography (EMG) and powerline noise [25]. We used eight-level DWT (db4 as the chosen mother wavelet function) decomposition with soft thresholding technique to eliminate high frequency noise [26]. The R peaks were further identified using moving window to detect the slope and local maximum, and with the availability of R peaks, we could then derive various physiological features (termed as heart rate variability features) by using the extracted R-R intervals. Figure 3 shows an example of the actual raw ECG signal with the identified R peaks after our post-processing procedure.

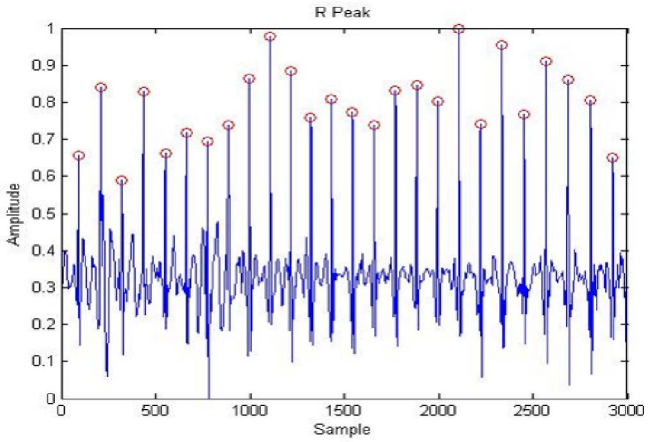

Figure 3. An example of the collected raw ECG signal with the identified R peaks in one of our sessions.

3.4. Emotion Attributes Annotations

The NNIME provides a rich set of perception-based emotion annotations. The evaluations can be broadly categorized into two different types based on the annotated time granularities, i.e., *discrete* of *continuous* emotion annotations. We present a brief description of each with the annotation protocols in this section.

3.4.1. Discrete Emotion Annotation. Discrete emotions were annotated for each individual actor at the *session*-level. We instructed the rater to annotate the dimensional emotion attributes of valence and activation ranging from 1 to 5 (a graphical illustration for the rater on the 5-levels is shown in Figure 4) and categorical emotion states (six categories: angry, happy, sad, neutral, frustration, surprise), which summarized the particular attribute over the entire recording for each actor in the scene.

There were a total of four different types of raters: 1) annotators with drama background (*peer-report*), 2) professional director (*director-report*), 3) actors of the interaction (*self-report*), and 4) naive observers with no explicit training in drama (*observer-report*). There were 44 peer-reports, 1 director-report, 1 self-report, and 4 observer-reports (a total of 49 unique raters) for every actor in each session. This large pool of raters from different background of view points is another unique novelty of this database.

3.4.2. Continuous Emotion Annotation. While discrete annotations are standard labels to be used for assessing emotion contents, the spontaneous spoken-dialog centric of the NNIME makes it contain a variety of multimodal expressions and interaction dynamics that would continuously unfold during the improvisation. Hence, we additionally performed continuous (i.e,. continuous-in-time) evaluation to describe accurately and precisely the emotional flow within each interaction. The evolution of the perceived emotional state for each participant were annotated in terms of dimensional attributes of activation and valence.

We recruited four naive raters with no drama background. We used the Feeltrace software for continuous

TABLE 2. IT SUMMARIZES INTER-EVALUATOR AGREEMENT ON THE ATTRIBUTES WITH MULTIPLE RATERS' RATINGS AND PRESENTS THE CORRELATION BETWEEN VARIOUS TYPES OF DISCRETE EMOTION ANNOTATIONS (*inter-types* CORRELATION). FOR DIMENSIONAL ATTRIBUTES, WE PRESENT BOTH SPEARMAN AND CONCORDANCE CORRELATIONS (CCC)

Spearman/CCC	**Labels**	**Agreement**	**Labels**		*peer-report*	*director-report*	*self-report*
Discrete: *peer*	Act.	0.58/0.49	Act.	*director-report*	0.62/0.46		
	Val.	0.61/0.56	Val.		0.77/0.65		
	Categ.	0.81					
Discrete: *observer*	Act.	0.69/0.63	Act.	*self-report*	0.63/0.53	0.76/0.67	
	Val.	0.78/0.74	Val.		0.77/0.74	0.62/0.47	
	Categ.	0.89					
Continuous	Act.	0.36/0.26	Act.	*observer-report*	0.85/0.81	0.56/0.46	0.58/0.55
	Val.	0.50/0.40	Val.		0.85/0.81	0.72/0.70	0.65/0.67

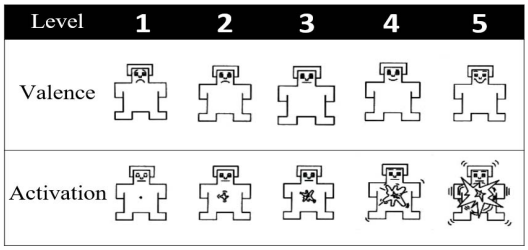

Figure 4. A graphical illustration given to annotator when rating valence and activation dimensional attributes ranging from level 1 to 5

emotion annotation [27]. Raters were given training prior to carrying out the actual annotation: they performed their first annotation multiple times to be familar with Feeltrace, and also in order know the type and the range of emotional manifestations that appear in the database, they had to watch in advance a few selected sessions. We instructed the annotators to rate the activation and valence (one at a time) of each actor (sampling rate is 0.1 second/frame) for every session. The value of annotations ranged from -1 to 1.

3.5. Emotion Annotations Analyses

There are 49 annotators in total that rated the discrete, i.e., session-level, affect attributes (activation, valence, and categorical emotions): 44 annotations from *peer-report*, 1 from *director-report*, 1 from *self-report*, and 4 from *observer-report*. Furthermore, there are 4 observer-based continuous annotations. In this section, we present the inter-evaluator agreement on the attributes with multiple raters and further present the correlation between various types of discrete emotion annotations (*inter-types* correlation).

Table 2 summarizes both inter-evaluator agreement and inter-type correlation. To compute inter-evaluator agreement for discrete annotations of valence and activation, we first calculate Spearman correlation and concordance correlation coefficient (CCC) between every N^{th} annotator's rating and the average rating of the rest of N-1 raters, we then report the average correlations on the entire database as the inter-evaluator agreement. To compute inter-evaluator agreement for discrete categorical attributes, we follow the exact same approach except that we use the percentage of the same categorical choice as the metric. For the continuous annotation, we first perform smoothing using moving average window and then follow the same approach; we finally report the median of the correlation coefficients (Spearman and CCC) computed within each session of the entire database. Furthermore the *inter-type* correlations are carried out on averages of the rater scores with both Spearman and CCC (reported in the right portion of Table 2).

Most of the emotion attributes rated achieve a reasonable inter-evaluator agreement and comparable to previous works. We observe that the four types of annotation, while correlated with each other, demonstrate certain variabilities, which may be reflective in the difference in the raters background. The highest correlation occur between *peer-report* and *observer-report*, which is also intuitive pleasing.

4. Baseline Emotion Recognition Results

We present our preliminary results on discrete emotion annotation (session-level attributes) recognition using acoustic, video, and ECG features separately in this paper. We use the average value of activation and valence and majority vote result of emotion categories from *peer-reports* as the ground truth to carry out regression and classification tasks, respectively. We use support vector machine for both regression and classification. The evaluation scheme is carried out via leave-one-subject-out cross validation. We briefly describe our acoustic, visual, and ECG feature extraction approach and the recognition results in this section.

4.1. Audio-Video Feature Extraction

4.1.1. Audio-Video Low-Level Descriptors (LLDs). We extract 45 dimensional LLDs from the segmented utterances at a framerate of 16.67ms. These 45 dimensional LLDs include 13 Mel Frequency Cepstral Cofficients (MFCCs), pitch, and intensity and their associated first and second derivatives computed using the Praat toolkit [28]. LLDs are further z-normalized with respect to each individual speaker.

Further, we extract a total of 426 dimension low-level video descriptors using dense trajectory extraction methods computed on the automatically-tracked bounding boxes for each individual speaker in the session. These descriptors include trajectory displacements (Traj), histogram of gradients (HOG), histogram of optical flow (HOF), motion boundary histogram in x and y directions (MBH_x and MBH_y) and their associated first and second derivatives [29].

4.1.2. Session-level Feature Encoding. In order to encode the sequence of LLDs to a fixed-length high-dimensional session-level features, we use the GMM based Fisher-vector encoding [30], [31], i.e., a generative model with discriminative power, to represent low-level information of an individual at the session level. A brief description is as follows: let $X = x_t, t = 1...T$ be a set of T LLD samples and p be a probability distribution function (pdf) chosen to be GMM with parameters $\lambda = w_i, \mu_i, \Sigma_i, i = 1...K$ where w_i , μ_i and Σ_i are respectively the weight, mean vector and covariance matrix for each mixture of Gaussian i. The gradient of log likelihood to characterize samples X can be derived by defining Fisher score function:

$$\nabla_\lambda \log p(X|\lambda)$$

where likelihood $p(x_t|\lambda) = \sum_{i=1}^{K} w_i p_i(x_t|\lambda)$ and thereby the posterior is given by the following:

$$\gamma_t(i) = p(i|x_t, \lambda) = \frac{w_i p_i(x_t|\lambda)}{\sum_{j=1}^{N} w_j p_j(x_t|\lambda)}$$

The gradient vector known as Fisher-vector, represents the direction for λ to better fit X with $p(X)$ by computing the first order and second order statistics,

$$g^X_{\mu_k} = \frac{1}{T\sqrt{w_k}} \sum_{i=1}^{T} r_t(k) \left(\frac{x_t - \mu_k}{\sigma_k} \right)$$

$$g^X_{\sigma_k} = \frac{1}{T\sqrt{2w_k}} \sum_{i=1}^{T} r_t(k) \left(\frac{(x_t - \mu_k)^2}{\sigma_k^2} - 1 \right)$$

the vector of $[g^X_{\mu_k}\ g^X_{\sigma_k}]$ is the Fisher-vector encoding.

4.2. Heart Rate Variability Feature Extraction

With the post-processing done in Section 3.3, we derive 16 dimensional session-level features of ECG for each participant based on a combination of both statistics and frequency analyses method using the extracted RR-intervals. These features are termed conventionally as HRV features. The list of 16 features is below:

1) **HF** : The power within 0.15 to 0.4Hz
2) **Hfnu** : HF divided by the sum of LF and HF
3) **LF/HF** : The power within 0.04Hz to 0.15Hz
4) **LFnu** : LF divided by the sum of LF and HF
5) **VLF** : The power within 0.003Hz to 0.04Hz
6) **Total power** : Total activity power
7) **meanNN** : The mean of all RR-intervals
8) **stdNN** : The standard deviation of all RR-intervals
9) **rmsSD** : The root mean square of differences of successive RR-intervals
10) **SDSD** : The standard deviation of differences of successive RR-intervals
11) **NN50** : The number of pairs of successive RR-intervals that differ more that 50ms
12) **NN50/min** : The number of pairs of successive RR-intervals that differ more that 50ms per minute
13) **pNN50** : Percentage of all sequential RR deviations exceeding 50 ms
14) **NN20** : The number of pairs of successive RR-intervals that differ more that 20ms
15) **NN20/min** : The number of pairs of successive RR-intervals that differ more that 20ms per minute
16) **pNN20** : Percentage of all sequential RR deviations exceeding 20 ms

TABLE 3. BASELINE ACCURACIES: SPEARMAN CORRELATION AND CONCORDANCE CORRELATION COEFFICIENT (CCC) OBTAINED ON VALENCE AND ACTIVATION REGRESSION AND UNWEIGHTED AVERAGE RECALL (UAR) ON CATEGORICAL EMOTION CLASSIFICATION

Valence & Activation Regression			
Spearman/CCC	*Audio*	*Video*	*ECG*
Activation	0.658/0.639	0.531/0.475	0.457/0.385
Valence	0.596/0.512	0.407/0.315	0.200/0.143
Categorical Emotions Classification			
UAR	*Audio*	*Video*	*ECG*
Categories	0.483	0.604	0.258

4.3. Baseline Recognition Results

A summary of our baseline recognition results is listed in Table 3. For session-level activation and valence regression task, we report both Spearman correlation and concordance correlation coefficient (CCC) for each modality of features separately. Unweighted average accuracy (UAR) is also reported on six-class emotion categories classification tasks. These recognition results provide baseline results for future algorithmic endeavor when using this database.

5. Conclusions and Future Works

We present a newly-collected Chinese interactive multimodal emotion corpus (NNIME) in this paper. The NNIME database is a result of a close collaboration between engineers and drama experts. The collection protocol is based on dyadic human-human communication with scenario setup resembling real-life interactions. The NNIME includes a variety of information: high quality multimodal audio-video data and ECG data collected simultaneously of both interlocutors, a complete rich set of emotion attributes annotated, and derived data (transcripts, non-verbal vocalizations, RR-intervals) after our post-processing.

This novel and newly publicly-available corpus opens up several new possibilities in future affective computing research[1]. For example, it naturally can be used to study cross-cultural (cross-language) multimodal emotion recognition, and detailed analyses on dyadic interaction dynamics. Furthermore, due to its similarity in the design with the IEMOCAP and the CreativeIT, the idiosyncratic factor in the database design can be much mitigated. Moreover, the availability of both recorded internal physiology and

1. Contact Huang-Cheng Chou or Chi-Chun Lee for the NNIME access

expressive multimodal behaviors offers a chance for obtaining quantitative insights into understanding between the *emotion-being-felt* and the *emotion-being-expressed*. Lastly, the rich set of emotion annotations, e.g., discrete emotion perception evaluations from raters of different backgrounds and continuous-in-time emotion annotations, completed by a large-pool of raters will be important assets in investigating subjective and diverse multimodal emotion perception, e.g., rater modeling. This large-scale mulitmodal Chinese interactive emotion corpus would be a valuable addition to the affective computing community that could open up a variety of new possibilities for future research.

Acknowledgments

The authors would like to thank the Ministry of Science and Technology for funding. Special thanks to teachers and students at the Drama Department of the National Taiwan University of Arts in the creation of the database.

References

[1] R. W. Picard and R. Picard, *Affective computing*. MIT press Cambridge, 1997, vol. 252.

[2] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[3] M. L. Knapp, J. A. Hall, and T. G. Horgan, *Nonverbal communication in human interaction*. Cengage Learning, 2013.

[4] D. N. Stern, "Mother and infant at play: The dyadic interaction involving facial, vocal, and gaze behaviors." 1974.

[5] D. Kuhn, V. Shaw, and M. Felton, "Effects of dyadic interaction on argumentive reasoning," *Cognition and instruction*, vol. 15, no. 3, pp. 287–315, 1997.

[6] J. N. Cappella, "Mutual influence in expressive behavior: Adult–adult and infant–adult dyadic interaction." *Psychological bulletin*, vol. 89, no. 1, p. 101, 1981.

[7] J. F. Benenson, N. H. Apostoleris, and J. Parnass, "Age and sex differences in dyadic and group interaction." *Developmental psychology*, vol. 33, no. 3, p. 538, 1997.

[8] K. Johnstone, *Impro: Improvisation and the theatre*. Routledge, 2012.

[9] T. Bänziger and K. Scherer, "Using actor portrayals to systematically study multimodal emotion expression: The gemep corpus," *Affective computing and intelligent interaction*, pp. 476–487, 2007.

[10] C. Busso and S. Narayanan, "Recording audio-visual emotional databases from actors: a closer look," in *Second international workshop on emotion: corpora for research on emotion and affect, international conference on language resources and evaluation (LREC 2008)*, 2008, pp. 17–22.

[11] I. S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a danish emotional speech database." in *Eurospeech*, 1997.

[12] E. Douglas-Cowie, R. Cowie, and M. Schröder, "A new emotion database: considerations, sources and scope," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[13] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner *et al.*, "The humaine database: addressing the collection and annotation of naturalistic and induced emotional data," *Affective computing and intelligent interaction*, pp. 488–500, 2007.

[14] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[15] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The usc creativeit database: A multimodal database of theatrical improvisation," *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, p. 55, 2010.

[16] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The semaine corpus of emotionally coloured character interactions," in *Multimedia and Expo (ICME), 2010 IEEE International Conference on*. IEEE, 2010, pp. 1079–1084.

[17] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[18] Y. Li, J. Tao, L. Chao, W. Bao, and Y. Liu, "Cheavd: a chinese natural emotional audio–visual database," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–12, 2016.

[19] D. Morrison, R. Wang, and L. C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech communication*, vol. 49, no. 2, pp. 98–112, 2007.

[20] A. Nogueiras, A. Moreno, A. Bonafonte, and J. Mariño, "Speech emotion recognition using hidden markov models, speech prosody 2002," in *an International Conference, France*, 2002, pp. 11–13.

[21] L. Fu, X. Mao, and L. Chen, "Speaker independent emotion recognition based on svm/hmms fusion system," in *Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on*. IEEE, 2008, pp. 61–65.

[22] J. Zhou, G. Wang, Y. Yang, and P. Chen, "Speech emotion recognition based on rough set and svm," in *Cognitive Informatics, 2006. ICCI 2006. 5th IEEE International Conference on*, vol. 1. IEEE, 2006, pp. 53–61.

[23] W. Bao, Y. Li, M. Gu, M. Yang, H. Li, L. Chao, and J. Tao, "Building a chinese natural emotional audio-visual database," in *Signal Processing (ICSP), 2014 12th International Conference on*. IEEE, 2014, pp. 583–587.

[24] J.-W. Jhuang and H.-P. Ma, "A patch-sized wearable ecg/respiration recording platform with dsp capability," in *E-health Networking, Application & Services (HealthCom), 2015 17th International Conference on*. IEEE, 2015, pp. 298–304.

[25] C. Cai and P. d. B. Harrington, "Different discrete wavelet transforms applied to denoising analytical data," *Journal of chemical information and computer sciences*, vol. 38, no. 6, pp. 1161–1170, 1998.

[26] Q. Haibing, L. Xiongfei, and P. Chao, "Discrete wavelet soft threshold denoise processing for ecg signal," in *Intelligent Computation Technology and Automation (ICICTA), 2010 International Conference on*, vol. 2. IEEE, 2010, pp. 126–129.

[27] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder, "'feeltrace': An instrument for recording perceived emotion in real time," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[28] P. P. G. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, 2002.

[29] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3551–3558.

[30] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods." in *BMVC*, vol. 2, no. 4, 2011, p. 8.

[31] T. S. Jaakkola, D. Haussler *et al.*, "Exploiting generative models in discriminative classifiers," *Advances in neural information processing systems*, pp. 487–493, 1999.