Computational Analyses of Thin-Sliced Behavior Segments in Session-Level Affect Perception

Wei-Cheng Lin, *Student Member, IEEE* and Chi-Chun Lee, *Member, IEEE*

Abstract—The ability to accurately judge another person's emotional states with a short duration of observations is a unique perceptual mechanism of humans, termed as the thin-sliced judgment. In this work, we propose a computational framework based on mutual information to identify the thin-sliced emotion-rich behavior segments within each session and further use these segments to train the session-level affect regressors. Our proposed thin-sliced framework obtains regression accuracies measured in Spearman correlations of 0.605, 0.633, and 0.672 on session-level attributes of activation, dominance, and valence, respectively. It outperforms framework using data of the entire session as baseline. The significant improvement in the regression correlations reinforces the thin-sliced nature of human emotion perception. By properly extracting these emotion-rich behavior segments, we obtain not only an improved overall accuracy but also bring additional insights. Specifically, our detailed analyses indicate that this thin-sliced nature in emotion perception is more evident for attributes of activation and valence, and the within-session time distribution of emotion-salient behavior is located more toward the ending portion. Lastly, we observe that there indeed exists a certain set of behavior types that carry high emotion-related content, and this is especially apparent in the extreme emotion levels.

Index Terms—Emotion recognition, multimodal behaviors, thin-sliced affect perception, mutual information

1 Introduction

Understanding the underlying human perception and decision-making mechanism has been a popular area of research in psychology for a long time (e.g., [1], [2], [3]). Various studies have pointed out a particular powerful human perceptual capability, i.e., the ability to integrate information from multiple perceived time events in order to come up with a single *overall-global* holistic judgment of higher-level (e.g., preferences, emotions, personalities, etc) attributes. This human perceptual mechanism is not only evident in daily life, but also is being leveraged as an important ability in aiding research across fields in behavior sciences, where human evaluation is repeatedly used as the core methodology for carrying out evidence-based analyses. For example, coding distressed couples behaviors to analyze the effectiveness of therapy sessions [4], [5], assessing the emphatic quality of the therapists in drug addiction rehabilitation [6], [7], and measuring the atypical socio-communicative behaviors of autistic children during Autism Diagnostic Observation Schedule (ADOS) interviews [8].

Researchers have previously proposed a theory stating that there exists an unique property of this particular human perceptual mechanism, i.e., an accurate perception of another person's higher-level attributes, e.g., personality [9], intelligence [10], affect [10], and even negotiation outcome [11], can in fact be obtained with a short duration of observations. This mechanism is termed as the *thin-slice* theory of judgment [12], [13]. A vast number of psychological experiments have investigated and corroborated this hypothesis in various contexts. For example, Fowler et al. conducted a study and demonstrated that lay raters were indeed capable of reliably and validly detecting features of psychopathy from small excerpts (a couple seconds long) of recorded interviews from 97 maximum-security inmates [14]. Naumann et al. additionally showed that human's first impression can achieve reliable accuracy on assessing other people's ten personality factors by merely observing the static full-body photographs [15]. The same phenomenon applies in cases of perceiving personality disorder; Oltmanns et al. designed an experiment and demonstrated that by observing a minimal 30-seconds of video-taped interviews, untrained undergraduate students were in fact capable of accurately rate the personality characteristics associated with different personality disorders [16]. Lastly on judging interpersonal affective interaction styles, Ovies et al. pointed out the feasibility of utilizing thin-sliced information, i.e., snippets of smile intensity and tactile contact, in order to gain an understanding of the kindergarteners and their family members' affective interaction styles at home [17].

Computationally modeling this thin-slice theory of judgment has also becoming more relevant for research fields that are at the cross-cutting between behavior science and engineering, e.g., behavioral signal processing (BSP) is one such fields. BSP aims at providing objective computational frameworks, e.g., those derived based on direct modeling of audio-video recordings and/or physiological sensory data, for behavior science experts in order to facilitate a more informed decision making [18], [19]. In BSP-related application domains such as healthcare, education, and performing art, researchers develop computational frameworks in order to mitigate issues centered around subjectivity of manual human annotation. Examples of applications are listed below: in couple therapy, trained experts annotate

• *The authors are with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan.*
E-mail: winston810719@gmail.com, cclee@ee.nthu.edu.tw.

Manuscript received 17 June 2017; revised 4 Mar. 2018; accepted 12 Mar. 2018. Date of publication 16 Mar. 2018; date of current version 25 Nov. 2020. (Corresponding author: Chi-Chun Lee.)
Recommended for acceptance by A. A. Salah.
Digital Object Identifier no. 10.1109/TAFFC.2018.2816654

couples' behaviors after watching 10-minute long interactions [20], [21], in education setting, expert coaching principals grade the candidate principals' speech after listening to their 3-minute long impromptu talk as part of the principal-ship certification program [22], [23], and in performing art, audiences and/or director often assess the quality of actors' affective expressions at the level of minute-long even hour-long complete plays [24], [25]. Knowing that there exists a thin-sliced nature in the human's perceptual mechanism, when deriving appropriate signal-based analytics for these humans global perceptual ratings, identifying which *salient* slices of behaviors that contribute to the overall perception becomes essential in the development of robust algorithms. This would help in advancing the robustness of diagnostic instruments or training materials' design, even in possibly bringing quantitative evidence into humans perceptual mechanism [26].

While there is already a vast amount of evidence in showing the existence of thin-sliced perception in the psychology literature, the engineering works involved in deriving framework that aims at automatically identifying emotionally thin-sliced behavior segments remain limited, if any. A few notable related works in other application domains are listed below: Ozlim et al. leveraged auditory attention model in order to derive salient low-level acoustic features to improve the prominent syllable detection algorithm [27]. Gunes found that by using temporal segmental data, it improved the overall system of affect recognition from visual modality [28]. Han boosted the performance of emotion recognition by choosing the segments with highest energy in an utterance as the training samples, i.e., considering those as containing the most prominent emotional information in the audio modality [29]. Gibson et al. proposed a multiple instance learning-based framework to identify salient multimodal local behavior events in order to perform session-level behavior coding in couple therapy [30], [31], and this particular application domain has also been carried out using sequential probability ratio test [32].

Furthermore, while there is a large body of multimodal emotion recognition works that have been previously proposed in the past decades (e.g., [33], [34], [35], [36]), most of these works deal with scenarios of *utterance*-level (often seconds-long) recognition. In this work, we present a computational framework in identifying within-session *thin-sliced* behavior segments that bear *emotion-rich* information about the perceived *session*-level (minutes-long) affect. Specifically, we use the USC CreativeIT database to carry out this work [37]. The USC CreativeIT is a dyadic affective database includes multimodal behavior data (audio and full-body motion capture recordings) and the *session*-level emotion attributes annotations. Each session lasts approximately three minutes long. With the availability of session-level emotion annotations and multimodal behavior data, the USC CreativeIT database presents an ideal opportunity for systematic analyses of humans' thin-sliced affective perception at the session-level.

In this work, we present a mutual information based framework in identifying within-session thin-sliced emotion-rich behavior segments for global session-level emotion attributes. The framework first introduces the use of computing mutual information between discretized session-level attribute and quantized multimodal behavior clusters as criterion in selecting segments containing high emotion information. Then, we propose an automatic session-level emotion attribute regression framework by learning feature representation using Fisher-vector encoding (FV) [38]. FV encoding is done by computing the Fisher scoring of data samples with respect to the first and second order statistics of a background Gaussian Mixture Model, resulting in a fixed-length high dimensional vector representation on these thin-sliced emotion-rich behavior segments.

Specifically, the two major study contributions of the paper are listed below:

- *Thin-sliced Behavior Segments Analyses*: detailed analyses of the identified thin-sliced *emotion-rich behavior segments* in terms of their time allocation within a session, their effect in changing the original behavior types' distributions, and finally a detailed session-level emotion-dependent analysis
- *Global Emotion Recognition using the Thin-Sliced Segments*: incorporating thin-sliced perception by utilizing the identified *emotion-rich behavior segments* to enhance the session-level emotion recognition.

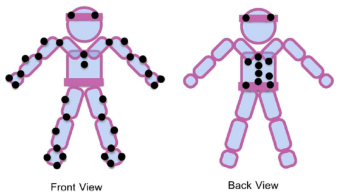
Our detailed analyses bring insights into demonstrating how streams of multimodal behavior manifestation may affect humans integrative perception of different emotion constructs when they make a judgment on the interaction as a whole. Further our proposed session-level emotion regressor obtain Spearman correlations of 0.605, 0.633, and 0.672 on attributes of activation, dominance, and valence by leveraging the identified within-session thin-sliced *emotion-rich behavior segments*. To the best of our knowledge, this work is one of first in providing a systematic analysis and computational method in identifying *thin-sliced emotional-rich behavior segments* in a spontaneous large-scale corpus and further utilize them to improve predictive power in tasks of session-level affect regressions.

The rest of the paper is organized as follows: research method, including database, multimodal feature extraction, the proposed emotion-rich behavior segment identification method are in Section 2. Detailed thin-sliced behavior segments analyses are provided in Section 3. Recognition results from the automatic thin-sliced based session-level affect regressors are described in Section 4. Finally, conclusions and future works are given in Section 5.

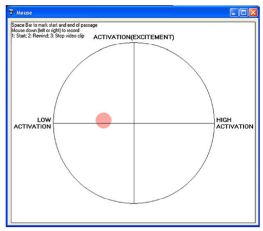
2 Research Methodology

2.1 The USC CreativeIT Database

We use the USC CreativeIT database for the present work [37]. The USC CreativeIT is a multimodal dyadic interaction database collected as a result from the combined expertise and effort from engineers and theatrical professionals. This database includes performances of dyadic improvisations based on an established theatrical acting technique (the Active Analysis [39]) in order to help elicit natural expressive behaviors. There are two types of performance sessions: the first one is called the *2-sentence* play, where each actor is limited lexically to repeat the same sentence while carrying out a scene in order to emphasize nonverbal, bodily gesture, and interaction dynamics; the second type is termed the *paraphrase* play, which the actors are directed to perform based on a scripted play although they are free to use their own words and interpretation. Each play lasts approximately 3–5 minutes long. The behavior modalities included in the database are audio recordings from lapel microphones and full body motion capture data of each actor (i.e., a recording of 45 markers' $(x; y; z)$ positions using 12 Vicon cameras collected at 60 frames per second). The



(a) The placement of 45 markers for full body motion capture



(b) Feeltrace interface used for time-continuous emotion annotation

Fig. 1. The USC CreativeIT database.

marker placement is shown in Fig. 1a. There are a total of 8 pairs of actors (16 actors in total) with 50 dyadic interaction sessions in the database.

The USC CreativeIT database adopts two different schemes for annotating three emotion attributes (valence, activation, dominance). Each rater is asked to rate every actor in a play with both local time-continuous and global session-level emotion labels. The rater is instructed to rate the local time-continuous label as he/she watches the interaction in real time (using a modified Feeltrace [40] interface shown in Fig. 1b). The global emotion attributes (i.e., rated on a scale of 1 to 5 per actor on the session-level) is then annotated afterward. In this present work, we concentrate on the global emotion label since it provides the most natural human annotation for the study of *session*-level emotion perception. Each actor's emotion within a session is annotated by at least three different naive raters, and the average value is utilized as the ground truth throughout the work. The inter-evaluator agreement for global emotion label is 0.72, 0.78, 0.67 measured by Cronbach's α, for activation, valence, and dominance respectively.

In summary, we use a total of 90 samples of global emotion annotation, full body motion capture, and audio data (due to device failure during part of the data collection) in this work. Table 1 further summarizes the correlations computed between the three global emotion attributes. The result indicates a clear trend that activation correlates moderately with dominance, and valence does not correlate with either dominance or activation.

2.2 Proposed Study of Thin-Sliced Emotion Perception

A complete schematic of the present work is illustrated in Fig. 3. Essentially, our computational study includes: 1) identifying the *emotion-rich behavior segments* with respect to the session-level emotion attributes within each interaction (upper portion of Fig. 3), and 2) leveraging these identified segments of *emotion-rich behavior segments* to enhance multimodal global emotion attributes regression results (bottom portion of Fig. 3). The procedure in deriving our thin-sliced based session-level emotion regressors can be summarized into the following steps:

- *Step1:* Extract multimodal behavior features and quantize representations using Gaussian Mixture Model, GMM (denoted as X_a and X_b for audio and body language respectively)
- *Step2:* Discretize the global emotion annotation into 10 equally-spaced levels (denoted as $Y_i, i \in [0, 1, 2 \ldots 9]$) for the three emotion attributes (denoted as $E_j, j \in$ {activation, valence, dominance})
  - Level Y_i: $E_j \in [min = 1, max = 5, step = 0.4]$
- *Step3:* Select top k-segments to form the global emotion-rich behavior segments (i.e., relevant thin-sliced behaviors) by computing mutual information between X and Y over a sub-segment within a session
- *Step4:* Utilize GMM-based Fisher Vector encoding to generate high-dimensional behavior feature vectors from the selected k-segments per session
- *Step5:* Perform automatic global emotion attribute regression by leveraging the identified thin-sliced behavior segments

TABLE 1
The Inter-Attributes (Session-Level Emotion Attributes) Correlation in the USC CreativeIT Database

	Activation	Dominance	Valence
Activation	1		
Dominance	0.568	1	
Valence	-0.076	-0.044	1

In the following sections, we will first focus on describing the detailed approach in the extraction of the k thin-sliced behavior segments (Step1-3). Section 3 will first present analyses on various characteristics of these behavior segments. The detailed implementation of using the identified behavior segments in the process of building automatic regressors (Step4-5) will be elaborated later in Section 4.

2.2.1 Multimodal Behavior Feature Extraction

For the audio modality, since the actors' audio recordings are collected using close-talk lapel microphones, we first apply a simple voice activity detection method based on short-term energy. In specific, if a frame's $(1/60$ seconds) energy intensity falls below 80 percent of the mean intensity computed over a session, it is considered as a silence frame; then a median filter of frame size 10 is applied afterward to smooth the voicing decision. We extract a total of 45 low-level acoustic descriptors over the speaking portion at 60 frames per second, including 13 MFCCs, 1 fundamental frequency, 1 intensity and their delta and delta-delta. We additionally perform speaker-wise z-normalization on these acoustic features to remove individual differences. This normalization is done for each session separately, i.e., no a-priori knowledge about the speaker identity is required. The features are extracted using the Praat toolbox [41].

For the body language modality, we use a similar feature extraction approach inspired from a previous work done by Metallinou et al. on the CreativeIT database [25]. The method designs a set of psychologically-inspired body language features using global and local coordinate system (Fig. 2): the origin of the global coordinate system is the center of the recording space, and the local coordinate system for each actor is defined using the four waist markers' average position. With the availability of 45 motion capture markers' 3D position (x, y, z), Metallinou et al. designed 70 features categorized into four different types: distance, angle, position, and velocity. Aside from these 70 features, we compute 25 extra body language features including 14 acceleration-based features (computed for all velocity-based features), 5 distance-based features (left/right hand to head and to torso distances and left leg to right leg distance), 3 features indicating the actor's head (x, y, z) coordinates,

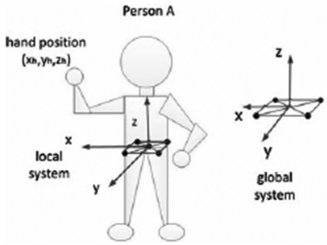


Fig. 2. Depiction of local and global coordinate system used for deriving body language features.

3 angle-based features (computed between left and right leg and between left/right leg and global origin). In summary, we extract a total of 95 frame-level body language low-level descriptors at 60 frames per second. A complete list of these 95 features is shown in the Appendix, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TAFFC.2018.2816654.

2.2.2 Global Emotion-Rich Behavior Segments Selection

We propose to select *global emotion-rich behavior segments* based on identifying within-session segments (each segment is defined as 1 percent of the entire interaction) possessing the largest mutual information computed between session-level emotion annotation and quantized behavior features. The use of mutual information in measuring variable dependencies have been found to be useful across applications, e.g., feature selection for classification tasks [42], [43], prosodic entrainment measures between interacting dyads [44], and even gene expression clustering [45]. Mutual information provides a general information-theoretic measure of joint dependencies (i.e., those could be non-linearly related) between two random variables as compared to measures such as correlations and euclidean distances [46].

We compute mutual information between discretized emotion attribute and quantized behavior features on segments within each interaction session. We first quantize behavior LLD features into m clusters from a trained GMM with M mixtures. The quantization is done by assigning each frame to the mixture m with the largest posterior data likelihood. The conventional quantization is often carried out using k-means clustering. However, k-means clustering can be regard as a special case of GMM if we make the covariance matrices a constant shared by all the components and take the limit to be 0. Further, the k-means quantization assumes the clusters to be spherical, which makes it less robust to complex geometrically-shaped data such as acoustic or body language features. Hence, we use GMM to perform behavior quantization in this work.

Each behavior modality can then be represented as discrete random variables, i.e., X_a, X_b, denoted for acoustic and body language behaviors respectively, where each variable takes on a discrete value between 1 to m. With the quantized behavior types, X_a and X_b, and discretized global emotion levels of each attribute, Y_j, we can carry out our proposed framework to select *global emotion-rich behavior segments*. The method is the following:

1) *Session segmentation*. We first split each actor's data within each session into 100 equally-space segments (i.e., each segment corresponds to 1 percent of the entire session—roughly 2 to 3 seconds long). Each segment is indexed as the lth segment with approximately 120 to 180 frames.
2) *Top k% thin-sliced behavior segments selection*. Each actor's sequence of quantized behavior for each segment l is denoted as X_l, and discretized sequence of global emotion label is denoted as Y_l (Y_l equals to the global emotion annotation for all l). We first obtain joint global emotion-behavior probability mass

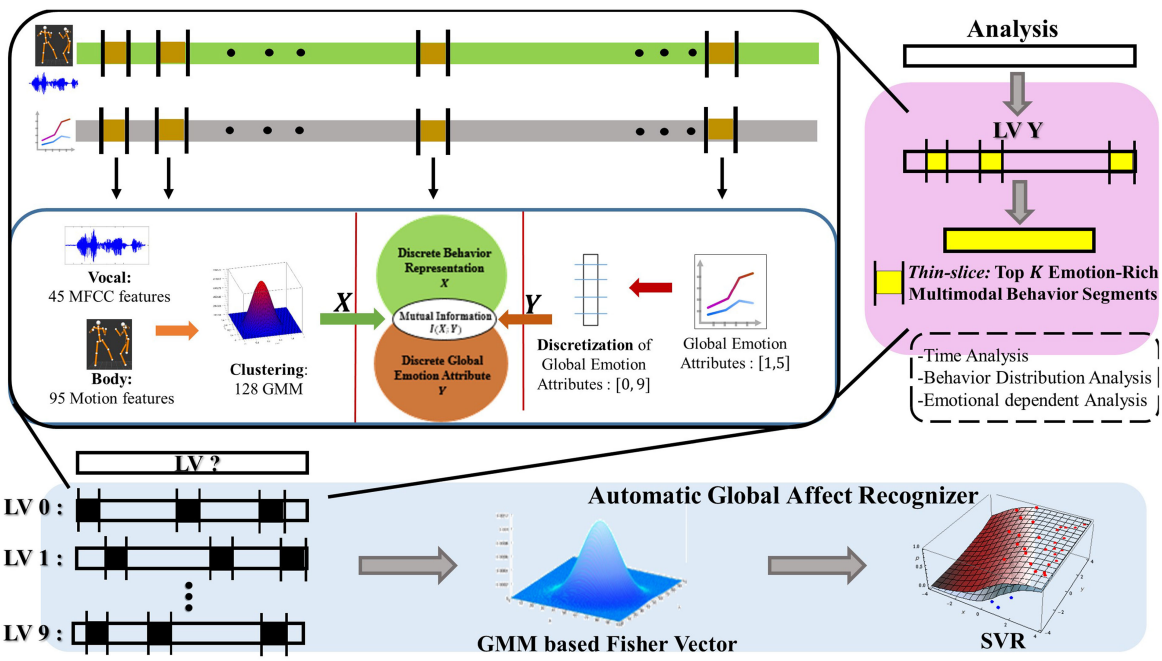


Fig. 3. A complete workflow of our proposed global emotion regression framework: 1) Identification of global emotion-rich behavior segment as our *thin slice* behavior segments for every session using mutual information, 2) session-level feature encoding of these emotion-salient behavior segments, and 3) support vector regression trained on these features to perform multimodal recognition of activation, valence, and dominance.

TABLE 2
The Total Number of the Identified Thin-Sliced *Global Emotion-Rich Behavior Segment* Selected within Each of the Session for the Database

Thin-sliced Behavior Segments Time Distribution						
	Audio			Body Language		
# of Total Segments	T1	T2	T3	T1	T2	T3
Act.	2,069	2,086	**2,145***	1,455	1,510	**1,535**
Dom.	1,212	1,168	**1,220**	2,377	2,310	**2,513***
Val.	584	607	**609**	1,147	1,190	**1,263***

*The * indicates that mean on the number of segments selected in that particular modality is significantly higher in T3 than in T1 ($p < 0.05$).*

TABLE 3
The Choice of % of Segments Used in the Analyses Section (Section 3) and Automatic Recognition Results (Section 4)

Activation	**percentage**
Audio	70%
Body Language	50%
Dominance	**percentage**
Audio	40%
Body Language	80%
Valence	**percentage**
Audio	20%
Body Language	40%

function, $P(X = m, Y = y)$, and the marginal distributions, $P(X = m)$ and $P(Y = y)$ using maximum likelihood estimation. We then compute the per-segment mutual information, I_l, between X_l, Y_l using the following equation for each lth segment

$$I_l = \sum_{Y_l} \sum_{X_l} P(X_l, Y_l) \log \frac{P(X_l = m, Y_l = y)}{P(X_l = m) P(Y_l = y)}.$$

We add up per-frame information computed over the segment length (∼120-180 frames) to obtain individual segment-level I_l. There are a total of of 100 $I_l's$ for each actor computed for each behavior modality of every session. I_l can be regarded as a quantification on how much *information* exists between the actor's behavior and annotator's emotion perception within the designated segment l. Hence, we can select top k segments (the thin-sliced segments) with the highest mutual information as the thin-sliced segments.

In summary, with this procedure, we can select k segments (each segment is 1 percent of the interaction) within each session for each emotion attribute, i.e., activation, valence, and dominance, of each behavior modality, i.e., acoustics and body language, separately. These k-percent of segments are terms as the thin-sliced behavior segments in this work. Since our study scheme uses leave-dyad-out cross validation, the GMM model for behavior quantization, marginal PMF and joint PMF are all computed on training set of within each cross-validation fold.

The number of behavior clusters used in quantization is 128, and the number of discretized emotion levels is 10. In general, we observe that a more granular quantized representation would result in a more robust estimate of mutual information. However, too many clusters of either behaviors or emotion levels would lead to sparsity issue (detailed parameter analyses are presented in Section 4.2.1. The percentages of thin-sliced segments selected within each session used in our analyses are shown in Table 3 (i.e., correspond to the best accuracies obtained in the automatic regression experiments described in Section 4.2.2).

3 Thin-Sliced Behavior Segment Analyses

3.1 Analysis Setup

In the section, we present detailed analyses on the properties of these thin-sliced emotion-rich behavior segments (Section 2.2.2). Our analyses include three major parts: 1) time distribution 2) behavior distribution and 3) emotion level-dependent analysis. In *time distribution* analysis, our aim is to understand whether the selected behavior segments are likely to be in the beginning, middle, or ending portion of the session. In *behavior distribution* analysis, we investigate the distribution of the behavior types on these selected segments as compared to the behavior distribution in the original database. Lastly, we perform analysis to understand the change of behavior types distribution as a function of the discretized session-level emotion attributes.

All analyses are carried out with a cross-validation (leave-dyad-out) setting, i.e., in every given session of a testing dyad, we identify the per-session thin-sliced *emotion-rich behavior segments* using the computed PMF from the training set. The three different analyses results are then aggregated over all testing sessions and reported in the following section. This particular cross validation based analyses scheme help support more robust results, since each dyad interacts in multiple interactions, by performing leave-dyad-out cross validation, it ensures the generalization of our analyses (i.e., speaker-independent setup in dyad settings).

3.2 Time Analysis

We first examine where in each session are these behavior segments being selected. By splitting every session into three parts, i.e., the beginning $1/3$ (T1), the middle $1/3$ (T2), and the ending $1/3$ (T3), where each part corresponds to approximately one minute in duration. There is a total of 100 segments per session. In each of the 90 data samples, we select 70, 40, 20 segments in the audio modality for activation, dominance, and valence respectively, and for the body language modality, we select 50, 80, 40 segments for activation, dominance, and valence, respectively (Table 3).

Table 2 lists the total number of segments selected for each part of time duration of each behavior modality for the three emotion attributes. We observe that for all of the emotion attributes, the selected thin-sliced behavior segments are located more toward the end of the session (the T3 portion of each session). We additionally carry out t-test to assess whether there are statistically differences in the average number of segments selected between each portion of time ($p < 0.05$). Results indicate that for body language modality, the number of segments selected in T3 is significantly larger than in T1 for dominance and and valence; for audio modality, the number of segments selected in T3 is significantly larger than in T1 for activation (Table 2 with *).

It is quite interesting to see that by computationally identifying these thin-sliced behavior segments, it seems to

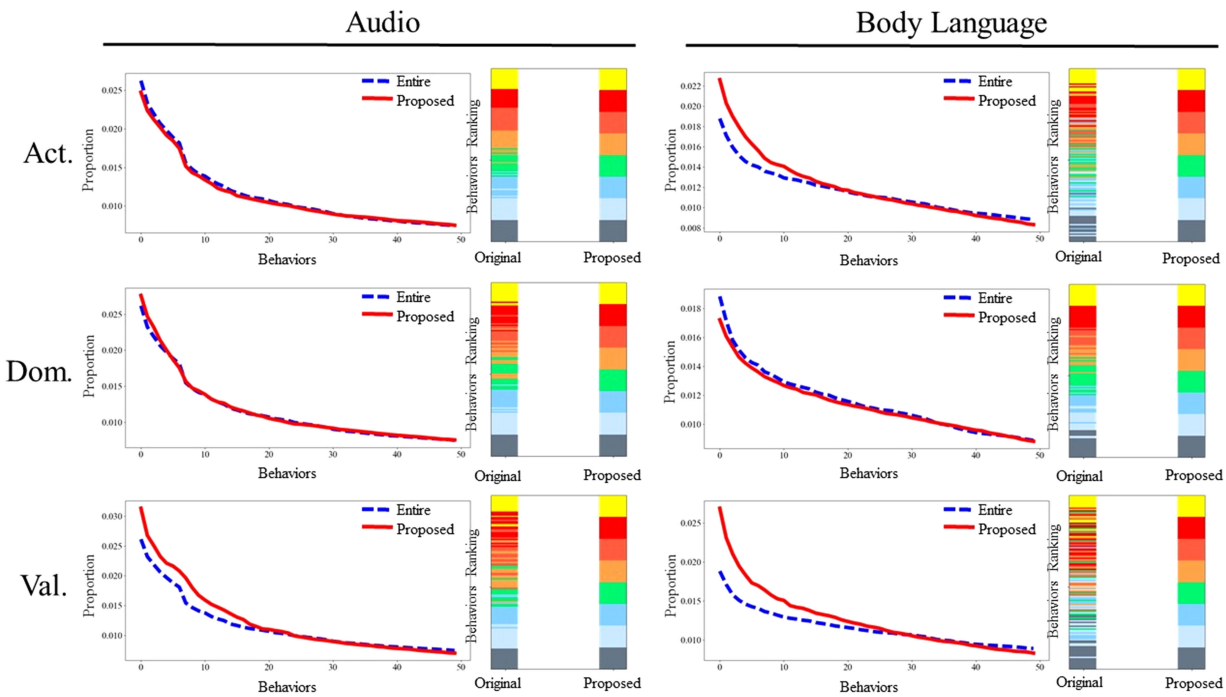


Fig. 4. Figure plots the percentages of behavior clusters in the descending order to the 50th-ranked (in terms of the frequeny of occurrences) clusters (the left plot) for the original *entire* database (in blue) and our *proposed* thin-sliced segments (in red) of each behavior modality. Also, the *Proposed* color bar in this figure (right) shows the sorted ranking of behavior types from top (yellow) to bottom (dark gray) using eight different colors, i.e., each color corresponds to a total of 16 behavior clusters. The same color in both bars (the *Proposed* and the *Original*) indicate the same exact behavior clusters.

corroborate with findings concluded from the past controlled psychology experimentations and theory in relating perception, attention and emotion [47]. The recent experience tend to exert a larger effect on human's overall perceptual assessment due to the continuous shaping of emotion perception when human attentional-effect is integrated with the sensory stimuli exposure across time.

3.3 Behavior Types Distribution Analysis

The behavior quantization process results in a total of 128 unique data-driven behavior clusters for audio and body language separately. We then carry out our framework to identify the salient subsegments of each session as the thin-sliced *emotion-rich* behavior segments. In this part, while the exact physical interpretation of each behavior type can be difficult to discern intuitively, we attempt to analyze how the distribution of these behavior types changes as a result of this selection process.

In Fig. 4, for each modality, the left plot shows the percentages of behavior types in the descending order to the 50th-ranked (ranking in terms of the frequeny of occurrences) for the original entire database (in blue) and our proposed thin-sliced emotion-rich behavior segments (in red). Note that the x-axis indicates the rank not the actual behavior cluster. From Fig. 4, as an example, we observe that by comparing the distribution of these top-50 ranked behavior types between the original database and the selected subset, it is evident that the procedure effectively emphasizes some of the behavior types. The top most occurring behavior types take up a larger percentage of the entire database after the selection process. This change is especially evident in the valence dimension of audio modality and activation and valence dimension of body language modality. By performing our behavior segment selection, it seems to reduce part of data within each session that carry little emotion-related information and retain (emphasize) those segments with high-emotional content.

We further demonstrate exactly how the ranking of each behavior cluster changes after the thin-sliced behavior segment selection procedure in Fig. 4. The *Proposed* bar in Fig. 4 (right) shows the sorted ranking of behavior types from top (yellow) to bottom (dark gray) using 8 different colors, i.e., each color corresponds to 16 different behavior types. The same color in both bars (the *Proposed* and the *Original*) indicate the same exact behavior types. By referencing the same color in the *Original* bar, which is also sorted in terms of ranking from top to bottom, we could see then the changes of behavior distribution from the original entire database to the *thin-sliced emotion-rich* behavior segments.

Overall, we can see that our proposed thin-sliced selection method does not alter the rank order on behavior clusters' frequency of occurrences too drastically, i.e., most of the 'yellow'-behavior clusters still remain on the top and the 'dark gray'-behavior types still stay on the bottom more or less. Although we do observe that the distribution in the mid-ranked behaviors can be re-arranged and mixed. By examining the two different plots in Fig. 4, we note that our thin-sliced behavior segment selection framework essentially performs emphasis on certain behavior clusters though it does not alter the rank order (the most to least occurring behavior types) too much. The effect of emphasizing certain behavior types with high emotion-information is likely to underscore the reason of our improved global emotion recognition correlations (Section 4.2).

3.4 Global Emotion Level-Dependent Analysis

In Section 3.3, we present an analysis on the change of behavior type distribution after we perform thin-sliced behavior segments selection. In this section, we analyze this change as a function of each discretized *global* emotion level

TABLE 4
Table Lists the the Unique Number of Behavior Clusters in Each of the Global Emotion Level (*Uniq. #*)

		Audio										
		LV0	**LV1**	**LV2**	**LV3**	**LV4**	**LV5**	**LV6**	**LV7**	**LV8**	**LV9**	**All LV**
Act.	**Entire**(*Uniq*. #):	N/A	128	128	128	128	128	128	128	128	128	128
	Proposed(*Uniq*. #):	N/A	126	128	128	128	128	128	128	128	128	128
Dom.	**Entire**(*Uniq*. #):	N/A	128	128	128	128	128	128	128	128	128	128
	Proposed(*Uniq*. #):	N/A	127	128	128	128	128	128	128	128	127	128
Val.	**Entire**(*Uniq*. #):	128	128	128	128	128	128	128	128	128	128	128
	Proposed(*Uniq*. #):	117	128	128	128	128	128	128	104	125	111	128
		Body Language										
		LV0	**LV1**	**LV2**	**LV3**	**LV4**	**LV5**	**LV6**	**LV7**	**LV8**	**LV9**	**All LV**
Act.	**Entire**(*Uniq*. #):	N/A	88	109	109	110	120	114	122	109	77	128
	Proposed(*Uniq*. #):	N/A	48	81	85	101	114	99	116	100	54	126
Dom.	**Entire**(*Uniq*. #):	N/A	99	106	116	114	120	120	116	100	88	128
	Proposed(*Uniq*. #):	N/A	93	100	115	111	118	118	114	94	81	127
Val.	**Entire**(*Uniq*. #):	102	106	119	118	115	123	112	97	90	52	128
	Proposed(*Uniq*. #):	75	88	114	105	94	114	88	59	53	20	125

'Entire' and 'Proposed' denotes the numbers of unique behavior types in the original database and after our proposed thin-sliced selection process

(the ten levels) for the three emotion attributes in both behavior modalities

While we observe that our selection method essentially alters the composition of behavior types that seem to bear little emotion information and emphasizes those that are emotion-rich, this does not correspond to directly choosing a reduced number of behavior types, i.e., decreasing behavior diversity overall for each emotion attribute. In fact, in Table 4, we list the total number of unique behavior types (denoted as Uniq #) for the original entire database and our proposed framework for each discretized level of the three emotion attributes. Under the *All LV* column, we can see that even after applying our thin-sliced segment selection methodology, the total number of unique behaviors for each emotion attributes of individual behavior modality remains to be around 128. It shows that our proposed emotion-rich behavior segments still contain the complete variety of behavior manifestation as the original database just with altered distributions.

There is, however, a very interesting point that we observe in Table 4. If we break down the total number of unique behavior types across each of the discretized global emotion level (LV0 - LV9), each level of emotion attribute in the original database contains pretty much the complete variety of the these behavior clusters. After the selection, we observe that our methodology reduces the total Uniq # at the extreme emotion levels, i.e., LV0 and LV9, across all emotion attributes in the body language modality. Specifically, in the body language modality, the total number of unique behavior types goes from 102 to 75 at the LV0, and 52 to 20 at the LV9 for valence dimension—the same trend also holds for activation and dominance. It implies that the raters are likely be influenced by a more limited types of behaviors among a wide variety of behaviors when annotating the emotions of those parts of the session at the level of extreme. For the mid-level emotions, this effect is minimal. The analysis also implies even though the total number of behavior manifestations is still large for the portions of data that annotators decide to be at the extreme level, there in fact is only a relatively limited number of behavior manifestations that actually possess meaningful emotion information—as evident by the reduction of Uniq # in those extreme levels.

Furthermore, in Fig. 5, we also present the total proportions of data from the top five most occurring behavior types within the *Original* and the *Proposed* database (denoted as Top5 Proportion). Across all individual levels for all emotion attributes and behavior modalities, by comparing to the original database, the top 5 most occurring behaviors covers much larger percentage of data—a result reinforcing the finding summarized in Fig. 4.

In summary, we present detailed analyses of the proposed thin-sliced emotion-rich behavior segments in Section 3. We obtain several insights about their properties:

1) *Time distribution*. Our analyses indicate that the within-session time distribution of these sliced behavior segments for both audio and body language modalities tend to be located toward the ending portion of an interaction.
2) *Behavior distribution*. The behavior distribution analysis shows that our proposed framework is not simply choosing a subset of behavior clusters for each emotion attribute. It effectively changes the overall distribution by emphasizing those that may carry more emotion information and not alter the rank order (relative occurrence frequencies) too drastically.
3) *Emotion level-dependent analysis*. In analyzing the behavior cluster distribution as a function of each discretized emotion level, our analyses indicate that the change in the distribution occurs across all individual levels. Furthermore, the method does not simply select a subset of behaviors as the total unique number of behavior types remain consistent except at the extreme levels of emotion, where there seems to be a smaller set of behavior types that carry meaningful session-level emotion information.

4 Session-Level Emotion Regression

In this section, we will describe our automatic session-level emotion regression framework by leveraging the identified thin-sliced emotion-rich behavior segments (Step 4-5 in Section 2.2.2). Furthermore, we also detail our experimental setup and discuss the results obtained.

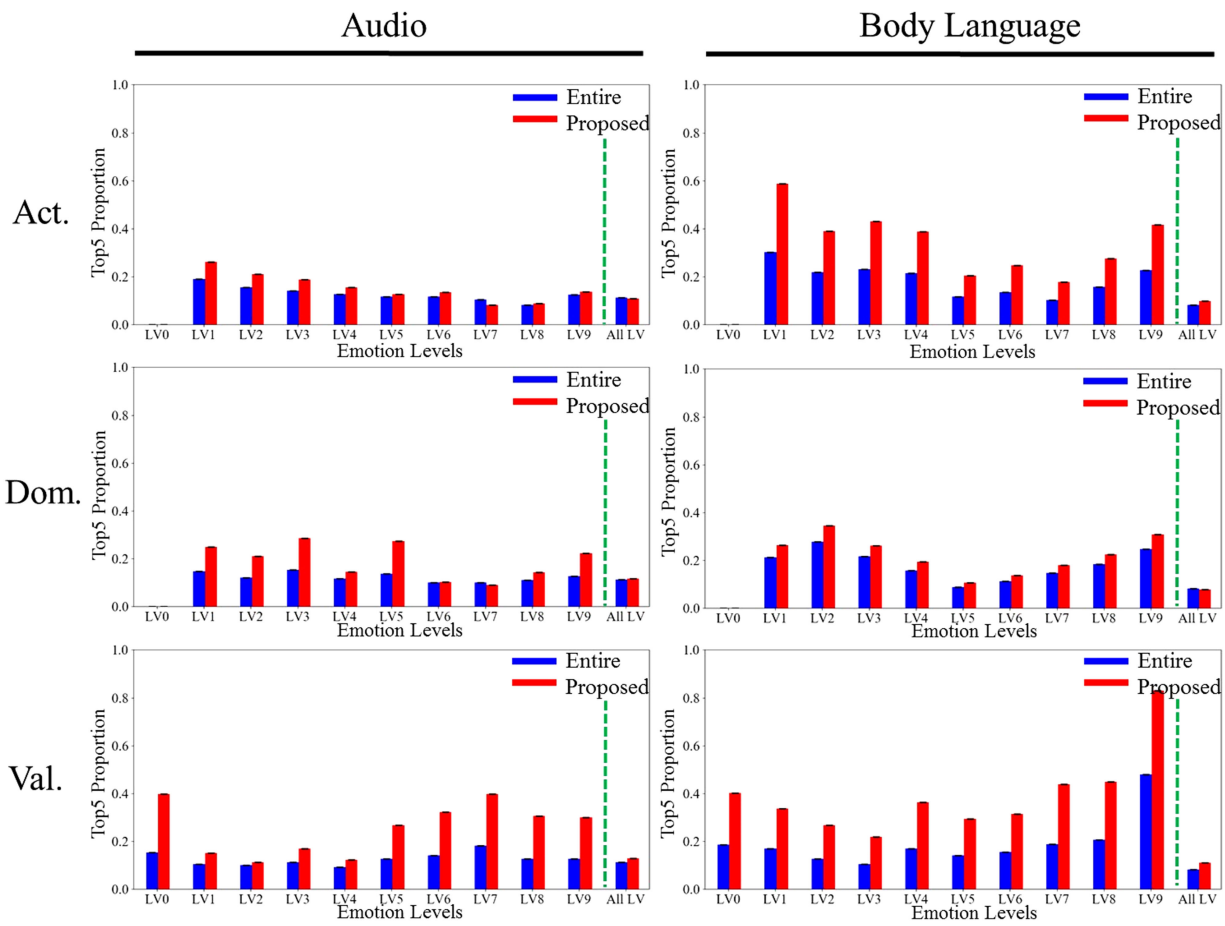


Fig. 5. Figure plots the percentage of data that the top five behavior types account (Top5 proportion) for each emotion attribute of each behavior modality in each of the global emotion level.

4.1 Thin-Sliced Session-Level Emotion Regressor

Our proposed automatic session-level emotion regression is to first build a regressor based on the identified *thin-sliced* behavior segments. The overall architecture (i.e., training and testing) procedure is described as follows (the first two steps are the same as described in Sections 2.2.1 and 2.2.2):

- *Step1:* Extract multimodal behavior features and quantize representations using GMM and discretize the global emotion annotation into 10 equally-spaced levels
- *Step2:* Select top k-segments to form the global emotion-rich behavior segments by computing mutual information between X and Y (all done only on the training set)
- *Step3:* Utilize GMM-based Fisher Vector encoding approach to generate a fixed session-level high-dimensional behavior feature vector from the selected k-segments per session in the training set
- *Step4:* Train an emotion regressor for each of the three emotion attributes using support vector regression on these thin-sliced FV behavior feature vectors (Step 3).
- *Step5:* At testing, since we do not have knowledge of the true global emotion levels of that particular session, we generate 10 different emotion-level dependent thin-sliced behavior segments, (i.e., by assuming the testing session is one of the 10 emotion levels, carrying out Step 2-3 to derive the behavior segments, then iterating over ten levels). These segments are further encoded using the FV approach. Finally, we average the regression output result by passing each of these 10 emotion-level dependent thin-sliced FVs to the trained session-level emotion recognizer (Step 4) to be our regressed value.

The use of GMM-based Fisher-vector encoding on the low-level features within each session data is carried out to learn a high-dimensional feature representation as a vector input to the machine learning algorithm, i.e., support vector regression (SVR). FV encoding has been shown to be successful in computer vision tasks (e.g., [48], [49], [50]), and has recently been demonstrate to possess competitive modeling power in speech-based paralinguistic and emotion recognition tasks (e.g.,[51], [52], [53]). FV has the advantage of both being a generative and discriminative feature representation model. It encodes both first and second order statistics. We will briefly describe the FV encoding below:

FV encoding can be derived as a special case of constructing Fisher kernel. The use of Fisher kernel is to measure the similarity between the two sets of data samples. Let's define a scoring function to measure similarity

$$\mathbf{G}_{\lambda}^{\mathbf{X}} = \nabla_{\lambda} \log \mathbf{u}_{\lambda}(\mathbf{X}), \tag{1}$$

where $\mathbf{u}_{\lambda}(\mathbf{X})$ denotes the likelihood of data $\mathbf{X}$ given the probability distribution function PDF, $\mathbf{u}_{\lambda}$. Here the choice of PDF is Gaussian Mixture Model (GMM), and λ represents the parameters of GMM, i.e., $(\bar{w}, \bar{\mu}, \bar{\Sigma})$. $\mathbf{G}_{\lambda}^{\mathbf{X}}$ is the direction where λ has to move to provide a better fit between $\mathbf{u}_{\lambda}$ and $\mathbf{X}$. Hence, if we can imagine the behavior feature over the

entire database is distributed as a GMM. By measuring the fit between each session's local emotion-rich behavior segments' data to this specified GMM, we then can encode the sequence of low-level features for each session into $\mathbf{G}_{\lambda}^{\mathbf{X}}$.

Furthermore, we would like to obtain a normalized $\mathbf{G}_{\lambda}^{\mathbf{X}}$ better suited for SVM. Fisher Information Matrix (FIM), $\mathbf{F}_{\lambda}$, from the theory of information geometry is hence utilized

$$\mathbf{F}_{\lambda} = E_{\mathbf{X}\sim\mathbf{u}_{\lambda}}[\mathbf{G}_{\lambda}^{\mathbf{X}}\mathbf{G}_{\lambda}^{\mathbf{X}'}],$$

and the normalized $\mathbf{G}_{\lambda}^{\mathbf{X}}$, denoted as $\mathbf{g}_{\lambda}^{\mathbf{X}}$, can be defined as

$$\mathbf{g}_{\lambda}^{\mathbf{X}} = \mathbf{F}_{\lambda}^{-\mathbf{1/2}}\mathbf{G}_{\lambda}^{\mathbf{X}} = \mathbf{F}_{\lambda}^{-\mathbf{1/2}}\nabla_{\lambda}\log\mathbf{u}_{\lambda}(\mathbf{X}). \tag{2}$$

In Equation (2), the term $\mathbf{g}_{\lambda}^{\mathbf{X}}$ is the so called Fisher vector. Let $\mathbf{X} = \bar{x}_t, t = 1 \ldots T_1$ and assume each $\bar{x}_i$ is i.i.d. then

$$\mathbf{g}_{\lambda}^{\mathbf{X}} = \sum_{t=1}^{T}\mathbf{F}_{\lambda}^{-\mathbf{1/2}}\nabla_{\lambda}\log\mathbf{u}_{\lambda}(\mathbf{x_t}),$$

we additionally know that $\mathbf{u}_{\lambda}(\mathbf{x})$ is a GMM with k mixtures expressed as

$$\mathbf{u}_{\lambda}(\mathbf{x}) = \sum_{k=1}^{K} w_k\mathbf{u}_{\mathbf{k}}(\mathbf{x}),$$

with $\lambda = \{w_k, \mu_k, \Sigma_k, k = 1, \ldots, K\}$ correspond to mixture weight, mean, and covariance matrix for each mixture of Gaussian. These parameters are of the following form:

$$\sum_{k=1}^{K} w_k = 1$$

$$\mathbf{u}_{\mathbf{k}}(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_k|^{1/2}}e^{\left(-\frac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k)\right)},$$

covariance matrices are set to be diagonal, i.e., $\Sigma_k = diag(\sigma_k^2)$.

We can put together Equation (2) with $\mathbf{u}_k(X)$ using the parameters described above. First, we define a probability $\gamma_t(k)$ as

$$\gamma_t(k) = \frac{w_k u_k(x_t)}{\sum_{j=1}^{K} w_j u_j(x_t)}.$$

From this, we can derive the gradient with respect to μ_k, σ_k of a data point x_t

$$\nabla_{\mu_k}\log\mathbf{u}_{\lambda}(x_t) = \gamma_t(k)\left(\frac{x_t-\mu_k}{\sigma_k^2}\right)$$

$$\nabla_{\sigma_k}\log\mathbf{u}_{\lambda}(x_t) = \gamma_t(k)\left(\frac{(x_t-\mu_k)^2}{\sigma_k^3} - \frac{1}{\sigma_k}\right).$$

With these, we derive the Fisher encoding of $\mathbf{X}$ for the first and second order statistics below

$$\mathbf{g}_{\mu_{\mathbf{k}}}^{\mathbf{X}} = \frac{1}{T\sqrt{w_k}}\sum_{t=1}^{T}\gamma_t(k)\left(\frac{x_t-\mu_k}{\sigma_k}\right) \tag{3}$$

$$\mathbf{g}_{\sigma_{\mathbf{k}}}^{\mathbf{X}} = \frac{1}{T\sqrt{2w_k}}\sum_{t=1}^{T}\gamma_t(k)\left(\frac{(x_t-\mu_k)^2}{\sigma_k^2} - 1\right). \tag{4}$$

This results in a fixed dimension vector at the session-level by concatenating the output from Equations (3) & (4), i.e.,

$$FV = \left[\mathbf{g}_{\mu_1}^{\mathbf{X}}, \mathbf{g}_{\sigma_1}^{\mathbf{X}}, \ldots, \mathbf{g}_{\mu_{\mathbf{k}}}^{\mathbf{X}}, \mathbf{g}_{\sigma_{\mathbf{k}}}^{\mathbf{X}}, \ldots, \mathbf{g}_{\mu_{\mathbf{K}}}^{\mathbf{X}}, \mathbf{g}_{\sigma_{\mathbf{K}}}^{\mathbf{X}}\right]. \tag{5}$$

At last, we employ the method of improved FV by performing L2 normalization [54]. In summary, we encode the sequences of individual behavior modality feature, i.e., those gathered from the *k-global emotion-rich behavior segments*, with the FV encoding (defined in Equation (5)). The use of FV-encoding is similar in principal to the use of GMM supervector [55] and i-vector approach [56] in tasks of speaker verification. These different methodologies, however, differ in that GMM-supervectors and i-vectors are generative models where the encoded representations for a data sample are mainly derived using maximum a-posterior adaption on GMM mean supervectors (i-vectors further perform dimensional reduction). The FV encoding is both a generative and discriminative approach due to its use of the Fisher scoring criteria to encode a data sample; it also encodes both parameters of means and variances.

4.1.1 Experimental Setup

We compare our proposed emotion regression method (Section 4.1) to two other different methodologies listed below:

1) *Baseline*. We use 100 percent of behavior data (i.e., the entire session) to derive the FV representation used for training the session-level emotion regressor. This method can also regard as the conventional method without selecting the *emotionally-salient* portion.
2) *Random*. We perform within-session random subsampling to generate the thin-sliced behavior segments used for deriving FV representation in the training of the session-level emotion regressor (i.e., randomly select k segments from 100 sub-partitions in each interaction without using the mutual information based selection).

All of the experiments are carried out in a leave-dyad-out cross validation (speaker-independent setup in dyad settings). The metric of evaluation is the average spearman correlation obtained for each testing dyad.

Major parameters used: GMM cluster number used for each behavior modality quantization is 128. The mixture number used in the session-level Fisher-vector encoding is 8 for audio and 2 for body language for all three emotion attributes. The parameter $k\%$ of the thin-sliced behavior segment selection of each modality for each emotion attribute is listed in Table 3. The SVR is trained using $C = 1$.

Finally, since the SVR is trained on each behavior modality, the multimodal fusion is carried out using a simple linear late fusion technique

$$S_{\text{Multimodal}} = \alpha \times S_{\text{Audio}} + (1-\alpha) \times S_{\text{BodyLanguage}},$$

where S_* indicates the regressed session-level emotion score from individual modality respectively, and α is determined using greedy search on the training set within the interval $[0, 1]$, and we set this number fixed across all folds in order to reduce the potential issue in over-fitting where every fold can have a very different weight.

4.2 Results and Discussions

A summary of recognition accuracies is presented in Table 5. The data proportion k is chosen empirically and its effect will be analyzed in Section 4.2.2. It is evident that the proposed *thin-sliced*-based global emotion regressor, i.e., focusing on the *emotionally-salient* behavior segments, outperforms the *baseline* model, i.e., using the entire sessions. In specifics, for the audio modality, our proposed thin-sliced global emotion regression obtains 16.27, 4.4, and 15.72 percent relative improvement

TABLE 5
Summary of Global Emotion Prediction Results for Baseline, Oracle, Proposed, and Random Selection

	Audio		
	baseline(k)	**proposed**(k)	**random**(k)
Activation	0.497 ($k = 100\%$)	***0.578*** ($k = 70\%$)	0.429 ($k = 70\%$)
Dominance	0.317 ($k = 100\%$)	***0.331*** ($k = 40\%$)	0.339 ($k = 40\%$)
Valence	0.426 ($k = 100\%$)	***0.493*** ($k = 20\%$)	0.354 ($k = 20\%$)
	Body Language		
	baseline(k)	**proposed**(k)	**random**(k)
Activation	0.402 ($k = 100\%$)	***0.436*** ($k = 50\%$)	0.368 ($k = 50\%$)
Dominance	0.609 ($k = 100\%$)	***0.603*** ($k = 80\%$)	0.595 ($k = 80\%$)
Valence	0.497 ($k = 100\%$)	***0.622*** ($k = 40\%$)	0.494 ($k = 40\%$)
	Fusion Model		
	baseline(α)	**proposed**(α)	**random**(α)
Activation	0.527 ($\alpha = 0.6$)	***0.605*** ($\alpha = 0.8$)	0.493 ($\alpha = 0.6$)
Dominance	0.652 ($\alpha = 0.1$)	***0.633*** ($\alpha = 0.3$)	0.604 ($\alpha = 0.4$)
Valence	0.591 ($\alpha = 0.6$)	***0.672*** ($\alpha = 0.6$)	0.511 ($\alpha = 0.4$)

The accuracy is measured using the average Spearman correlation of each cross validation fold. The table present all details for audio, body language, and fusion model of the recognition accuarcy, data proportion k and fusion weight α.

over the baseline model on attributes of activation, dominance, and valence, respectively. In the body language modality, our framework improves 8.45 and 25.15 percent over the baseline model on activation and valence respectively. Finally, by fusing audio and body language modality, we obtain promising Spearman correlations of 0.605, 0.633, and 0.672 on activation, dominance, and valence. The only exception is in the body language modality for the dominance attribute. The baseline model itself already achieves 0.609 correlation, which may indicates that the perception of global dominance, when influenced by body language behaviors, could largely be based on the entire session. We hypothesize this could due to the nature in the design of Active Analysis in eliciting behaviors. Specifically in the 2-sentence exercises, actors' body language behavior need to be highly utilized when playing a force and counter-force interaction (e.g., continuous holding, grabbing, getting in the face of the other person) with highly restricted lexical content.

From Table 5, we also provide a comparison to the *random selection* baseline. For the random experiment, we select the within-session behavior segments randomly instead of using our proposed method. We report results by selecting the same portion of data used in the *proposed* method. Clearly from the results, the improvement in our method is not due to a simple reduction in the amount of data; the selected thin-sliced behavior segments are those with higher emotional content with respect to the global affect perception.

In summary, our proposed multimodal thin-sliced based global emotion regressor obtains promising correlations of 0.605, 0.633, and 0.672 on attributes of activation, dominance, and valence respectively. The amount of data required for audio is 70, 40, and 20 percent for activation, dominance, and valence respectively, and by selecting these thin-sliced segments, we obtain improvement in the prediction correlations for all three attributes.

4.2.1 Analyses of Parameters Choice

We will present analyses and discussions on the emotion recognition correlations obtained as a function of the two major parameter choices for our proposed framework in this section:

- *Number of discretized emotion levels (Y):* Table 6 shows different emotion regression accuracies obtained as a function of different discretized levels of each emotion attribute. We observe that the accuracy improves with an increased number of discretized levels. However, if we further discretize each emotion attribute to a even finer granularity (e.g., 15 levels), the sparsity becomes an issue where only few or even none of the samples would fall in certain emotion levels. Fig. 6 shows a histogram distribution of our samples as we set different numbers of discretized levels. The appropriate granularity in the discretized emotion annotation is necessary in robust identification of emotion-rich behavior segments.
- *Number of quantized behavior clusters (X):* The mutual information is computed between discrete representation of both behavior and emotion attributes. We further report emotion regression results obtained as a function on the number of clusters used (mixtures of GMM) for the extracted low-level descriptors in Table 7. A similar trend is observed in the number of quantized behavior clusters and the number of discretized emotion levels. The accuracy improves as the number of quantized behavior clusters increases; however learning instability is observed if we continue to increase the number of mixtures used.

In summary, we observe a trend that the correlations improve as we increase the number of discretized emotion levels and the quantized behavior clusters. This is potentially due to a more reliable and robust estimation of mutual information used in our framework. However, the benefit of increasingly granular quantization often plateaus due the amount of data available.

4.2.2 Effect of Different Data Proportion

In this section, we further provide analysis on the prediction correlations obtained as a function of the $k\%$ within-session data selection. The modality-wise results are summarized in Fig. 7. In the body language part, for all three attributes, we observe a tendency that the prediction accuracy slightly improves (especially evident in the valence attribute) as the percentage of thin-sliced behavior segments reduces. This phenomenon, however, would plateau and start to show a

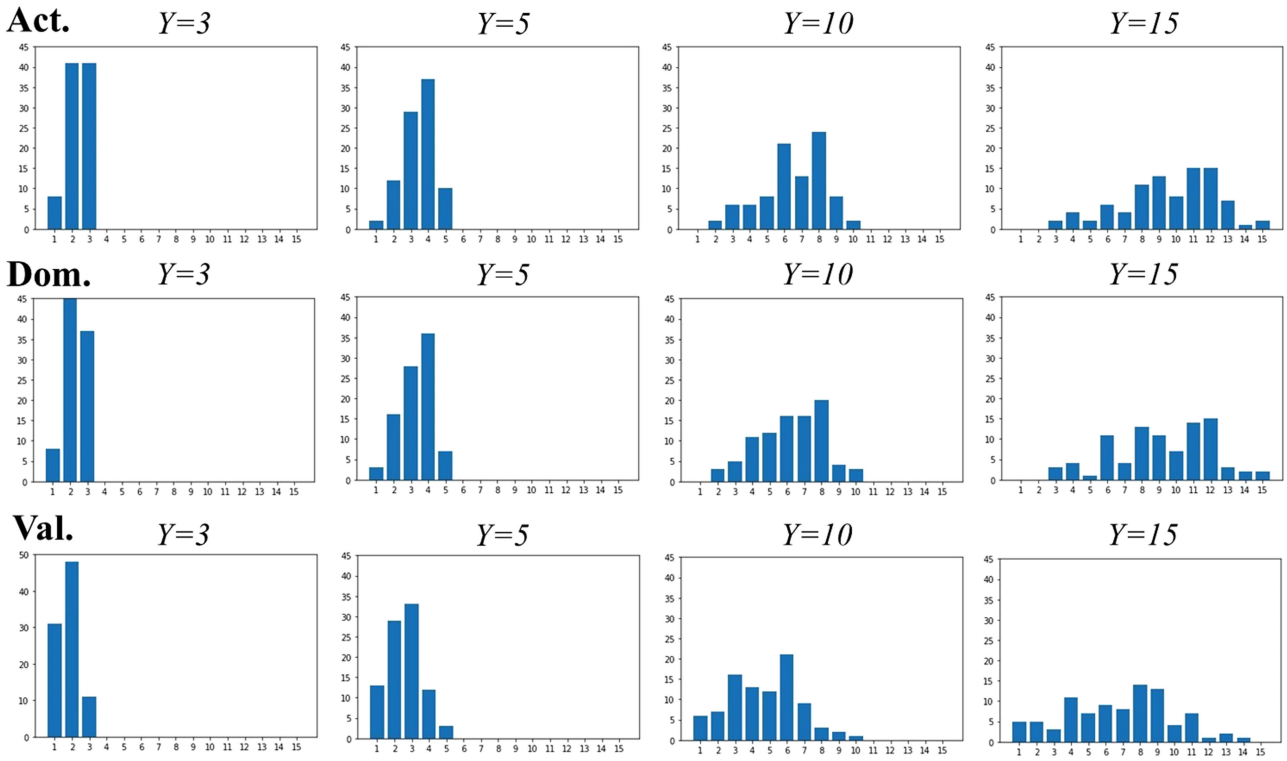


Fig. 6. Figure shows the a histogram of sample distribution for different number of emotion discretization (3, 5, 10, and 15 levels) for activation, dominance, and valence.

detrimental effect when the percentages of behavior data remain are too little (e.g., 10–30 percent).

A similar trend also holds in the audio modality. In general, in the audio modality, the amount of thin-sliced behavior segments required is fewer than the body language modality. When comparing between the three emotion attributes, we observe that valence attributes use lesser amount of data compared to activation; dominance requires a close to 100 percent (entire session) worth of data, i.e., reinforcing the finding in Section 4.2.

Furthermore, we report various fusion results by combining different k% segments from audio and body language modalities (Table 8). The results indicate that the appropriate choice of k% from each behavior modality is essential in obtaining results above baseline accuracies. This composition seems to be attribute-dependent, i.e., differs between activation, valence, and dominance. For example, in activation dimension, our proposed method uses 70 percent audio and 50 percent body language. Keeping audio at around 70 percent seems to be important in reaching accuracy above baseline, and the added benefit in incorporating body language arises from choosing k to be approximately 50 percent. In terms of valence dimension, our proposed method uses 20 percent within-session segments for audio and 40 percent for body language. The 40 percent for body language seems to be critical in obtaining improved correlation beyond baseline. In case where 90 percent audio and 40 percent of body language is used, the α needs to be set at 0.3 indicating that the majority of the regression contribution come from the body language modality.

TABLE 6
Correlations Obtained with Different Number of Quantized Emotion Attributes

Activation	Correlation
Y=10 (proposed)	0.605 (α = 0.8)
Y=5	0.552 (α = 0.7)
Y=3	0.540 (α = 0.8)
Dominance	**Correlation**
Y=10 (proposed)	0.633 (α = 0.3)
Y=5	0.617 (α = 0.1)
Y=3	0.609 (α = 0.3)
Valence	**Correlation**
Y=10 (proposed)	0.672 (α = 0.6)
Y=5	0.603 (α = 0.4)
Y=3	0.500 (α = 0.4)

While the exact relationship between different percentages of modality-specific data needed for each emotion attribute is difficult to rigorously tease apart due to the intertwining effect between the controlling factors of α fusion parameter and the $k\%$ parameter. However, our analyses indicate ranges of data proportion that is likely to be effective may differ for each emotion attribute in the CreativeIT database. The generalization of these insights to other emotion databases will be required to further substantiate these interpretations on humans perception.

5 Conclusions and Future Works

Human perception is thin-sliced in nature. In this work, we present computational analyses by identifying *session-level emotion-rich behavior segments* using a framework based on mutual information. With the thin-sliced identification framework, our proposed multimodal (audio and body language) global emotion regressor achieve an accuracy of 0.605, 0.633, and 0.672 on attributes of activation, dominance, and valence respectively, which outperforms

TABLE 7
Correlations Obtained with Different Numbers of Quantized Behavior Clusters

Activation	Correlation
128-GMM (proposed)	0.605 ($\alpha = 0.8$)
64-GMM	0.535 ($\alpha = 0.9$)
32-GMM	0.524 ($\alpha = 0.6$)
16-GMM	0.507 ($\alpha = 0.9$)
Dominance	**Correlation**
128-GMM (proposed)	0.633 ($\alpha = 0.3$)
64-GMM	0.616 ($\alpha = 0.2$)
32-GMM	0.577 ($\alpha = 0.6$)
16-GMM	0.606 ($\alpha = 0.3$)
Valence	**Correlation**
128-GMM (proposed)	0.672 ($\alpha = 0.6$)
64-GMM	0.585 ($\alpha = 0.5$)
32-GMM	0.536 ($\alpha = 0.6$)
16-GMM	0.567 ($\alpha = 0.6$)

framework using the entire session. The amount of thin-sliced data needed varies with the type of emotion attribute, and we observe that valence requires the least amount. We also present detailed analyses on these thin-sliced segments in order to bring additional insights about the potential human affect perception mechanism. We demonstrate that the time distributions of data within each session are located skewed toward the end of the session for both audio and body language thin-sliced segments. Furthermore, we show that our proposed thin-sliced selection technique effectively change the overall behavior distribution, i.e., emphasizing those behavior parts that are more emotion-information rich. This phenomenon is especially noticeable at the extreme emotion level. It seems to indicate that for those extreme emotion levels, the underlying human perception mechanism is likely to be affected by a few salient behavior manifestations.

There are multiple threads of future work. On the technical side, we observe in this work that when comparing to valence and dominance attributes, the overall accuracies tend to be lower for activation in general. The thin-sliced emotion perception mechanism may be fundamentally different between these three emotion attributes, we will develop computational frameworks that can better model the thin-sliced perception specifically for activation. Second, the current *emotion-rich behavior segments* are selected per modality separately. However, when annotators are being exposed to these data, they observe and make a judgment holistically and multimodally. We will develop algorithms to identify these segments jointly between audio and gestural information. Third, our proposed automatic thin-sliced affect regressor is done by averaging emotion-level dependent regression outputs. We will further investigate a joint confidence-weighted combination of these regression outputs multimodally to improve our framework. With additional insights obtained about the properties and characteristics of these *emotion-rich* behavior segments within each interaction session, we hope to inspire and develop

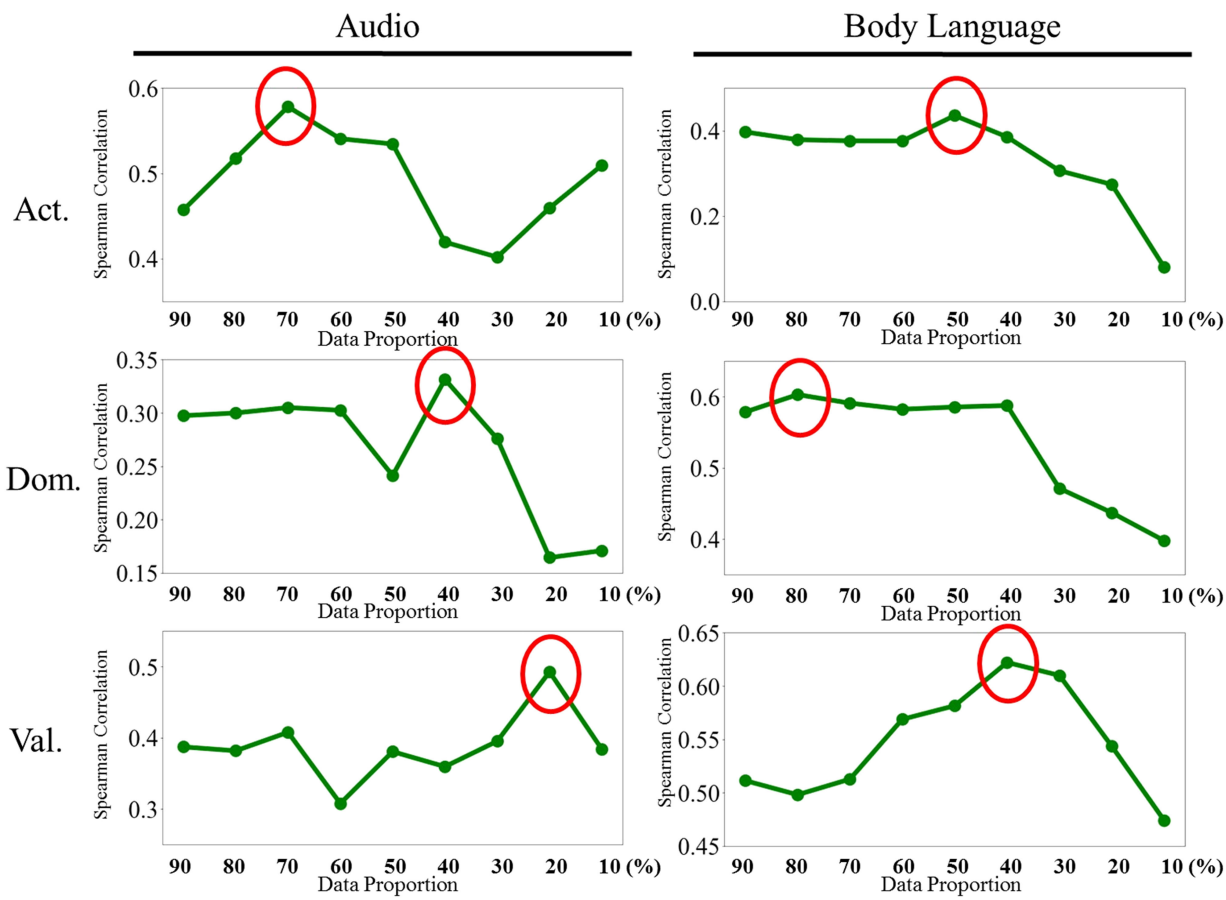


Fig. 7. Figure shows the prediction correlations obtained for different percentage (k) of thin-sliced behavior data selected using our proposed framework for each of the two behavior modalities separately for activation, dominance, and valence.

TABLE 8
Fusion Results Obtained with Different k% Selection for Both Modalities

Activation	Correlation
Audio 100% , Body 100% (baseline)	0.527 (α = 0.6)
Audio 70% , Body 50% (proposed)	0.605 (α = 0.8)
Audio 70% , Body 90%	0.596 (α = 0.8)
Audio 90% , Body 40%	0.487 (α = 0.7)
Audio 50% , Body 60%	0.543 (α = 0.7)
Audio 30% , Body 70%	0.522 (α = 0.6)
Dominance	**Correlation**
Audio 100% , Body 100% (baseline)	0.652 (α = 0.1)
Audio 40% , Body 80% (proposed)	0.633 (α = 0.3)
Audio 90% , Body 90%	0.640 (α = 0.3)
Audio 90% , Body 40%	0.652 (α = 0.2)
Audio 50% , Body 60%	0.615 (α = 0.3)
Audio 30% , Body 70%	0.642 (α = 0.3)
Valence	**Correlation**
Audio 100% , Body 100% (baseline)	0.591 (α = 0.6)
Audio 20% , Body 40% (proposed)	0.672 (α = 0.6)
Audio 90% , Body 40%	0.678 (α = 0.3)
Audio 70% , Body 90%	0.608 (α = 0.7)
Audio 50% , Body 60%	0.630 (α = 0.7)
Audio 30%, Body 70%	0.598 (α = 0.4)

further advanced and sophisticated algorithms to improve much less-investigated session-level (i.e., long durational behavior data) emotion recognition.

Furthermore, we will immediately initiate a human perceptual experiment protocol, i.e., recruiting additional annotators to judge the emotional content on the extracted thin-sliced behavior segments in order to provide further validity of this data-driven framework. By cross-referencing this data-driven framework in capturing emotion-rich behavior segments to the rigorous design of human perceptual experiment, we aim at knowing the limitation and potential of our proposed framework. In the long run, our aim is to understand the underlying thin-sliced perception mechanism of various emotion attributes that human possess in tasks of performing holistic judgment when observing continuous affective multimodal behavior displays. We hope to advance both scientific understanding of human perception of affect and also inspire novel technical algorithms of affect recognition framework that can supplement the emerging research efforts into deriving analytics for cross-domain human-centered applications.

Acknowledgments

The authors would like to thank the Ministry of Science and Technology, Taiwan (MOST-103-2218-E-007-012-MY3) for funding.

References

[1] B. De Martino, D. Kumaran, B. Seymour, and R. J. Dolan, "Frames, biases, and rational decision-making in the human brain," *Sci.*, vol. 313, no. 5787, pp. 684–687, 2006.

[2] J. S. Lerner and D. Keltner, "Beyond valence: Toward a model of emotion-specific influences on judgement and choice," *Cognition Emotion*, vol. 14, no. 4, pp. 473–493, 2000.

[3] I. Blanchette and A. Richards, "The influence of affect on higher level cognition: A review of research on interpretation, judgement, decision making and reasoning," *Cognition Emotion*, vol. 24, no. 4, pp. 561–595, 2010.

[4] G. Margolin, P. H. Oliver, E. B. Gordis, H. G. O'hearn, A. M. Medina, C. M. Ghosh, and L. Morland, "The nuts and bolts of behavioral observation of marital and family interaction," *Clinical Child Family Psychology Rev.*, vol. 1, no. 4, pp. 195–213, 1998.

[5] R. E. Heyman, "Observation of couple conflicts: Clinical assessment applications, stubborn truths, and shaky foundations," *Psychological Assessment*, vol. 13, no. 1, 2001, Art. no. 5.

[6] J. C. Norcross and B. E. Wampold, "Evidence-based therapy relationships: Research conclusions and clinical practices," *Psychotherapy*, vol. 48, no. 1, 2011, Art. no. 98.

[7] L. Sanders, C. Trinh, B. Sherman, and S. Banks, "Assessment of client satisfaction in a peer counseling substance abuse treatment program for pregnant and postpartum women," *Eval. Program Planning*, vol. 21, no. 3, pp. 287–296, 1998.

[8] C. Lord, S. Risi, L. Lambrecht, E. H. Cook Jr, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, "The autism diagnostic observation schedulegeneric: A standard measure of social and communication deficits associated with the spectrum of autism," *J. Autism Develop. Disorders*, vol. 30, no. 3, pp. 205–223, 2000.

[9] P. Borkenau, N. Mauer, R. Riemann, F. M. Spinath, and A. Angleitner, "Thin slices of behavior as cues of personality and intelligence," *J. Personality Social Psychology*, vol. 86, no. 4, 2004, Art. no. 599.

[10] N. Ambady and H. M. Gray, "On being sad and mistaken: Mood effects on the accuracy of thin-slice judgments," *J. Personality Social Psychology*, vol. 83, no. 4, 2002, Art. no. 947.

[11] J. R. Curhan and A. Pentland, "Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes," *J. Appl. Psychology*, vol. 92, no. 3, 2007, Art. no. 802.

[12] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis," *Psychological Bulletin*, vol. 111, no. 2, 1992, Art. no. 256.

[13] N. Ambady and R. Rosenthal, "Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness," *J. Personality Social Psychology*, vol. 64, no. 3, 1993, Art. no. 431.

[14] K. A. Fowler, S. O. Lilienfeld, and C. J. Patrick, "Detecting psychopathy from thin slices of behavior," *Psychological Assessment*, vol. 21, no. 1, 2009, Art. no. 68.

[15] L. P. Naumann, S. Vazire, P. J. Rentfrow, and S. D. Gosling, "Personality judgments based on physical appearance," *Personality Social Psychology Bulletin*, vol. 35, no. 12, pp. 1661–1671, 2009.

[16] T. F. Oltmanns, J. N. Friedman, E. R. Fiedler, and E. Turkheimer, "Perceptions of people with personality disorders based on thin slices of behavior," *J. Res. Personality*, vol. 38, no. 3, pp. 216–229, 2004.

[17] C. Oveis, J. Gruber, D. Keltner, J. L. Stamper, and W. T. Boyce, "Smile intensity and warm touch as thin slices of child and family affective style," *Emotion*, vol. 9, no. 4, 2009, Art. no. 544.

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

[19] D. Bone, C.-C. Lee, T. Chaspari, J. Gibson, and S. Narayanan, "Signal processing and machine learning for mental health research and clinical applications [perspectives]," *IEEE Signal Process. Mag.*, vol. 34, no. 5, pp. 196–195, Sep. 2017.

[20] M. P. Black, A. Katsamanis, B. R. Baucom, C.-C. Lee, A. C. Lammert, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Toward automating a human behavioral coding system for married couples interactions using speech acoustic features," *Speech Commun.*, vol. 55, no. 1, pp. 1–21, 2013.

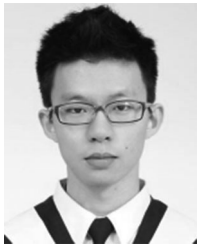
[21] P. G. Georgiou, M. P. Black, A. C. Lammert, B. R. Baucom, and S. S. Narayanan, "Thats aggravating, very aggravating: Is it possible to classify behaviors in couple interactions using automatically derived lexical features?" in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2011, pp. 87–96.

[22] S.-W. Hsiao, H.-C. Sun, M.-C. Hsieh, M.-H. Tsai, H.-C. Lin, and C.-C. Lee, "A multimodal approach for automatic assessment of school prinicipals' oral presentation during pre-service training program," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 2529–2533.

[23] W.-Y. Huang, S.-W. Hsiao, H.-C. Sun, M.-C. Hsieh, M.-H. Tsai, and C.-C. Lee, "Enhancement of automatic oral presentation assessment system using latent N-grams word representation and part-of-speech information," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 1432–1436.

[24] Z. Yang, A. Metallinou, E. Erzin, and S. Narayanan, "Analysis of interaction attitudes using data-driven hand gesture phrases," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 699–703.

[25] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image Vis. Comput.*, vol. 31, no. 2, pp. 137–152, 2013.
[26] D. R. Carney, C. R. Colvin, and J. A. Hall, "A thin slice perspective on the accuracy of first impressions," *J. Res. Personality*, vol. 41, no. 5, pp. 1054–1072, 2007.
[27] O. Kalinli and S. Narayanan, "Prominence detection using auditory attention cues and task-dependent high level information," *IEEE Trans. Audio Speech Language Process.*, vol. 17, no. 5, pp. 1009–1024, Jul. 2009.
[28] H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)*, vol. 39, no. 1, pp. 64–84, Feb. 2009.
[29] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 223–227.
[30] J. Gibson, A. Katsamanis, F. Romero, B. Xiao, P. Georgiou, and S. Narayanan, "Multiple instance learning for behavioral coding," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 81–94, Jan.–Mar. 2017.
[31] A. Katsamanis, J. Gibson, M. P. Black, and S. S. Narayanan, "Multiple instance learning for classification of human behavior observations," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2011, pp. 145–154.
[32] C.-C. Lee, A. Katsamanis, P. G. Georgiou, and S. Narayanan, "Based on isolated saliency or causal integration? Toward a better understanding of human annotation process using multiple instance learning and sequential probability ratio test," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 619–622.
[33] T. Bänziger, D. Grandjean, and K. R. Scherer, "Emotion recognition from expressions in face, voice, and body: The multimodal emotion recognition test (MERT)," *Emotion*, vol. 9, no. 5, 2009, Art. no. 691.
[34] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. 6th Int. Conf. Multimodal Interfaces*, 2004, pp. 205–211.
[35] A. Metallinou, M. Wöllmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 184–198, Apr.–Jun. 2012.
[36] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.
[37] A. Metallinou, Z. Yang, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC creativeIT database of multimodal dyadic interactions: From speech and full body motion capture to continuous emotional annotations," *Language Resources Eval.*, vol. 50, no. 3, pp. 1–25, 2015.
[38] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[39] S. M. Carnicke, "The Knebel technique: Active analysis in practice," *Actor Training*, pp. 99–116, 2010.
[40] 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 *Proc. ISCA Tut. Res. Workshop Speech Emotion*, 2000, pp. 19–24.
[41] P. Boersma, et al., "Praat, a system for doing phonetics by computer," *Glot Int.*, vol. 5, no. 9/10, pp. 341–345, 2002.
[42] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[43] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Trans. Neural Netw.*, vol. 20, no. 2, pp. 189–201, Feb. 2009.
[44] C.-C. Lee, M. Black, A. Katsamanis, A. C. Lammert, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 793–796.
[45] I. Priness, O. Maimon, and I. Ben-Gal, "Evaluation of gene-expression clustering via mutual information distance measure," *BMC Bioinf.*, vol. 8, no. 1, 2007, Art. no. 111.
[46] T. M. Cover and J. A. Thomas, "Entropy, relative entropy and mutual information," *Elements Inf. Theory*, vol. 2, pp. 1–55, 1991.
[47] R. M. Todd, W. A. Cunningham, A. K. Anderson, and E. Thompson, "Affect-biased attention as emotion regulation," *Trends Cogn. Sci.*, vol. 16, no. 7, pp. 365–372, 2012.
[48] X. Peng, C. Zou, Y. Qiao, and Q. Peng, "Action recognition with stacked fisher vectors," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 581–595.
[49] C. Sun and R. Nevatia, "Large-scale web video event classification by use of fisher vectors," in *Proc. IEEE Workshop Appl. Comput. Vis.*, 2013, pp. 15–22.
[50] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Comput. Vis. Image Understanding*, vol. 150, pp. 109–125, 2016.
[51] H. Kaya, A. Karpov, and A. A. Salah, "Fisher vectors with cascaded normalization for paralinguistic analysis," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 909–913.
[52] H. Kaya and A. A. Karpov, "Fusing acoustic feature representations for computational paralinguistics tasks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 2046–2050.
[53] Z. Yang and S. Narayanan, "Analyzing temporal dynamics of dyadic synchrony in affective interactions," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 42–46.
[54] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 143–156.
[55] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–311, May 2006.
[56] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio Speech Language Process.*, vol. 19, no. 4, pp. 788–798, May 2011.



Wei-Cheng Lin (S'15) received the BS degree in communication engineering from National Taiwan Ocean University (NTOU), Taiwan, in 2014. He is working toward the MS degree in the Electrical Engineering Department, National Tsing Hua University (NTHU), Taiwan. His research interests include human-centered behavioral signal processing (BSP), machine learning, and multimodal signal processing. He is a student member of the IEEE and IEEE Signal Processing Society.

Chi-Chun Lee (M'13) received the BS and PhD degrees in electrical engineering from the University of Southern California, in 2007 and 2012, respectively. He is an assistant professor with the Electrical Engineering Department, National Tsing Hua University (NTHU), Taiwan. He was a data scientist with id:a lab at ID Analytics in 2013. His research interests include the human-centered behavioral signal processing (BSP) and affective computing. He was awarded with the USC Annenberg Fellowship (2007-2009). He led a team that participated in and won the Emotion Challenge Classifierer Sub-Challenge in Interspeech in 2009. He is a coauthor on the best paper award in Interspeech 2010, and the most cited paper published in 2013 in the *Journal of Speech Communication* on automatic modeling of couples' behaviors during therapeutic sessions. He recently served as an area chair for Interspeech 2016 and 2018, respectively, senior program committee for ACII 2017 and ICMI 2018, and a guest editor of the *Journal of Computer Speech and Language* special issue on speech and language processing for behavioral and mental health. He has been involved in multiple granted interdisciplinary research projects, including aspects on education, psychology, neuroscientific, and health-related applications, with a focus of modeling human behaviors using signal processing and machine learning. He is a member of the IEEE.