Recognizing Social Signals with Weakly Supervised Multitask Learning for Multimodal Dialogue Systems

Yuki Hirano, Shogo Okada
Japan Advanced Institute of
Science and Technology
Nomi, Ishikawa, Japan
s1810154@jaist.ac.jp,okada-s@jaist.ac.jp

Kazunori Komatani
The Institute of Scientific and Industrial Research
Osaka University
Ibaraki, Osaka, Japan
komatani@sanken.osaka-u.ac.jp

ABSTRACT

Social signal processing is a methodology that is used to infer human inner states, including attitudes, sentiments and impressions, from verbal and nonverbal multimodal information. The difficulty in training a social signal recognition model is that the ground-truth (target) labels given by multiple coders often disagree because the annotation of social signals such as sentiments is a subjective and ambiguous task. We introduce weakly supervised learning (WSL) algorithms to such an inaccurate supervision setting in which the target label is not necessarily accurate. The novel challenge in this paper is to explore an effective WSL strategy for recognizing social signals. The strategy is verified through two multimodal datasets including audio, visual, and linguistic data collected in a human-agent dialogue setting. First, we clarify that the proposed WSL strategy for deep neural networks (DNNs), called tri-teaching works well in almost all classification tasks. Second, we demonstrate the effectiveness of integrating WSL and multitask learning (MTL), which exploits several label types in the datasets. Third, we show that our proposed approach achieves less accuracy degradation than an existing training algorithm for a DNN (curriculum learning) in a cross-corpus setting, with a maximum improvement of 7.2%.

CCS CONCEPTS

• **Computing methodologies → Discourse, dialogue and pragmatics**; *Multitask learning*.

KEYWORDS

Multimodal interaction, Social signal processing, Weakly supervised learning

ACM Reference Format:
Yuki Hirano, Shogo Okada and Kazunori Komatani. 2021. Recognizing Social Signals with Weakly Supervised Multitask Learning for Multimodal Dialogue Systems. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3462244.3479927

1 INTRODUCTION

A multimodal dialogue system should be able to naturally communicate with users by recognizing their current statuses, such as emotions and engagement statuses, to generate empathetic and adaptive responses. Developing such a system is one of the main challenges in the dialogue system domain. Social signal processing (SSP) [33] is a key technique that is used to recognize user inner states, including spoken language and nonverbal behaviors (facial expression, gesture, prosody, etc.), from multimodal data.

This study presents an effective machine learning (ML) algorithm for recognizing user social signals in human-machine communication by addressing the following problems. (i) It is difficult to determine ground-truth labels of human inner states. Such labels are usually determined by aggregating annotated labels given by coders (e.g., through majority voting or simple averaging). However, the annotation of a social signal such as sentiment is a subjective and ambiguous task, so the annotation results often disagree, and the results are unreliable. Training samples with such unreliable labels tend to degrade the performance of supervised learning models. (ii) It is expensive to collect multimodal communication data such as human-robot interaction data, so it is not easy to collect a large-scale corpus in such a setting. We need to train multimodal recognition models with a limited quantity of data. (iii) Developing a generalized social signal recognition model is an essential challenge because multimodal features vary due to differences in individual personalities and conversational situations (e.g., casual or formal). The recognition accuracy of a model should be maintained even when a model trained with data in a certain situation is used in another situation for testing.

We propose a weakly supervised multitask learning algorithm to mitigate the influence of unreliable labels in training datasets. Weakly supervised learning (WSL) and multitask learning (MTL) are appropriate for addressing unexplored problems (i) and (ii) simultaneously. Therefore, we explore the effectiveness of integrating WSL algorithms with MTL to recognize social signal labels. We use multimodal features, including language (words and question types of system utterances), vision (gestures and expressions), and acoustics (prosodic aspects and changes in vocal tones), as inputs.

The target task is to recognize multiple labels (interests and sentiments) for representing the user states observed in multimodal human-computer dialogue. In particular, we regard the annotated labels of user inner states as inaccurate supervision [45] because these labels sometimes disagree among coders. The aim of utilizing WSL is to improve the recognition accuracy of the user social signal labels by removing the effects of such noisy labels. For this purpose, we use two shared multimodal dialogue datasets from a publicly

ICMI '21, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8481-0/21/10…$15.00
https://doi.org/10.1145/3462244.3479927

available dialogue corpus that we released [20] [1], which includes the multimodal (audio and visual) behaviors of participants talking with a virtual agent. These two datasets were collected with different dialogue strategies and include three labels that were independently annotated during each exchange, which consisted of a system utterance followed by a user utterance. We summarize the main contributions in this work below.

Weakly supervised learning for SSP: We utilize a WSL algorithm in the inaccurate supervision setting for SSP tasks and demonstrate its effectiveness in social signal recognition. To our knowledge, applying WSL for multimodal SSP tasks has yet to be explored. We also propose an ensemble-based weakly supervised learning algorithm called tri-teaching, which aims to remove noisy samples by aggregating the learning results obtained from multiple deep neural networks. In addition, to avoid the effect of inaccurate supervision (Problem (i)), we modify the co-teaching algorithm [13] for the SSP domain as an efficient weakly supervised learning strategy.

Integrating multitask learning and WSL for SSP: We show that integrating multitask learning and weakly supervised learning is effective for recognizing multiple social signals. Multitask learning alleviates the data sparsity problem, in which a limited number of labeled data are given (Problem (ii)). This type of integration was not explored in past studies.

Effectiveness of WSL in cross-corpus settings: Weakly supervised multitask learning prevents accuracy degradation in cross-corpus settings, where the training and test corpora are different (Problem (iii)). This setting, in which a model trained with one corpus is used in the recognition module of another situation, is assumed to be more realistic.

2 RELATED WORK

2.1 Multimodal social signal processing

The fusion of multimedia or multimodal data, including audio, visual, and linguistic features, is a promising method for recognizing social signals, including emotions and engagement. Many existing studies have focused on developing social signal recognition models for human-robot/agent interaction settings using multimodal machine learning techniques [4]. In human-agent or robot interaction settings, some studies [7, 27] have focused on engagement recognition based on users' multimodal behaviors. Agent systems with social signal sensing [16, 32] have recently been developed for training interpersonal communication skills. Several studies have focused on detecting user interests [10, 15]. Weber et al. [35] developed a dynamic user modeling approach based on reinforcement learning, which analyzed a person's reactions while a robot told jokes. Nasihati et al. [28] presented dialogue management routines for a system to engage in multiparty agent-infant interactions.

As another important direction, many studies have focused on modeling social signals in multimedia settings. Biel et al. [6] presented a multimodal analysis approach for predicting personality impressions based on what was said in YouTube videos. Brilman et al. [8] proposed a prediction model for successful debaters using multimodal information. Recently, it was shown that deep neural network (DNN) techniques contribute to accurate multimodal modeling [4] when applied to SSP. A temporally selective attention model [37], multi-attention recurrent network [41], memory fusion network [40], and tensor fusion network [39] were proposed for multimodal sentiment analysis. In another setting, group detection during natural social gatherings based on standing conversations was conducted using ubiquitous and multimodal sensing [1, 11].

Many previous studies focused on modeling one annotation label, such as engagement, communication skills, personality, or humor. However, several labels annotating the same data help improve the recognition accuracy of the resulting model. Hirano et al. [14] presented a multimodal modeling method with multitask learning to recognize multiple labels, such as interest levels, sentiment levels and next-action decisions, to implement adaptation strategies for multimodal dialogue systems. In many previous works, ground truths (labels) were carefully determined by aggregating annotation results obtained using multiple coders or self-reported annotation results. However, the influence of disagreement among coders on the accuracy of SSP cannot be ignored. We propose a WSL algorithm that can train robust SSP models with inaccurate labels on human-system dialogue interactions.

2.2 ML with subjectively annotated labels

Social signal perception is subjective and thus often results in disagreement among coders (annotators). It is difficult to improve model accuracy by removing samples with disagreeing labels from the training dataset. There are two existing approaches for solving this problem. One approach involves focusing on how to define reliable labels from subjectively annotated labels. The other approach explicitly trains the differences in the labels produced by multiple coders [17, 23, 29].Earlier studies explored approaches for avoiding coder disagreement and merging different labels given by the multiple coders recruited through crowdsourcing via majority voting [36]. Ozkan et al. [29] proposed a two-step conditional random field (CRF), which was designed for a backchannel prediction task. Inoue et al. [17] proposed a recognition model for user engagement in human-robot interactions using a hierarchical Bayesian model that estimates both the engagement level and the characteristics of each coder as latent variables. Kumano et al. [23] proposed a probabilistic model for integrating labels of empathy as perceived by coders. The model was used to capture how gazes and facial expressions co-occur between a pair of participants. Lotfian and Busso [25] proposed a method for designing a machine learning curriculum to maximize efficiency during the DNN training process for emotion recognition in speech by considering disagreements in crowdsourced labels.

Recent research has shown that curriculum learning plays a similar role to weakly supervised learning (WSL) in terms of inaccurate supervision. If we assume that label disagreement is noise for supervised learning, the WSL approach can avoid the effect of label disagreement. Although WSL is promising for improving image segmentation [43] and image retrieval [24] accuracy, recent research has not focused on utilizing WSL for SSP tasks.

2.3 Weakly supervised learning

The setting in which some labels are not reliable is called inaccurate supervision. Weakly supervised learning (WSL) is a research area

[1] https://doi.org/10.32130/rdata.4.1

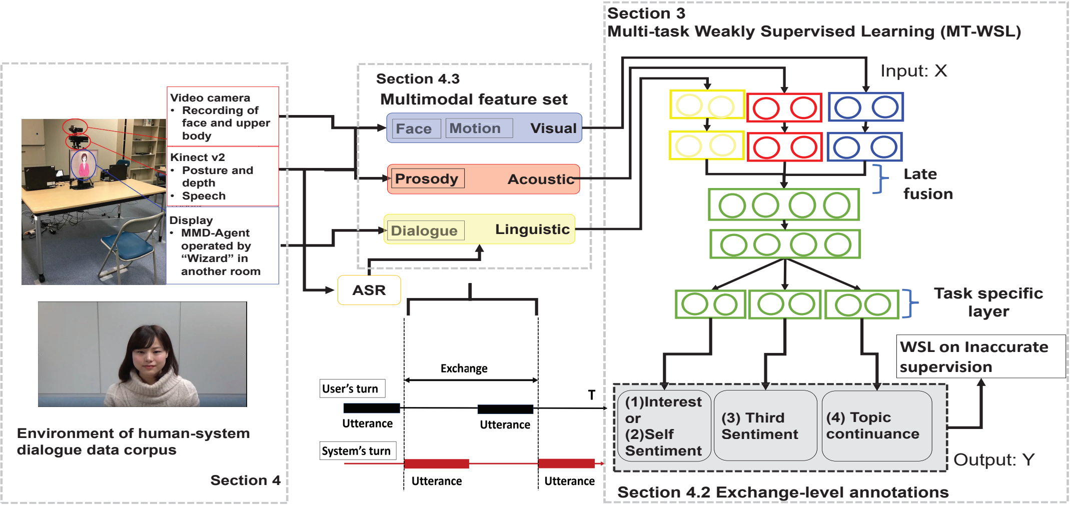

Figure 1: Overview of weakly supervised multitask learning for SSP

that aims to build a robust prediction model for inaccurate supervision [45]. Many algorithms have been proposed to train robust models with noisy labels. Several studies have attempted to estimate noise transition matrices and correct loss functions [12, 30]. However, it is difficult to estimate a noise transition matrix. An alternative approach is to train a model by using selected samples, e.g., via MentorNet [18], decoupling [26], and co-teaching [13]. In addition, many loss functions that are robust to noise labels have been proposed, e.g., the symmetric cross-entropy [34] and generalized cross-entropy functions [44]. Han et al. [13] showed that co-teaching yields better experimental results than decoupling and MentorNet, so we modify the co-teaching algorithm using three networks (we call this approach tri-teaching). We require a large quantity of labeled data to achieve high performance with a large-scale DNN. However, it is not easy to collect reliable annotated data for dialogue systems. The utilization of a multitask learning (MTL) algorithm is an efficient approach for mitigating the data limitation problem. To mitigate both the inaccurate label problem and the data limitation problem, which are inherently caused in SSP tasks, we integrate MTL and WSL algorithms. We show that integrating MTL and state-of-the-art WSL can improve accuracy.

3 PROPOSED APPROACH

An overview of the proposed algorithm for the human-agent dialogue setting is shown in Figure 1. We propose four weakly supervised multitask learning (MT-WSL) algorithms as follows:

MT+WSL(1): Multitask learning with co-teaching [13]
MT+WSL(2): Multitask learning with co-teaching+ [38]
MT+WSL(3): Multitask learning with tri-teaching
MT+WSL(4): Multitask learning with tri-teaching+

Tri-teaching and tri-teaching+ are extensions of co-teaching and co-teaching+, respectively. In Section 3.2, we propose a new weakly supervised learning method, tri-teaching, which aims to overcome the data limitations of the SSP task. In Section 3.3, we combine WSL methods with MTL by considering situations where three similar labels are annotated for each sample x_i.

3.1 Preliminaries

We frame the social signal recognition task as a classification task; that is, we consider a K-class dataset with a feature vector for sample x_i and its associated reference label $y_i \in \{1, 2, \cdots, K\}$. Let $f(\cdot)$ denote a neural network model. We denote the probability of each label for sample x_i, which is output by the trained model, as $p(k|x_i)$, where $k \in \{1, 2, \cdots, K\}$, and the ground-truth distribution for sample x_i as $q(k|x_i)$, where $\sum_{k=1}^{K} q(k|x_i) = 1$, $q(y_i|x_i) = 1$, and $q(k|x_i) = 0$ for all other $k \neq y_i$. As preliminaries of the proposed method, we employ two types of ensemble methods: co-teaching and co-teaching+.

Co-teaching: Co-teaching maintains and trains two networks, $f^{(1)}$ and $f^{(2)}$, simultaneously [13]. In each minibatch, we use the model $f^{(1)}$ (resp. $f^{(2)}$) to select small-loss samples $\overline{D}^{(1)}$ (resp. $\overline{D}^{(2)}$). The number of selected samples is determined by $\lambda(E)$, where E is the current epoch. We select $\lambda(E)$ percent of the small-loss samples in each minibatch. Then, the selected samples in $f^{(1)}$ (resp. $f^{(2)}$) are sent to the peer network $f^{(2)}$ (resp. $f^{(1)}$) as training data to update the parameters. A DNN can filter out noisy samples at the beginning of the training process because DNNs learn easy/clean data first [2, 42]. However, the problem is that a DNN starts to fit noisy data as the training process proceeds. We gradually decrease $\lambda(E)$ so that we can remove noisy data before the DNN fits noisy labels. $\lambda(E)$ is defined as $1 - \min\{\tau, \frac{E}{E_k}\tau\}$, where τ is the estimated noise ratio and E_k is a hyperparameter for controlling $\lambda(E)$. $\lambda(E)$ is a decreasing function in terms of the number of epochs, and $\lambda(1) = 1$. After E_k epochs, $\lambda(E)$ is fixed to τ, and $(1 - \tau)$ samples are used to update the parameters. A small E_k (large E_k) reduces $\lambda(E)$ gradually (rapidly).

Co-teaching+: Co-teaching+ is an extension of the co-teaching method [38], which has the problem by which two models converge to similar statuses with an increase in the number of epochs. To

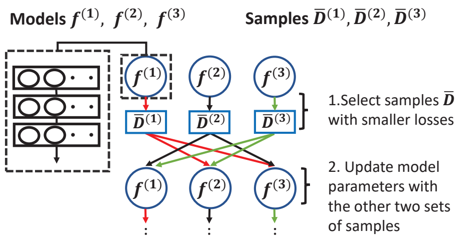
Models $f^{(1)}$, $f^{(2)}$, $f^{(3)}$
Samples $\overline{D}^{(1)}, \overline{D}^{(2)}, \overline{D}^{(3)}$
$f^{(1)}$
$f^{(2)}$
$f^{(3)}$
$\overline{D}^{(1)}$
$\overline{D}^{(2)}$
$\overline{D}^{(3)}$
1.Select samples $\overline{D}$ with smaller losses
2. Update model parameters with the other two sets of samples

Figure 2: Overview of the tri-teaching algorithm

overcome this drawback, co-teaching+ combines the "update by disagreement" strategy proposed in the decoupling algorithm of [26] with co-teaching. In the decoupling algorithm of [26], the disagreeing samples are selected as follows. Let two DNN networks $f^{(1)}$ and $f^{(2)}$ have the same architecture but different weight initializations. In each minibatch, the algorithm selects samples where the prediction labels disagree by the following rule: $D^{'}$=$\{(x_i, y_i)$: $\overline{y}_i^{(1)} \neq \overline{y}_i^{(2)}\}$, where $\overline{y}_i^{(1)}$ and $\overline{y}_i^{(2)}$ are the labels predicted by $f^{(1)}$ and $f^{(2)}$, respectively. Among the disagreeing samples, each model selects small-loss samples and sends them to its peer models for the parameter update phase in the same manner as co-teaching.

3.2 Enhancement of WSL

We propose tri-teaching, which is an extension of co-teaching. The overview of this method depicted in Figure 2. Co-teaching excludes some training samples, but this can reduce the generalization ability of the resulting model because the size of the dataset decreases. Tri-teaching can exploit training data more efficiently than co-teaching, as union sets are obtained from the other two models.

Tri-teaching: The algorithm is shown in Algorithm 1. We maintain three neural network models: $f^{(1)}$, $f^{(2)}$, and $f^{(3)}$. In each minibatch, each of the three models selects small-loss samples $\overline{D}^{(1)}$, $\overline{D}^{(2)}$, and $\overline{D}^{(3)}$, respectively (lines 5-7 in Algorithm 1). Here, $Loss(D^{'}; f^{(i)})$ denotes the sum of the losses in minibatch $D^{'}$ calculated by model $f^{(i)}$. To update model $f^{(1)}$, we use $\overline{D}^{(2)} \cup \overline{D}^{(3)}$. Similarly, we use $\overline{D}^{(1)} \cup \overline{D}^{(3)}$ and $\overline{D}^{(1)} \cup \overline{D}^{(2)}$ to update $f^{(2)}$ and $f^{(3)}$, respectively (lines 8-10 in Algorithm 1). After the three models are trained, a result is obtained by averaging the model outputs.

Tri-teaching+: The concept of tri-teaching+ is the same as that of co-teaching+. We maintain three networks: $f^{(1)}$, $f^{(2)}$ and $f^{(3)}$. In each minibatch $\overline{D}$, we select disagreeing samples: $\overline{D}^{'} = \{(x_i, y_i) : \neg(\overline{y}_i^{(1)} == \overline{y}_i^{(2)} == \overline{y}_i^{(3)})\}$, where $\overline{y}_i^{(m)}$ is a label predicted by the estimates of model $f^{(m)}$. From the disagreeing samples, each model selects small-loss samples $\overline{D}^{(1)}$, $\overline{D}^{(2)}$, and $\overline{D}^{(3)}$, and these data are used for updating, similarly to the process used in tri-teaching.

Algorithm 1: Tri-teaching algorithm

```
    input :initial models $f^{(1)}, f^{(2)}, f^{(3)}$; learning rate $\eta$; epoch
           $E_k$; $E_{\max}$; training dataset $D$; batch size $B$;
           estimated noise rate $\tau$
 1  for $E = 1, 2, \cdots, E_{max}$ do
 2      Shuffle training dataset $D$ into $\frac{|D|}{B}$ minibatches;
 3      for $N = 1, \cdots, \frac{|D|}{B}$ do
 4          Fetch a minibatch $\overline{D}$ from $D$;
 5          Select $\overline{D}^{(1)} = \mathrm{argmin}_{D^{'}:|D^{'}| \geq \lambda(E)|\overline{D}|} Loss(D^{'}; f^{(1)})$;
 6          Select $\overline{D}^{(2)} = \mathrm{argmin}_{D^{'}:|D^{'}| \geq \lambda(E)|\overline{D}|} Loss(D^{'}; f^{(2)})$;
 7          Select $\overline{D}^{(3)} = \mathrm{argmin}_{D^{'}:|D^{'}| \geq \lambda(E)|\overline{D}|} Loss(D^{'}; f^{(3)})$;
 8          Update $f^{(1)} \leftarrow f^{(1)} - \eta \nabla Loss(\overline{D}^{(2)} \cup \overline{D}^{(3)}; f^{(1)})$ ;
 9          Update $f^{(2)} \leftarrow f^{(2)} - \eta \nabla Loss(\overline{D}^{(1)} \cup \overline{D}^{(3)}; f^{(2)})$ ;
10          Update $f^{(3)} \leftarrow f^{(3)} - \eta \nabla Loss(\overline{D}^{(1)} \cup \overline{D}^{(2)}; f^{(3)})$ ;
11      end for
12      Update $\lambda(E) = 1 - \min\{\tau, \frac{E}{E_k}\tau\}$
13  end for
    output: $f^{(1)}, f^{(2)}, f^{(3)}$
```

3.3 Weakly supervised multitask learning

To address the limitations regarding the quantity of available training data, we use multitask learning (MTL), which incorporates information from other tasks. MTL is a machine learning approach that simultaneously solves multiple learning tasks while exploiting the commonalities and differences across the tasks [9]. Generally, an MTL model consists of shared layers and task-specific layers. The lower layers are shared across all tasks, while the higher layers are maintained as task-specific layers, as shown in Figure 1.

The loss function per task L_t for sample x_i is defined as

$$L_t(x_i, y_{i,t}) = -\sum_{k=1}^{K} q_t(k|x_i) \log(p_t(k|x_i)) \qquad (1)$$

where $y_{i,t}$ is the associated label, $p_t(k|x_i)$ is output of the trained model, and $q_t(k|x_i)$ is the ground-truth distribution for task $t \in T$. Hereafter, we explain the process with a case containing three tasks, i.e., $T = \{task1, task2, task3\}$. Let the loss functions of the three tasks be L_{task1}, L_{task2}, and L_{task3}; then, the loss function for MTL L_{multi} is defined as $L_{multi} = L_{task1} + L_{task2} + L_{task3}$. The subtasks, $task1$ to $task3$, correspond to the classification tasks of the three labels (e.g., the sentiment labels) in each corpus (see "Output: Y" on the right side of Figure 1 and Section 4.2). In our MTL setting, our motivation is to learn all tasks equally, and the equally weighted loss of each task is used as the loss function.

We combine MTL with WSL methods in the following procedure. We use L_{multi} to select the small-loss samples. In co-teaching+ and tri-teaching+, which use prediction disagreement for the update step, if one of the three tasks' predictions is different, we use the sample to update the parameters.

4 DATA

We used two shared multimodal dialogue datasets: Hazumi1712 and Hazumi1902, which are a part of the corpus in [20].

4.1 Recording settings and datasets

A virtual agent, MMD-agent[2], was used as the interface, as shown on the left side of Figure 1. The participants talked with the agent

[2] http://www.mmdagent.jp/

on a display. The behaviors of the participants were recorded using a video camera and a Microsoft Kinect V2 sensor. The virtual agent was manipulated by the WoZ method. The operator pretended to be a dialogue system. No specific task was defined; chat dialogues were utilized. The Hazumi1712 dataset was collected to analyze users interest levels and sentiment states. The operator selected questions or responses from the wizard interface for six topics that were selected by the participant in a random order [21]. The number of participants was 30 (15 males and 15 females), and their ages ranged from 20 to 50 years. The Hazumi1902 dataset was collected to analyze an adaptive dialogue strategy so that users could enjoy talking. The operator selected utterances and appropriately changed topics. For example, if participants were not interested in the topic, the operator changed the topic. If the participants seemed to enjoy talking with the system, the operator served the role of a good listener. The number of participants was 30 (10 males and 20 females), and their ages ranged from 20 to 70 years. The Hazumi1902 recording settings were the same as that of Hazumi1712, but the dialogue strategies of the WoZ operator and participants were different.

4.2 Definitions and statistics of labels

We used four labels per exchange: (1) the interest level (IN), (2) self-sentiment levels felt by participants (SS), (3) sentiment levels annotated by third-party coders (TS), and (4) topic continuance level (TC). (1) IN helped the system determine the timing for changing topics. (2) SS and (3) TS helped the system recognize whether the user was enjoying the dialogue and adapt its utterances to the users. (4) TC captured whether the system should continue with or change the current topic. (1) IN was annotated with three choices: interest (o), unknown (t), and no interest (x). (2) SS, (3) TS, and (4) TC were annotated on a seven-point scale [14]. Each exchange had three labels: (1) IN, (3) TS, and (4) TC in Hazumi1712 and (2) SS, (3) TS, and (4) TC in Hazumi1902.

4.2.1 Statistics of annotations. The multiple labels annotated for the exchanges can be used for multitask learning. The agreement values (Cronbach's alpha values) were 0.86 and 0.86 for (3) TS and (4) TC in Hazumi1712 and 0.86 and 0.82 for (3) TS and (4) TC in Hazumi1902, respectively. The average agreement value was 0.50 (Fleiss' kappa) for (1) IN in Hazumi1712[3]. These values showed high agreement but were not identical. The correlations between labels are shown in Table 1. From the left table, three labels in Hazumi1712 were highly correlated with each other, as the correlation values ranged from 0.62 to 0.87. Conversely, from the right table for Hazumi1902, the correlation values between (2) SS and the other two labels, (3) TS and (4) TC, were lower than those between the other pairs. This means that the self-sentiment labels annotated by the participants were different from the sentiment labels produced by third-party coders or the topic continuance labels in Hazumi1902.

4.2.2 Ternary labels. Averaged annotated scores (AASs) were calculated for each sample and were converted into ternary labels (high, neutral, and low). More specifically, for (1) IN, we regarded interest (o) as 1, unknown (t) as 0, and no interest (x) as −1 and calculated the corresponding AAS. Exchanges with AAS values higher than $1/N_A$ and lower than $-1/N_A$ were given high and low labels, respectively, where N_A is the number of annotators. The other exchanges were neutral. In the cases of the (2) SS, (3) TS, and (4) TC labels, which were annotated on 7-point scales, exchanges with AAS values higher than 4.5 and lower than 3.5 were given high and low labels, respectively. The other exchanges were neutral. The label distributions are shown in Table 2.

Table 1: Correlation matrices between labels

Hazumi1712

	(3) TS	(4) TC
(1) IN	0.62	0.87
(3) TS	-	0.63

Hazumi1902

	(3) TS	(4) TC
(2) SS	0.40	0.28
(3) TS	-	0.69

Table 2: Label distribution (%)

	Hazumi1712			Hazumi1902		
Class	(1) IN	(3) TS	(4) TC	(2) SS	(3) TS	(4) TC
High	32.6	43.7	38.1	49.1	49.8	44.7
Neutral	28.8	37.8	34.2	30.5	42.7	39.4
Low	38.6	18.5	27.7	20.4	7.5	15.9
Total	2,261 samples			2,337 samples		

4.3 Multimodal features

The multimodal feature set [4] was automatically extracted in the same manner as in [14][19].

4.3.1 Acoustic features. Acoustic features were extracted per utterance using openSMILE software. The features used in [31] were used. The types of utilized acoustic features were the root mean square frame energy (RMSenergy), mel-frequency cepstral coefficients, zero-crossing rate, voice activity probability based on VAD (VoiceProb), and fundamental frequency (F0). For each of these features, delta coefficients were also calculated.

4.3.2 Visual features. Visual features are composed of facial features and motion features. We used OpenFace [3] to extract facial features. We used 10 facial feature points around both eyes and the mouth. The velocities between frames and the absolute values of the between-frame accelerations were calculated. In addition, we used the occurrence proportions of 18 action units.

The participants joint positions with 3D coordinates were obtained from Microsoft Kinect V2. The positions were used to extract the corresponding motion features.

4.3.3 Linguistic features. We extracted linguistic features from the utterances of the participants and dialogue log data. Linguistic features were extracted from text data, which were the results of automatic speech recognition (ASR) for the recorded audio data. The frequencies (denoted as bag-of-words (BoW)) and parts of speech (PoSs) of words were obtained as linguistic features after the morphological analysis using MeCab [22]. The features of the system utterances were extracted from dialogue logs. We used the numbers of words in a system utterance (*lenS*) and a user utterance (*lenU*), their difference ($lenS - lenU$), the duration of a system utterance and the types of the system utterances.

[3] This dataset was composed of 3 discrete labels, so kappa was used.

[4] The detail is shown in
https://www.jaist.ac.jp/~okada-s/paper/MultimodalFeatures_ICMI2021.pdf

5 EXPERIMENTAL SETTINGS

5.1 Experiment types and measurements

Two types of experiments were conducted: one took place in a "one-corpus setting", where the training and test data came from the same dataset, and the other took place in a "cross-corpus setting", where we used one dataset for training and the other dataset for testing. The cross-corpus setting was important for assessing the generalization performance of the model. The two datasets had two common labels: (3) TS and (4) TC. Thus, we used these two labels for the cross-corpus setting. We did not use linguistic features because the BoW dictionaries were different in the two corpora. We report the average accuracy values obtained by the methods on the test dataset in Section 6.4. In the one-corpus setting, 10-fold cross validation was conducted. In the cross-corpus setting, we trained and tested the models 3 times with random initialization and reported their average accuracies.

5.2 Comparative methods

We prepared multiple comparative algorithms to evaluate the performance of the proposed MT-WSL algorithms.

5.2.1 Single-task DNN (ST-DNN). We used a DNN with multiple fully connected layers and dropout after each layer. The unimodal model consisted of one hidden layer with 100 units. After concatenating the output units of the three unimodal models, the layer with the concatenated units was connected to one hidden layer that also had 100 units. Finally, we added an output layer with three dimensions.

5.2.2 Multitask DNN (MT-DNN). The architectures of the lower layers were the same as that of the single-task DNN. After concatenating the output units of the three unimodal models, the layer with the concatenated units was connected to add one task-specific layer with 100 units for the three tasks. Finally, we added an output layer with three dimensions after each task-specific layer. In both the single-task and multitask DNNs, we set the batch size to 100, the total number of epochs to 80, and the dropout rate to 0.5. We used the Adam optimizer and set the learning rate to 0.001.

5.2.3 Multitask DNN with curriculum learning (MT-CL). The strategy that learns simpler samples before more difficult samples is called curriculum learning [5]. It is well known that the effectiveness of curriculum learning is partially similar to that of weakly supervised learning. We used two curriculum learning strategies proposed in [25].

Pretraining method (MT-CL-pre): This strategy used a pretrained model to determine a curriculum. We trained a model using all training samples for E_{CLP} epochs. After training, the models were tested with the same training samples. Equation (2) was used to define the difficulty d_i of sample x_i,

$$d_i = \begin{cases} -p(\overline{y}_i|x_i), & \text{if } y_i = \overline{y}_i \\ p(\overline{y}_i|x_i), & \text{if } y_i \neq \overline{y}_i \end{cases} \tag{2}$$

where $\overline{y}_i$ is a label predicted by the model. An easy sample x_i has a smaller d_i than a difficult sample. The idea behind this strategy is that a sample is easy to predict if it can be predicted correctly even with a small number of training epochs because the DNN learns simple patterns first [2]. We denote this strategy as MT-CL-pre.

Intercoder agreement method (MT-CL-agree): The second strategy used the level of disagreement between coders to determine a curriculum. If the coder agreement was low for a sample, we considered the sample ambiguous (difficult). The difficulty d_i is defined as

$$d_i = \frac{\sum_j^{A_i}[y_{ij} \neq \hat{y}]}{A_i} \tag{3}$$

where $\hat{y}$ is the consensus label obtained by majority voting, A_i is the number of coders, and y_{ij} is the label assigned by coder j for sample x_i. We denote this strategy as MT-CL-agree.

Based on these two strategies, we divided the training samples into three bins, where the first bin contained the easiest samples. During the first E_{CLA1} epochs, we trained the models using only the easiest samples. Between epochs E_{CLA1} and E_{CLA2}, we added medium-difficulty samples to the training process. After E_{CLA2} epochs, we used all of the training samples. We combined MTL with CL in the following procedure. First, we calculated the difficulty $d_{i,t}$ of sample x_i for task t using Equations (2) and (3). We defined the difficulty $d_{i,multi}$ of sample x_i in the multitask setting as a summation of the three task difficulties: $d_{i,multi} = d_{i,task1} + d_{i,task2} + d_{i,task3}$.

5.3 Hyperparameters and fusion method

The parameter settings for WSL are described here. For the proposed WSL algorithms, we set the estimated noise ratio τ to 0.2 in the one-corpus setting and to 0.6 in the cross-corpus setting; E_k was set to 25 for both settings. For CL, in the case of the self-sentiment label in Hazumi1902, which was annotated by only one person, we defined the disagreement difficulty $d_{i,multi}$ using only the other two labels. For MT-CL-pre and MT-CL-agree, we set the number of epochs $E_{CLP} = 30$, $E_{CLA1} = 30$, and $E_{CLA2} = 60$.

Early fusion and late fusion were used to fuse different modalities. In early fusion, the feature vectors from different modalities were concatenated into one feature vector. Late fusion combined the results of the trained unimodal models into a final output, as shown on the right side of Figure 1. We observed that late fusion outperformed early fusion, so we report the results of late fusion.

6 RESULTS

Three research questions (RQs) were addressed in the experiments, each of which corresponded to the subsections after Section 6.2.

RQ1: How does WSL contribute to tasks involving unreliable annotated labels?

RQ2: How does combining WSL and MTL contribute to task performance compared with other algorithm?

RQ3: How does MTL+WSL contribute to preventing performance degradation in a cross-corpus setting?

6.1 Effectiveness of multimodal fusion

We investigated the effectiveness of multimodal fusion for a multitask weakly supervised learning model. Table 3 shows the classification accuracies of three unimodal (audio, visual and linguistic) models and multimodal models with the three modality features obtained by using tri-teaching. From Table 3, the best accuracies for TS and TC (these impression scores) on Hazumi1902 (77.5, 73.3)

Table 3: Effectiveness of multimodal fusion for the proposed MT-WSL models with tri-teaching

[%]	Hazumi1712 corpus			Hazumi1902 corpus		
Modality	IN	TS	TC	SS	TS	TC
Audio	55.6	58.1	57.4	52.5	74.6	69.2
Visual	45.5	53.6	47.5	52.9	70.4	66.0
Linguistic	55.6	56.6	56.5	56.1	74.4	70.7
Multimodal	**61.4**	**62.4**	**63.1**	**59.7**	**77.5**	**73.3**

Table 4: The contribution of weakly supervised learning (WSL) and multitask learning (WSL model: tri-teaching)

[%]	Hazumi1712 corpus			Hazumi1902 corpus		
	IN	TS	TC	SS	TS	TC
Majority	38.6	43.7	38.1	49.1	49.8	44.7
ST-DNN	59.4	61.8	60.9	58.6	76.2	72.5
MT-DNN	60.5	62.1	62.4	57.8	77.2	73.1
ST-WSL-DNN	60.4	61.2	61.0	**60.2**	77.3	**73.7**
MT-WSL-DNN	**61.4**	**62.4**	**63.1**	59.7	**77.5**	73.3

were significantly better than those on Hazumi1712 (62.4, 63.1). From the finding that the number of utterances with low-level sentiments in Hazumi1712 was greater than that in Hazumi1902 (Table 2), it was found that estimating low-level sentiment is a more difficult task for models. In the comparison among the unimodal and multimodal models, the multimodal models generated by the MT-WSL algorithm obtained the best accuracies on all six tasks, so multimodal fusion promises to improve the classification accuracy. In Sections 6.2 to 6.4, we evaluate the performance of the algorithms through a multimodal model comparison.

6.2 Contribution of WSL (RQ1)

We investigated whether MTL or WSL contributed more to the recognition tasks by conducting ablation tests, in which the MTL or WSL strategy was removed from MT-WSL. The MT-WSL is an algorithm that integrates multitask learning (MTL) and weakly supervised learning (WSL). Table 4 shows the classification accuracy comparison among the single-task learning model (ST-DNN), multitask learning model (MT-DNN), weakly supervised single-task learning model (ST-WSL-DNN), and weakly supervised multitask learning model (MT-WSL-DNN). The tri-teaching model was used as the WSL algorithm. MT-WSL-DNN obtained the best accuracies on four tasks (all labels in Hazumi1712 and TS in Hazumi1912) at 61.4%, 62.4%, 63.1%, 77.5%. ST-WSL-DNN obtained the best accuracy on two tasks (SS and TC in Hazumi1912) at 60.2% and 73.7%, respectively. The results show that an ensemble-based sample selection strategy with weakly supervised learning is effective in improving model accuracy.

To investigate the difference in effectiveness between MTL or WSL, we compared the accuracies of MT-DNN (without WSL) and ST-WSL-DNN (without MTL). The accuracies of MT-DNN on all tasks in Hazumi1712 were better than those of ST-WSL-DNN, so the multitask strategy was more effective than the WSL strategy for Hazumi1712. Conversely, ST-WSL-DNN obtained better results for all tasks in Hazumi1902, so MTL and WSL contributed to improving the accuracy of the model for different tasks. From these results, it can be seen that MT-WSL-DNN took advantage of both MTL and WSL because it obtained the best accuracies on four tasks and the second-best accuracies on the other two tasks.

Table 5: Classification accuracies of multitask learning with weakly supervised learning

[%]	Hazumi1712 corpus			Hazumi1902 corpus		
	IN	TS	TC	SS	TS	TC
Majority	38.6	43.7	38.1	49.1	49.8	44.7
ST-DNN	59.4	61.8	60.9	58.6	76.2	72.5
MT-DNN	60.5	62.1	62.4	57.8	77.2	73.1
MT-CL-pre	60.6	62.3	62.8	57.5	76.7	72.8
MT-CL-agree	60.5	62.5	62.9	57.7	77.5	73.0
MT-WSL (Proposed)						
Co-teaching	61.0	**62.9**	**63.2**	58.4	77.3	74.1
Co-teaching+	59.0	61.6	61.9	59.4	77.5	**74.5**
Tri-teaching	**61.4**	62.4	63.1	**59.7**	77.5	73.3
Tri-teaching+	60.3	61.1	62.2	59.6	**77.9**	**74.5**

Table 6: Classification accuracies of multitask learning with weakly supervised learning in cross-corpus settings

[%]	train1712 → test1902		train1902 → test1712	
	TS	TC	TS	TC
Majority	49.8	44.7	43.7	38.1
MT-DNN	63.2	56.2	56.5	51.9
MT-CL-pre	64.1	56.4	56.7	**52.4**
MT-CL-agree	63.0	55.5	56.3	51.6
MT-WSL (Proposed)				
Co-teaching	63.6	58.6	56.8	52.2
Co-teaching+	64.4	58.0	55.9	51.5
Tri-teaching	**66.8**	**63.6**	**57.2**	52.2
Tri-teaching+	63.9	58.6	55.5	51.6

6.3 Comparison with other algorithms (RQ2)

We compared the accuracies of the multitask WSL strategies and comparative algorithms, including multitask curriculum learning algorithms. Table 5 shows the accuracies of multitask learning for the two datasets. In the comparisons between the multitask WSL algorithms (MT-WSL) and non-WSL algorithms, the best accuracy across all tasks was obtained by MT-WSL. In the experiments with the Hazumi1712 corpus, the best performances of the MTL+WSL models for the three labels, (1) IN, (3) TS, and (4) TC, were 61.4% (tri-teaching), 62.9% (co-teaching) and 63.2% (co-teaching), respectively, which were better than the best accuracy of MT-CL (60.6%, 62.5% and 62.9%, respectively). In the experiment with the Hazumi1902 corpus, MTL+WSL outperformed the best accuracy of MT-CL on all tasks. The best (2) SS, (3) TS and (4) TC performances were 59.7% (tri-teaching), 77.9% (tri-teaching+) and 74.5% (co-teaching+ and tri-teaching+), respectively. Although the best WSL model differed depending on the given task, the tri-teaching-based models obtained the best accuracy in four tasks, so increasing the number of ensemble models (from two to three) tended to contribute to the recognition tasks. However, in the comparison between the MT-WSL algorithms and non-WSL algorithms, the improvements in accuracy achieved by the best MT-WSL algorithm were in the range from 0.3% (TC in Hazumi1712) to 1.4% (TC in Hazumi1902), which were not significant. The MT-WSL algorithms are appropriate for tasks with small amounts of data, so we need to further analyze how large of a data sample is required for these algorithms. The exploration of these issues will be addressed in future work.

6.4 Cross-corpus experiment (RQ3)

We compared the classification accuracies of the models in a cross-corpus setting, which is more realistic, because a model trained with one corpus is often used in the recognition modules of different dialogue systems (with different dialogue scenarios or participants). Table 6 shows the classification accuracies obtained in the cross-corpus setting. Although the results in Table 6 were worse than the results in Table 5, this was expected because the cross-corpus setting was more difficult. Columns 2 and 3 show the results of training with Hazumi1712 and testing with Hazumi1902 (train1712→test1902). Columns 4 and 5 show the results of the opposite case (train1902→test1712).

In the case of train1712→test1902, the accuracies of the MT (MT-DNN) were 63.2% and 56.2% for (3) TS and (4) TC, respectively, while the MT+WSL method obtained the best accuracies: 66.8% (tri-teaching) and 63.6% (tri-teaching). MT+WSL outperformed the simple MT method in terms of accuracy by 3.6% and 7.4% and outperformed MT-CL by 2.7% and 7.2%. In the case of train1902→test1712, the accuracies of the simple MT method were 56.5% and 51.9%, while the best accuracies were 57.2% (tri-teaching) and 52.4% (MT-CL-pre). The improvement in train1902→test1712 was smaller than that in train1712→test1902. We speculate that the reason for this is that, as shown in Table 2, the label distribution of Hazumi1902 was biased and the quantity of low-class data was insufficient.

7 DISCUSSION AND CONCLUSION

The advantages and limitations of the proposed methods are discussed based on the experimental results (Tables 5 and 6). After the discussion, we conclude this study.

Difference between tri-teaching and curriculum learning: We discuss the reason why the accuracies of the WSL-based algorithm were better than those of other algorithms. In particular, MT+WSL obtained better accuracy than the curriculum learning (CL-based) algorithms in the cross-corpus setting. The main difference between WSL- (tri-teaching) and CL-based algorithms lies in how to select the samples that have priority for use as training data. On the one hand, CL-based algorithms (MT-CL-pre and MT-CL-agree) determine the order of training samples before the model is trained on the basis of the output of the pretraining model or the agreement of annotators. On the other hand, ensemble-based WSL algorithms (tri-teaching or co-teaching) select the training samples per epoch for training the model based on majority voting with respect to the outputs of ensemble models (Algorithm 1). From the results, it can be seen that the approach of sample selection per epoch in the ensemble-based WSL algorithms is effective for obtaining better accuracy in the cross-corpus setting.

Co-teaching vs tri-teaching: We discuss the contribution of tri-teaching and tri-teaching+ methods by comparing them with co-teaching and co-teaching+. In the cross-corpus setting, co-teaching+ and tri-teaching+ did not work well, so the strategy in these algorithms that selects samples where the prediction labels disagree was not effective. From Table 5, the performance of co-teaching and that of tri-teaching were quite similar, so tri-teaching using three models did not improve the accuracy in the one-corpus setting. Conversely, tri-teaching outperformed co-teaching, with a maximum improvement of 5% in the cross-corpus setting, as shown in Table 6, so increasing the number of ensemble models has the potential to select effective samples for training a task-independent robust model.

Conclusion: In this paper, we defined three critical problems with respect to the recognition of user social signals in human-machine communications: (i) the unreliable annotated labels in training samples, (ii) the high cost of collecting multimodal interaction data, and (iii) the dependency of models on training data. We first incorporated the two methods, multitask learning (MTL) and weakly supervised learning (WSL), into the SSP task. We have demonstrated their potential through our experiments. The experimental results clarified the contributions designed to mitigate the three problems in this paper. First, the WSL algorithms slightly improved the classification accuracies of both single-task and multitask DNNs in almost all tasks for the two tested corpora. This result indicates that the unreliable label problem can be alleviated by proposed WSL (thereby addressing Problem (i) and **RQ1**). Second, combining WSL with MTL worked well even with a small amount of multimodal interaction training data (thus addressing Problem (ii) and **RQ2**). Third, among the proposed WSL strategies, tri-teaching obtained the best accuracies under the cross-corpus setting for the three tasks (thus addressing Problem (iii), **RQ3**). MTL+WSL could prevent accuracy degradation more effectively than other algorithms (a maximum of 7.2%). However, the improvement in the one-corpus setting was less than 2%, so the model needs to be refined further.

There are several future research directions to enhance MT-WSL. The first direction is to improve the performance of MT-WSL by integrating ensemble-based methods such as tri-teaching and robust loss methods (e.g., [44]). The second direction is to extend tri-teaching to N-teaching with N ensemble models and to validate the robustness of the extended version. We need to explore how to exchange (smaller loss) samples to enhance N-teaching. Investigating a suitable fusion method and exploring the successful conditions of WSL with various kinds of data corpora will be addressed in future work. The third direction is to validate the robustness of the proposed algorithm on a more difficult cross-corpus setting, with corpora that include participants who have different cultural backgrounds and speak different languages. The fourth direction is to implement an adaptive dialogue system with a SSP recognition module based on the proposed learning method.

ACKNOWLEDGMENTS

This work was partially supported by "Five-star Alliance", the Cooperative Research Program of NJRC Materials and Devices, KAKENHI: Grant-in-Aid for Scientific Research (A) (grant no. 19H01120), and JST AIP Trilateral AI Research (grant no. JPMJCR20G6), Japan.

REFERENCES

[1] Xavier Alameda-Pineda, Yan Yan, Elisa Ricci, Oswald Lanz, and Nicu Sebe. 2015. Analyzing Free-Standing Conversational Groups: A Multimodal Approach. In *Proc. ACM International Conference on Multimedia*. 5–14.

[2] Devansh Arpit, Stanis Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. 2017. A Closer Look at Memorization in Deep Networks. In *Proc. International Conference on Machine Learning (ICML)*. 233–242.

[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. 59–66.

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 41, 2 (Feb 2019), 423–443.

[5] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum Learning. In *Proc. International Conference on Machine Learning (ICML)*. 41–48.

[6] Joan-Isaac Biel, Vagia Tsiminaki, John Dines, and Daniel Gatica-Perez. 2013. Hi YouTube! Personality Impressions and Verbal Content in Social Video. In *Proc. ACM International Conference on Multimodal Interaction (ICMI)*. 119–126.

[7] Dan Bohus and Eric Horvitz. 2009. Learning to Predict Engagement with a Spoken Dialog System in Open-world Settings. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. 244–252.

[8] Maarten Brilman and Stefan Scherer. 2015. A Multimodal Predictive Model of Successful Debaters or How I Learned to Sway Votes. In *Proc. ACM International Conference on Multimedia*. 149–158.

[9] Richard Caruana. 1993. Multitask Learning: A Knowledge-Based Source of Inductive Bias. In *Proc. International Conference on Machine Learning (ICML)*. 41–48.

[10] Yuya Chiba, Masashi Ito, Takashi Nose, and Akinori Ito. 2014. User Modeling by Using Bag-of-Behaviors for Building a Dialog System Sensitive to the Interlocutor's Internal State. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. 74–78.

[11] Ekin Gedik and Hayley Hung. 2018. Detecting Conversing Groups Using Social Dynamics from Wearable Acceleration: Group Size Awareness. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technology* 2, 4, Article 163 (2018), 24 pages.

[12] Jacob Goldberger and Ehud Ben-Reuven. 2017. Training deep neural-networks using a noise adaptation layer. In *Proc. International Conference for Learning Representations (ICLR)*.

[13] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Proc. Conference on Neural Information Processing Systems (NeurIPS)*. 8527–8537.

[14] Yuki Hirano, Shogo Okada, Haruto Nishimoto, and Kazunori Komatani. 2019. Multitask Prediction of Exchange-Level Annotations for Multimodal Dialogue Systems. In *Proc. ACM International Conference on Multimodal Interaction (ICMI)*. 85–94.

[15] Takatsugu Hirayama, Yasuyuki Sumi, Tatsuya Kawahara, and Takashi Matsuyama. 2011. Info-concierge: Proactive multi-modal interaction through mind probing. In *The Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2011)*.

[16] Mohammed Ehsan Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W Picard. 2013. Mach: My automated conversation coach. In *Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. ACM, 697–706.

[17] Koji Inoue, Divesh Lala, Katsuya Takanashi, and Tatsuya Kawahara. 2018. Latent character model for engagement recognition based on multimodal behaviors. In *Proc. International Workshop on Spoken Dialogue Systems (IWSDS)*.

[18] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels. In *Proc. International Conference on Machine Learning (ICML)*. 2304–2313.

[19] Shun Katada, Shogo Okada, Yuki Hirano, and Kazunori Komatani. 2020. Is She Truly Enjoying the Conversation?: Analysis of Physiological Signals toward Adaptive Dialogue Systems.. In *Proc. ACM International Conference on Multimodal Interaction (ICMI)*. 315–323. https://doi.org/10.1145/3382507.3418844

[20] Kazunori Komatani and Shogo Okada. 2021. Multimodal Human-Agent Dialogue Corpus with Annotations at Utterance and Dialogue Levels. In *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII)*.

[21] Kazunori Komatani, Shogo Okada, Haruto Nishimoto, Masahiro Araki, and Mikio Nakano. 2019. Multimodal Dialogue Data Collection and Analysis of Annotation Disagreement. In *Proc. International Workshop on Spoken Dialogue Systems (IWSDS)*.

[22] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 230–237.

[23] Shiro Kumano, Kazuhiro Otsuka, Dan Mikami, Masafumi Matsuda, and Junji Yamato. 2015. Analyzing Interpersonal Empathy via Collective Impressions. *IEEE Trans. on Affective Computing* 6, 4 (2015), 324–336.

[24] Zechao Li and Jinhui Tang. 2015. Weakly supervised deep metric learning for community-contributed image retrieval. *IEEE Trans. on Multimedia* 17, 11 (2015), 1989–1999.

[25] Reza Lotfian and Carlos Busso. 2019. Curriculum Learning for Speech Emotion Recognition From Crowdsourced Labels. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 27, 4 (2019), 815–826.

[26] Eran Malach and Shai Shalev-Shwartz. 2017. Decoupling "when to Update" from "How to Update". In *Proc. Conference on Neural Information Processing Systems (NeurIPS)*. 961–971.

[27] Yukiko I. Nakano and Ryo Ishii. 2010. Estimating User's Engagement from Eye-gaze Behaviors in Human-agent Conversations. In *Proc. ACM International Conference on Intelligent User Interfaces (IUI)*. 139–148.

[28] Setareh Nasihati Gilani, David Traum, Arcangelo Merla, Eugenia Hee, Zoey Walker, Barbara Manini, Grady Gallagher, and Laura-Ann Petitto. 2018. Multimodal Dialogue Management for Multiparty Interaction with Infants. In *Proc. ACM International Conference on Multimodal Interaction (ICMI)*. 5–13.

[29] Derya Ozkan and Louis-Philippe Morency. 2013. Latent Mixture of Discriminative Experts. *IEEE Trans. on Multimedia* 15, 2 (2013), 326–338.

[30] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. 2017. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[31] Björn Schuller, Stefan Steidl, and Anton Batliner. 2009. The INTERSPEECH 2009 emotion challenge. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*. 312–315.

[32] Hiroki Tanaka, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. 2017. Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. *PloS one* 12, 8 (2017), e0182151.

[33] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and vision computing* 27, 12 (2009), 1743–1759.

[34] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. 2019. Symmetric Cross Entropy for Robust Learning With Noisy Labels. In *The IEEE International Conference on Computer Vision (ICCV)*.

[35] Klaus Weber, Hannes Ritschel, Ilhan Aslan, Florian Lingenfelser, and Elisabeth André. 2018. How to Shape the Humor of a Robot - Social Behavior Adaptation Based on Reinforcement Learning. In *Proc. ACM International Conference on Multimodal Interaction (ICMI)*. 154–162.

[36] Qianli Xu, Liyuan Li, and Gang Wang. 2013. Designing Engagement-Aware Agents for Multiparty Conversations. In *Proc. ACM CHI Conference (CHI '13)*. 2233–2242.

[37] Hongliang Yu, Liangke Gui, Michael Madaio, Amy Ogan, Justine Cassell, and Louis-Philippe Morency. 2017. Temporally Selective Attention Model for Social and Affective State Recognition in Multimedia Content. In *Proc. ACM International Conference on Multimedia*. 1743–1751.

[38] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. 2019. How does Disagreement Help Generalization against Label Corruption?. In *Proc. International Conference on Machine Learning (ICML)*. 7164–7173.

[39] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1103–1114.

[40] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory Fusion Network for Multi-view Sequential Learning. In *Proc. Association for the Advancement of Artificial Intelligence (AAAI)*.

[41] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention Recurrent Network for Human Communication Comprehension. In *Proc. Association for the Advancement of Artificial Intelligence (AAAI)*.

[42] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. *CoRR* abs/1611.03530 (2016).

[43] Luming Zhang, Yue Gao, Yingjie Xia, Ke Lu, Jialie Shen, and Rongrong Ji. 2013. Representative discovery of structure cues for weakly-supervised image segmentation. *IEEE Trans. on Multimedia* 16, 2 (2013), 470–479.

[44] Zhilu Zhang and Mert R. Sabuncu. 2018. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. In *Proc. Conference on Neural Information Processing Systems (NeurIPS)*. 8792–8802.

[45] Zhi-Hua Zhou. 2017. A brief introduction to weakly supervised learning. *National Science Review* 5, 1 (08 2017), 44–53.