Exploring Semi-Supervised Learning for Predicting Listener Backchannels

Vidit Jain
vidit17370@iiitd.ac.in
IIIT-Delhi
New Delhi, India

Maitree Leekha
maitreeleekha_bt2k16@dtu.ac.in
Delhi Technological University
New Delhi, India

Rajiv Ratn Shah†‡
rajivratn@iiitd.ac.in
IIIT-Delhi
New Delhi, India

Jainendra Shukla†
jainendra@iiitd.ac.in
IIIT-Delhi
New Delhi, India

ABSTRACT
Developing human-like conversational agents is a prime area in HCI research and subsumes many tasks. Predicting listener backchannels is one such actively-researched task. While many studies have used different approaches for backchannel prediction, they all have depended on manual annotations for a large dataset. This is a bottleneck impacting the scalability of development. To this end, we propose using semi-supervised techniques to automate the process of identifying backchannels, thereby easing the annotation process. To analyze our identification module’s feasibility, we compared the backchannel prediction models trained on (a) manually-annotated and (b) semi-supervised labels. Quantitative analysis revealed that the proposed semi-supervised approach could attain 95% of the former’s performance. Our user-study findings revealed that almost 60% of the participants found the backchannel responses predicted by the proposed model more natural. Finally, we also analyzed the impact of personality on the type of backchannel signals and validated our findings in the user-study.

CCS CONCEPTS
• Human-centered computing → User studies: Empirical studies in HCI. • Computing methodologies → Semi-supervised learning settings.

KEYWORDS
Conversational Agents, Backchanneling, Multimodal analysis.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:
Vidit Jain, Maitree Leekha, Rajiv Ratn Shah, and Jainendra Shukla. 2021. Exploring Semi-Supervised Learning for Predicting Listener Backchannels. In CHI Conference on Human Factors in Computing Systems (CHI ’21), May 8–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3411764.3445449

1 INTRODUCTION & RELATION TO PRIOR WORK
Human conversations, even the most casual ones, have a lot of complexity associated with them. Two people actively engaged in a conversation frequently respond to each other, not only with respect to the content of the conversation, but also to behavioral aspects, such as facial expressions and prosody. Developing Embodied Conversational Agents (ECAs) [9] and spoken dialogue systems capable of incorporating these complex elements to converse naturally is a challenging task and has been a constant focus of the Artificial Intelligence and Human Computer Interaction research communities. Of these complex conversational constructs, dyadic components like listener backchannels are among the most crucial for modeling virtual humans and are also the main focus of this study1.

In a peer-to-peer conversation, a backchannel occurs when one of the participants is speaking, and the other (the listener) interjects a short response to the former [41]. These responses do not interrupt the flow of the conversation; rather, they convey the listener’s state-of-mind about the speaker’s dialog. They also reflect cooperation and understanding between the two parties [17]. Backchannels can be verbal, non-verbal (visual) or both. Vocalisations like ‘hmm’ or ‘uh-huh’, gestures such as head nods or head shakes, and a combination of verbal and non-verbal responses are common examples of backchannels.

In the past few years, the research community has shown a keen interest in modeling the listener’s backchanneling behavior. A large number of such studies on backchannel prediction have focused on the use of rule-based classifiers. Ward [38], Truong et al. [36], Ward and Tsukahara [39] utilized different acoustic features of the speaker such as pitch, pausal information, etc. to predict backchannel opportunities. A recent study by Park et al. [29] on backchannel

1A teaser video demonstrating our work can be found in the video figure section of the submission. It illustrates a virtual listener that emits backchannels to the speaker’s context using the models proposed with this work.
prediction for children also used similar prosodic features and hand-crafted rules. Expanding the feature set, Moubayed et al. [1] used both visual and prosodic speaker features.

Transitioning towards data-driven automatic prediction from hand-crafted rules, Solorio et al. [34] used prosodic features with locally weighted linear regression to predict backchannel opportunities. Morency et al. [27] used multimodal (visual and acoustic) speaker features and a hidden markov model to predict backchannels. More recently, many researchers also used deep learning techniques for predicting backchannels. Ruede et al. [32] used Long short-term memory (LSTM) based model with acoustic features. Hara et al. [16] also used LSTMs to predict turn-taking, filler words, and backchannels in a multitask learning paradigm. In a recent study, Goswami et al. [15] used state of the art machine learning and deep learning-based time-series classification techniques to model backchannel opportunities in children with multimodal features. Inspired by prior work, we too explore the use of machine learning and deep learning based time series models and include several similar multimodal features in our analysis.

All these studies, although well-founded, have the following limitations:

1. Most of them have focused only on the backchannel opportunity prediction task, with a very few involving the next step of the problem i.e., predicting the type of backchanneling signals.

2. All studies in the literature have relied on data with annotated backchannel instances during their modeling phase, where they develop a backchannel predictor using the speaker context. For instance, the datasets used by prior works, including the Switchboard Dialog Act Corpus (SwDA) [23], Iraqi Arabic [40], P2PStory [33], data collected by Morency et al. [27], and many others, had to be manually annotated for listener backchannels by multiple coders. This annotation process can be extremely time-consuming, depending on the amount of data present. Furthermore, this approach also does not scale well when trying to expand the scope of development by collecting more data. For instance, in many studies, including the ones cited above, the datasets used cover a relatively small population size. The size of the dataset used is primarily constrained by the amount of time it will take to annotate it, which further impacts the generalizability of the study. Similar challenges are faced when considering conversations from low-resource languages, which may be harder to annotate by virtue of their limited resources. This is one of the major issues we address in the present study i.e., can we automate the process of identifying when and how the listener backchannels from the data, thereby easing the annotation process?

Our novel contributions include:

1. Use of self-training based semi-supervision for labeling the instances in a dataset for the presence or absence of listener backchannels, and therefore, exploring computational techniques to guide the development of conversational agents. This identification model partly replaces the human annotator, and therefore uses the listener’s multimodal features to detect his/her backchannel responses. In addition to identifying backchannel opportunities, this step also identifies the type of signals (verbal, visual or both) associated with the backchannels. To analyze the feasibility of automating the labeling process via semi-supervision, we compare the backchannel prediction models trained on the labels assigned by the semi-supervised identification process, with the models trained on the ground truth labels (from the annotators).

2. Inspired by Bevacqua et al.’s work [4] on personality contingent listener backchannels, we statistically analyzed how people with varying personalities emit different types of backchannel responses. In particular, we study the impact of the extraversion trait of a subject on their preference of modality [5] for their backchannel response.

3. Finally, unlike most prior works, in addition to predicting the backchannel opportunities, we also predict the signal for the listener agent. The signal prediction task itself is way more challenging than the opportunity prediction task as the signals may vary significantly from person to person. We approach this task by first predicting the type of signal to emit (visual, verbal, or both). Then, based on our findings in (2) and depending on what personality we want our virtual listener to embody, we select the exact signal combination for it to express.

With this work, unlike most of the past studies that have focused on English datasets, we are also amongst the first to use peer-to-peer conversations in Hindi, which is a low-resource language, for listener backchannels\(^2\). In particular, we use the conversational dataset collected by Khan et al. [24], and annotate it for analyzing backchannels. Our quantitative and subjective evaluations reveal that:

1. By leveraging semi-supervision for identification of listener backchannels, we were able to detect the presence of backchannels \(\sim 90\%\) of the times, and the type of signals associated \(\sim 85\%\) of the times, with only a small subset (25\%) of labeled data.

2. Comparing the prediction models trained on the labels generated by the identification models, with those trained on manually-annotated labels: the former setting is able to reach \(\sim 93\%\) of the latter’s performance in case of opportunity prediction, and \(\sim 96\%\) for signal category prediction. Note that the cost performance is significantly less for the former as it needs only a small amount of labeled data, thereby substantially reducing the efforts required in annotating the data.

3. Subjective analysis in the form of a user study with twenty seven participants supported our quantitative observations. Approximately 75\% of the subjects found the backchannels produced by our proposed model more or equally natural to the responses by the model trained on annotated labels. The participants also confirmed our observations of personality impacting the preference of modality for backchannel responses.

\(^2\)However, this does not entail that the techniques we propose cannot be used for datasets from other languages.
The rest of this paper has been organized as follows: Section 2 discusses the dataset used in this work, the annotation process, and the initial data analysis. In Section 3, we formally describe the problem statements for the identification and prediction tasks and discuss how we model them. Additionally, this section also briefly discusses the features used for modeling these tasks. Section 4 begins by detailing the complete experimental setup and follows with a quantitative evaluation of all the tasks. The section also discusses the observations of a user study, performed to analyze the efficacy of our models in real-time. Finally, we conclude the paper with Section 6, which discusses the future scope of this work.

2 DATASET

In this work, we use Vyaktitv, a peer-to-peer Hindi conversations dataset, curated by Khan et al. [24]. The dataset provides audio and video recordings of participants involved in a dyadic conversation. There are a total of 25 conversations (50 individual recordings) with each one lasting 16 minutes and 6 seconds on average. A total of 38 subjects (24 Male, 14 Female) were a part of the dataset. It also provides the Big Five personality traits [10] for all the subjects. Note that, to the best of our knowledge, no work has used Hindi conversations so far for analysing listener backchannels.

2.1 Annotating Listener Backchannels

The audio and visual feeds for the individual speakers were annotated for verbal and visual activity based backchannels. Specifically, three annotators used the ELAN annotation software [8] with a custom tier template to mark the onset and offset time for different backchannel signals including nod, head-shake, mouth, eyebrow, and short-utterance (Table 1). The overall agreement amongst the annotators with respect to the presence of backchannels (i.e., if any backchannel activity was present in a particular time range) was near perfect with a Fleiss’ $\kappa$ of 0.86. For the individual signals ($\kappa$): substantial agreement was observed for nod (0.71), head-shake (0.64), and mouth (0.63), and moderate agreement in case of eyebrow (0.45) and short-utterance (0.49). Similar values were also reported by other dataset annotators, like [1, 33]. A primary reason for relatively low values of annotator agreement in such datasets is the subjective nature of backchannels.

A consensus strategy was adopted to combine the three sets of annotations from different coders. In particular, we assigned a positive label to an instance, provided at least two annotators agreed upon the existence of backchannel activity during that time frame. The feedback signals associated with these backchannel instances (like a nod, smile, etc.) were also decided in a similar fashion using consensus voting, where a particular feedback signal is included in the final annotation if at least two annotators observed that signal’s presence. Figure 1 depicts this strategy using an example where two annotators agree upon the existence of backchannel activity.

Note that the onset and offset time annotated by the three coders were bound to have slight variation even for the same instance. Therefore, to resolve this issue while combining the annotations, any two instances from different annotation sets (labeled by different coders) whose onset and offset time were less than 1 second apart, were considered to point to the same instance. The final value for the onset and offset time was taken as the average of the timestamps by different annotators.

2.2 Data Analysis

A total of 2781 backchannel instances were observed across all the participants in the dataset. In Table 1, we present a descriptive analysis for these instances in terms of the type of feedback signals. In particular, we record the total number (N), average frequency per participant (Mean Freq.), and the average duration (Mean Dur.) for each signal type. Note that a backchannel instance could have multiple feedback signals (eg., nod, and smile), and therefore, has been included in each of the possible signal categories. Observations from the analysis reflect how all the participants frequently used nods and short utterances. On the other hand, backchanneling via head-shakes, mouth, and eyebrow movements was not common. In particular, instances of backchanneling through eyebrow movements were scarce in the dataset (only 27), and therefore, we refrained from considering them in our predictive analysis. Furthermore, the average duration for all these backchanneling signals was around 1.8 seconds.

Figure 1: Sample depicting the consensus strategy adopted for combining the annotations from different coders.

Table 1: Backchannel Signals and their Descriptive Statistics. (after taking consensus) N represents the total number of a particular feedback signal observed across the complete dataset. Mean Freq. is the average number of times a participant emits a particular signal during a conversation. Mean Dur. is the average duration of the signals in seconds (s).
were amongst the top frequent signals emitted. Furthermore, most of the frequent multimodal instances included at least either a nod or an utterance. Multimodal signals where more than three types of signals co-occurred were infrequent.

Finally, we utilize the Big Five traits provided in the dataset to analyze if personality traits influenced the type of signals (multimodal or unimodal) emitted by the subjects. For our analysis, we defined the variable \( r \) as the ratio of the number of multimodal to unimodal backchannel instances emitted by a subject. Then, we used a two-sample Kolmogorov-Smirnov (K-S) test, with the null hypothesis that the distribution of \( r \) for the two categories of participants (based on a particular personality trait) was similar. The test presented an interesting insight in terms of the subjects’ preference for multimodal and unimodal feedbacks. Of the five traits, the tests suggested a significant difference in the distribution of \( r \) when considering the extraversion trait. The probability density distribution in Figure 2 (ii) shows how subjects with low extraversion scores (introverts) tended to have a lower value for \( r \), indicating their preference for unimodal signals, while extroverts used multimodal feedback more often. Quantitatively, the average probabilities of an extrovert emitting multimodal or unimodal backchannels were 0.51 and 0.49, respectively. The same values for an introvert were 0.35 and 0.65, respectively. This observation has been utilized later in this study while deploying the backchannel prediction models to a virtual-agent, thereby lending it a ‘personality’ based on which it can choose to emit multimodal or unimodal feedbacks in a probabilistic fashion (Section 4.3).

2.3 Negative Samples

Negative samples are the instances from the conversations where the listener did not emit any backchannel. Specifically, we evaluated the following two conditions while extracting such instances:

1. the listener is not speaking, and
2. the listener is not backchanneling in that time frame.

We used the Audacity tool\(^3\) to extract the voice-activity of the listener. Combining the voice-activity and the annotations, we extracted regions from the conversations that met the above two conditions. We sampled disjoint instances from these regions, where the length of each instance (in seconds) was taken as a random floating-point number in the range [1.06, 5.43]. This range also determined the lengths of the backchannel instances, and hence the choice. Using this approach, we sampled a total of 2670 negative instances.

3 METHODOLOGY

This section elaborates on the two-step methodology followed in the present study. We begin with the backchannel opportunity and signal identification module, which utilizes the listener’s features to automatically classify the instances into different categories based on the presence or absence of backchannel activity. This step employs a semi-supervised training paradigm and aims to simplify the manual annotation task. Next, we discuss the prediction module, which uses the speaker’s contextual features to predict these backchannels. The training step of this latter module makes use of the labels generated by the former semi-supervised step. Figure 3 summarises this workflow.

3.1 Semi-Supervision for Backchannel Opportunity and Signal Identification

Task Formulation. We begin this module by first formally defining the task of identifying backchannel opportunities and signals. Consider a time frame \( T_{ij} \), which starts at the \( i \)th second and ends at the \( j \)th second, and let \( L_{ij} \) represent the listener’s visual and acoustic features in that time frame. Then, our aim for the backchannel opportunity identification task is to learn a function \( \mathcal{F}_{bc} \) mapping a time series in the listener’s feature space to the corresponding backchannel opportunity label \( BO_{ij} \) (binary label signifying the presence or absence of backchannels in the time \( T_{ij} \)), i.e.,

\[
\mathcal{F}_{bc}(L_{ij}) \Rightarrow BO_{ij} \quad \text{(1)}
\]

We model (identify and predict) the backchannel signals differently from the literature. Instead of identifying the different signals (like nod, head-shake, etc.) individually, we categorize them into two types—visual (nod, mouth, head-shake) and verbal backchannels (utterances). This is done because the task of predicting the exact backchannel signal that a listener must emit based on the speaker’s context (i.e., the signal prediction task) is challenging, primarily because of the subjective nature of these signals. For instance, subject A may emit a smile to a particular speaker context, whereas another subject B may emit a nod in response to the same. Grouping signals together into visual and verbal categories simplifies the tasks at hand.

A backchannel instance could be associated with visual, verbal, or both kinds of signals. Therefore, the signal identification task aims at finding the type of signal the listener emits whenever s/he backchannels. The goal is to learn a mapping function \( \mathcal{F}_{sig} \) from the listener’s feature space to one of the three signal categories (verbal, visual, both) \( BS_{ij} \), i.e.,

\[
\mathcal{F}_{sig}(L_{ij}) \Rightarrow BS_{ij} \quad \text{(2)}
\]

\(^3\)https://www.audacityteam.org/
Exploring Semi-Supervised Learning for Predicting Listener Backchannels

Figure 3: Methodology: (i) Semi-supervised learning for identifying backchannels and type of signals emitted using a subset of labeled data. (ii) Learning to predict these instances and signals using the speaker’s context.

$$F_{\text{sig}}(L_{ij}) \mapsto BS_{ij}$$ (2)

It is important to note that both the identification tasks make use of only the listener’s features. This is a crucial distinction from the prediction module. The identification module identifies backchannel opportunities and signals much like a human annotator, by paying attention to the listener. On the other hand, as expected, the prediction module would use only the speaker’s contextual features to predict these backchannels.

**Modelling: Semi-Supervised Learning.** Annotating large datasets is a tedious task, and researchers have long been exploring ways to ease out and automate this manual process [2, 6, 19, 31]. Even in the context of the present study, the annotation process took around 90 hours, where the three annotators viewed and labeled all the conversations. This indeed is a bottleneck! To the best of our knowledge, this challenge in developing ECAs, specifically for modeling the backchannel behavior of an active listener, is an open research gap that has not been investigated by prior literature. The identification module of our workflow is a step towards tackling this challenge.

Several AI techniques based on Semi-Supervision [14, 21] and Weak-Supervision [11, 26] have been utilized to decrease the annotation costs in different application domains. Here we explore self-training based semi-supervised learning paradigm [37] for identifying listener backchannel instances and the signals associated based on the listener’s features.

(2) $C$ is then used to predict the labels for the unlabeled data ($\mathcal{U}$).

(3) Of these predictions, the instances which meet a specific selection criterion are removed from $\mathcal{U}$ and added, along with their predicted pseudo labels, to the training set. This updated training set comprising the initially labeled ($\mathcal{L}$) and the newly added pseudo labeled instances are used to train the classifier $C$ again.

(4) The cycle continues until no new instance matches the selection criteria.

In the present study, we have the labels for the complete dataset ($\mathcal{D}$). Therefore, in our context, $\mathcal{L}$ and $\mathcal{U}$ are disjoint subsets of $\mathcal{D}$. Furthermore, as selection criteria for Step-3, we use a high threshold value of 0.90 on the predicted class probability. Note that after Step-4, we use the trained classifier $C$ to predict the pseudo labels for all the remaining instances from $\mathcal{U}$. Our experiments revealed that such instances were very few in number.

### 3.2 Backchannel Opportunity and Signal Prediction

The formal definition for the backchannel opportunity prediction task is similar to the one proposed in prior literature [15]. Consider a time window $T_{ij}$, and let $S_{ij}$ be the speaker’s visual and acoustic features for that period. The backchannel opportunity prediction task entails predicting whether the listener would backchannel after $T_{ij}$ (i.e., the label $BO_{ij+}$) using only the speaker’s features (context). Similarly, the signal prediction task aims to predict the type of feedback signal (visual, verbal, or both) (i.e., $BS_{ij+}$) that will be emitted by the listener using the speaker’s context. The following function mappings represent these tasks:

$BO_{ij+}$ represents the backchannel opportunity after $T_{ij}$. $BS_{ij+}$ represents the type of feedback signal after $T_{ij}$.
We now discuss the experimental settings for our work. Since each widely used Label Spreading algorithm (we follow a more structured pattern to limit the number of settings window before each instance in our dataset and use the speaker’s visual and acoustic features for that time frame to predict listener backchannels.

The prediction tasks are performed in a supervised fashion, with the labels derived from the semi-supervision based identification module. Since the identification module used the listener’s channel (features), the prediction module is essentially learning based on cross-channel semi-supervised labels.

### 3.3 Feature Extraction

This section elucidates all the features extracted from the dataset for modeling the identification and prediction tasks. Table 2 summarises these visual and prosodic features. In particular, inspired by prior work [15, 22] we use OpenFace [3] to extract- 18 facial action units (FAUs), velocity and acceleration of gaze, translational and rotational head velocities and accelerations, blink rate, pupil location, and smile ratio. Additionally, we also find the gaze state as a categorical variable, taking up three values- left, right, or blinks. These features from the listener’s channel are used for backchannel opportunity, and signal identification tasks and the prediction tasks utilize the corresponding features from the speaker’s channel.

As prosodic features, in addition to the voice activity (discussed in Section 2.3), we also extract the fundamental frequency (F0), the energy, and the first 13 Mel-Frequency Cepstral Coefficients (MFCC) using the pyAudioAnalysis library [15]. All of them are used for the identification as well as the prediction tasks.

### 3.4 Experimental Setup

We now discuss the experimental settings for our work. Since each task can be analyzed and evaluated based on several dimensions, such as the type of features used (multimodal vs. unimodal), and the amount of initial ‘labeled’ data in semi-supervision (L), we could theoretically have had a combinatorial number of settings corresponding to all pairs of values for these variables. However, we follow a more structured pattern to limit the number of settings and draw focus on the most crucial elements.

#### 3.4.1 Backchannel Opportunity and Signal Identification

Separate models were trained for backchannel opportunity and signal identification tasks. We tried several different machine learning algorithms for training the classifier C in a semi-supervised paradigm described in Section 3.1. Specifically, we experimented with Random Forests (RF), Support Vector Machine Classifier (SVC), K-Nearest Neighbour Classifier (KNN), AdaBoost (ADA), and ResNet (ResNet). ResNet is the state-of-the-art deep learning model for time series classification [12]. Additionally, we also experimented with the widely used Label Spreading algorithm (LSpread, implemented as a part of the scikit-learn python library [30]) for semi-supervised learning as a baseline. All but for the ResNet model were trained using the mean and standard deviation aggregates of the time series based listener features. ResNet was trained using the detailed time series features for the complete time window. Finally, for the proportion (x) of the dataset D taken as the initial ‘labeled’ data (L) for training C, we experimented with all values in the range (5%, 100%) with a step size of 5%. Although semi-supervision is applicable only when the amount of unlabelled data exceeds the labeled, experimenting with the whole range helps in the analyzing the models’ sensitivity, i.e., how the performance changes when we approach the fully supervised setting (by increasing x).

Manifold evaluation for the identification tasks can be easily understood from Figure 4. First, we randomly create 5 folds from the data (D). Four of these folds are used for training C via semi-supervision; i.e., a random x% sample of the data from these four folds serve as the initial ‘labeled’ set (L), while the rest is termed as ‘unlabelled’ (U). Once the model has been trained, it is evaluated on the 5th fold, i.e., the pseudo labels produced by the identification models are compared against the ground truth labels. To ensure that our models do not over-fit to a particular random split, we run 10 simulations of training and evaluation for each pair of values of C and x, and report the average results across all the simulations. As evaluation metrics, we use the weighted average precision, recall, F1-score, and overall accuracy, for both the opportunity and signal identification tasks.

---

6The following GitHub repository was used for the same: https://github.com/antoinelame/GazeTracking

7The results for each simulation were taken as the average of all 5 folds.
models for signal prediction, including Random Forests (RF), Support Vector Machine Classifier (SVC), AdaBoost (ADA), K-Nearest Neighbors (KNN), and a Multi-Layered Perceptron model (MLP)\(^{11}\).

The next set of experiments analyze the feasibility and the performance to cost trade-off when adopting semi-supervision to label the dataset. For this, we compare the prediction tasks in two paradigms:

1. A complete supervised setting, using true labels provided by the annotators during the training and evaluation phases. Note that previous evaluations of the prediction tasks already produced the results for this setting.
2. The proposed setting using labels generated by the semi-supervised identification modules for training, while evaluating on the true labels. For this, we use only the best feature set (unimodal/multimodal) found via the previous evaluations, along with the pseudo labels produced by the best-found pair of \(C\) and \(x\) from the identification tasks.

With these two paradigms, we assessed how far semi-supervised backchannel identification fetches in comparison to human annotations for the downstream prediction tasks.

For evaluating the backchannel prediction models, several recent works have used the leave-one-subject-out approach [22]. In the present study, we use a slightly modified version of this technique to evaluate our prediction models in different paradigms. Instead of having the data from each subject comprise a test fold, we divide the subjects into 6 groups, and data from each of these groups form a test fold, i.e., leave-one-group-out. This modification was primarily done because some subjects had very few backchannel instances in their conversations. Combining instances from multiple subjects to form a test set helped in better analyzing the model predictions and keeping the number of folds tractable. As metrics for the backchannel opportunity prediction task, we report the positive class’ precision, recall, and F1-score, as well as the overall accuracy. For signal prediction, we report the weighted average precision, recall, and F1-score, the overall accuracy, and the confusion matrices.

4 RESULTS

4.1 Identification Tasks

In Figure 6, we have the sensitivity analysis in terms of accuracy of all the models we tried for the backchannel opportunity and signal identification tasks. Each curve in these plots represents how the performance of a particular classifier (\(C\)) for the identification task changes as we increase the amount of initial labeled data available (\(x\)) for semi-supervision. Using this information, we looked for the best possible pair of values for \(C\) and \(x\). However, in addition to achieving high performance, we also wanted to limit \(x\). For this, we manually found the \textit{elbow} point for the best performing classifier in each task. The elbow value for \(x\) is one where decreasing it further would cause a drastic drop in performance, while increasing it would not change the performance significantly. Ideally, we also wanted \(x\) to be less than 50\%, where we could have more unlabelled than labeled data for semi-supervision, thereby improving the annotation cost trade-off.

\(^{11}\)Please refer the supplementary material for details on the hyper-parameters used for all the different models in the identification and prediction tasks.
Fig. 6: Sensitivity Analysis using accuracy as the metric for semi-supervised backchannel opportunity and signal identification models. The shaded portion represents standard deviation.

Table 3: Backchannel opportunity and signal identification: Detailed results for $x = 25\%$.

Table 4: Backchannel opportunity prediction model trained trained across different sets of features using supervised (manually-annotated) labels for training. Overall, the multimodal feature set performed the best.

Table 5: Backchannel opportunity prediction: comparison of model trained using supervised manually-annotated labels with the one using labels generated by the identification module.

Table 6: Backchannel signal prediction: comparison of models trained using supervised manually-annotated labels, and those using labels generated by the semi-supervised signal identification Random Forest model (with 25% initial labeled data).

Table 7: Confusion matrices for the best backchannel signal prediction model (i) trained on manually-annotated labels, (ii) trained using labels generated by the signal identification models. Here, C1, C2, C3 refer to the ‘visual’, ‘verbal’, and ‘both’ class labels, respectively.

Since this set has overall the best performance for the baseline opportunity prediction task with supervised labels, we use the same to train and analyze models for subsequent tasks as well.

4.2 Prediction Tasks

We begin discussing the prediction models with Table 4, which records the results of our backchannel opportunity prediction task performed using the annotated labels for training, while experimenting with different subsets of input features. The unimodal-audio model, with positive class F1 of 0.74 and overall accuracy of 0.70, outperforms the unimodal-video model, with the corresponding metric values of 0.71 and 0.66, respectively. The multimodal model utilizing both the video and audio features beats the audio model by a small margin, attaining an F1 of 0.75 and 0.72 accuracy.
Moving on to backchannel signal prediction for determining the category of signals for the listener to emit, Table 6 shows the results obtained for all the models trained in different settings. When using the manually-annotated labels for training, we observe that the MLP model has the best accuracy of 0.87, however in terms of the F1-Score, the ADA model outperforms all the others. Given the data imbalance in the signal prediction task, we use F1-score as the deciding metric, and choose ADA as the best performing model, with an F1-score of 0.81. We observe a similar performance trend when using the semi-supervised signal identification model to generate the training labels. Here as well, ADA is the best performing model in terms of F1-score, with a value of 0.78, which is 96% of the corresponding metric for the supervised setting. In Table 7, we report the confusion matrices for the best (ADA) signal prediction models found for the two settings. We used the worst-case confusion matrix [25] computed using all the matrices generated across manifold evaluation of the model. The performance for the ‘visual’ (C1) class remained almost unchanged in the two settings. The false positives between the ‘visual’ and the ‘both’ (C3) classes reduced when using semi-supervised labels. The ‘verbal’ (C2) class performance dropped slightly. On the other hand, we performed better for the class C3, but that came at the cost of a slight increase in the false positives from C2. Overall, the performance of the model trained using semi-supervised labels was indeed comparable with the one trained on manually-annotated labels.

4.3 Subjective Evaluation: User Case Study

For problems like predicting listener backchannels, quantitative evaluations based on different metrics may not be sufficient to analyze the models’ efficacy. For such cases, qualitative assessments become extremely important. Many prior studies have used ECAs or robots to deploy their models and assess the backchannel predictions. However, in the present study, we follow a slightly different approach, which is partly inspired by [15] and [28]. Specifically, using Apple’s Memoji feature, we created two virtual avatars - Arjun (introvert) and Karan (extrovert). Note that the probabilities used here are calculated from the data and are mentioned in Section 2.2.

• Now, we choose the exact signals to emit from Table 8 using inverse transform sampling, where the probabilities for each signal combination are calculated from the data.

We recorded six response videos, three for each Arjun and Karan, using the following different prediction policies:

(1) Random Prediction Model (Random policy): A baseline model that predicted backchannel opportunities and the signal categories at random.

(2) Supervised Model (MA policy): The best opportunity (multimodal) and signal category (ADA) prediction models trained on the manually-annotated labels.

(3) Semi-supervised Model (SSL policy): The opportunity (multimodal) and signal category (ADA) prediction models trained with the labels generated from the semi-supervised identification models (ResNet and RF, respectively with 25% data) used in this work.

With these short clips recording the backchannel responses as predicted by different models for Arjun and Karan, we wanted to test the following three hypotheses:

[H1]: As perceived by a human watching the response videos, the backchannels predicted by our models look more natural, and therefore, better than those predicted by the random policy model.

[H2]: From a user’s perspective, there is no significant difference in the quality of backchannel responses emitted by models in the settings 2 and 3 above, i.e., using semi-supervised learning with a small subset of labelled data does not impact the quality of backchannel responses generated by the final prediction models.

[H3]: Finally, to assess the extent to which backchannel responses depend on personality, and the utility of using such personality contingent signal sampling: A human can judge the extraversion personality trait for Arjun and Karan based on their response videos.

Table 8: Signal combinations from different categories used in the present study. The corresponding normalized probabilities inferred from the data and used while sampling the signal combinations are also shown.

|          | Unimodal | Multimodal |
|----------|----------|------------|
| Visual   | nod (0.83), head-shake (0.09), smile (0.08) | nod + smile (0.60), head-shake + smile (0.40) |
| Verbal   | utter: top short utterances like okay, hmm, haan (yes), etc., were probabilistically sampled | nod+utter (0.80), smile+utter (0.13), head-shake+utter (0.05), head-shake+utter+smile (0.01), nod+utter+smile (0.01) |
| Both     | -        | -          |

**Note:** The only step that is contingent on personality is deciding on the unimodal vs. multimodal signals. The second step of signal sampling uses common probabilities for both characters.

22The names used are hypothetical, and do not compromise anonymity.

23Backchannel opportunity prediction was done at an interval of 3 seconds. Depending on the output, the signal categories were predicted. We also ensured that the models used for different speakers in the compilation were the ones where the speaker belonged to the test set.
of listener backchannels. They were instructed to watch all six response videos and carefully observe the virtual listeners’ backchannel responses in each one. It was a blind study, i.e. the participants knew neither about the prediction policies nor about the listeners’ personality traits in the videos.

Furthermore, to ensure that the appearance of the two avatars did not bias the participants’ judgement, we switched the avatars (with everything else as is) for half the participants. In other words, for half of the participants, Arjun acted as an extrovert, and vice versa. With this, we will be able to analyze whether backchannels indeed depend on the extraversion personality trait.

After watching all the videos, the participants were presented with a questionnaire meant to test the three hypotheses mentioned above. In particular, we asked the participants to:

- Rank the three policies in order based on the quality of backchannels, taking into consideration factors like the placement, frequency, and selection of backchannel signals.
- Rate each one on a Likert scale based on the above parameters.
- Identify the personality traits of the avatars based on the videos.

Post survey analysis of the participants’ responses lead to some exciting findings which supported our hypothesis:

1. **Models from both the manually-annotated and semi-supervised settings performed better than random:** 95% of the participants ranked our prediction models to be better than the random prediction policy. In addition, the Likert scale ratings suggested that our models performed significantly better than the random policy, both in terms of the placement & frequency, as well as the signal selection. We used the Wilcoxon signed-rank test, and the corresponding p-values are [P - placement & frequency, S - signal selection, * indicates significant difference at 95% confidence interval]:
   - **MA > Random**: 0.0019 (P*); 0.0013 (S*)
   - **SSL > Random**: 0.0001 (P*); 0.0001 (S*)

A box-plot for these ratings is shown in Figure 8. Notice how the boxes for MA and SSL do not overlap with Random. This also indicates that the former two certainly received better ratings than the later. This validates our first hypothesis [H1] that with a data-driven approach, we were able to emit more natural and human-like backchannels.

2. **Prediction models trained on semi-supervised labels produced more natural backchannels:** Comparing the ratings provided for the MA and the SSL policies, we found that in terms of the frequency and placement (P), there was no significant difference between the two (p-value=0.1024). For signal selection (S), the ratings suggested that SSL was significantly better than MA (p-value=0.0122). The box-plot in Figure 8 also aligns with these results. In case of Placement & Frequency, the boxes for MA and SSL overlap, with the median line for SSL slightly above MA's. This indicates that there is likely to be a difference between the two sets of ratings, even though it may not be significant (as found from the test). For Signal Selection, the boxes do not overlap at all, indicating a difference.

Furthermore, as a part of the questionnaire, we also asked the participants to compare the response videos generated by these two prediction models. Of the total, 60% of the participants found the backchannel responses emitted by the SSL model more natural than the MA model (SSL > MA), and 15% observed no perceptible difference between the two (SSL ~ MA). This also aligns with our quantitative prediction results where the proposed SSL model was able to reach ~ 95% of the latter’s performance. Thus, the two models were, both qualitatively and quantitatively, very similar. This confirms our second hypothesis [H2] with most of the participants finding the proposed model (SSL) similar (or more natural) to the MA model.

3. **Karan and Arjun’s extraversion traits were perceptible:** 80% of the participants were able to accurately identify which of the two virtual-human characters was introvert and extrovert. This confirms our third and final hypothesis [H3] that the type of backchannel signals emitted by an individual indeed depends on their personality.

5 DISCUSSION

Relation to prior work and some interesting findings: Our quantitative and qualitative evaluations in the previous section greatly validate how semi-supervision can be extremely useful in designing human-like ECAs, by focusing on the task of listener backchannel prediction. We found that with just 25% of manually annotated data (175 minutes), we were able to train a backchannel prediction system that performed comparably well, and even better
in terms of some parameters, as the one trained using 100% data. Most of the prior works, even the most recent ones, including [15, 16, 29, 32], have depended on a large amount of annotation.

We believe that this observation can significantly benefit the HCI community. Furthermore, studies have shown that backchannel responses vary greatly with culture [18, 42], and most of the prior studies have focussed primarily on the American and European population in this regard. In this work, we worked with subjects of Indian origin, and therefore, our work holds cultural significance as well.

A particularly interesting finding from Section 4.3 was that most participants found the backchannel responses generated by the SSL policy more ‘natural’ than the MA policy. We hypothesise this could be attributed to some form of label noise: After annotations, we only took those positive instances where at least two raters agreed, and the rest were discarded. Our negative samples overlapped with those instances where only 1 of the raters had annotated a positive BC. The SSL model could be learning to predict these backchannels. In fact, in our demo videos, we found two such instances. Starting with a small amount of seed data, the SSL model could have been learning these instances as well (given there are some hints of BC), which could explain the observation that even though SSL performed comparably with MA quantitatively, but seemed more natural to the participants in the subjective study.

**Limitations:** We would also like to highlight some limitations of our work, which can form the basis for future studies. First, we only used the participants’ visual and acoustic features and did not include the content of the conversation itself for predicting backchannels. The main reason for this was that the primary language used in the Vyaktitv dataset was Hindi. The English translations for the dialogues were not available, making the use of state-of-the-art NLP techniques non-trivial. In another aspect, our qualitative user case study involved a short 4-minute long compilation of speakers from the Vyaktitv dataset. A more rigorous analysis could follow by deploying the models and the avatars as a real time-system. Finally, we did not annotate eye blinks/gaze [20] as backchannel signals, primarily because they were not as apparent in the dataset. This could be a cultural difference as well.

**Ethical Consideration:** When developing data-driven systems leveraging data from human-subjects, it becomes imperative that we respect their privacy boundaries. We want to state that no Personally Identifiable Information (PII) was used while training or evaluating the system. We also complied with the agreements in the Vyaktitv dataset to ensure the safe use of data.

**6 CONCLUSION**

In this work, we confirmed the feasibility of using semi-supervised learning to (semi-) automate the process of identifying and labeling listener backchannel instances (both opportunities and the associated signals) from conversations. We used a Hindi peer-to-peer conversation-based multimodal dataset Vyaktitv, for our experiments. However, the methodology proposed in the study is general and can be adapted for other conversational datasets as well. Quantitative evaluation alongside a subjective analysis in the form of a user study strongly validated our hypothesis that prediction models trained using semi-supervised labels performed comparably with those using manually annotated labels. Furthermore, we statistically and qualitatively confirmed that the type of backchannel signals emitted are intimately linked to an individual’s personality (extraversion in particular).

Future work directions include validating the scope of semi-supervised learning for listener backchannel prediction on other datasets. Other parallel tasks, like listener disengagement prediction, can also be similarly performed. Methodologically, future studies could explore by devising heuristics and using weak supervision based techniques to identify backchannels. Our observations from analyzing the impact of personality on the type (modality) of backchannel response also open some new research questions; for instance- can we also analyze if the frequency of backchannels emitted by different individuals depend on personality? Furthermore, can we use these findings to embed personality into a virtual human further? These are some exciting lines future researchers can look into by conducting more extensive analysis.

**ACKNOWLEDGEMENTS**

Jainendra Shukla and Rajiv Ratn Shah are partly supported by the Infosys Center for AI and Center for Design and New Media at IIIT Delhi. In addition, Rajiv Ratn Shah is also partly supported by the ECRA Grant (ECR/2018/002776) by SERB, Government of India. Finally, we would like to thank the annotators, and all the participants who took part in our case study for their time and efforts.

**REFERENCES**

[1] Samer Al Mouhayed, M. Baklouti, Mohamed Chetouani, Thierry Dutoit, Ammar Mahdhaoui, Jean-Claude Martin, Stanislav Onďal, Catherine Pelachaud, Jerome Urbain, and M. Yilmaz. 2009. Generating Robot/Agent backchannels during a storytelling experiment. Proceedings – IEEE International Conference on Robotics and Automation ICRA’09 (06 2009), 3749 – 3754. https://doi.org/10.1109/ROBOT. 2009.5152572

[2] Stéphane Ayache and Georges Quénot. 2008. Video corpus annotation using active learning. In European Conference on Information Retrieval. Springer Berlin Heidelberg, Berlin, Heidelberg, 187–198.

[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) 1 (2018), 59–66.

[4] Elisabetta Bevacqua, Etienne De Sevin, Sylwia Julia Hyniewska, and Catherine Pelachaud. 2012. A listener model: introducing personality traits. Journal on Multimodal User Interfaces 6, 1-2 (2012), 27–38.

[5] Elisabetta Bevacqua, Maurizio Mancini, and Catherine Pelachaud. 2008. A listening agent exhibiting variable behaviour. In International Workshop on Intelligent Virtual Agents. Springer Berlin Heidelberg, Berlin, Heidelberg, 262–269.

[6] Patricia Bota, Joana Silva, Duarte Folgado, and Hugo Gamboa. 2019. A semi-automatic annotation approach for human activity recognition. Sensors 19, 3 (2019), 501.

[7] Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. 2011. SMOTE: Synthetic Minority Over-sampling Technique. CoRR abs/1106.1813 (2011), 321—357. arXiv:1106.1813 http://arxiv.org/abs/1106.1813

[8] Hennie Brugman, Albert Russel, and Xd Njmegen. 2004. Annotating Multimedia/Multi-modal Resources with ELAN. LREC Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04) (2004), 2065–2068.

[9] Justine Cassell, Joseph Sullivan, Elizabeth Churchill, and Scott Prevost. 2000. Embodied conversational agents. MIT press, Cambridge, MA, United States.

[10] John M. Digman. 1990. Personality structure: Emergence of the five-factor model. Annual review of psychology 41, 1 (1990), 417–440.

[11] Jared A Dunham, Alexander J Ratner, Khaled Saad, Nishith Khandwala, Matthew Markert, Hersh Sagriya, Roger Goldman, Christopher Lee-Messer, Matthew P Lungren, Daniel L Rubin, et al. 2020. Cross-modal data programming enables rapid medical machine learning. Patterns 1 (2020), 100019.

[12] John A. Dunnmon, Alexander J Ratner, Khaled Saad, Nishith Khandwala, Matthew Markert, Hersh Sagriya, Roger Goldman, Christopher Lee-Messer, Matthew P Lungren, Daniel L Rubin, et al. 2020. Cross-modal data programming enables rapid medical machine learning. Patterns 1 (2020), 100019.

[13] John A. Dunnmon, Alexander J Ratner, Khaled Saad, Nishith Khandwala, Matthew Markert, Hersh Sagriya, Roger Goldman, Christopher Lee-Messer, Matthew P Lungren, Daniel L Rubin, et al. 2020. Cross-modal data programming enables rapid medical machine learning. Patterns 1 (2020), 100019.

[14] John A. Dunnmon, Alexander J Ratner, Khaled Saad, Nishith Khandwala, Matthew Markert, Hersh Sagriya, Roger Goldman, Christopher Lee-Messer, Matthew P Lungren, Daniel L Rubin, et al. 2020. Cross-modal data programming enables rapid medical machine learning. Patterns 1 (2020), 100019.
