Artificial Intelligence for Human-Robot Interaction
AAAI Technical Report FS-17-01

Topic and Prosodic Modeling for Interruption Management in Multi-User Multitasking Communication Interactions

Nia Peters
Electrical & Computer Engineering
Carnegie Mellon University
Pittsburgh, PA - USA
nbradley@ece.cmu.edu

Bhiksha Raj
Language Technology Institute
Carnegie Mellon University
Pittsburgh, PA - USA
bhiksha@cs.cmu.edu

Griffin Romigh
Battlespace Acoustic Branch
Air Force Research Laboratory
Dayton, OH - USA
griffin.romigh@us.af.mil

Abstract

When to send system-mediated interruptions within collaborative multi-human-machine environments has been widely debated in the development of interruption management systems. Unfortunately, these studies do not address when to send interruptions in multi-user, multitasking scenarios or predictors of interruptibility within communication tasks. This paper addresses the issue of predicting interruptibility within these interactions with special attention to which users are engaged in which tasks or *task engagement* and where users are within a current task or *task structure* as predictors of interruptibility. Using natural human speech from these interactions, we attempt to model task engagement and task structure to predict candidate points of interruptions. The motivation for these models and their performance in a multi-user, multitasking environment are discussed as proposals in developing communication interruption management systems. To model task structure, a task breakpoint model is proposed which performs with a 90% accuracy within a multi-user, multitasking dataset. Integrating this task breakpoint model into a real-time interaction results in an average accuracy of 93% using the proposed task breakpoint model and a rule-based model. To determine the current task in which users are engaged or task engagement, a proposed task topic model performs with an accuracy between 76-88% depending on the topic within the dataset. Closely examining task structure and task engagement as predictors of interruptibility sheds new light on a rarely explored area for system-mediated interruption timings within multi-user, multitasking communication tasks.

Introduction

Within human-human-machine collaborative communication tasks, it is imperative that machines adhere to appropriate communication strategies that do not hinder the overall task goals. One form of communication within these tasks are *interruptions*. The work in interruption management system development primarily explores interruptions in human-machine tasks from the perspective of single-user, multitasking and multi-user, single task environments. A single-user, multi-task interaction is one in which one user is engaged in a primary task while interrupted with information relevant to a secondary task. A multi-user, single-task environment is one in which multiple users are engaged in a primary task and interrupted with information related to that primary task. Interruption management systems within these interactions leverage information from task interactions and apply rule-based or machine learning techniques to disseminate information at appropriate times. This research area is motivated by the reality that as users increasingly multitask among proactive systems, their tasks are being interrupted more often.

Appropriate interruption timings within multi-user, multitasking communication interactions is the primary focus of this paper. In contrast to previously mentioned interactions, a multi-user multitasking interaction is one in which multiple users are engaged in multiple unrelated tasks. An *interruption* within these interactions can be defined as an unanticipated request for task switching from a person, an object, or an event while multitasking (Arroyo and Selker 2011). Examples of multi-user, multitasking interactions include human-robot teams, air traffic control stations, unmanned aerial vehicle (UAV) operations, commercial and military pilots in cockpits, and human-computer technical support teams. Within these exchanges, humans are not only multitasking, but collaborating within a communication task as well. In multitasking environments humans are simultaneously working on one or more unrelated tasks. While collaborating, switching tasks could affect interdependencies with other teammates (human or machine). Providing awareness information to machine collaborators could be beneficial in helping align their tasks and interactions.

Though proactive delivery of information can benefit users, studies show that interrupting primary tasks can negatively impact productivity (Bailey and Konstan 2006; Czerwinkski, Cutrell, and Horvitz 2000; Monk, Boehm-Davis, and Trafton 2002; Cutrell and Horvitz, 2000) and affective state (Adamczyk and Bailey 2004; Zijlstra et al. 1999). Within these contexts there have been proposed methods of intelligent system-mediated interruptions.

There is empirical research dedicated to manipulating time on the delivery (Bailey and Konstan, 2006; Czerwinkski, Cutrell, and Horvitz 2000; Monk, Boehm-Davis, and Trafton 2002) of system-mediated interruptions (McCrickard, Chewar, Somervell, and Ndiwalana 2003) in multi-task environments (McFarlane and Latorella 2002). There is also literature that explores immediate interruption dissemination (Czerwinski, Cutrell, and Horvit, 2000; Dabbish and Kraut 2004; Latorella,1996) within dual-task scenarios. Studies have shown that delivering interruptions at random times can result in a decline in performance on primary tasks (Bailey and Konstan 2006; Czerwinski, Cutrell, and Horvitz 2000; Kreifeldt and McCarthy 1981; Latorella 1996; Robinstein, Meyer, and Evans 2001). Other studies show similar results (Altmann and Trafton 2004; Czerwinski, Cutrell, and Horvitz 2000; McFarlane D. A., 1999; Zijlstra et al.1999) and the differences in cost of interruptions are typically attributed to differences in workload at the point of interruption (Bailey and Konstan, 2006). Additionally, studies have illustrated that interrupting users engaged in tasks has a considerable negative impact on task completion time (Cutrell, Czerwinski, and Horvitz, 2001; Czerwinski, Cutrell, and Horvitz 2000; Czerwinkski, Cutrell, and Horvitz 2000; Kreifeldt and McCarthy 1981; McFarlane D. C. 1997; Bailey and Iqbal 2008). Studies in (Peters, Romigh, Raj, and Bradley, 2017) explores the benefits of providing a form of intelligent interruption dissemination within multi-user, multitasking interactions in terms of task performance across multiple tasks. Although there has been considerable research in developing systems that use intelligent methods to disseminate interruptions in multitasking environments, these studies do not address when to send system interruptions in multi-user, multitasking scenarios and do not address predictors of interruptibility within communication tasks for these interactions. The aim of this work is to augment this area of research by investigating appropriate interruption timings and their effect within multi-user, multitasking communication environments.

Issues in developing interruption management systems for multi-user, multitasking interactions include 1) extracting and modeling acoustic and speech information from noisy communication channels innate to these interactions 2) modeling the turn-tasking variability of different teams even for the same or similar teaming activities 3) Measuring the effect of these interruptions on team performance and separating interruption effect from other factors that influence team performance. The long-term goal is to address various issues associated with the development of an interruption management system for multi-user, multitasking communication tasks. The scope of this paper is focused on the exploration of predicting appropriate interruptions times using speech and acoustic information which could potentially provide some insight into modeling turn-taking variability within these interactions to predict interrupability and make optimal interruption timing decisions.

One proposed method in determining appropriate interruption timings is via task structure or more specifically task breakpoints. Task breakpoint modeling has been used in single-user, multitasking interruption management systems to indicate appropriate points of interruptibility (Adamczyk and Bailey 2004; Bailey and Konstan 2006; Czerwinkski, Cutrell, and Horvitz 2000; Isbal and Bailey 2006) and shown that deferring delivery of notifications until a breakpoint is reached can meaningfully reduce costs of interruptions. Conversely, interrupting tasks at random moments can cause users to take up to 30% longer to resume the tasks, commit up to twice the errors, and experience up to twice the negative affect than when interrupted at boundaries (Adamczyk and Bailey 2004; Bailey and Konstan 2006; Isbal and Bailey 2005). Within single-user, multitasking interruption systems the primary modality used to model task breakpoints is system-state information (Adamczyk and Bailey 2004; Bailey and Konstan 2006; Czerwinkski, Cutrell, and Horvitz 2000; Isbal and Bailey 2006). Using task breakpoints as candidate points of interruptibility within single-user multitasking interactions is the primary motivation for similar modeling techniques in multi-user, multitasking interactions. The primary difference in our approach is the modality used to build the task boundary models. Since the interruption management system in these interactions is making decisions within a collaborative communication task, we explore the use of acoustic and speech information to predict the presence or absence of task breakpoints.

Another proposed interruption timing strategy is based on task engagement via task topic modeling. For multi-user, single task interactions, task engagement has been explored as a useful predictor of interruptibility (Fogarty, Ko, and Anug 2005; Fogarty, Hudson, and Lai 2004; Horvitz and Apacible 2003) via multimodal cues to predict users' engagement in the current task and make inferences on appropriate interruption moments within the task. Within these interactions, the interruption information is primarily related to the current task which differs in this work where the information could be related or unrelated. The motivation to use a task engagement such as those proposed in multi-user, single task interactions for our purposes is that if the system has information related to the current topic in which users are engaged, this could augment the priority of the interruption timings. In contrast, if the information the system has is unrelated to the topic, other interruption strategies such as deferring until the end of the current task may be more appropriate. Finally, if the system has information for a user that is currently "disengaged", this could provide an optimal opportunity to interrupt. Since the task engagement models are being integrated into collaborative communication interactions and the primary input stream in natural human speech, we pro-

pose a task engagement model via speaker, addressee, and topic modeling to determine who is speaking (speaker), who they are speaking to (addressee), and what they are speaking about (task topic). Task engagement results indicate that within a multi-user multi-tasking interaction, standard topic modeling techniques using automatic speech recognition (ASR) transcriptions have potential to be useful in determining task topics in these interactions.

In evaluating the performance of the task breakpoint and task engagement model in a dual-user, dual-task interaction dataset, the proposed models are not only promising in predicting interrupability within multi-user, multitasking scenarios, but also seem to be suitable for real-time systems based on the low cost of information processing via the raw audio stream and ASR transcriptions. Closely examining task boundaries and task engagement as predictors of interruptibility sheds new light on a rarely explored area of predictors of interruptibility for system-mediated interruption timings within multi-user, multitasking communication tasks.

Multi-user Multitasking Interaction

Prior to developing an interruption management system for multi-user multitasking interactions, a simulation of such an interaction is necessary to test the proposed models. For simplicity, a dual-user, dual-task interaction is simulated as illustrated in Figure 1 which shows two users not only engaged in a primary human-human task, but simultaneously in an orthogonal human-machine task where the primary information stream is speech, hence a dual-user, dual-task communication scenario.

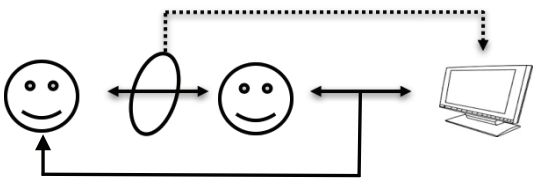

Figure 1: Dual-user, dual-task interaction

The proposed interaction for our experiments is a simulation of unmanned aerial vehicle (UAV) operators and ground troop teammates collaborating on a target identification alignment task. In such collaborations, the UAV operator and ground teammates have the same information from two different perspectives and are tasked with communicating over a push-to-talk communication network to align their knowledge and perspective, and confirm they are talking about the same thing. In most military missions, troops are multitasking and some of their tasks can be offloaded to machine teammates, but as proactive systems are integrated into the entire exchange, the human tasks are being interrupted more often. The objective of this data collection is to simulate a dual-task comprised of a primary human-human task like the alignment task previously described in conjunction with a secondary human-machine task. The machine must listen to the human interaction and make decisions on when to interject information related to a secondary task such that it is least disruptive to the overall exchange.

The primary human-human task or Tangram task involves two users corresponding over a push-to-talk communication network to arrange abstract shapes (Tangrams) within a column into corresponding order to simulate aligning knowledge from two perspectives. To introduce more complexity into the primary task, the set of abstract objects generated for each trial are similar in appearance. For instance, objects that looked like humans were generated together to avoid dialogue exchanges such as, "dog, person, boat, square." The five categories that the abstract shapes can be extracted from include: birds, people, miscellaneous, boats/miscellaneous, and animals.

Figure 2 is an example graphical user interface (GUI) of the primary task for both participants. Figure 3 is a corresponding sample dialogue of the participants describing the shapes to one another.

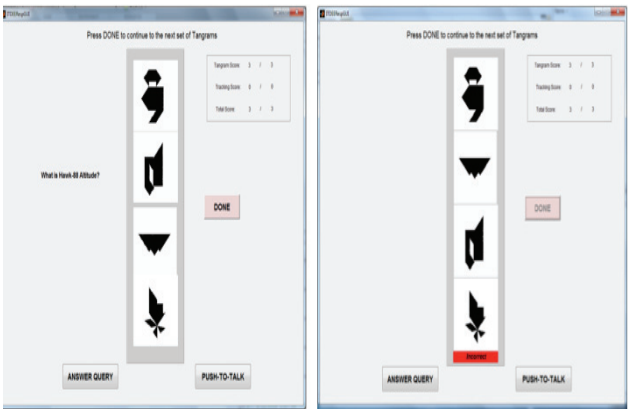

Figure 2: Multi-user, multitasking Interface (L: participant 1, R: participant 2)

Player 1: Seventy-percent. Ok I gotta sideways teapot, sideways iron, upside down mountain, and upside down head **(A1)**

Player 2: Got it **(A2)**

***Both users press the Done button to move on to the next set of shapes*

Player 1: Aw man, I thought we had it. Ok I have a uh a fox I have like a dog running to the left then I have a um ram I think it's been called and then I have this weird animal that's on one leg **(A3)**

Figure 3: Sample dialogue within Multi-user, Multi-tasking Communication Interaction

In Figure 3, for A1 Player 1 begins by rehearsing an interruption (UAV state update) from the secondary task then proceeds to describing the current columns (

Figure *2*). In A2, Player 2 confirms that he/she understands the shapes Player 1 described. Both users press DONE then another set of abstract shapes is generated for the players to describe. In A3 Player 1 makes a reference to the feedback that potentially illustrates the player's inability to correctly arrange the Tangram shapes in the corresponding order. Player 1 then proceeds to describing the next column of shapes. The dialogue within the primary task seems to follow the theory of *Conversational Games* where conversations consist of a series of GAMES, each of which involves an INITIATING MOVE (such as an instruction or query), followed by either a RESPONSE MOVE (such as acknowledgment or reply) or possibly an embedded game (e.g. a query may be followed by a clarification sub dialogue) (Carletta et al. 2007).

The secondary human-machine task involves the same two users simultaneously receiving interruptions or synthesized audio of status updates and queries related to varying UAV states. The interruption timing decisions are based on modeling the audio stream of the primary human-human task and predicting the time to send UAV updates or related queries. Both users receive 3 -5 updates about various UAV states before being queried about the current state of a UAV previously presented. For example, a user is only required to keep track of 3-5 UAV states prior to a query. After a user receives a query and responds, a different set of 3-5 UAV states is presented. Below is an example of an Update/Query block:

{Update I}: Hawk-88' LOCATION is Point Bravo
{Update II}: Raven-3's FUEL-LEVEL is 30%
{Update III}: Falcom-11's ALTITUDE is 1900 ft.
{Query}: What is Raven-3's current FUEL-LEVEL?

Once a pre-specified number of UAV queries is sent to both users, the experiment is over. For both tasks, a correct response results in a point and an incorrect response results in a deduction. The total points for the primary and secondary task are summed as a total team score. A detailed description of the entire experimental design and corpus is described in (Peters et al. 2017). The aim of this data collection is to best simulate an alignment task where users may have differing vocabulary for the same object and must take turns to align their knowledge while simultaneously simulating a human-machine task that monitors the interaction of the primary task to make interruption timings decisions and send information related to a human-machine task.

Task Breakpoint Model

A preliminary proposition for predicting interruption timings is via task breakpoint detection. Since these decisions are being made within a collaborative communication task, it is necessary for the system to leverage acoustic and speech information to predict the presence or absence of a task breakpoint. Within the literature there has been exploration of the use of prosody to detect boundaries in sentences, discourse structure, and grounding (Mushin et al. 1999; Mixdroff 2004; Syrdal and Kim 2008; Hasti, Poesio, and Isard 2002) and we aim to use similar techniques in using prosody to detect task breakpoints. Here the intended task breakpoint model leverages prosodic content of push-to-talk utterances and predicts whether an utterance will be followed by a task breakpoint.

From this interaction, the interrupted task is the primary human-human task. A task breakpoint is defined as a timestamp associated with both users pressing the DONE button indicating they are done with the current abstract shape or Tangram column and proceeding to the next. Utterances preceding these points are labeled as breakpoints. A non-breakpoint is an utterance that does not precede such points. The entire dataset contains 6590 potential breakpoint candidates with a 60-40 distribution of 3875 non-breakpoints and 2715 task breakpoints.

The use of only prosodic information to infer task breakpoints is an exploratory measure of how well one can do in predicting task breakpoints using only derived features from the raw audio. Making predictions from the raw information within a communication channel has potential for quick data processing and modeling, but may hinder detection accuracies. Several binary classifiers that use supervised learning techniques can be used to solve this problem. Here we compare three as potential candidate models to integrate into a real-time system: Naïve Bayes (Russel and Norvig 1995) , Support Vector Machines (Cortes and Vapnik 1995), and Random Forests (Ho 1995). Although more sophisticated models may be employed, these simple classifiers have previously been found to be effective in such scenarios, and are particularly well suited to real-time implementation. The predictors in this classification problem are prosody features extracted

as a 989-dimensional feature vector from the emotion detection feature set in OpenSmile (Eyben, Weninger, and Schuller 2013). These features are derived from the raw audio of the push-to-talk utterance preceding candidate breakpoints. These features are a composition of the utterance signal energy, loudness, Mel-/Bark-/Octave-spectra, MFCCs, PLPs, Pitch, voice quality (jitter, shimmer), formants, LPCs, Linear Spectral Pairs (LSPs), and spectral shape descriptors. Additionally, statistical functions or feature summaries are included in the feature set: means/extremes, moments, segments, samples, peaks, linear and quadratic regressions, percentiles, durations, onsets, DCT coefficients, and zero-crossing. Each utterance is sampled at 32K and the features are extracted from the audio partitioned in 25ms windows with 10ms in overlap, common in audio processing and modeling.

We hypothesize that classification techniques that model predictor dependencies will perform better at predicting task breakpoints like other boundary detection work (Swerts 1997; Shriberg, Stolcke, and Hakkani-Tur 2000) using prosody which illustrates the combination of pitch, energy, and their contours are useful at sentence and discourse boundary detection. Additionally, we hypothesize that utterances preceding breakpoints will be shorter in duration and lower in energy corresponding to confirmation utterances such as "done," "got it," "ok," "finished," "copy that". These utterances can be confused with backchannels, shorter descriptions of shapes, or turn-tasking confirmations in the middle of a task. With this confusability, there is the potential for more false positives.

The task breakpoint model performance is explored from two different perspectives, how well the model performs on the dataset (offline) and how well the model performs when integrated into a real-time system (real-time). The performance of the *offline* model is based solely on the data from the data collection. Conversely the *real-time* model performance is derived from integrating the task breakpoint model into the interaction described in the Multi-user Multitasking Interaction section and assessing how well the model performs within the interaction at predicting breakpoints. Overall the objective is to see how well a task breakpoint detection algorithm can perform via a prosody-only model by validating the model offline and finally integrating the model into a live system and evaluating its performance.

Offline Model

Prior to testing the system on the real-time dual-user, dual-task interaction, the model was first evaluated offline. From the data and task breakpoint model described in the Task Breakpoint Model section, Weka (Frank, Hall, and Witten 2016) is used to discriminate task breakpoint and non-breakpoints using prosodic features of utterances preceding these points. A 10-fold cross-validation method is used to generalize each model. The performance of the classifier is evaluated using the metrics precision and recall evaluating how well the system correctly classifies a task breakpoint with respect to all classified task breakpoints (precision) and with respect to all the present task breakpoint within an interaction (recall). Since the classes are relatively balanced (60-40), accuracy is also used as an evaluation criterion to see how well the overall output predictions are correctly classified.

Table 1: Offline Task Breakpoint Modeling Results

	Precision	*Recall*	*Accuracy*
Naïve Bayes	0.88	0.86	0.87
SVM	0.91	0.90	0.90
Random Forest	***0.91***	***0.91***	***0.91***

Table 1 illustrates the final results in task breakpoint detection. From these results, we can infer that overall the prosody-only task breakpoint model performs with accuracies between 87-91%. Although there is not a large difference between the classifier performances, we select Random Forest as the model to integrate into the real-time system since it has that highest precision, recall, and accuracy values.

Real-time Model

We integrate the task breakpoint model into the interaction described in the Multi-user Multitasking Interaction section and evaluate its performance where the interruption timing decisions (when to send an update or query) are based on the detection of a task breakpoint by leveraging acoustic information from the primary Tangram task. The system extracts the push-to-talk utterances from the primary task, generates a 989-dimentional feature vector from OpenSmile (Eyben, Weninger, and Schuller 2013), and makes decisions on the probability of a task breakpoint using a Random Forest (Banerjee 2016) model presented in the Offline Model section. The operation threshold of the classifier is 0.50 so a classification output probability of greater or equal to 0.50 is a task breakpoint, otherwise a non-breakpoint. The only difference between the current interaction and the one described in the Multi-user Multitasking Interaction section is only 10 original participants in the original data collection are retained and 5 additional participants are added for the current experiment. Random teams are generated from this pool of 15 participants.

From this data collection, there are a total of 2149 candidate (1285 non-breakpoints and 864 breakpoints) breakpoint decisions. We first evaluated the performance of the prosody-only offline model, then augment the system with a rule-based duration cue to evaluate if this increases overall performance. The augmented task boundary detection and rule based model states that if an utterance is less than

3 seconds or a task boundary is detected, send an interruption otherwise don't send an interruption.

Table 2: Online Task Breakpoint Modeling Results

	Precision	*Recall*	*Accuracy*
Prosody-Only	0.74	0.81	0.83
Pros-Duration	***0.96***	***0.86***	***0.92***

Table 2 illustrates the system's performance using the same metrics as the offline model for comparison. Accuracy is also considered because the classes balance in the overall interaction.

Discussion

From the results in

Table 1 one can surmise that task breakpoint detection via prosodic cues could be useful in detecting task breakpoints in a simulated dual-task, dual-user interaction. This could be illustrative of our previous hypothesis suggesting that one discriminating factor in breakpoint and non-breakpoint detection is related to duration or the amount of energy in the raw audio signal. In analyzing the important detection features in the Random Forest algorithm, the l*oudness_pcm_minPos* feature is a key indicator in discriminating these classes where 61.3% of classification decisions were based on this cue. This does not necessarily mean that how loudness is perceived is a predictor of task breakpoints, but could suggest that information correlated with energy and potentially duration could be useful in predicting task breakpoints. This suggestion is corroborated by the results from Table 2 that show that augmenting the prosody-only model with duration rules results in a 10.84% improvement in overall accuracy.

From these results duration and energy based cues seem to be potential indicators of predicting task breakpoints in multi-user, multitasking interactions, but if the interaction changes slightly and there is more variability in turn-taking behavior, are these still reliable predictors? We cannot draw any conclusions from this since the preliminary models are tested on a single interaction, but future work will provide an opportunity to test these prosodic information streams on similar interactions.

Task Engagement via Topic Modeling

A second proposed method of predicting interruptibility within multi-user, multitasking interactions is via task engagement. The motivation behind topic modeling as an indicator of task engagement is the assumption that appropriate interruption timings can be made if the system knows which speakers are engaged in which tasks. Within a communication task, engagement can be defined by who is speaking (speaker), to whom (addressee), and about what (task topic). Assuming the speaker is known via the communication channel, one objective of inferring task engagement is to identify the addressee and topic. For complex interactions, it may be useful to know who is talking to whom about what. For example, if the interruption management system can infer speaker 1 is speaking to speaker 2 about a topic, it can assume that speaker 1 and 2 are engaged in the task and a) subsequently offer information relevant to this task topic or b.) provide information to another participant who is not engaged. Prior to developing such a model for more complex interactions, we are interested in evaluating the inference of topic modeling performance in a simulated dual-user, dual task interaction where the speaker and addressee are already known so we need only infer the task topic.

Latent Dirichlet Allocation

We use Latent Dirichlet Algorithm (LDA) (Blei, Ng, and Jordan 2003) which is an unsupervised generative statistical model for a collection of discrete data. It aims to describe how a set of observations are explained by unobserved groups (latent variables) to capture the similarities between observed data. For topic modeling via words LDA assumes a document is a mixture of topics and that each word in the document is attributed to one of the document's topics.

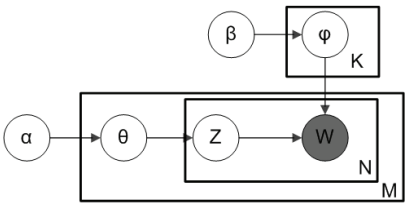

Figure 4: Plate notation for LDA with Dirichlet-distribution topic-word distributions

The graphical representation of the LDA model is illustrated in Figure 4, capturing the dependencies among the variables:

$\boldsymbol{\alpha}$ − *parameter of the Dirichlet prior on the per-document topic distributions*
$\boldsymbol{\beta}$ − *parameter of the Dirichlet prior on the per-topic word distribution*
$\boldsymbol{\theta_m}$ − topic distribution for document m,
$\boldsymbol{\phi_k}$ − word distribution for topic k,
$\boldsymbol{z_{mn}}$ − the n-th word in document m
$\boldsymbol{w_{mn}}$ − specific word.

In Figure 4 the outer plate represents the document and the inner plate represents the repeated choice of topic and words within the document. The nodes are the variables where the shaded nodes are the observed variables and the

unshaded are the latent variables. The total probability of the model is:

$$P(W, Z, \theta, \phi; \alpha, \beta) = \prod_{i=1}^{K} P(\phi_i; \beta) \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{t=1}^{N} P(Z_{j,t}; \theta_j) P(W_{j,t} | \phi_{Z_{j,t}})$$

Learning the various distributions for the set of topics, their associated word probabilities, the topic of each word, and the topic mixture of each document is a problem of Bayesian inference. In this paper, a document and utterance are synonymous and the LDA is used to uncover the hidden thematic structure of each utterance. The inference technique used to infer the posterior distribution in this paper is Gibbs sampling although other techniques may also be employed (Blei, Ng, and Jordan 2003).

Data and Method

To model task topics, the push-to-talk utterances are extracted from the interaction described in the Multi-user Multitasking Interaction section. From these utterances, automatic speech recognition (ASR) transcriptions via Pocketsphinx (Huggins-Daines et al. 2006) are used to model the task topics using Mallet (McCallum 2002) which implements the LDA algorithm (Blei, Ng, and Jordan 2003). The output from the LDA model is a 5-dimensional feature vector which represents the probability that the utterances belongs to one of the 5 categories: people, miscellaneous/boats, boats, birds, and animals.

In

Figure *2*, a user must describe the illustrated four Tangram shapes. These shapes are usually similar in appearance to add complexity and uncertainty into the interaction for richer dialogue. Since the images are similar in appearance and carry some topic based information, the labels of the utterances are based on the group the images are associated with. For example, users may be describing a column of birds which was generated from a subset of images illustrated in

Figure 5.

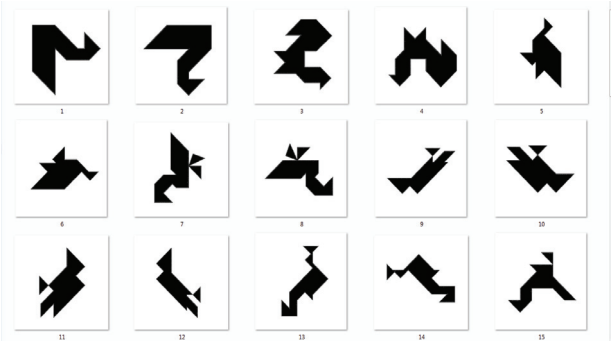

Figure 5: Subset of images related to bird category

If the users are discussing a column of images generated from the bird category, all the utterances associated with this exchange are labeled as "bird." The issue with this labeling process is that there are several utterances within the exchange that do not contain topical information. For instance, backchannels and confirmations like "ok" or "got it" have no topical information so we compare two different topic modeling strategies:

1) Use all the utterances in the dataset and the labeling process described above (*All*)
2) Only use utterances that are longer than 3 seconds in duration. (*Partial*)

Since the task breakpoint model considers utterances that are less than 3 seconds -- potential indicators in inferring interruptibility via task breakpoint detection -- topic modeling may be useful for utterances longer than 3 seconds. The feature vector output from the LDA model and corresponding label are modeled using a Naïve Bayes (Russel and Norvig 1995) classifier, generalized using a 10-fold cross-validation and evaluated using the accuracy metric area under the curve (AUC) which illustrates the trade-off between the true positive rate and false positive rate at various detection thresholds. Table 3 illustrates the performance of the topic model of the two datasets *All (N = 5499)* and *Partial (N=3412)* with a uniform distribution across categories.

Table 3: Topic Model Results

	people	birds	animal	boats	misc.
All	0.78	0.76	0.72	0.70	0.68
Partial	***0.88***	***0.88***	***0.83***	***0.79***	***0.76***

There are a few takeaways from these results. The first clear observation is that the AUC results from the *Partial* dataset are larger than those of the *All* dataset. This could be attributed to shorter utterances (less that 3 seconds) not having as much topic information because they are more associated with backchannels and confirmation utterances. As illustrated in the sample dialogue in *Figure 3*, A1 and A3 are longer in duration and contain words that could better illustrate which set of shapes users are describing. Conversely shorter utterances such as "got it" do now contain such information.

Secondly although AUC results for the *Partial* dataset range between 76-88%. The categories with larger AUC results (people, bird, and animals) may have words that are more indicative of the categories they are describing since people, birds, and animals are actual categories. The boat category is mainly boats, but also contains other abstract shapes and the miscellaneous category has shapes that didn't fit a category so the vocabulary may not have as

much topic information. Below are the top 10 words the topic model selected for each category:

People: Person, man, guy, triangle, upside-down, we're, running, lady, body, sideways

Birds: Monster, upside-down, crab, turkey, big, ness, pat, raptor, humming bird, back

Animal: it's, dog, left, square, top, bottom, head, cat, thing, side

Misc./Boat: sailboat, house, tree, sideways, boat, speedboat, arrow, christmas, apple, i've

Misc.: yeah, i'm, good, trial, honor, added, thing, ship, shape, there's

Finally, the topic modeling results are promising especially because they are based on ASR transcriptions which affords error propagation through the system. The baseline word error rate (WER) for Pocketsphinx in (Huggins-Daines, et al., 2006) is 9.73% and without fine tuning any of the acoustic or language models, we assume that this error is propagated into the LDA model. ASR transcriptions are tested here because to develop a real-time system, we won't have access to hand-transcribed data when making the actual topic classification decisions.

Overall task topic modeling using LDA seems to be a good starting point in predicting interruptibility via task engagement. In comparison to the task breakpoint model, the topic model was only evaluated on offline data since topic modeling only makes sense for more complex interactions where one could infer the addressee from topic predictions and make inferences on who is engaged in which tasks.

Conclusion

In conclusion, two potential models were proposed as predictors of interruptibility within multi-user, multitasking communication interactions: task breakpoint and task engagement. The results from the task breakpoint models give some indication that utterance energy and duration may be good predictors of a task breakpoints because these utterances could be characteristics of confirmations or knowledge aligning indicators that a user is ready to continue to the task. Although task breakpoint confirmation utterances can be confused with backchannels and mid-task confirmations, experimentation of this task breakpoint model in other interactions will give us a better illustration of how useful these prosodic and duration cues can be at detecting task breakpoints and, in turn, interruptibility.

Additionally, using ASR transcriptions, we illustrated a preliminary strategy for inferring task topics within a simplified multi-user, multitasking interaction. Even though ASR transcriptions can result in error propagation through the system, the results of higher AUC results being attributed to categories with more defined topics provides a promising first pass at the use of topic modeling as an indicator of task engagement. Both show promise for an efficient real-time system.

The combination of both models could be useful in predicting interruptibility within multi-user multitasking interactions by using shorter duration utterances as indicators of where a potential breakpoint could be and longer utterances as an indicator of which participants are engaged in which tasks. Overall this work offers a first pass at simulating an interaction that has been rarely explored in the past and experimenting with two modeling schemes that have promising potential of predicting interruptibility. For future work, there is a need to evaluate these models in other interactions to get a more illustrative idea of how robust these models could be in the overall design of a multi-user, multitasking communication interruption management system.

Bibliography

Adamczyk, P.D., and Bailey, B.P. 2004. If not now, when? The effects of interruptions at different moments within task execution. *Proceedings of the SIGCHI conference on Human factors in computer systems*: 271-278

Altmann, E.M., and Trafton, J.G. 2004. Task interruption: Resumption lag and the role of cues. *Cog Sci*.

Andy, L., and Weiner, M. 2002. Classification and regression by Randomforest. *R News*: 18-22.

Arroyo, E., and Selker, T. 2011. Attention and intention goals can mediate disruption in human-computer interaction. *Proceedings of the 13th IFIP TC 13 International Conference on Human-computer Interaction.*

Bailey, B.P., and Iqbal, S.T. 2008. Understanding changes for interruption management. *ACM Transactions on Computer Human Interaction (TOCHI).*

Bailey, B.P., and Konstan, J.A. 2006. On the need for attention aware systems: Measuring effects of interruption on task performance, error rate, and affective state. Computers in human behavior(22): 685-708.

Banerjee, S. 2016. Random forest classifier.

Blei, D.; Ng, A.; Jordan, M,; Bohus, D.; and Horvitz, E. 2009. Latent dirichlet allocation. *Journal of Machine Learning.*

Bohus, D., and Horvitz, E. 2009. Learning to predict engagement with a spoken dialog system in openworld settings. *Proceedings of the SIGDIAL*: 993-1022.

Bohus, D., and Horvitz, E. 2009. Models for multiparty engagement in open-world dialog. *Proceedings of the SIGDIAL.*

Carletta, J.; Isard, S.; Doherty-Sneddon, G.; Isard A.; Kowtko J.C.; and Anderson, A. 1997. The reliability of a dialogue structure coding scheme. *Computational linguistics*.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine Learning*.

Cutrell E.M., and E. Horvitz, E. Cutrell. and M. Czerwinski. 2000.Effects of Instant Messaging Interruptions on Computing Tasks. *CHI.*

Cutrell E.; Czerwinski, M.; Horvitz, E. 2001. Notification disruption and memory: Effects of messaging interruptions on memory and performance. *INTERACT'01: IFIP TC. 13 International Conference on Human-Computer Interaction.*

Czerwinski, M.; Cutrell, E.; and Horvitz, E. 2000. Instant messaging: Effects of relevance and timing. *People and computers XIV Proceedings of HCI British Computer Society*: 71-76.

Czerwinski, M.; Cutrell E.; and Horvitz, E. 2000. Instant messaging and interruption: Influence of task type on performance. *OZ-CHI conference proceedings*: 361-367.

Dabbish, L., Kraut, R.E. 2004. Controlling interruptions: Awareness displays and social motivation for coordination. *Proceedings of the 2004 ACM Conference on Computer supported cooperative work*.

Eyben, F.; Weninger F.; Schuller, B.; Fogarty, J.; Hudson S.; and Lai J. 2013. Recent Developments in openSMILE, the Munich OpenSource Multimedia Feature Extractor. *Proceedings ACM Multi-Media (MM)*

Fogarty, J.; Hudson, S.; and Lai, J. 2004. Examining the Robustness of Sensor-Based Statistical Models of Human Interruptibility. *Proceedings of the ACM Conference on Human Factors in Computer Systems*: 182-191.

Fogarty, J.; Ko, A.; and Anug H. 2005. Examining task engagement in sensor-based statistical models of human interruptibility. *CHI.*

Frank, E.; Hall, M.; Witten, I. 2016. The WEKA Workbench. Online appendix for "Data mining: Practical machine learning tools and techniques. *Morgan Kaufman*.

Hasti, H.W.; Poesio, M.; and Isard, S. 2002. Automatically predicting dialogue structure using prosodic features. *Speech Communication* (36): 63-79.

Ho, T.H. 1995. Random decision forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition*: 278-282

Horvitz, E., and Apacible J. 2003. Learning and reasoning about interruption. *Proceedings of the International Conference on Multimodal Interfaces.*

Huggins-Daines, D.; Kumar, D.M.; Chan, A.; Black, A.; Ravishankar, M.; abd Rudnicky, A. 2006. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. *IEEE Acoustics Speech and Signal Processing Conference*.

Isbal, S.T., and Bailey, B.P. 2005. Investigating the Effectiveness of Mental Workload as Predictor of Opportune Moments for Interruption. *CHI*

Isbal, S.T., and Bailey, B.P. 2006. Leveraging Characteristics of Task Structure to Predict Costs of Interruptions. *Proceedings CHI*

Kreifeldt J.G., and M.E. McCarthy. 1981. Interruption as a test of the user-computer interface. *JPL Proceedings of the 17th Conference of Manual Control.*

Latorella, K.A. 1996. Investigating Interruptions Implications for flightdeck performance.

McCallum, A.K. 2002. Mallet: Learning for language toolkit. *Online: http://mallet.cs.umass.edu.*

McCrickard, S.D.; Chewar, C.M.; Somervell, J.P.; and Ndiwalana A. 2003. A model for notification systems evaluation assessing user goals for multitasking activity. *ACM Transactions on Computer-Human Interaction TOCHI*: 312-338.

McFarlane, D.C. 1999. Coordinating the interruption of people in human-computer interaction. *Human-computer interaction INTERACT.*

McFarlane, D.C. 1997. Interruption of people in human-computer interaction: Unifying definition of human interruption and taxonomy. *Arlington Office of Naval Research*.

McFarlane, D.C., and Latorella K.A. 2002. The scope and importance of human interruption in human-computer interaction design. Human Computer Interaction (17): 1-61.

Mixdroff, H. 2004. Quantitative analysis of prosody in task-oriented dialogs. *Speech Prosody International Conference*.

Miyata, Y., and Norman, D.A. Psychological issues in support of multiple activities. *User-centered System Design*: 265-284

Monk, C.A.; Boehm-Davis, D.A.; Trafton, G.J.; and Sage. 2002. The attentional costs of interrupting tasks performance at various stages. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (46): 1824-1828.

Mushin, I., Stirling, L.; Fletch, J.; and Wales, R. 1999. Discourse structure, grounding, and prosody in task-oriented dialogue. *Discourse* (35):1-31.

Peters, N.S.; Romigh, G.; Raj, B.; and Bradley, G. 2017. When to interrupt: Analysis of interruption timings within collaborative tasks. *Advances in Human Factors and System Interactions*.

Robinstein, J.S.; Meyer, D.E.; and Evans, J.E. 2001. Executive control of cognitive processes in task switching. *Journal of Experimental Psychology Human Perception and Performance*

Rumelhart, D.E.; Hinton G.; and William R.J. 1986. Learning internal representations by error propagation. *Parallel distributed processing Explorations in the microstructure of cognition*.

Russel, S., and Norvig, P. 1995. *Artificial intelligence: Approach*.: Prentice-hall.

Shriberg, E.; Stolcke A.; and Hakkani-Tur, D. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech communication* (32): 127-154.

Swerts, M. 1997. Prosodic features at discourse boundaries of different strength. *The Journal of the Acoustical Society of America* (101): 514-521.

Syrdal, A.K., and Kim Y. 2008. Dialog speech acts and prosody: Consideratin for TTS. *Proceedings of Speech Prosody.*

Zijlstra, F.; Roe, R.; Leonora B.; and Krediet, I. 1999. Temporal factors in mental work: *Effects of interrupted activities. Journal of Occupational and Organizational Psychology*: 163-185