WHAT SPEECH TELLS US ABOUT DISCOURSE: THE ROLE OF PROSODIC AND DISCOURSE FEATURES IN DIALOGUE ACT CLASSIFICATION

A Thesis

Presented for the

Master of Science

Degree

The University of Memphis

Mohammed Ehasanul Hoque

August 2007

DEDICATION

To my parents ……

One of them survived cancer with the hope to see her son excel, whereas the other one financially burdened himself to ensure his son's proper education.

ACKNOWLEDGEMENTS

I would like to first thank all my thesis committee members, Drs. Mohammed Yeasin, Max Louwerse, and David Russomanno, for proving invaluable guidance and suggestions towards this work.

I would like to extend my sincere appreciation towards my colleagues from Multiple Aspects of Discourse (MAD) Lab for their help with the experiment set up, data collection, and data analysis. I also want to acknowledge my colleagues from the Computer Vision, Pattern and Image Analysis (CVPIA) Lab for being very supportive of my research throughout my tenure in Memphis.

I would like to thank Nafisa Alam who worked diligently on speech segmentation from Chicago over the weekends. Without her contribution, this thesis would not have been finished on time.

My parents have endured the pain of being away from their son for eight years to ensure my proper education. I sincerely acknowledge the sacrifices that they have made throughout the years.

ABSTRACT

Hoque, Mohammed E. MS. The University of Memphis. August 2007.
What Speech Tells us About Discourse: The Role of Prosodic and Discourse Features in Dialogue Act Classification.
Major Professor: Mohammed Yeasin, Ph.D.

This thesis investigates the automatic dialogue acts classification in multimodal communication using prosody, discourse features, and their fusion. From an experiment investigating multimodal communication, eight hours of natural audio data was collected. Prosodic and discourse features, which were believed to be strong correlates of dialogue acts, were extracted and the best features were selected using a combination of feature selection algorithms. A variety of classifiers, including traditional and ensemble, were designed and evaluated on a dialogue act classification to compare their performance. The results show that the ensemble feature selection based classification performs consistently across all the models with high validation scores. The final results demonstrated 55% accuracy on classifying 14 dialogues based on prosody, 75% accuracy with discourse and 74% with their fusion, for the best classifier. The unexpected reduction of performance due to fusion is possibly due to the lack of proper normalization of data coming from two different sources and is subject to further exploration.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Understanding and producing multimodal communication in humans and agents requires an understanding not only of the semantic meaning of an utterance, but also of the intended meaning behind that utterance. Take for instance an utterance like “go between those.” This utterance could be interpreted as an instruction (“you should go between those!”), as a yes/no question (“should I go between those?”), as an acknowledgment (speaker just stated “go between those” and the respondent confirms acknowledging the utterance by repeating “got it, go between those”). In all three cases the semantic meaning of the utterance is the same (there is an event of going and an implied patient is undergoing this event). What differs is the pragmatic meaning behind each of these utterances, typically expressed through speech acts.

The concept of speech act was first introduced by Austin [1]. He argued that the intention behind an utterance may be different from the structured sequence of words that the utterance contains. For example, “Can you please pass me the salt?” does not necessarily inquire whether person is capable of passing the salt or not, but rather indirectly asks for the salt. Austin described three aspects of speech acts: locutionary act, illocutionary act, and perlocutionary act. The locutionary act is referred as the meaning of the utterance itself in respect with the correct grammar and syntax. The illocutionary act is the meaning or intention behind the utterance in context. The perlocutionary acts pertain to the effects that an utterance has on the attitude of the hearer. In this study, the focus is mainly on illocutionary acts which are also referred as dialogue act. Searle [2] further elaborated on the illocutionary

act by stating that whenever we speak or write, we express intentions for something. We do not just talk to each other to exercise our vocal cords, but rather we express an intention or meaning through our speech. Those intentions or meanings are conveyed through various ways, such as, by making assertions, declarations, questions, expressions, etc. Austin and Searle mainly explored dialogue acts from the speech perspective within certain social context. It was claimed that intentions behind dialogues are very much context dependent and language has little role to play. However, it has been argued in [3], [4] that a significant amount of information can be derived about dialogue acts from language alone. Based on that assertion, Carletta *et al.* [3] proposed 13 dialogue acts [5] for Map-task scenarios, such as, explanation, instruction, query, reply, clarify, check, align. Besides Map-task, a few other taxonomies [2], [6], [7] of dialogue act were also derived.

1.1. WHY STUDY DIALOGUE ACTS?

Dialogue acts are known to shape the structure of the dialogue and intonational pattern. Studies have shown that the sequence of dialogue acts and the association between such acts and observed intonational contours can significantly help the performance of speech recognition engines [8], [9]. For example, possible knowledge of the intention of an utterance can be helpful in constricting the word hypothesis for speech recognition system. Dialogue acts have even proven to be useful in predicting eyebrow movements [10]. Dialogue acts have also proven to be helpful as a unit of analysis in multimodal communication [11]. For example, in multimodal communication, analyzing and correlating heterogeneous multimodal data, such as eye gaze, hand gesture, and facial expression are still considered a

difficult problem. Even though time seems to be a feasible unit of analysis, it may not be very effective as some of the human behaviors could evolve over time. Dialogue acts have proven to be an excellent substitution [11] for time as a unit of analysis in multimodal communication. Knowing what happens to modalities such as eye movement, facial expression and gestures during a specific dialogue act can be extremely helpful in the design of Embodied Conversational Agents (ECA). The existing ECAs have very limited ability to communicate using multimodal channels. Using dialogue acts as a unit of analysis to analyze multimodal data is promising in design of more appealing and effective natural ECAs. Dialogue acts could be useful in a call center environment as well, where users are first prompted into automated response systems to get their questions answered. Due to the far from optimal performance of the existing speech recognition systems, the interaction between real callers and the automated response system often results in customer dissatisfaction. Automated recognition of dialogue acts could be helpful to bridge this gap between callers and automated response systems. For example, a simple system with the ability to track pitch contours and boundary cues could be helpful to differentiate between questions and declarative statements. This information could be useful to tailor a more personable response to prevent callers from being frustrated.

The typical linguistic features of dialogue acts are useful in the domain of computer animated tutoring systems as well [12]. In tutoring systems, autonomous computer animated agents play the role of the tutor as they interact with human learners. The student learning progresses as the tutors ask questions and provide useful clues to the learner to get to the correct answer. However, an effective tutor should not only understand the semantics, but also the intention behind an utterance. For example, the tutor asked the question “What is the

value of gravity?", a learner can respond by saying, "Isn't it 9.8 m/sec^2?", or "Can you repeat the question?" or "gravity equals 9.8 m/sec^2, right?" or "No idea". Being able to understand the pragmatics or speech acts of those utterances would enable the tutor to tailor a more customized response. For example, it has been shown [13] that longer turns or statements (explanations, instructions) positively correlate with learning. Dialogue acts such as questions and feedback [14], [15] also known to maximize learning when used in appropriate context by the tutor.

Dialogue acts consist of speech (sound files) and text (transcription of the sound files) data. The text data can be automatically captured using speech recognition systems. However, due to the below-optimal performance of speech recognition systems, the text data are normally carefully transcribed by human experts. In this study, the transcription of the conversations, as well as the acoustic data, is used to model dialogue acts. The text data is used to capture discourse-related features using a bag of words as well as syntactical models. The acoustic data is used to capture the intonation patterns rather than thes semantic meaning of the utterance. This concept is termed prosody (as explained in the next section). In the following section, the importance of prosody and discourse features in dialogue acts is discussed.

1.2. IMPORTANCE OF PROSODY IN DIALOGUE ACTS

Prosody contains speech related information, which is not entirely predictable at word or sentence level, by analyzing phoneme sequences [16]. Speech features like pitch, energy, pauses, rhythm, formant, and intensity are called prosodic features. These features are independent of words and can not be deduced from lexical channels. Prosody, therefore,

provides valuable information about various dialogue acts that are difficult to disambiguate with only text. For example, declarative statements (you will go) often have similar word structures and order as questions (you will go). This can be primarily distinguished using prosodic cues.

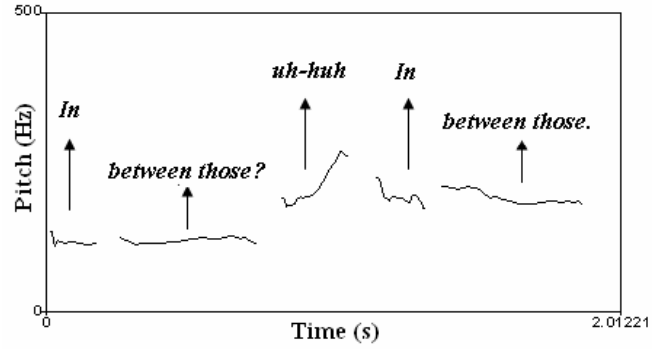
Discourse features rely heavily on carefully transcribed text data from speech. Due to the far-from-optimal performance of existing speech recognition systems, it is not practical to build a real-time dialogue act classifier based only on discourse. Studies [17] show that even the best speech recognizer can introduce up to 30% word error rate for a large vocabulary. Also speech recognition systems are expensive and may be overkill for systems where high accuracy of automated classification of dialogues is not a requirement. Prosody can be extremely useful in addressing those limitations. Introducing prosody in dialogue act classification can also help aid the research of speech synthesis.

1.3. IMPORTANCE OF DISCOURSE IN DIALOGUE ACTS

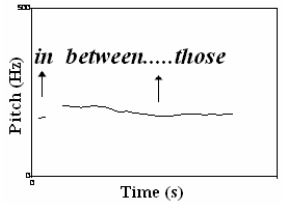
It may seem easy to identify prosodic features in dialogue act classification. In the example used earlier (“go between those”), analyzing the intonation pattern (e.g. rising or falling pitch), the utterance can be classified as a question or an instruction. Natural conversations, however, turn out to have little variation in pitch contour and intonation pattern for many dialogue acts.

Let us illustrate this with an example from a large corpus of natural multimodal communication to be discussed below. Figure 1(a) shows the pitch contour of a small segment of a conversation between two dialogue partners, where one speaker initially asks a

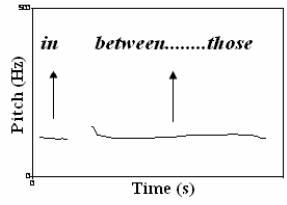
question (“in between those?”) and the other reaffirms by responding (“uh-huh, in between those.”). Figure 1(b) and 1(c) shows the pitch contour of the same statement (*in*



(a) A segment of a map-task conversation
Speaker A: in between those?
Speaker B: uh-huh..in between those



(b) Speaker A: in between those?



(c) Speaker B: in between those.

Figure 1. Pictorial description (pitch) of a case where prosody fails to distinguish between a question and statement (a) The overall conversation in context; (b) Question made by Speaker A; (c) Response made by Speaker B.

between those), used in two different ways, a question and statement. From Figure 1, it is evident that there are a few noticeable differences between the pitch contours despite the fact that the two utterances mark different dialogue acts (instruction and yes/no question).

The little variation in pitch contours perhaps explains the relatively low accuracy in dialogue act classification obtained only through prosody, ranging from 43% [18] for 12 categories and 47% [19] for 8 categories of dialogue acts.

Discourse provides context information often not available through prosodic channels. For example, a question is normally followed by a reply, whereas, a properly executed instruction or explanation yields an acknowledgement, as shown in Figure 2. These patterns of dialogue are extremely helpful to disambiguate intentions even though they may contain similar lexical information.

IG (Question): Do you agree with me?
IF (Reply): Yeah.

IG (Explain): As you move to your left, you should see a house.
IF (Acknowledge): Yeah.

Figure 2. An example of one particular word ("Yeah") being used in two different contexts with two different intentions.

Syntactical structure of an utterance, the sequences or repetition of certain parts of speech could provide useful clues about the intentions of an utterance. The number of words in an utterance is also considered a crucial factor.

1.4. PREVIOUS WORK ON DIALOGUE ACT CLASSIFICATION

Prosody has been initially introduced in dialogue act classification to segment speech. A real-time dialogue act classifier is expected to able to segment spontaneous speech into dialogues first. Often pitch range, pause patterns, speaking rate, energy patterns, utterance duration, and patterns of the pitch contour provide useful clues about utterance segmentation. Most studies [20], [21], [22] have focused on hand-coded utterances since automating the

process with reasonable accuracy still remains a difficult problem. However, in [23], it has been shown how features derived from F0 tracker can be used to approximate the intonational phrase boundaries, and thus help the automated segmentation process.

Previous studies have used various machine learning algorithms to correlate prosodic and discourse features to various dialogue acts (for a complete reference, see [24]). Examples are the Markov Model, Hidden Markov Model, Neural Networks, Self-Organizing Map Kohonen Networks, Support Vector Machine, Transformation-Based Learning, word-N-gram modelling, Polygram language model, Decision Tree, and Bayesian Networks.

Even though the dialogue act classification has been extensively explored, it is not easy to compare the studies as the feature sets, algorithms and datasets tested in the previous studies are significantly different. The cultural and language differences in different corpora, for example, English [25], German [26], Spanish [27], Japanese[28], determines the feature sets and methods to build the classifier. The variation in tasks, for example, map task, and phone conversation, also dominates the dialogue act taxonomies.

1.5. PROPOSED HYBRID APPROACH

In this thesis, it is hypothesized that the performance of automatic classification of dialogue acts can be improved by fusing prosody and discourse information together, as shown in Figure 3. The classifier should not only be capable of disambiguating discourse information, but should also compensate for the low word recognition rate of the speech engines by using prosody.

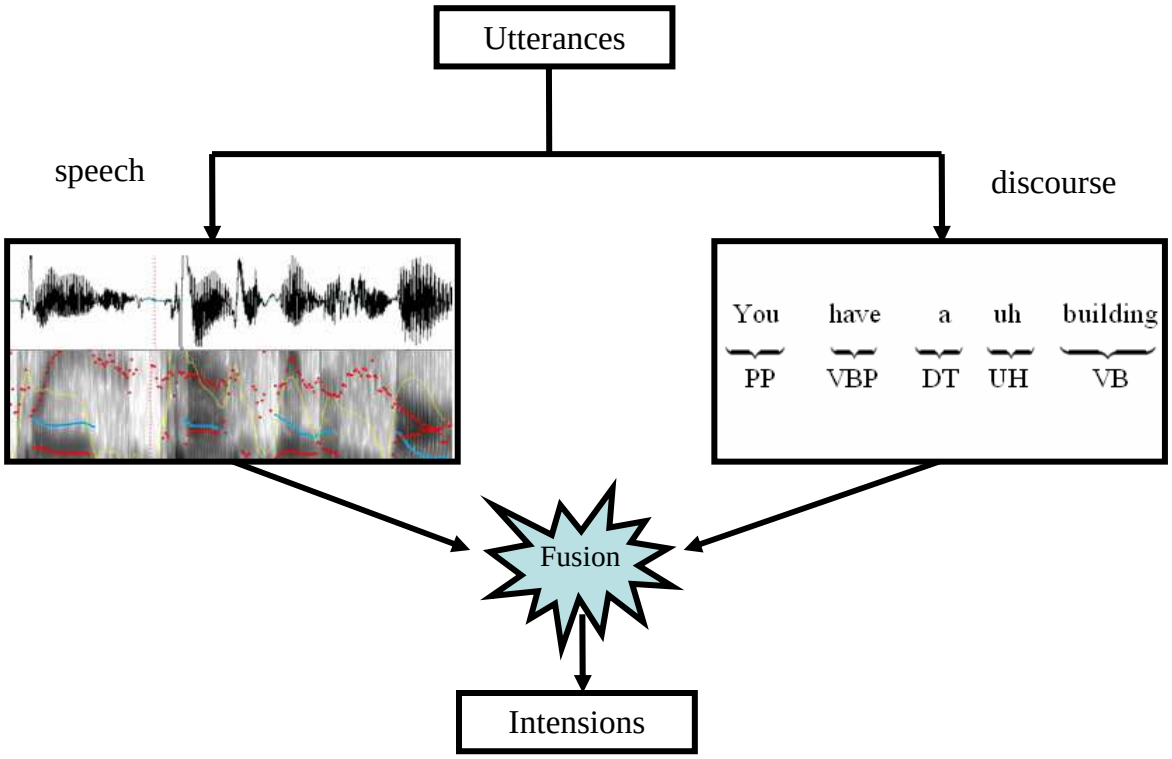


Figure 3. The overall pictorial description of how intentions are detected from utterances by fusion of prosody and discourse.

In this study, novel and distinct prosodic and discourse features were extracted. The feature extraction aspects were mainly stimulated and hypothesized by intelligent observations and assertions. For example, the patterns of pitch in instructions and explanations are expected to have a higher percentage of falling edges, whereas queries are supposed to have higher percentage of rising edges. Therefore, pitch characteristics related to rising and falling of edges were examined and taken into consideration. Patterns of pauses were also investigated with the intention of correlating those patterns with certain dialogue acts. For example, it was predicted that instructions and explanations should have a higher number of pauses, or more long pauses, as opposed to a firm reply or acknowledgment with fewer and shorter pauses. Similarly, number of words in an utterance could provide useful charactertics about

particular dialogue acts. An utterance with more number of words is like to be an explanation or instruction, rather than a quick acknowledgement, or answer. Parts of speech sequences and tagging could also add semantic meaning to the utterance. In this study, all those features were studied, extracted and evaluated.

Empirical studies [29][30] have demonstrated that discourse features provide satisfactory accuracy in classification of dialogue acts. However, the successful performance of the discourse model in real-time environment is contingent upon the 100% success rate of the speech recognition engines. Studies [17] show that even the best speech recognition system can have up to 30% error for a large vocabulary of conversation part. Therefore, discourse, even though can provide better performance given carefully transcribed data, may not be a practical approach towards building a real time dialogue act classifier. Prosodic features, on the other hand, can be computed in a real-time environment [31]. Thus, in this study, more emphasis was put on careful extraction of novel and unique prosodic features which may boost the performance of the prosody based dialogue act classifier.

One of the aims of this study is to identify a set of features that are effective across a variety of classifiers. This goal is motivated by the assertion that effectiveness of a feature set is dependent on the characteristics of classifiers. While a certain feature set may work well with one classifier, it may fail for others. The proposed feature selection framework in this study is not only expected to provide useful cues regarding which features are more relevant for a particular dialogue act, but also helps to reduce the dimensionality of the feature set by eliminating collinear features.

It is argued in this study that training multiple classifiers on the same training data and the combining their predictions on test data can potentially improve the classification accuracy

[32]. This method is called Ensemble based classification. In this study, a few ensemble based classifiers have been designed, and evaluated for dialogue act classification. A novel framework for ensemble feature selection [33] has also been proposed, implemented and evaluated in this study. The main motivation was to vary the feature subsets to enhance diversity and to produce individual classifiers that span different sub-areas of the instance space (a detail explanation of this approach is provided in Chapter 2). Combination of ensemble based classifier predictions by using majority voting, average of probabilities, maximum probabilities, and stacking are explored.

The remainder of the thesis is organized as follows. Chapter 2 provides details about the experimental setting that was designed and employed to collect data for the empirical analysis. Chapter 3 represents the big picture of the proposed solution with a detail description of prosodic and discourse features. Details on the classifications models are also illustrated in this chapter 3. Chapter 4 presents the experimental results, evaluation of various models to classify dialogue acts and future research direction.

CHAPTER 2

DATA COLLECTION METHODS

Our interest in dialogue acts stems from a large multimodal communication project [11][34]. This research project explores how different modalities in face-to-face dialogues align with each other and tries to implement these rules extracted from human experiments in an Embodied Conversational Agents (ECA). The ECA is expected to interact with humans more naturally as a validation of the study. To engage human participants into a natural, task-oriented conversation, the Map Task scenario [36] has been chosen as the general setup for study.

The Map Task is a map-oriented experimental setting in which two participants work together to achieve a common goal through the conversation. One of the participants is arbitrarily denoted as Instruction Giver (IG) who collaborates with the other partner, known as Instruction Follower (IF), to reproduce on the IF's map a route printed on IG's map (Figure 4). However, the maps of the IG and IF are not identical. Different landmarks or features of landmarks are used to order to elicit dialogues. Moreover, the color of some landmarks on IF's map are obscured by an ink blot. The differences are intentionally designed to elicit dialogue in a controlled environment based on common ground and differences in their maps. These inconsistencies between the maps are expected to be resolved through multimodal communication between the IG and IF. Speech and language are the most obvious modalities available to the participants, and therefore, focus of this study.

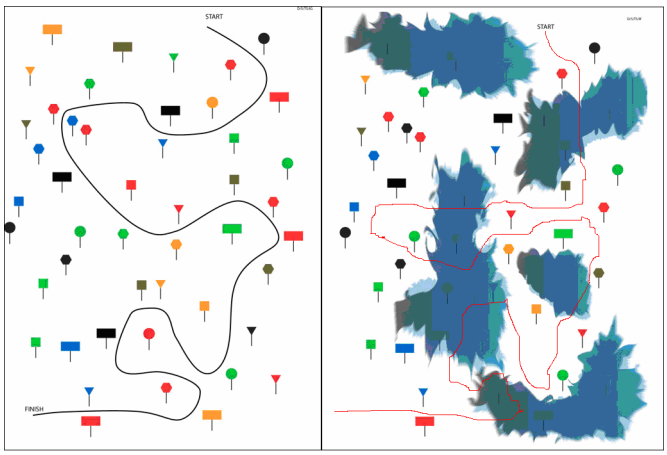


Figure 4. Example of maps. IG map presented on left, IF's map (with route drawn by IF) on right.

The current multimodal Map Task corpus includes a total of 256 conversations from 64 participants totaling 35 hours of data. Data from each conversation consists of recordings of participants' facial expressions, gestures, speech, eye gaze patterns (for IG), and map drawings (for IF). All participants performed the role of IG (4 conversations) and the role of IF (4 conversations). In each conversation, different maps were used that varied in terms of homogeneity of objects. An example of maps for the IG and IF are given in Figure 4.

2.1. DATA SIZE AND PARTICIPANTS

For the current study, 50 conversations were randomly sampled from 256 conversations totaling 8 hours of dialogue with different maps for each conversation. The 50 conversations had 56 participants in total. The gender distribution of the participants is 60% female and 40% male. The ethnic distribution of the participants is 42% African American, and 53% Caucasian and 5% others.

2.2. APPARATUS

A Marantz PMD670 speech recorder was used to record speech of IG and IF on two separate (left and right) channels using two AKG C420 headset microphones, producing optimal quality audio.

2.3. PROCEDURE

Participants, seated in front of each other, were separated by a divider to prevent any direct communication between them that could not be recorded. They could only communicate through microphones and headphones, while they viewed both the upper torso of the dialogue partner and the map on a computer monitor in front of them. A colored map was presented to IG with a route drawn on it (similar to the one presented in Figure 4). The IG was supposed to communicate the route information to the IF as accurately as possible. The 14 dialogue acts that are typically used for Map Task coding were used [3]. Table 1 presents an overview of these dialogue acts with necessary descriptions and examples.

Among the 50 conversations, the first 16 conversations for thirteen dialogue acts were manually coded by Coder A and Coder B. Inter-rater reliability between the coders in terms of Cohen's Kappa was satisfactory at .67. Next, Coder A, C and D coded the next 10 conversations. The agreement between A and C, A and D, and, C and D, were .82, .67 and .65, respectively. Due to the high inter-rater reliability, Coder A and C coded the remaining 24 conversations. Coders resolved the conflicts, primarily relating to the ACKNOWLEDGMENT, CLARIFY, ALIGN, CHECK dialogue acts, and coded the remaining transcripts for dialogue acts.

TABLE 1. THE 14 MOVE TYPES USED IN THE MAP TASK, A DESCRIPTION, AN EXAMPLE, AND FREQUENCIES FOR THE SELECTED CONVERSATIONS

Dialogue Act	Description	Example	IG	IF
INSTRUCT	Commands partner to carry out action	*Go down between the blue and the red car.*	**2150**	**9**
EXPLAIN	States information not directly elicited by partner	*Ok I went the wrong way.*	**688**	**212**
CHECK	Requests partner to confirm information	*So, between the black and the grey one?*	**29**	**106**
ALIGN	Checks attention, readiness, agreement of partner	*Ok, do you see those two blue cars?*	**174**	**4**
QUERY-YN	Yes/no question that is not CHECK or ALIGN	*Do you see that?*	**503**	**213**
QUERY-W	Any query not covered by the other categories	*If I'm at the red car what do I do there?*	**22**	**84**
ACKNOWL	Verbal response minimally showing understanding	*Uh huh.*	**319**	**1336**
REPLY-Y	Reply to any yes/no query with yes-response	*Yeah, start at the top.*	**240**	**396**
REPLY-N	Reply to any yes/no query with no-response	*No, go like above the puddle.*	**33**	**39**
REPLY-W	Reply to any type of query other than 'yes or 'no'	*It goes below.*	**63**	**26**
CLARIFY	Reply to question over and above what was asked	*So you'll be between the blue and red car.*	**120**	**88**
READY	Preparing conversation for new dialog game	*Alright. We're going to move to the left.*	**305**	**15**
UNCODBL	Laughing, sneezing		**112**	**68**
INTERJECTIONS	Resumption of discourse or surprise.	*Uh, um, ah*	**156**	**37**

CHAPTER 3

THE PROPOSED APPROACH

The proposed approach consists of five main components (as shown in Figure 5), namely, i) segment the conversation automatically into dialogues using pauses, ii) manual inspection to verify the automated segment of dialogues and then label them using human experts, iii) feature selection from the text and speech data, iv) combine prosodic and discourse features, and v) model dialogue acts using various machine learning techniques and their fusion. Subsequent subsections briefly discuss each module of the proposed speech act classification system.

3.1. DATA SEGMENTATION

As explained in Chapter 1, segmenting spontaneous speech into dialogue acts based on prosody is a difficult problem. In this study, a semi-automated technique to time stamp speech act boundaries was employed. The pauses in spoken words were used as the feature to detect the beginning and end of a turn in a natural conversation. Pauses were detected on each audio channel using the upper intensity limit and minimum duration of silences. In the measurement of intensity, minimum pitch specifies the minimum periodicity frequency in any signal. In this case, 75 Hz for minimum pitch yielded a sharp contour for the intensity. Audio segments with intensity values less than its mean intensity were classified as pauses. Thereby, mean intensity for each channel rather than a pre-set threshold was used, enabling our pause detection system to properly adapt to the diverse set of voice properties of the participants. Any audio segments with silences more

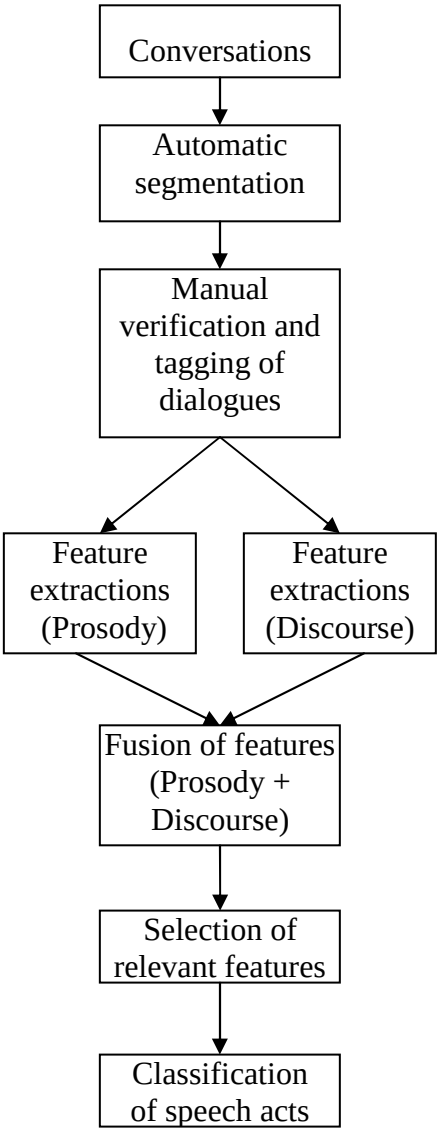


Figure 5. Schematic diagram of the proposed hybrid dialogue act classification system.

than 0.4 seconds were denoted as pauses, based on intuition. The speech processing software *Praat* [35] was used to perform all calculations to identify these pause regions.

Figure 6 shows an example of automatic segmentation of conversation into turns based on pauses. This conversion, however, is imperfect, as an utterance such as, “um, okay, then, go

straight", with no significant amount of pauses in between, can contain multiple dialogue acts. This imperfection of our automated segmentation system is resolved by manual inspection of humans by using an interface with only mouse clicks as inputs.

IG: Go right……... um,okay, then..go straight.

IF: ……..okay……..

Automatically Segmented speech acts:
IG - *Go right*
IF - *Okay*
IG - *um, okay, then..go straight.*

Segmented speech acts after manual inspection:
IG - *Go right* (Instruction)
IF - *Okay* (Acknowledgment)
IG - *um* (interjections)
IG - *okay* (Acknowledgement)
IG - *then...go straight* (Instruction)

Figure 6. Example of how turns are segmented from conversations.

3.2. PROSODIC FEATURES

To compute the prosodic information from utterances from various dialogue acts, features related to segmental and suprasegmental information, which are believed to be correlates of dialogue acts, were calculated. Computed features were utterance level statistics related to pitch, [37][38][39], an attribute of auditory sensation. In scientific terms, pitch means the frequency of a sine wave that listeners judge to be equally as high as the signal. However, others have defined pitch as the "vocal-fold vibration frequency" [35]. The moving patterns of pitch in a speech file often provide useful information not available through any other

communication channels. For example, abrupt changes of pitch patterns are correlated with frustration or anger. Low average pitch value and slightly narrower pitch range of an utterance indicate disappointment [40]. It is also well known that rising patterns of pitch correlates with questions or curiosity, whereas falling patterns of pitch indicate a belief or statement. Therefore, high occurrences of falling edges in an utterance are very likely to fall into the category of an instruction or explanation.

The intensity of speech, or energy flow, is also an important charactertic of the speech, as it is correlated with sensation of loudness. It is expected that more energy of speech, or loudness, would have a relationship with clarification of a concept that has already been introduced. Formant, which corresponds to the resonance frequencies of the vocal cords, is also an important aspect of prosody. The concentration of energy in particular frequencies are called formants. At different frequencies, there could be several formants corresponding to the resonance frequencies of the vocal cords.

The patterns of pauses are one of the important aspects of prosody, and could have important implication towards modeling of dialogue acts. For example, there are more possibilities for long pauses in an utterance when one is explaining or instructing rather than providing a firm reply or acknowledgement. More details on the patterns of pauses that were extracted are as follows:

Voice breaks are denoted as number of distances between consecutive pulses that are longer than 1.25 divided by the pitch floor. For this study, the pitch floor was set to 75 Hz, and, therefore, all inter-pulse intervals longer than 16.67 are considered as voice breaks. Degree of voice breaks is described as the total duration of voice breaks divided by the total

duration of the analyzed part of the signal. In this study, speaking rate was defined as 1/number of voiced frames.

Jitters and shimmers are characteristics of speech that contain information about the voice quality. The following features per utterance were computed for developing the prosodic model:

Pitch: Minimum, Maximum, Mean, Mode, Standard Deviation, Absolute Value, Unvoiced/Voiced frames of pitch, differences between Maximum pitch and Mean/Mode/Minimum pitch

Edges: Magnitude of the highest rising edge, magnitude of the highest falling edge, average magnitude of all the rising edges, average magnitude of all the falling edges, number of rising edges, number of falling edges.

Intensity: Minimum, Maximum, Mean, Mode, Standard Deviation, Standard Deviation, differences between Maximum intensity and Mean/Mode/Minimum intensity

Formant: Average value of first formant, second formant, third formant, average bandwidths of first, second and third formants, mean of first (f1mean), second (f2mean), third formants (f3mean), f2mean/ f1mean, f3mean/f1mean, Standard deviations of first(f1STD), second (f2STD) and third (f3STD) formants, f2STD/f1STD, f3STD/f1STD

Duration: Duration of the speech act (d1), ε_{time}, ε_{height} [41]

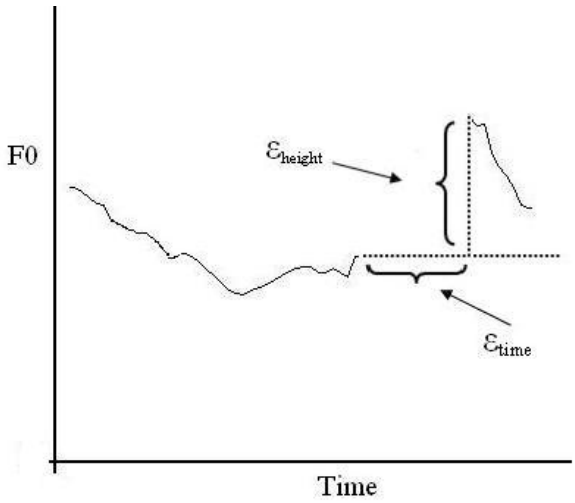


Figure 7. Measures of F0 for computing parameters (ε_{time}, ε_{height}) which corresponds to rising and lowering of intonation.

ε_{height} and ε_{time} features are related to phenomenon when pitch breaks down in utterance levels. ε_{time} refers to the pause time between two disjoint segments of pitch, whereas ε_{height} refers to the vertical distance between the segments symbolizing voice breaks as shown in Figure 7. Inclusion of *height* and *time* accounts for possible low or high pitch accents [42].

Pauses: Number of pauses, maximum duration of pauses, average duration of pauses, total duration of Pauses.

Voice Breaks: percent of Unvoiced Frames, Number of Voice Breaks, degree of Voice Breaks.

Speaking Rate: 1/voiced frames

Misc.: jitter, shimmer, energy, power

The speech processing software *Praat* [35] was used to calculate the prosodic features of speech. ε_{time}, ε_{height} features, which are part of duration, are prominence measures.

3.3. DISCOURSE FEATURES

Discourse features consisting of syntax and context information, namely, parts of speech tagging and sequence [43] , and dialogue history, were extracted.

Figure 8 below provides a visual example of how an utterance is tagged with parts of speech sequence.

You	have	a	uh	building	and	a	house
PP	VBP	DT	UH	VB	CC	DT	NNS

Figure 8. Parts of speech sequence example. PP=Prepositional Phrase, VBP= verb - present tense, DT=Determiner, UH =Interjections, VB= Verb - base form, CC=Coordinating conjunction, NNS= Noun-singular.

Along with parts of speech sequence, frequency of parts of speech in an utterance was also considered, as shown in Figure 9. The Figure 9 shows the number of times a particular part of speech has occurred in a given utterance.

Utterance	You	have	a	uh	building	and	a	house
POS	CC	DT	NNS	PP	UH	VB	VBP	
# of Occurrences	1	2	1	1	1	1	1	

Figure 9. Parts of speech tagging. PP=Prepositional Phrase, VBP= verb - present tense, DT=Determiner, UH =Interjections, VB= Verb - base form, CC=Coordinating conjunction, NNS= Noun- singular.

The number of words in an utterance was used as a feature. Also, given a dialogue act, the previous five dialogue acts were also used as relevant features.

3.4. FEATURE SELECTION ALGORITHMS

It is often recommended to reduce the dimensions of the feature set to prevent curse-of-dimensionality, which can often paralyze the performance of the classifiers. The common dimension reduction techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) have been found to be useful [41] to remove collinear features by projection of original feature sets onto the low dimensional subspace. In it argued on this study that even though the subspace projection adds values in improving the performance of model, it often fails to answer important questions such as which set of features carry most information.

To solve this problem, a variety of feature mining algorithms are used to identify optimal feature sets. In this process, optimal feature subsets are first identified and then evaluated using search methods and evaluation techniques. There are two stages to feature selection algorithms. The first stage being using search methods to identify optimal subset of features and the second stage is to evaluate the subset using different measures. Three search techniques were used: best search, greedy stepwise, and ranker. Best search method uses

greedy hill-climbing amplified with a backtracking ability to search the space of attribute subsets. It can either start with an empty set and search forward, or start with the full of set of attributes and search backward or start at any point and search at both directions. Greedy step wise performs a greedy forward or backward search through the space of attribute subsets. It starts with no/all attributes or from an arbitrary point in the space. The algorithm terminates when the addition/deletion of any remaining attributes results in a no improvement or decrease in evaluation. It is also capable of producing a ranked list of attributes by traversing the space from one side to the other and recording the order that attributes are selected. Ranker takes individual evaluations into consideration while ranking the attributes.

The algorithms that were considered for evaluation of feature sets yielded by the search technique are Cfs Subset Evaluator [44], Consistency Subset Evaluator [46] and Chi Squared Attribute evaluator.

Cfs Subset Evaluator evaluates a subset of attributes by considering its predictive ability. It also takes the degree of redundancy into consideration while evaluating a feature set. In other words, subsets of features that are highly correlated with a class while having low inter-correlation are preferred. For example, if A, and B are nominal attributes, the correlation between them can be measured using symmetric uncertainty [45].

$$U(A,B) = 2\frac{H(A) + H(B) - H(A,B)}{H(A) + H(B)}$$

H is the entropy function, which is defined for A as.

entropy $(A_1, A_2, \ldots A_n) = - A_1 \log A_1 - A_2 \log A_2 \ldots - A_n \log A_n$

The joint entropy of A and B, H (A,B), can be calculated from joint probabilities of all combinations.

```
Input: MAX_TRIES
       D – dataset
       N – number of attributes
       γ – allowable inconsistency rate
Output: sets of M features satisfying the
 inconsistency criterion

C_best = N;
for i=0 to MAX_TRIES
    S = randomSet (seed);
    C= numOfFeatures (S);
    if (C <C_best)
        if (InconCheck (S,D) < γ)
            S_best = S; C_best  = C;
            Print_Current_best (S);
    else if ((C= C_best) and
          (InconCheck (S,D) < γ))
         Print_current_best (S)
end for
```

Figure 10. Algorithmic details of Consistency Subset Evaluator.

Consistency Subset Evaluator asserts that the performance of the optimal subset of attributes can never be lower than the full set of attributes. Therefore, the usual practice is using the subset evaluator in conjunction with a Random or Exhaustive search which looks for the smallest subset of attributes with consistency equal to that of the full set of attributes.

Figure 10, shows algorithmic details [46] of how the Consistency Subset Evaluator works.

Chi-Squared Attribute Evaluator evaluates an attribute by computing the value of the chi-squared statistic with respect to the class.

The selected set of features retrieved through the feature selection algorithms are used as input to various machine learning techniques to model different dialogue acts.

3.5. CLASSIFIERS

In Chapter 1, it was showed that training multiple classifiers on the same training data and then combining their predictions on test data could potentially improve the classification

accuracy. However, generating accurate and diverse set of ensemble classifiers to improve the classification remains a difficult problem in machine learning.

There are various ways to combine the output prediction of the individual classifiers; the most popular and simplest one being *majority voting* [47]. In *majority voting*, each classifier is provided equally weighted vote towards a particular classification. The classification which gets the highest number of votes from all the classifiers is ultimately selected. A similar advanced method is called the *weighted voting* where classifiers are assigned weights according to their generalized performance towards a particular classification task. It has been found in [48] that weighted voting is more effective than majority voting. Generalized stacking [49] is also used to combine classifiers. In stacking, cross validation is used to produce output from a set of level-0 (base) classifiers, which are then used to learn a level-1 (=meta) classifier which gives the final prediction.

Among all the tree-based ensemble classifiers, RandomForest was chosen. RandomForest ensembles many decision trees and outputs the mode of the decisions of the individual trees. RandomForest works well with large number of input variables and provides an estimation of importance of variables in determining classification.

Support Vector Machine (SVM) was also taken into consideration for its robust performance in speech act classification based on previous studies [50]. SVM is a discriminative method of creating classification or regression function from the labeled training data set. Training SVM requires solving a very large scale quadratic programming problem, which, in this case would have been impractical due to the large dataset. Sequential Minimal Optimization (SMO) [51] is a fast method to train SVMs. SMO can handle a large amount of training data in linear and quadratic time with a linear amount of memory in

proportional to the training set size. SMO breaks down a large scale quadratic problem into smaller groups and solves them analytically, avoiding the time consuming inner-loop executions of the quadratic problems. Therefore, in this study, SMO was used to train the SVM. However, it has been reported in [52] that SMO's rate of convergence slows down if the data is not very sparse and many support vectors are listed in the solution.

Bagging proposed by Breimen [53] was also used in this study. Bagging is an ensemble based classifier which contains n number of classifiers in them. For each classifier, a training set is generated by randomly sampling data from the original training set without replacement. At the end, the individual decisions are fused. In Bagging, individual training sets are randomly generated. This could result in a few instances being part of the training set multiple times and instances not being part of the training set at all. Therefore, one could argue that a classifier trained on a subset of the original training set can yield more test-set error than the classifier being trained on the original data set. However, in practice, it is often not the case. It is expected that when multiple classifiers are being trained on different set of data, the diversity among those classifiers can compensate for the high error rate of an individual classifier.

Boosting, another ensemble based classifier, builds on similar concept of bagging. In boosting training sets are also initially randomly sampled from the original training data set. However, boosting presents the "hard" or "difficult to classify" examples at the later part of the training sets, focusing more on misclassified training samples. In this study, logitboost, which builds on the concepts of Boosting, is used. LogitBoost performs classification using a linear regression scheme as the base learner. LogitBoost not only works well on high dimensional data, but also shrinks the dynamic range of training data set. The monotonic

logarithmic mapping and the combination of classifiers focusing more on misclassified examples, thus makes LogitBoost consistent and robust [50].

Multi-scheme is another type of ensemble based classifier which employs multiple classifiers on the training data set. For each classifier, an X-fold cross validation is performed to determine the error rate and the classifier with the lowest error rate is chosen to be used for test data. Performance is measured based on percent correct (classification) or mean-squared error (regression).

In this study, a novel ensemble feature selection classification has been implemented as shown in Figure 11. The idea is similar to other ensemble based classification scheme, bagging, for example. However, in this approach, the sub training sets from the original training set are generated by feeding the original training set through a variety of feature selection algorithms discussed in the previous section. As the internal mechanisms of the feature selection algorithms are different, it is expected that they will produce different sets of optimal feature sets. This will not only promote diversity, but also will make ensemble classifiers disagree with each other. This disagreement among classifiers is utilized by using statistical methods, such as average of probabilities, maximum probabilities, majority voting and stacking [54].

The Waikato Environment for Knowledge Analysis (WEKA) [45] was used to build the ensemble classification framework.

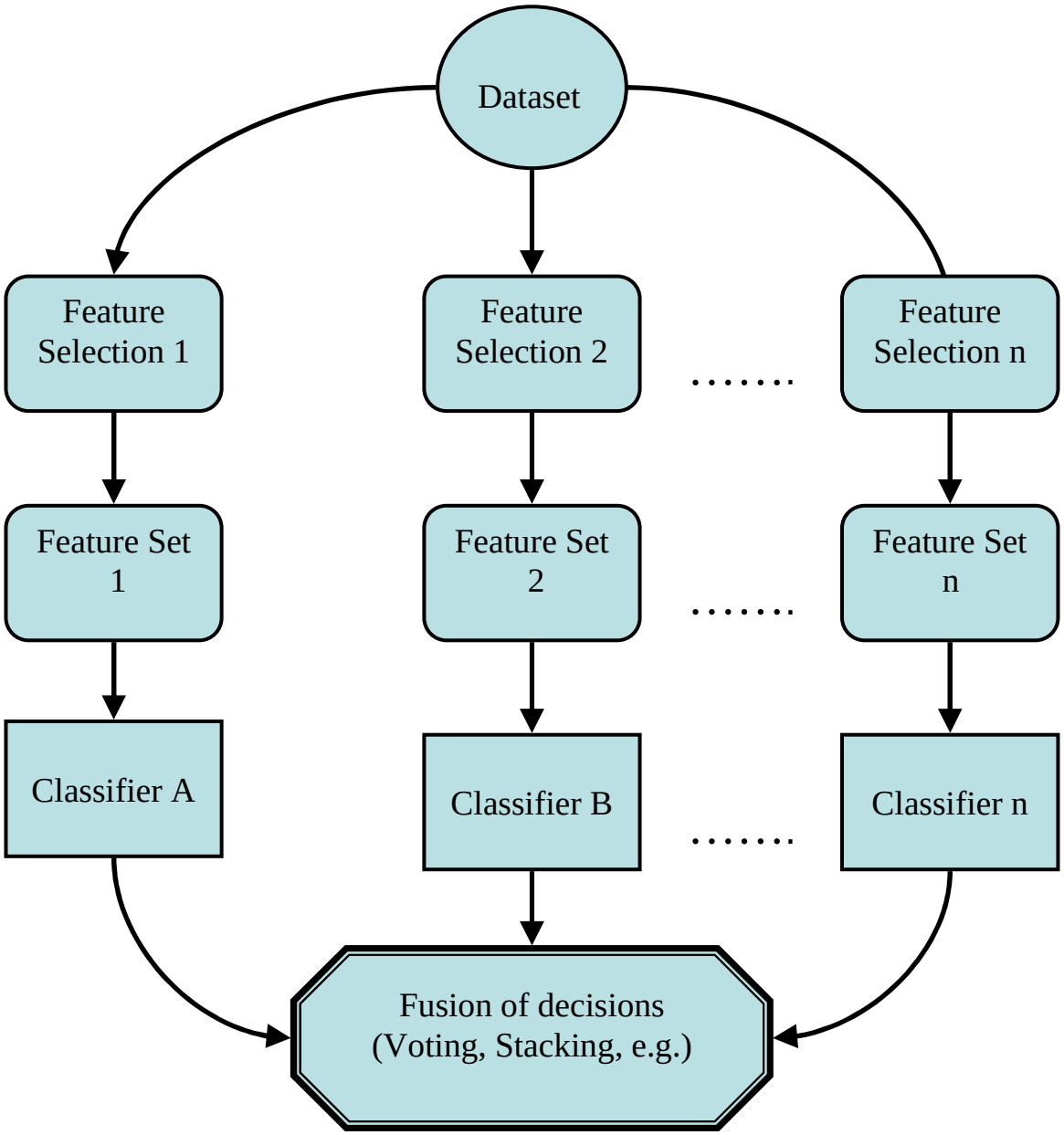


Figure 11. The proposed ensemble feature selection based classifier

CHAPTER 4

RESULTS AND INTERPRETATION OF THE PROPOSED MODELS

A number of analyses are conducted to illustrate the efficacy of the proposed approach. The database described in the Chapter 2 and feature set explained in the Chapter 3 are used to conduct the empirical analyses. In particular, classification accuracies of dialogue acts are performed using prosody, discourse and their combined feature models. In addition to the over all recognition, the performance measure such as precision, recall, true positive fraction (TP), false positive fraction (FP), F-measure and ROC area are computed. The precision corresponds to the reproducibility of the model and the recall corresponds to the generalizability (external validity) of the model. The F-measure (the ratio of geometric mean and arithmetic mean of precision and recall) provides the coherence between the precision and recall values of the model and is very good indicator of the reliability (higher F-measure implies better and more reliable model) of the predicted values. ROC curve, also known as Receiver Operating Characteristic curve, is a plot of true positive rate and false positive rate. The area under the curve is called the ROC area. The area measures discrimination, and the discrimination corresponds the ability of the test to correctly classify a category from the rest of them. In addition, confusion matrices for each model are also reported. Subsequent subsections reports the performances of various dialogue act models created using a diverse set of features and classification strategies.

4.1. MODEL 1: RESULTS AND EVALUATION

Dialogue act classification was performed on the Map task taxonomy proposed by Carletta *et al.* [3]. On top of the 13 existing categories, one extra category called INTERJECTIONS was introduced. INTERJECTIONS are stand alone fillers or sounds or words *("uh", "ah", "um")* that are spoken to fill the gaps in an utterance. In Map task environment, interjections are often used as transition stages between one dialogue acts to another. The task in Model 1 was to use machine learning techniques to distinguish among the 14 categories of dialogue acts using prosody, discourse and both. Details on the results and evaluation scores for classification using prosody only, discourse only and their combined feature set are given below.

4.1.1. Fourteen different dialogue acts using prosody

In Model 1, only prosodic correlates of dialogue acts were used to classify the 14 categories of dialogue acts.

TABLE 2. ACCURACY TO CLASSIFY 14 DIALOGUE ACTS BASED ON PROSODY (%).

Tree based classifier	Function based classifier	Ensemble based Classifiers			
Random Forest	*SMO*	*LogitBoost*	*Bagging*	*Ensemble selection (forward selection+ Backward elimination)*	*Multi scheme*
50.53	51.94	51.61	52.35	52.92	52.52

Table 2 provides classifier performances to classify 14 dialogue acts based on prosody only. It is evident from Table 2 that the all the classifiers perform in the range of 50%-53%, with the lowest of 50.53% with RandomForest and the highest of 52.92% with ensemble selection. Three classifiers, namely RandomForest, SVM and Bagging from Table 2 were selected as part of the feature selection ensemble classification framework based on their internal structure and intrinsic performances, as shown in Table 3. It was expected that the differences in internal structures of a tree based classifier (RandomForest), a function based classifier (SMO) and an ensemble based classifier (Bagging) would yield disagreement into the decision process. The disagreement among classifiers in an ensemble can be utilized by using various methods to combine the individual classifier's outcome. Four such methods, average of probability, majority voting, maximum probability, and stacking were used in this study. Feature selection algorithms such as Consistency Subset Evaluator using greedy stepwise search algorithm, Chi Squared Attribute Evaluator using Ranker search Algorithm and Cfs Subset Evaluator using Best First search Algorithm were used to rank the features in order of importance. The three distinct optimal feature sets generated from three separate feature selection algorithms were used an inputs to the three classifiers part of the ensemble. The feature sets were rotated per classifier to evaluate the robustness of the proposed ensemble feature selection classification framework. This framework of feature set generation, rotation and evaluation was consistent throughout this study for rest of the models.

Careful inspection of the Table 2 and the Table 3 indicates that the best accuracy in classifying dialogue acts using prosodic features is obtained using the Ensemble Based Feature Selection framework employing majoring voting and stacking.

Table 2 and Table 3 also indicate that the best accuracy in classifying dialogue acts using prosodic features is obtained using Ensemble Based Feature Selection framework. The models with majority voting and stacking to combine the predictions of individual classifiers yielded the best performances.

TABLE 3. ACCURACY TO CLASSIFY 14 DIALOGUE ACTS BASED ON PROSODY USING EMSEMBLE FEATURE SELECTION FRAMEWORK (%).

Classifiers	Feature Selection Algorithm	Fusion Technique	Accuracy %
Random Forest	Consistency Subset Evaluator using Greedy Stepwise search algorithm	Average of Probability	53.13
Bagging	Chi Squared Attribute Evaluator using Ranker search Algorithm		
SMO	Cfs Subset Evaluator using Best First search Algorithm		
Random Forest	Cfs Subset Evaluator using Best First search Algorithm	Majority Voting	55.67
Bagging	Consistency Subset Evaluator using Greedy Stepwise search algorithm		
SMO	Chi Squared Attribute Evaluator using Ranker search Algorithm		
Random Forest	Chi Squared Attribute Evaluator using Ranker search Algorithm	Maximum Probability	51.59
Bagging	Cfs Subset Evaluator using Best First search Algorithm		
SMO	Consistency Subset Evaluator using Greedy Stepwise search algorithm		
Random Forest	Chi Squared Attribute Evaluator using Ranker search Algorithm	Stacking	53.94
Bagging	Cfs Subset Evaluator using Best First search Algorithm		
SMO	Consistency Subset Evaluator using Greedy Stepwise search algorithm		

Table 4 provides the validation measures including true positive, false positive, precision, recall, F-measure and ROC area, on recognition of 14 categories of dialogue acts using prosody only. It is evident from Table 4 that the categories with less number of samples have low evaluation scores. For example, categories with less than 3% of total instances, such as

CHECK, REPLY-Y, MISC., ALIGN, CLARIFY, QUERY-W, AND REPLY-W have F-measures less than 0.1.

TABLE 4. VALIDATION DETAILS ON DIALOGUE CLASSIFICATION USING PROSODY ON 14 CATEGORIES. TP=TRUE POSITIVE, FP=FALSE POSITIVE. n = NUMBER OF INSTANCES, % = PERCENTAGE OF INSTANCES (THE CATEGORIES WITH LESS THAN OR EQUAL TO 3% OF OVERALL INSTANCES ARE INDICATED IN BOLD)

Class	n	%	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
QUERY-YN	716	9	0.205	0.046	0.325	0.205	0.251	0.787
REPLY-Y	636	8	0.305	0.032	0.467	0.305	0.369	0.863
CHECK	**135**	**2**	**0.016**	**0.002**	**0.133**	**0.016**	**0.028**	**0.844**
ACKNOWL	1655	21	0.82	0.145	0.618	0.82	0.705	0.928
INSTRUCT	2159	27	0.938	0.291	0.569	0.938	0.708	0.877
REPLY-N	**72**	**1**	**0**	**0**	**0**	**0**	**0**	**0.7**
EXPLAIN	900	11	0.091	0.032	0.28	0.091	0.137	0.703
READY	320	4	0.389	0.027	0.393	0.389	0.391	0.911
MISC.	**180**	**2**	**0.028**	**0.001**	**0.357**	**0.028**	**0.051**	**0.775**
ALIGN	**178**	**2**	**0**	**0**	**0**	**0**	**0**	**0.842**
CLARIFY	**208**	**3**	**0**	**0**	**0**	**0**	**0**	**0.701**
QUERY-W	**106**	**1.5**	**0.009**	**0.001**	**0.111**	**0.009**	**0.017**	**0.793**
REPLY-W	**89**	**1**	**0**	**0**	**0**	**0**	**0**	**0.655**
INTERJECT	636	8	0.387	0.013	0.452	0.387	0.417	0.918

Table 5, on the other hand, provides useful clues about what categories of dialogue acts get confused with other dialogue acts using prosody only. It has been observed that CLARIFY gets confused with INSTRUCTIONS; READY with ACKNOWLEDGEMENT; EXPLAIN with INSTRUCTIONS; and ACKNOWLEDGEMENT with REPLY-Y. The most surprising outcome was the confusion between QUERY-Y and INSTRUCT, as it was expected that the distinctive rising edges (with statements) and falling edges (with questions) of pitch patterns would be helpful to discriminate between them. Explanations need to be further investigated, but one possible justification could be the existing prosodic features do not take local features of the pitch patterns into consideration. Extracting local pitch features from a few selective places rather than the whole utterance could prove be more valuable. One selective place could be the end

of an utterance, which may provide clues about the emphasis factors of the syllables at the end of the utterance.

TABLE 5. CONFUSION MATRIX FOR CLASSIFICATION USING PROSODY ON 14 CATEGORIES OF DIALOGUE ACTS.

a	b	c	d	e	f	g	h	i	j	k	l	m	n	classified as
147	12	1	76	409	0	60	7	0	0	0	2	0	4	a=QUERY-YN
12	194	1	323	63	0	9	23	1	0	0	1	0	10	b=REPLY-Y
48	2	2	29	12	0	27	2	1	0	0	1	0	2	c=CHECK
21	109	1	1357	47	0	16	71	0	0	0	0	0	33	d=ACKNOWL
26	18	0	39	2025	0	6	36	0	0	0	0	0	10	e=INSTRUCT
5	6	2	32	18	0	2	1	1	0	0	0	0	6	f=REPLY-N
86	15	5	65	618	0	82	19	1	0	1	2	0	7	g=EXPLAIN
1	21	0	112	52	0	0	125	1	0	0	0	0	9	h=READY
31	5	1	18	93	0	20	5	5	0	0	1	0	2	i=MISC.
3	10	0	28	16	0	1	9	0	0	0	0	0	2	j=ALIGN
30	8	0	23	101	0	40	4	1	0	0	0	0	2	k=CLARIFY
30	4	2	26	18	0	21	0	3	0	0	1	0	2	l=QUERY-W
12	3	0	9	50	0	9	3	0	0	0	0	0	2	m=REPLY-W
1	8	0	59	37	0	0	13	0	0	0	1	0	75	n=INTERJECT

4.1.2. Fourteen different dialogue acts using discourse only

In this part of Model 1, only discourse, namely context and syntax, features of dialogue acts were used to classify the 14 categories of dialogue acts. The syntax features corresponding to the parts of speech sequences and tagging were used in this model. The number of words in an utterance and dialogue act history were also part of the discourse feature sets.

The same list of classifiers from the previous model of prosody was used, as shown in Tables 6 and 7. From Table 6, ensemble based classifiers like bagging and ensemble selection performed comparatively better than other classifiers. Majority voting, from Table

10 had the highest accuracy of 75.95%. Other classifiers, in Table 7, also were able to classify the 14 categories of dialogue acts more than 70% of the time.

TABLE 6. ACCURACY TO CLASSIFY 14 CATEGORIES OF DIALOGUE ACTS BASED ON DISCOURSE (%).

Tree based classifier	Function based classifier	Ensemble based Classifiers			
Random Forest	*SMO*	*LogitBoost*	*Bagging*	*Ensemble selection (forward selection+ Backward elimination)*	*Multi scheme*
65.78	63.37	68.44	72.41	71.29	66.07

TABLE 7. ACCURACY TO CLASSIFY 14 DIALOGUE ACTS BASED ON DISCOURSE USING EMSEMBLE FEATURE SELECTION FRAMEWORK (%)

Classifiers	Feature Selection Algorithm	Fusion Technique	Accuracy %
RandomForest	Consistency Subset Evaluator using Greedy Stepwise search algorithm	Average of Probability	72.41
Bagging	Chi Squared Attribute Evaluator using Ranker search Algorithm		
SMO	Cfs Subset Evaluator using Best First search Algorithm		
RandomForest	Cfs Subset Evaluator using Best First search Algorithm	Majority Voting	75.95
Bagging	Consistency Subset Evaluator using Greedy Stepwise search algorithm		
SMO	Chi Squared Attribute Evaluator using Ranker search Algorithm		
RandomForest	Chi Squared Attribute Evaluator using Ranker search Algorithm	Maximum Probability	71.47
Bagging	Cfs Subset Evaluator using Best First search Algorithm		
SMO	Consistency Subset Evaluator using Greedy Stepwise search algorithm		
RandomForest	Chi Squared Attribute Evaluator using Ranker search Algorithm	Stacking	72.36
Bagging	Cfs Subset Evaluator using Best First search Algorithm		
SMO	Consistency Subset Evaluator using Greedy Stepwise search algorithm		

Table 8 provides the evaluation scores of classification using discourse model. The lowest two F-measures in Table 8 occurred for ALIGN and CHECK and are indicated in bold. In the previous model of prosody, it was observed that categories with less number of instances had the lowest F-measures. The similar phenomenon, however, was not prevalent in this model of discourse. The categories with lowest number of instances, REPLY-N, REPLY-W, and QUERY-W, indicated in bold in Table 8, had F-measures of .53, .39 and .30 respectively, compared to 0, 0, 0.017 with prosodic model. There are a few possible reasons behind the improved recognition rate of those categories with discourse model. It is expected that A QUERY-YN is likely to be followed by REPLY-N, just the way QUERY-W is often followed by REPLY-W. In prosodic model, the internal structures of speech are analyzed without having any knowledge about the dialogue history, e.g., what dialogue act is likely to follow after another dialogue act. Therefore, not having enough instances of a category entails less familiarity about the variability of the speech data for that particular category. Thus, distinguishing between REPLY-N and REPLY-W with very few samples of extremely natural speech data could result in low recognition rate. However, with discourse, the availability of dialogue history helps to improve the recognition rate of REPLY-W, REPLY-N, QUERY-W, even though they have less number of samples. Therefore, it can be inferred that adding dialogue history and syntax into the classification process can improve the recognition rate of less represented and ambiguous categories like REPLY-N, REPLY-W, and QUERY-W.

One may argue that separation between ACKNOWLEDGE and INSTRUCTION can easily be done by simply counting the words per utterances, as INSTRUCTION categories are supposed to have more words than ACKNOWLEDGE. However, a counter argument would state that ACKNOWLEDGE does not always have less number of words compared to INSTRUCTION. For

example, we often acknowledge an instruction by simply repeating it. An example of that event is shown below:

IG (INSTRUCT) : *Go between those.*
IF (ACKNOWLEDGE) : *Got it, I will have to go between those.*

In the above example, ACKNOWLEDGE had more words compared to INSTRUCT, which is a strong indication that word count itself is not always a reliable source to distinguish between INSTRUCT and ACKNOWLEDGE dialogue acts.

TABLE 8. VALIDATION DETAILS ON CLASSIFICATION USING DISCOURSE ON 14 CATEGORIES. TP=TRUE POSITIVE, FP=FALSE POSITIVE, N= NUMBER OF INSTANCES, % = PERCENTAGE OF INSTANCES (THE CATEGORIES WITH THE LOWEST TWO F-MEASURES ARE INDICATED IN BOLD)

Class	n	%	TP Rate	FP Rate	Precision	Recall	F-Measure
QUERY-YN	716	9	0.617	0.054	0.552	0.617	0.583
REPLY-Y	636	8	0.834	0.021	0.787	0.834	0.809
CHECK	**135**	**2**	**0.151**	**0.009**	**0.224**	**0.151**	**0.18**
ACKNOWL	1655	21	0.88	0.049	0.836	0.88	0.857
INSTRUCT	2159	27	0.905	0.09	0.804	0.905	0.851
REPLY-N	72	1	0.466	0.003	0.63	0.466	0.535
EXPLAIN	900	11	0.511	0.057	0.554	0.511	0.531
READY	320	4	0.704	0.017	0.648	0.704	0.675
MISC.	180	2	0.171	0.004	0.525	0.171	0.258
ALIGN	**178**	**2**	**0.014**	**0.001**	**0.125**	**0.014**	**0.026**
CLARIFY	208	3	0.22	0.014	0.313	0.22	0.258
QUERY-W	106	1.5	0.196	0.002	0.6	0.196	0.296
REPLY-W	89	1	0.318	0.004	0.519	0.318	0.394
INTERJECT	636	8	0.598	0.007	0.695	0.598	0.643

Table 9, the confusion matrix of 14 categories of dialogue act classification using discourse, reveals additional findings. In discourse model, performance increases in terms of QUERY-YN getting confused with INSTRUCTION, which is opposite of what was seen in the prosodic model. However, performance degrades, compared to the prosodic model, to differentiate

between EXPLAIN and QUERY-YN. The rationalization for this incident need to be further explored with larger dataset. More examples of dialogue history among the three categories, QUERY-YN, EXPLAIN, and INSTRUCTION, would be helpful to understand the interaction trend among them. It was interesting to notice that confusion between EXPLAIN and INSTRUCT (total of 3059 instances of EXPLAIN AND INSTRUCT) is higher compared to the confusion between INSTRUCT and QUERY-YN (total of 2371 instances of INSTRUCT and QUERY-YN). It is expected that a pair of dialogue acts with more number of instances would be confused less compared to another pair of dialogue acts with less number of instances. The above assertion was made with the usefulness of dialogue history in mind. This is an example where dialogue history or context information was not very helpful with the purpose of disambiguation of dialogues. One possible explanation is the relationship between INSTRUCT and QUERY-YN, and INSTRUCT and EXPLAIN. Given an Instruction, if understood correctly, one is likely to acknowledge it, or ask questions, otherwise. It is very unlikely for someone to respond with an explanation to an instruction, especially in a task oriented environment. This is exactly the reason why INSTRUCT and EXPLAIN get confused a lot despite the availability of huge number of examples presented to the classifier.

Relative improvement to distinguish between ACKNOWLEDGEMENT and REPLY-Y, ACKNOWLEDGMENT and READY has also been observed. Once again, dialogue history or context information might have instigated that improvement. For example, ACKNOWLEDGE, REPLY-Y and READY can often same identical linguistic units in discourse. But the context of when one is used is different. A proper validation of the observations made in this section could be adding another discourse model without the context information and the compare the confusion matrices of the two models; one with context and one without.

TABLE 9. CONFUSION MATRIX FOR CLASSIFICATION OF 14 CATEGORIES OF DIALOGUE ACTS USING DISCOURSE.

a	b	c	d	e	f	g	h	i	j	k	l	m	n	<--classified as
443	13	26	15	77	5	106	3	6	0	13	6	3	2	a=QUERY-YN
17	531	1	49	11	1	3	2	6	0	14	0	1	1	b=REPLY-Y
48	6	19	12	1	2	26	0	1	0	7	1	1	2	c=CHECK
8	56	3	1456	14	1	21	53	1	2	4	0	3	33	d=ACKNOWL
48	9	2	13	1955	0	92	21	3	1	10	0	4	2	e=INSTRUCT
3	13	0	4	2	34	7	1	0	0	8	1	0	0	f=REPLY-N
84	9	11	40	246	1	460	10	2	0	28	3	3	4	g=EXPLAIN
0	9	0	46	25	1	2	226	2	4	1	0	1	4	h=READY
79	3	11	5	19	5	19	4	31	0	3	1	0	1	i=MISC.
3	4	0	19	20	0	1	18	1	1	0	0	0	2	j=ALIGN
25	11	6	16	36	1	55	1	3	0	46	1	8	0	k=CLARIFY
29	2	5	14	4	1	24	0	3	0	3	21	1	0	l=QUERY-W
15	1	0	7	14	1	11	0	0	0	10	1	28	0	m=REPLY-W
1	8	1	45	8	1	3	10	0	0	0	0	1	116	n=INTERJECT

4.1.3. Fourteen different dialogue acts using prosody and discourse

This is the final analysis of Model 1 where prosody and discourse information are fused in feature level to classify the 14 categories of dialogue acts. Once again, it has been hypothesized that the model with prosody and discourse would be more robust compared to the model with only prosody or only discourse.

TABLE 10. ACCURACY TO CLASSIFY 14 DIALOGUE ACTS BASED ON PROSODY + DISCOURSE (%).

Tree classifier	Function classifier	Ensemble based Classifiers			
Random Forest	*SMO*	*Logit Boost*	*Bagging*	*Ensemble selection (forward selection+ Backward elimination)*	*Multi scheme*
61.19	64.79	68.61	70.52	70.11	61.02

Tables 10 and 11 help us to understand that when RandomForest (61.19% of accuracy), SMO (64.79 % of accuracy) and Bagging (70.52% of accuracy) are combined with the ensemble

feature selection classification framework, they are capable of performing better than their individual performance. In the feature selection classification framework, majority voting performed the best. This once again supports our hypothesis that that an individual classifier, (e.g., bagging) may work the best for one particular model, but the combination of the diverse set of features and fusion of classifiers shows more consistence.

TABLE 11. ACCURACY TO CLASSIFY 14 DIALOGUE ACTS BASED PROSODY AND DISCOURSE USING EMSEMBLE FEATURE SELECTION FRAMEWORK (%).

Classifiers	Feature Selection Algorithm	Fusion Technique	Accuracy %
RandomForest	Consistency Subset Evaluator using Greedy Stepwise search algorithm	Average of Probability	70.57
Bagging	Chi Squared Attribute Evaluator using Ranker search Algorithm		
SMO	Cfs Subset Evaluator using Best First search Algorithm		
RandomForest	Cfs Subset Evaluator using Best First search Algorithm	Majority Voting	74.38
Bagging	Consistency Subset Evaluator using Greedy Stepwise search algorithm		
SMO	Chi Squared Attribute Evaluator using Ranker search Algorithm		
RandomForest	Chi Squared Attribute Evaluator using Ranker search Algorithm	Maximum Probability	69.84
Bagging	Cfs Subset Evaluator using Best First search Algorithm		
SMO	Consistency Subset Evaluator using Greedy Stepwise search algorithm		
RandomForest	Chi Squared Attribute Evaluator using Ranker search Algorithm	Stacking	71.78
Bagging	Cfs Subset Evaluator using Best First search Algorithm		
SMO	Consistency Subset Evaluator using Greedy Stepwise search algorithm		

Table 11 provides the validation scores for the 14 categories of dialogue act classification using prosody and discourse. The accuracies for model with prosody and discourse are comparable with the previous model with discourse only. However, evaluating the F-

measures does indicate an improvement of recognizing QUERY-YN with the fusion of prosody and discourse.

TABLE 12. VALIDATION RESULTS ON CLASSIFICATION ON 14 CATEGORIES USING PROSODY + DISCOURSE. TP=TRUE POSITIVE, FP=FALSE POSITIVE, N= NUMBER OF INSTANCES, % = PERCENTAGE OF INSTANCES.

Class	n	%	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
QUERY-YN	716	9	0.655	0.054	0.565	0.655	0.606	0.92
REPLY-Y	636	8	0.794	0.023	0.764	0.794	0.779	0.969
CHECK	135	2	0.04	0.003	0.179	0.04	0.065	0.89
ACKNOWL	1655	21	0.876	0.055	0.821	0.876	0.847	0.972
INSTRUCT	2159	27	0.911	0.099	0.791	0.911	0.847	0.964
REPLY-N	72	1	0.397	0.003	0.558	0.397	0.464	0.91
EXPLAIN	900	11	0.514	0.063	0.53	0.514	0.522	0.87
READY	320	4	0.71	0.018	0.64	0.71	0.674	0.964
MISC.	180	2	0.238	0.005	0.558	0.238	0.333	0.909
ALIGN	178	2	0	0	0	0	0	0.859
CLARIFY	208	3	0.105	0.006	0.349	0.105	0.162	0.855
QUERY-W	106	1.5	0.121	0.001	0.565	0.121	0.2	0.925
REPLY-W	89	1	0.295	0.004	0.481	0.295	0.366	0.886
INTERJECT	636	8	0.608	0.006	0.72	0.608	0.659	0.968

The confusion matrix, shown in Table 12, for the model with prosody and discourse to classify among 14 categories of dialogue acts does not show a huge improvement or radical disagreement. Similar patterns of confusion are noticed here well, with very less improvement from the previous models in terms of distinguishing between EXPLAIN - QUERY-YN, EXPLAIN – INSTRUCTION, and READY - ACKNOWLEDGEMENT.

The model based on combined discourse and prosodic features is expected to yield relative higher performance compared to the individual feature set. Empirical analyses show that it was not the case. One possible explanation could be the normalization scheme employed in this study, which is required to utilize the fusion of prosody and discourse in feature level. Future efforts may investigate proper normalization of disparate feature sets from two

different sources. It is also a possibility to explore decision level fusion of prosody and discourse features, along with feature level.

TABLE 13. CONFUSION MATRIX FOR CLASSIFICATION OF 13 CATEGORIES OF DIALOGUE ACTS USING PROSODY + DISCOURSE.

a	b	c	d	e	f	g	h	i	j	k	l	m	n	<--classified as
470	17	10	18	83	5	96	2	5	0	5	3	3	1	a=QUERY-YN
23	506	0	66	11	1	14	1	5	0	7	0	1	2	b=REPLY-Y
43	4	5	18	1	0	42	0	7	0	2	2	2	0	c=CHECK
9	59	0	1450	15	4	24	56	0	0	1	0	2	35	d=ACKNOWL
39	11	0	10	1967	0	94	25	5	0	2	0	5	2	e=INSTRUCT
6	16	1	5	3	29	9	1	0	0	2	0	0	1	f=REPLY-N
80	8	4	43	270	1	463	12	4	0	10	2	3	1	g=EXPLAIN
1	10	0	49	28	0	2	228	0	0	1	0	0	2	h=READY
63	5	5	8	20	5	24	4	43	0	1	2	0	1	i=MISC.
5	4	0	19	23	0	2	15	0	0	0	0	0	1	j=ALIGN
47	10	1	15	39	3	57	1	3	0	22	0	11	0	k=CLARIFY
35	1	2	14	1	2	30	0	3	0	5	13	1	0	l=QUERY-W
11	2	0	9	17	1	15	0	1	0	5	1	26	0	m=REPLY-W
0	9	0	43	9	1	2	11	1	0	0	0	0	118	n=INTERJECT

In Model 1, 14 categories from the Carletta *et al.* [3] dialogue act taxonomy were used. The model with prosody, discourse and the combination of prosody and discourse features yielded accuracies of 56%, 76% and 75%, respectively. The accuracies obtained in this study are higher than the accuracy reported in [29], [30] using Carletta *et al.* [3] taxonomy in map task environment. Surendran *et al.* [30] reported 43% of accuracy using prosody, 59% using discourse, and 66% with the fusion of both, where as Louwerse *et al.* [29] reported 58% using discourse features.

The next step of this study was to narrow down the dialogue act categories in a systemic way and observe its effects on the classification performance. Based on confusion matrices of the models with prosody, discourse features and their combination, it was evident that certain

categories of dialogue acts were consistently getting confused with some other categories of dialogue acts. For example, INSTRUCT getting confused EXPLAIN, ACKNOWLEDGE getting confused READY, CLARIFY getting confused with INSTRUCT and EXPLAIN, have been consistent in Model 1. Based on those observations, a second model of dialogue act classification was proposed.

4.2. MODEL 2: RESULTS AND EVALUATION

In Model 2, the Carletta *et al.* [3] dialogue act taxonomies were "collapsed" into four categories. Collapsing of the 14 dialogue act categories into 4 was done based on the synergy between two criteria. One criterion was the confusion matrices of Model 1, where observations were made relating to what categories of dialogue acts consistently get confused with other categories. The second criterion was the Carletta *et al*. [3] map-task taxonomy itself. Given an utterance, the first question that is asked in [3] whether it is an initiation (INSTRUCT, EXPLAIN, ALIGN, CHECK, QUERY-YN, QUERY-W), response (ACKNOWLEDGE, CLARIFY, REPLY-Y, REPLY-N, REPLY-W) or preparation (READY). Carletta *et al.* [3] then collapsed initiation into commands (INSTRUCT), statements (EXPLAIN) and questions (ALIGN, CHECK, QUERY-YN, AND QUERY-W). In this study, the questions categories were identical to what was proposed in [3]. Along with INSTRUCT and EXPLAIN, CLARIFY from response category was put into Statements category, mainly due to observations made in the confusion matrices. One major discrepancy between the proposed taxonomy and the Carletta *et al.* [3] taxonomy was the no separation between preparation (READY) and response categories. This was mainly done due to factors of READY dialogue act being confused with the reply

categories and Ready not having enough number of instances to be a category of its own. Table 14 shows the mapping between the Carletta *et al.* [3] Map task taxonomy and the proposed simplified taxonomy.

TABLE 14. SUB-GROUPING OF THE ORIGINAL 13 SPEECH CATEGORIES AND THEIR FREQUENCIES.

Original Categories	Simplified Categories	Total Utterances	IF	IG
CHECK + QUERY-YN + QUERY-W + ALIGN	Questions	1118	407	711
CLARIFY + EXPLAIN + INSTRUCT	Statements	3267	309	2958
REPLY-Y + REPLY-N + REPLY-W + ACKNOWL + READY	Reply	2712	1812	900
INTERJECT	Interjections	193	37	156

Another major motivation behind this "collapse" of categories into smaller subgroups is to explore synergy among dialogue act categories. Understanding the interaction patterns and relationship between the dialogue acts could be useful to computationally validate the original map-task taxonomy containing 13 dialogue acts proposed by Carletta *et al.* [3]. It may also be a possibility to examine whether the mistakes made by the computational algorithms are consistent with the mistakes made by humans. If not, it may provide useful insights towards future direction of research as to how the computational algorithms can be improved to adapt to the reasoning capabilities that we humans innately posses.

4.2.1. Four different dialogue acts using prosody

In this part of Model 2, only prosodic correlates of dialogue acts were used to classify the 4 categories of dialogue acts, namely, questions, statements, replies and interjections. The

narrowing down of 14 dialogue acts into 4 dialogue acts were done mainly the through the confusion matrices of prosody, discourse and their combinatory models from Model 1.

TABLE 15. ACCURACY TO CLASSIFY 4 DIALOGUE ACTS BASED ON PROSODY (%).

Tree classifier	Function classifier	Ensemble based Classifiers			
Random Forest	*SMO*	*LogitBoost*	*Bagging*	*Ensemble selection (forward selection+ Backward elimination)*	*Multi scheme*
75.04	75.11	74.99	76.42	75.81	75.00

Tables 15 and 16 show that, all the classifiers designed, implemented and used for this study had similar accuracy for this particular model of prosody for 4 categories of dialogue acts. In the previous models, the tree based classifier, RandomForest and function based classifier, SMO, had the lowest performances, compared to other classifiers. However, in this model, RandomForest and SMO provided similar performance metrics in comparison with the ensemble based classifiers. Therefore, it is intuitively evident that this model of dialogue act classification using prosody is very robust across classifiers. The combination of RandomForest, SMO and bagging, performed better in an ensemble, as shown in Table 16, compared to their individual performances. The highest accuracy to distinguish between 4 categories of dialogue act using prosody was noted to be around 77% with the majority voting and stacking methods.

TABLE 16. ACCURACY TO CLASSIFY 4 DIALOGUE ACTS BASED ON PROSODY USING EMSEMBLE FEATURE SELECTION FRAMEWORK (%).

Classifiers	Feature Selection Algorithm	Fusion Technique	Accuracy %
RandomForest	Consistency Subset Evaluator using Greedy Stepwise search algorithm	Average of Probability	76.16
Bagging	Chi Squared Attribute Evaluator using Ranker search Algorithm		
SVM	Cfs Subset Evaluator using Best First search Algorithm		
RandomForest	Cfs Subset Evaluator using Best First search Algorithm	Majority Voting	77.19
Bagging	Consistency Subset Evaluator using Greedy Stepwise search algorithm		
SMO	Chi Squared Attribute Evaluator using Ranker search Algorithm		
RandomForest	Chi Squared Attribute Evaluator using Ranker search Algorithm	Maximum Probability	75.90
Bagging	Cfs Subset Evaluator using Best First search Algorithm		
SMO	Consistency Subset Evaluator using Greedy Stepwise search algorithm		
RandomForest	Chi Squared Attribute Evaluator using Ranker search Algorithm	Stacking	77
Bagging	Cfs Subset Evaluator using Best First search Algorithm		
SMO	Consistency Subset Evaluator using Greedy Stepwise search algorithm		

Table 17 provides validation scores on this particular model of dialogue act classification for 4 categories. It shows that questions and interjections have the lowest F-measures, among the four categories of dialogue acts. Inspecting confusion matrix of this particular model, in Table 18, indicated that among 1118 instances of questions, 50% of them got confused with statements, and 20% of them got confused with replies, whereas remaining 30% instances get classified as questions properly. This statistics with the prosody model, for 4 categories of dialogue acts especially with questions, was not very encouraging. Prosody computes pitch level statistics of an utterance and should provide useful clues about the patterns of pitch fluctuations over time. Therefore, the variations of patterns of pitch for statements are

definitely going to differ from questions. The argument in support of that claim is that the rising patterns of pitch correspond of questions whereas falling edge patterns symbolize statements. Therefore, the intuition is that the grouping of questions with QUERY-YN, QUERY-W, ALIGN and CHECK was not very effective for prosodic model, as 50% of questions got confused with statements. Carletta *et al.* [3] defined CHECK as asking for confirmation for something that has already been stated previously. Any question asking for new information is not part of the CHECK category. Thus, it is evident that the identification of CHECK in discourse is solely dependent on the dialogue history, which is not available through the channel of prosody.

TABLE 17. VALIDATION DETAILS ON DIALOGUE ACT CLASSIFICATION OF 4 CATEGORIES USING PROSODY. TP=TRUE POSITIVE, FP=FALSE POSITIVE

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Questions	0.229	0.029	0.564	0.229	0.326	0.777
Replies	0.88	0.133	0.804	0.88	0.84	0.939
Statements	0.878	0.130	0.216	0.769	0.878	0.82
Interjections	0.216	0.004	0.592	0.216	0.317	0.932

It was also noticeable in Table 17 that interjections get confused with replies. Interjections are stand alone fillers *("uh", "ah", "um")* in conversations before a particular dialogue act is introduced. An example of interjections occurring in conversations is shown below:

IG (INSTRUCT): You want to go left.

IF (ACKNOWLEDGE): Okay

IG (INTERJECTIONS): umm.

IG (INSTRUCT): Then you turn around and go straight.

Figure 12. An example of interjections occurring in conversations.

However, when interjections occur in the middle of a dialogue act (e.g., "Yes..**umm**…go between those."), it is not considered an interjection. In the current map-task corpus, the occurrences of interjections into the reply type of dialogue acts (acknowledgement, for example) have been more prevalent, compared to other dialogue acts. This may be the potential reason as to why interjections are getting confused with replies. Also, for classification purposes, 2700 instances were used as examples for replies, whereas interjections had only 190 instances, which is only 6% of the total reply category. This explains why interjections are getting confused with replies more often, whereas the opposite has not been as frequent. Similar argument can be made to explain the confusion between statements and interjections. INSTRUCT and EXPLAIN, which are parts of statements, do contain interjections in them. The confusion between questions and replies and, questions and statements were also observed in Table 18. It is hypothesized based on the confusion patterns and the dialogue act taxonomy that a few of those confusions could be removed by introducing dialogue history and syntactical structures of discourse. Therefore, the obvious model to explore next was the discourse model with the existing 4 categories of dialogue acts.

TABLE 18. CONFUSION MATRIX FOR CLASSIFICATION OF 4 CATEGORIES OF DIALOGUE ACT USING PROSODY ONLY.

a	b	c	d	<-- classified as
42	1	122	29	a=interjections
3	234	210	573	b=questions
18	54	2442	260	c=replies
8	126	265	2871	d=statements

4.2.2. Four different dialogue acts using discourse

In this part of Model 2, only discourse features, mainly syntax and context, of dialogue acts were used to classify the 4 categories of dialogue acts, namely, questions, statements, replies and interjections.

TABLE 19. ACCURACY TO CLASSIFY 4 DIALOGUE ACTS BASED ON DISCOURSE (%).

Tree based classifier	Function based classifier	Ensemble based Classifiers			
Random Forest	*SMO*	*LogitBoost*	*Bagging*	*Ensemble selection*	*Multi scheme*
84.03	82.50	81.97	84.96	84.63	83.93

The classification accuracy for the 4 categories of dialogue acts look consistent across the tree based classifier, function based classifier and ensemble based classifiers, shown in Table 19. Even though, bagging itself provided approximately 85% accuracy in classifying dialogue acts using discourse, adding RandomForest, SMO with bagging in the ensemble was able to create the accuracy by little more than 1%, as shown in Table 20.

The measured precision, recall, f-measures, true and false positives rates in Table 21 were higher compared to the previous prosodic model, especially for questions and interjections.

TABLE 20. ACCURACY TO CLASSIFY 4 DIALOGUE ACTS BASED ON DISCOURSE USING EMSEMBLE FEATURE SELECTION FRAMEWORK.

Classifiers	Feature Selection Algorithm	Fusion Technique	Accuracy %
RandomForest	Consistency Subset Evaluator using Greedy Stepwise search algorithm	Average of Probability	85.61
Bagging	Chi Squared Attribute Evaluator using Ranker search Algorithm		
SMO	Cfs Subset Evaluator using Best First search Algorithm		
RandomForest	Cfs Subset Evaluator using Best First search Algorithm	Majority Voting	86.11
Bagging	Consistency Subset Evaluator using Greedy Stepwise search algorithm		
SMO	Chi Squared Attribute Evaluator using Ranker search Algorithm		
RandomForest	Chi Squared Attribute Evaluator using Ranker search Algorithm	Maximum Probability	84.61
Bagging	Cfs Subset Evaluator using Best First search Algorithm		
SMO	Consistency Subset Evaluator using Greedy Stepwise search algorithm		
RandomForest	Chi Squared Attribute Evaluator using Ranker search Algorithm	Stacking	85.35
Bagging	Cfs Subset Evaluator using Best First search Algorithm		
SMO	Consistency Subset Evaluator using Greedy Stepwise search algorithm		

TABLE 21. VALIDATION DETAILS ON CLASSIFICATION USING DISCOURSE ON 4 CATEGORIES.

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Questions	0.596	0.038	0.719	0.596	0.652	0.902
Replies	0.907	0.072	0.887	0.907	0.897	0.972
Statements	0.903	0.116	0.864	0.903	0.883	0.961
Interjections	0.608	0.006	0.752	0.608	0.672	0.962

Table 22 shows the confusion matrix for this particular model of classifying 4 categories of dialogue acts using discourse only. Similar patterns of confusion from the prosodic model using 4 categories of dialogue acts also persisted here. For example, INTERJECTIONS got confused with replies more than any other dialogue acts. Questions continued to get confused with statements, and replies. Statements were also seen to be confused with replies. However, all the confusions were comparatively less in number in discourse model compared

to the prosodic model. Questions were confused with statements 28% of the time, and 12% of the time replies, whereas 60% of the time questions would be recognized properly. Those numbers certainly suggest an improvement over the numbers generated using prosodic model, where 50% of the time questions would get confused with statements, 20% of the time with replies, and 30% of the time, it would classify questions properly. This improvement supports our hypothesis of introduction of discourse history and syntactical features would improve the accuracy as some of the categories are very context dependent.

TABLE 22. CONFUSION MATRIX FOR CLASSIFICATION OF 4 CATEGORIES OF DIALOGUE ACT USING DISCOURSE ONLY.

a	b	c	d	<-- classified as
118	2	67	7	a=interjections
2	608	122	288	b=questions
36	53	2516	169	c=replies
1	183	133	2953	d=statements

4.2.3. Four different dialogue acts using prosody and discourse

In the final part of Model 2, combination of prosodic and discourse features in feature level was used to classify the 4 categories of dialogue acts, namely, questions, statements, replies and interjections.

TABLE 23. ACCURACY TO CLASSIFY 4 DIALOGUE ACTS CATEGORIES BASED ON PROSODY AND DISCOURSE (%)

Tree based classifier	Function based classifier	Ensemble based Classifiers			
Random Forest	*SMO*	*LogitBoost*	*Bagging*	*Ensemble selection*	*Multi scheme*
81.75	83.45	83.93	84.91	84.27	81.99

The classification accuracies for the 4 categories of dialogue acts using the fusion of prosodic and discourse features are shown in Tables 23 and 24.

TABLE 24. ACCURACY TO CLASSIFY 4 DIALOGUE ACTS USING EMSEMBLE FEATURE SELECTION FRAMEWORK

Classifiers	Feature Selection Algorithm	Fusion Technique	Accuracy%
RandomForest	Consistency Subset Evaluator using Greedy Stepwise search algorithm	Average of Probability	84.40
Bagging	Chi Squared Attribute Evaluator using Ranker search Algorithm		
SVM	Cfs Subset Evaluator using Best First search Algorithm		
RandomForest	Cfs Subset Evaluator using Best First search Algorithm	Majority Voting	85.74
Bagging	Consistency Subset Evaluator using Greedy Stepwise search algorithm		
SVM	Chi Squared Attribute Evaluator using Ranker search Algorithm		
RandomForest	Chi Squared Attribute Evaluator using Ranker search Algorithm	Maximum Probability	84.18
Bagging	Cfs Subset Evaluator using Best First search Algorithm		
SVM	Consistency Subset Evaluator using Greedy Stepwise search algorithm		
RandomForest	Chi Squared Attribute Evaluator using Ranker search Algorithm	Stacking	84.78
Bagging	Cfs Subset Evaluator using Best First search Algorithm		
SVM	Consistency Subset Evaluator using Greedy Stepwise search algorithm		

The classification accuracies for 4 categories of dialogue act were very similar compared to the previous model of discourse. Bagging, as usual, had the highest classification performance of 84.91%. The combination of bagging, SMO and RandomForest, as expected, raised the classification accuracy to 85.74, particularly with majority voting and stacking methods. It is worth noting that the accuracy with discourse model with the best classifier was 86.11. Therefore, there was no performance improvement with the fusion of prosody and discourse in this model of dialogue act classification.

Table 25 provides the scores related to precision, recall, F-measures, true and false positives, and roc area for the classification model to recognize 4 different dialogue acts using prosody and discourse. Surprisingly the precision, recall, F-measures are very similar to what was observed in the previous model with discourse features, even though performance gain was expected. No performance boost with the fusion of prosody and discourse was also the case with Model 1. The reason is subject to further exploration with more data in the data. However, the lack of proper normalization techniques for data from two different sources might be the reason behind no performance enhancement after the fusion.

TABLE 25. VALIDATION RESULTS ON CLASSIFICATION ON 4 CATEGORIES USING PROSODY + DISCOURSE.

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Questions	0.553	0.036	0.718	0.553	0.625	0.896
Replies	0.912	0.078	0.879	0.912	0.895	0.972
Statements	0.899	0.123	0.857	0.899	0.878	0.957
Interjections	0.624	0.006	0.733	0.624	0.674	0.964

Table 26 provides insights about what categories of dialogue acts are getting confused with other ones for this particular model. A careful inspection of Table 26 suggests that the results are in very much synchronization with the previous model with discourse. Only noticeable difference is the system's performance degrades to recognize questions as prosody gets added to discourse. This once again motivates the problem of proper normalization schema of prosody and discourse information.

TABLE 26. CONFUSION MATRIX FOR 4 CATEGORIES OF DIALOGUE ACT CLASSIFICATION USING PROSODY AND DISCOURSE.

a	b	c	d	<-- classified as
121	1	66	6	a=interjections
2	564	132	322	b=questions
41	42	2530	161	c=replies
1	179	151	2939	d=statements

Apart from Model 1 and Model 2, a third model of classification was also proposed.

4.3. MODEL 3: RESULTS AND EVALUATION

In Model 3, the process from Model 2 is replicated with dialogue acts from IF only. As discussed earlier, the main motivation and interests in recognizing dialogue acts for IF stems from a current large multimodal communication project [34]. This research project explores how different modalities in face-to-face dialogues align with each other and tries to implement those rules extracted from human experiments in an ECA. In this project, the ECA is expected to interact with humans in a map-task environment, where the ECA plays the role of IG and a human plays the role of IF. Therefore, building models to recognize the dialogue acts of IF is significant in terms of development of the ECA. The final goal of this model was to apply machine learning techniques to classify 4 different dialogue acts for IF and then validate the results.

Table 27 provides distribution of 4 categories of dialogue acts instances for Model 2 (IG + IF) and Model 3 (IF). Statistical based classifiers are known to learn from the training examples. Therefore, the careful construction of training space with ample amount of data to capture all the variability is essential. It is also quintessential to have equal number of samples per category to assure a balanced classifier. However, maintaining equal number of

samples per category has a trade of not having enough samples per category. In this study, the models were built all the samples available per category. In model 2, it was evident that the precision, recall and F-measures for each category were correlated with categories with highest number of instances. For example, for Model 2, the highest precision, recall and F-measures were generated from the statements and reply categories. Therefore, it is expected that Model 3 would follow the same trend, by having the highest precision, recall and F-measures for replies and similar scores for the rest of the categories.

TABLE 27. THE OVERALL DISTRIBUTION OF INSTANCES IN MODEL 2 AND MODEL 3.

Categories	Model 2 (% instances)	Model 3 (% instances)
Questions	15.33	15.86
Statements	44.81	12
replies	37.20	70
interjections	2.64	1.44

4.3.1. Four different dialogue acts for followers using prosody

In this particular model, only prosodic correlates of dialogue acts were used to classify the 4 categories of dialogue acts for followers only.

TABLE 28: ACCURACY TO CLASSIFY 4 CATEGORIES OF DIALOGUE ACTS FOR FOLLOWER BASED ON PROSODY (%).

Tree classifier	Function classifier	Ensemble based Classifiers				
Random Forest	*SMO*	*Logit Boost*	*Bagging*	*Ensemble selection (forward selection+ Backward elimination)*	*Multi scheme*	*Multi-class Classifier using Random Forest as base classifier*
78.80	78.29	78.33	80.35	78.76	78.26	79.74

The classification accuracies for all the ensemble classifiers are comparable for recognizing the 4 categories of dialogue acts for followers only with an average of 79%. In particular, bagging, multi-scheme, majority voting and stacking provided the best results.

TABLE 29: ACCURACY TO CLASSIFY 4 DIALOGUE ACTS FOR FOLLOWERS BASED ON PROSODY USING EMSEMBLE FEATURE SELECTION FRAMEWORK.

Classifiers	Feature Selection Algorithm	Fusion Technique	Accuracy %
RandomForest	Consistency Subset Evaluator using Greedy Stepwise search algorithm	Average of Probability	79.50
Bagging	Chi Squared Attribute Evaluator using Ranker search Algorithm		
SVM	Cfs Subset Evaluator using Best First search Algorithm		
RandomForest	Cfs Subset Evaluator using Best First search Algorithm	Majority Voting	79.66
Bagging	Consistency Subset Evaluator using Greedy Stepwise search algorithm		
SVM	Chi Squared Attribute Evaluator using Ranker search Algorithm		
RandomForest	Chi Squared Attribute Evaluator using Ranker search Algorithm	Maximum Probability	78.72
Bagging	Cfs Subset Evaluator using Best First search Algorithm		
SVM	Consistency Subset Evaluator using Greedy Stepwise search algorithm		
RandomForest	Chi Squared Attribute Evaluator using Ranker search Algorithm	Stacking	78.88
Bagging	Cfs Subset Evaluator using Best First search Algorithm		
SVM	Consistency Subset Evaluator using Greedy Stepwise search algorithm		

Table 30 provides details on evaluation scores of recognizing 4 categories of dialogue acts for follower only using prosody and Table 31 provides the confusion matrix for that particular model. It is noticeable that interjections dialogue acts were misclassified as replies 86% of the time. The decrease in performance to recognize interjections and, instead, misclassify them with replies is, however, explainable. One reason being is the less number of samples, 37, for interjection in this model, whereas replies have 1812 number of instances.

Therefore, it is very likely for the classifier to misclassify interjections with the replies and not the other way around (0 instances from replies got confused with interjections as shown in Table 31). Also as expected, reply had the highest precision, recall and F-measures as they have highest number of instances in the distribution of categories.

Interjections are stand alone fillers in speech (e.g., *"uh", "um"*). They are not acknowledgements or replies in a conversation, rather are transition stages between one dialogue acts to another one. It is important to note that, acknowledgement or replies could contain similar linguistic units such as *"uh-huh", "aha", "um*" in different context of conversation. For example, for this current model of classifying 4 different dialogue acts of followers, IF have used "mhmm" 183 times and "uh-huh/um/uh" 83 times to acknowledge an instruction or explanation. It is very important to note the similar utterances were labeled as Interjections whenever they were used as stand alone fillers in a conversation with any context. Therefore, as prosody does not capture dialogue history information, the continuation of confusion of categories dependent on dialogue history or context, persisted.

TABLE 30: VALIDATION DETAILS ON CLASSIFICATION USING PROSODY ON FOLLOWER ON 4 CATEGORIES. TP=TRUE POSITIVE, FP=FALSE POSITIVE

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Questions	0.534	0.078	0.564	0.534	0.549	0.867
Replies	0.953	0.314	0.879	0.953	0.914	0.921
Statements	0.385	0.043	0.55	0.385	0.453	0.868
Interjections	0.026	0	1	0.026	0.051	0.845

In table 31, with prosodic model on follower only, it is shown that questions were getting confused with statements and replies 16% and 30% of the time, respectively. Statements also

continued to get confused with questions and replies; and replies getting confused with questions and statements.

TABLE 31: CONFUSION MATRIX FOR CLASSIFICATION OF 4 CATEGORIES OF DIALOGUE ACT FOR FOLLOWER USING PROSODY.

a	b	c	d	<-- classified as
219	67	124	0	a=questions
110	120	82	0	b=statements
56	30	1730	0	c=replies
3	1	33	1	d=interjections

4.3.2. Four different dialogue acts for followers using discourse

In this model, only discourse features related to syntax and context were used to classify the 4 categories of dialogue acts for followers only. It was expected to improve the accuracy achieved in the previous model by introducing discourse information related to syntactical features and dialogue history.

TABLE 32: ACCURACY TO CLASSIFY 4 CATEGORIES OF DIALOGUE ACTS FOR FOLLOWERS BASED ON DISCOURSE (%).

Tree based classifier	Function based classifier	Ensemble based Classifiers			
Random Forest	*SMO*	*LogitBoost*	*Bagging*	*Ensemble selection*	*Multi scheme*
83.81	84.05	85.24	84.51	84.51	84.25

Logitboost, a boosting method which uses linear regression, provided the highest accuracy among all the classifiers in Table 32 with an accuracy of 85.24%, whereas the average was 84.40%. The feature selection ensemble classification framework, with the combination of RandomForest, Bagging and SVM provided the highest accuracy for this mode. The majority

voting technique of predicting the outcome out from the individual classifier prediction yielded the highest accuracy of 86.36%, whereas the average was 85.43% in Table 33.

TABLE 33. ACCURACY TO CLASSIFY 13 DIALOGUE ACTS FOR FOLLOWERS BASED ON DISCOURSE USING EMSEMBLE FEATURE SELECTION FRAMEWORK.

Classifiers	Feature Selection Algorithm	Fusion Technique	Accuracy %
RandomForest	Consistency Subset Evaluator using Greedy Stepwise search algorithm	Average of Probability	84.97
Bagging	Chi Squared Attribute Evaluator using Ranker search Algorithm		
SVM	Cfs Subset Evaluator using Best First search Algorithm		
RandomForest	Cfs Subset Evaluator using Best First search Algorithm	Majority Voting	86.36
Bagging	Consistency Subset Evaluator using Greedy Stepwise search algorithm		
SVM	Chi Squared Attribute Evaluator using Ranker search Algorithm		
RandomForest	Chi Squared Attribute Evaluator using Ranker search Algorithm	Maximum Probability	85.52
Bagging	Cfs Subset Evaluator using Best First search Algorithm		
SVM	Consistency Subset Evaluator using Greedy Stepwise search algorithm		
RandomForest	Chi Squared Attribute Evaluator using Ranker search Algorithm	Stacking	85.48
Bagging	Cfs Subset Evaluator using Best First search Algorithm		
SVM	Consistency Subset Evaluator using Greedy Stepwise search algorithm		

This discourse model displayed similar trend of confusion among dialogue act categories, compared the prosodic model discussed in the previous section. One improvement, however, was noticed in terms of recognizing interjections. The F-measure for interjections was .452, which is much higher than the F-measure of .051 achieved using the prosodic model. Just like Model 2, introducing context features (using the previous five dialogue acts as features for a given dialogue act) was helpful to disambiguate the categories between replies and interjections, successfully.

TABLE 34. VALIDATION DETAILS ON CLASSIFICATION ON FOLLOWERS USING DISCOURSE ON 4 CATEGORIES.

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Questions	0.715	0.067	0.667	0.715	0.69	0.928
Replies	0.96	0.142	0.942	0.96	0.951	0.961
Statements	0.484	0.049	0.579	0.484	0.527	0.918
Interjections	0.368	0.004	0.583	0.368	0.452	0.899

Improvement was noticed in almost all categories using discourse compared to prosody. Performance degradation was noticed with replies getting confused with statements, and statements getting confused questions. Therefore, it was evident that prosody does help to differentiate between questions with a smaller subset of data, e.g., data used in Model 3.

TABLE 35. CONFUSION MATRIX FOR 4 CATEGORIES OF DIALOGUE ACT CLASSIFICATION FOR FOLLOWER USING DISCOURSE

a	b	c	d	<-- classified as
293	68	47	2	a=questions
121	151	39	1	b=statements
24	41	1744	7	c=replies
1	1	22	14	d=interjections

4.3.3. Four different dialogue acts for followers using prosody and discourse

In this model, prosody and discourse features were fused in feature level to classify the 4 categories of dialogue acts for followers only.

In Table 36, it is noticeable that SMO had the highest performance with 85.05%, whereas the average was 83.67%. However, the ensemble feature selection classification framework, as shown in Table 37, performed consistently with 85% of accuracy, by combining the predictions of individual classifiers using majority voting and stacking.

TABLE 36. ACCURACY TO CLASSIFY 4 CATEGORIES OF DIALOGUE ACTS OF FOLLOWERS BASED ON PROSODY AND DISCOURSE (%).

Tree based classifier	Function based classifier	Ensemble based Classifiers			
Random Forest	SMO	LogitBoost	Bagging	Ensemble selection	Multi scheme
82.95	85.05	83.69	84.27	83.81	82.29

TABLE 37. ACCURACY TO CLASSIFY 4 CATEGORIES OF DIALOGUE ACTS OF FOLLOWERS BASED ON PROSODY AND DISCOURSE USING EMSEMBLE FEATURE SELECTION FRAMEWORK.

Classifiers	Feature Selection Algorithm	Fusion Technique	Accuracy %
RandomForest	Consistency Subset Evaluator using Greedy Stepwise search algorithm	Average of Probability	83.92
Bagging	Chi Squared Attribute Evaluator using Ranker search Algorithm		
SVM	Cfs Subset Evaluator using Best First search Algorithm		
RandomForest	Cfs Subset Evaluator using Best First search Algorithm	Majority Voting	85.51
Bagging	Consistency Subset Evaluator using Greedy Stepwise search algorithm		
SVM	Chi Squared Attribute Evaluator using Ranker search Algorithm		
RandomForest	Chi Squared Attribute Evaluator using Ranker search Algorithm	Maximum Probability	83.85
Bagging	Cfs Subset Evaluator using Best First search Algorithm		
SVM	Consistency Subset Evaluator using Greedy Stepwise search algorithm		
RandomForest	Chi Squared Attribute Evaluator using Ranker search Algorithm	Stacking	85.36
Bagging	Cfs Subset Evaluator using Best First search Algorithm		
SVM	Consistency Subset Evaluator using Greedy Stepwise search algorithm		

The precision, recall, f-measure, roc curve areas for this model of recognizing 4 categories of dialogue acts for follower only are reported in Table 38. Replies, the category with highest number of instances, had the highest precision, recall and F-measure scores.

TABLE 38. VALIDATION RESULTS ON CLASSIFICATION FOR FOLLOWERS ON 4 CATEGORIES USING PROSODY + DISCOURSE.

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Questions	0.69	0.058	0.692	0.69	0.691	0.931
Replies	0.959	0.166	0.933	0.959	0.945	0.963
Statements	0.538	0.053	0.583	0.538	0.56	0.916
Interjections	0.184	0.002	0.583	0.184	0.28	0.94

The introduction of prosody with discourse for this model was helpful to increase the accuracy rate of statements (38% of accuracy with prosody, 48% of accuracy with discourse and 53% with the fusion). This model also helped to decrease the confusion between statements and questions compared to the previous models with prosody and discourse.

TABLE 39. CONFUSION MATRIX FOR 4 CATEGORIES OF DIALOGUE ACT CLASSIFICATION FOR FOLLOWER USING PROSODY + DISCOURSE

a	b	c	d	<-- classified as
283	72	53	2	a = questions
100	168	44	0	b= statements
24	48	1741	3	c = replies
2	0	29	7	d = interjections

4.4. OPTIMAL FEATURE SET EVALUATION

In this section, the outcome of the variety of feature selection algorithms and their evaluations are presented. The feature selection algorithms were the quintessential part of the proposed ensemble feature selection classification framework. In this framework, individual classifier in the ensemble was given distinct subset of the original feature sets. The different subsets of the original features were identified and evaluated through a variety of feature selection algorithms. A few such feature selection algorithms are Subset Evaluator (Best First), Chi Squared Attribute Evaluator (Ranker), and Consistency Subset Evaluator (Greedy

step wise) (for details on those, see section 3). Using SubSet evaluator, combination of four speech features, such as, *role (IG or IF), duration of the dialogue act, average value of the second formant, and speaking rate* were found to be the most important features and its performance was comparable to the model which employed more than 50 prosodic features. Chi Squared Attribute evaluator yielded features such as *speaking rate, duration of the dialogue act, ε_{time}, role, number of voice breaks in a speech act* as the optimal features with reasonable accuracy rate. Feature sets generated using Consistency Subset Evaluator were able to classify 14 different dialogue acts more than 50% of the time in average, using features such as *role, energy, F0 related statistics, statistics related to second and third formant, number of voice breaks, pauses, and number of rising and falling edges* in a dialogue act.

A similar procedure was employed to identify and evaluate the optimal discourse feature sets. For SubsetEvaluator feature selection algorithm, features such as, *role (IG or IF), number of words in each dialogue act, previous dialogue act, the first three sequences of the parts of speech of the dialogue act,* yielded comparable accuracy in compare to another model with more than 100 discourse features. Consistency Subset Evaluator, however, yielded the highest accuracy of distinguishing any of the 13 dialogue acts more than 70% of the time, using features such as *role, number of parts of speech (cardinal number, determiner, noun, verb, and adjective), number of words in each dialogue act, the first 5 sequence of the parts of the speech, previous two dialogue acts*.

A comprehensive list of all the features and their optimality is provided in Table 40, for completeness.

TABLE 40: THE PROSODY AND DISCOURSE FEATURES EXTRACTED FROM SPEECH ACTS AND THEN OPTIMAL FEATURES WERE IDENTIFIED

		Features	Optimal features
Prosody	Pitch	Minimum (pMin), Maximum (pMax), Mean (pMean), Standard Deviation (pSD), Absolute Value (pAV), Quantile (pQ), Unvoiced/Voiced frames of pitch (pUV).	pMin, pMax, pMean, pAB, pQ, pUV
	Intensity	Minimum (iMin), Maximum (iMax), Mean (iMean), Standard Deviation (iSD), Quantile (iQ)	iMin, iMax, iQ
	Formant	Average value of first formant (fVal1), second formant (fVal2), third formant (fVal3). Average bandwidth of first formant (fBand1) , second bandwidth (fBand2), third bandwidth (fBand3), Mean of first formant (fMean1), second formant (fMean2), third formant (fMean3), fMean2/fMean1, fMeanf3/fMean1, Standard deviation of first formant (f1STD), second formant (f2STD), third formant (f3STD), f2STD/f1STD, f3STD/f1STD	fVal2, fVal3, fBand1, fmean3, fMean3/fMean1, f1STD, f3STD
	Duration	duration of the speech act (d1), ε_{time}, ε_{height}	D1, ε_{time}
	Pauses	percent of Unvoiced Frames (pUF), # of Voice Breaks (#OVB), percent of Voice Breaks (pVOB), # of Pauses (nP), maximum duration of Pauses (mdp), average duration of pauses (adp) , total duration of Pauses (tdp)	#OVB, pVOB, nP, adp, mdp, tdp
	Rhythm	speaking rate or (1/voiced frames) (sr)	sr
	Edges	Magnitude of the highest rising edge (mhre), magnitude of the highest falling edge (mhfe), average magnitude of all the rising edges (amare) average magnitude of all the falling edges (amafe), # of rising edges (#re), # of falling edges (#fe).	#re, #fe
	Misc.	jitter (jt), shimmer (sh), energy (e), power (p), role	Role, energy
Discourse	Parts of speech sequence in each utterance [P1-P30]		P1, P2, P3, P4, P4, P5, P6, P7
	Number of words in an utterance [WC]		WC
	Previous speech acts (prev1, prev2)		Prev1, Prev2
	Parts of speech tagging		CD, DT, EX, IN, JJ, VB, VBN, WP, VBP

4.5. DISCUSSION AND FUTURE WORK

In this study, three different models of dialogue act taxonomies were studied, explored and evaluated using prosodic, discourse and their fusion, using Carletta et al. [3] map task taxonomy.

"What algorithm is going to be the most accurate for my classification problem?" is the classic question that researchers deal with for any classification problem. It has been argued in this study that the accuracy of a classifier is dependent on the dataset. For example, it has been shown in [50] that bagging classifier works well on discourse data, whereas boosting works well on prosodic data and when combined, SMO works the best. These inconsistencies of classifiers across different dataset do not add any significant value towards building a robust automatic classification system. In this experiment, therefore, purposefully, a set of classifiers was used across 3 different models of dialogue act classification. Model 1 contained the original Carletta *et al.* [3] map task taxonomy. In Model 2, the original map-task taxonomy was "collapsed" into a smaller subcategory, and Model 3 had the same "collapsed" sub-category of dialogue for instruction follower. For each model, evaluation scores of precision, recall, F-measures, true/false positives, roc curve area, and confusion matrices were reported for the best classifier. The main goal of setting up such a setting was to observe the performances of the classifiers across different taxonomies, and feature sets. Even though the accuracies of all the classifiers were comparable, ensemble feature selection classification framework with stacking method had the most consistent evaluation scores. This suggests that the ensemble classification framework suggested in Figure 11 provides the most consistent performance compared to any other classifiers used.

It was hypothesized in this study that the fusion of prosody and discourse would enhance the overall classification performance. But the actual numbers, after the experiment, did not fully support that hypothesis, as being shown in Table 41.

TABLE 41. COMPARISON OF PERFORMANCE FOR MODELS CREATED FOR PROSODY, DISCOURSE, AND DISCOURSE + PROSODY.

	Average Accuracy Across all the Classifiers (%)			Accuracy for the best Classifier (%)		
	Prosody	Discourse	Prosody + Discourse	Prosody	Discourse	Prosody + discourse
Model 1 (14 categories)	52.62	69.92	68.21	55.67	75.95	74.38
Model 2 (4 categories)	75.96	84.37	83.94	77	86.11	85.74
Model 3 (4 categories for IF	79.02	84.87	84.07	79.66	86.36	85.51

One possible explanation could be the lack of proper normalization schema to utilize the fusion of prosody and discourse information in feature level. For example, features generated from speech have completely different charactertics and dynamic ranges compared to discourse features. Discourse features, on the other hand, map context and syntactic properties related to syntax and dialogue history into numerals. Therefore, a proper normalization is absolutely necessary to properly synchronize the information coming from two different sources. However, proper normalization of features generated from two different modalities remains an open problem in the area of machine learning [55]. More experiments need to be conducted to understand how to best normalize speech and discourse features appropriately. Also, decision level fusion between prosody and discourse can also be explored in which normalization may not be as significant as it is in feature level fusion. It

has been noticed that discourse features are useful to disambiguate between certain categories of dialogue acts, where prosody fails, and vice versa. This once again motivates further exploration of this problem by fusing discourse and prosodic features in decision level by putting more weight on prosody on certain categories and discourse on other categories.

Empirical studies [29][30] have demonstrated that with an n-gram (uni-, bi- and tri-grams) model, approximately 60% accuracy was reported by using Support Vector Machine (SVM) and Hidden Markov Model, and rule based methods using discourse features only. Acoustic features of speech, however, only provided 43% accuracy in recognizing dialogue acts. Both the studies used Carletta *et al.* [3] map task taxonomy. The results reported in this study to classify the 14 dialogue acts using prosody was 55%, and with discourse 75%, for the best classifier. This definitely shows improvement of the proposed dialogue act classification schema over other prosodic models for dialogue act classification, previously implemented.

It is agreed that discourse alone yields better classification performance than prosody. However, using discourse features for a real time non-intrusive dialogue act classification system is impractical. One obvious reason is that effectiveness of discourse features in real time environment is contingent upon the performance of a speech recognition system. Studies [17] show that even the state-of-the-art speech recognizer could introduce up to 30% of word error rate for large vocabulary of conversational speech. Also computing discourse features requires computationally intensive syntactic and semantic search complexities [56][57], which may paralyze a real time system. Therefore, the performance boost of prosodic models for dialogue act classification certainly looks encouraging.

Model 3 of classifying 4 different dialogue acts for follower only was crucial in terms of building Embodied Conversational Agents (ECA). An ECA, in map-task environment, play

the role of IG as they interact with humans playing the role of IF. Therefore, being able to successfully recognize the dialogue act of IF would help the ECA to tailor a proper response. One surprising outcome was the better performance rate (79% of accuracy) of recognizing the dialogue acts for IF compared to both IG and IF (76% of accuracy). One possible explanation is the frequency distribution of dialogue acts for IF and IG is not uniform. For example, for IF, 50% of the dialogues contained ACKNOWLEDGE, whereas, dialogues of IG had 46% of INSTRUCTIONS. The average duration of INSTRUCTIONS for dialogues of IG was 2.34 seconds, and the average duration of ACKNOWLEDGEMENTS for IF was .52 seconds. More globally, the average duration of all the dialogues of IG was 1.85 seconds and for IF it was .85 seconds. Prosodic features captured from a smaller portion of speech often show more distinctive characteristics than features extracted from a bigger portion of the speech. Therefore, with utterances solely from IF, prosodic model was able to capture all the local variability, resulting in better classification accuracy to recognize dialogues of IF.

Dialogue act taxonomy anomaly for the working dataset could also be improved as part of future work. There were many cases where an utterance could not be labeled as one of the existing map task dialogue act category (for example, a statement like, “I am sorry” was labeled as miscellaneous). There were also cases where one utterance could potentially have multiple dialogue acts. An example of that is shown in Figure 13. In Figure 13, assuming that the speech file is not available, the fourth statement *“directly below them”*, when looked at with context seemed confusing as it could potentially be either CHECK, QUERY-YN or ACKNOWLEDGE.

IF (QUERY-W) : okay and i'm going in between them or below them

IG (REPLY-W) : below

IF (ACKNOWLEDGE): okay

IF (?) : directly below them (Could be CHECK? QUERY-YN? ACKNOWLEDGE?)

Figure 13. An ambiguous example of dialogue act coding without the speech.

Another interesting example could have been the ambiguity between EXPLAIN and INSTRUCT category. For example, *"You'll go in between a uh checkered car"* – could either be interpreted as EXPLAIN or INSTRUCT, depending on the perspective.

One major limitation of this study was the uneven distribution of instances per dialogue act. It is known to us that the statistical based classifiers learn from examples, and their performance is dependent upon the balanced training data set. Also, providing more data to train a classifier ensures learning all the variability of the data in order to provide robust performance. In this study, the number of instances used per dialogue act to train the three models was uneven. The main reason being some of the categories of dialogue act not occurring as frequent as others in the discourse. On the other hand, taking the minimum set of instances from each category to guarantee symmetry in the training data could have been an option. However, it would have come with the trade off of having less number of samples, incapable of spanning all the sub-areas of the training space. Future studies will incorporate adding more instances to the less represented categories towards a balanced classification scheme.

CHAPTER 5

CONCLUSION

This thesis investigates the automatic dialogue acts classification in multimodal communication using prosody, discourse and their fusion. The prosodic and discourse features, which were believed to be strong correlates of dialogue acts, have been extracted and the best features are selected using a variety of feature selection algorithms. To automatically classify the dialogue acts using machine learning techniques, three models of dialogue act classification were created. The models were created by "collapsing" the Carletta map-task taxonomy into smaller sub-groups. A variety of classifiers, including traditional and ensemble ones, were tested on the models (Model 1, 2 and 3) to compare their performances. A novel "Ensemble Feature Selection Classifier Fusion" technique has been implemented, to enhance diversity among the n-number of feature sets that were created. The n-number of feature sets was used as inputs to n-number of classifiers in the ensemble. The main motivation behind this framework was to make classifiers disagree in the decision making process and then using statistical methods to combine their predictions, e.g., majority voting. The results were validated by reporting by precision, recall, F-measures, roc area, true positive, and false positive for the best classifier for all the models created. A confusion matrix for each model was also reported.

The major claim behind this study is that a one-size-fits-all approach for algorithms and classifiers does not yield optimal performance. Instead, a combination of algorithms and classifiers is needed depending on the working dataset. This study also provides useful clues and a framework to ensemble multiple classifiers by varying the feature sets. Even though

the accuracies across the variety of classifiers did not vary a lot, the precision, recall, F-measures, true positive, false positive rates were comparatively better for ensemble based classifiers. Through the validation process, the claim has been made that "Ensemble Feature Selection based Classification" performs more consistently than the other classification models used in this study. Similarly, the results presented here show that discourse and prosodic features are intrinsically related, whereby for dialogue act classification, speech says as much about discourse, as discourse about speech.

References

[1] J. L. Austin, *How to Do Things with Words*. Oxford: Oxford University Press, 1962.

[2] J. Searle, "A Taxonomy of Illocutionary Acts," in *Minnesota Studies in the Philosophy of Language*, ed. K. Gunderson, pp. 334-369. J. Minnesota: Univ. of Minnesota Press, 1975.

[3] J. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, and A. Anderson, "The Reliability of a Dialogue Structure Coding Scheme," *Computational Linguistics*, vol. 23, no. 1, 1997, pp. 13-31.

[4] M. M Louwerse and S. Crossley, "Dialog Act Classification using n-gram Algorithms," *Proceedings of the Florida Artificial Intelligence Research Society International Conference (FLAIRS)*, Florida, USA, 2006. Menlo Park, CA: AAAI Press.

[5] J. C. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, and A. Anderson, HCRC Dialogue Structure Coding Manual (HCRC/TR-82). Edinburgh, Scotland: Human Communication Research Centre, Univ. of Edinburgh, 1996.

[6] M. Finke, M. Lapata, A. Lavie, L. Levin, L. M. Tomokiyo, T. Polzin, K. Ries, A. Waibel, and K. Zechner, "CLARITY: Inferring Discourse Structure from Speech," *Proceedings of the AAAI 98 Spring Symposium: Applying Machine Learning to Discourse Processing*, Stanford, CA, 1998, pp. 25-32.

[7] J. Alexandersson, N. Reithinger, and E. Maier, "Insights into the Dialogue Processing of Verbmobil," *Proceedings of the Fifth Conference on Applied Natural Language Processing*, *ANLP* 1997, Washington, DC, 1997, pp. 33-40.

[8] P. Taylor, S. King, S. Isard and H. Wright, "Intonation and Dialogue Context as Constraints for Speech Recognition," *Language and Speech*, vol. 41, no. 1-2, 1998, pp. 493-512.

[9] H. Hastie-Wright, M. Poesio and S. Isard. "Automatically Predicting Dialogue Structure Using Prosodic Features," *Speech-Communication*, vol. 36, pp. 63-79.

[10] M. L. Flecha-Garcia, “Eyebrow Raising and Communication in Map Task Dialogues,” *Proceedings of the 1st Congress of the International Society for Gesture Studies*, Univ. of Texas at Austin, TX, USA, 2002.

[11] M. M. Louwerse, N. Benesh, M. E. Hoque, P. Jeuniaux, G. Lewis, J. Wu, M. Zirnstein, “Multimodal Communication in Face-to-face Conversations,” *the 29th meeting of Cognitive Science Society*, Nashville, TN, 2007.

[12] J. Marineau, P. Wiemer-Hastings, D. Harter, B. Olde, P. Chipman, A. Karnavat, V. Pomeroy, A. Graesser, and the TRG. “Classification of Speech Acts in Tutorial Dialog,” *Proceedings of the workshop on modeling human teaching tactics and strategies at the Intelligent Tutoring Systems 2000 conference*, Montreal, Canada, 2000, pp. 65–71.

[13] D. J. Litman and K. Forbes-Riley, “Correlations between Dialogue Acts and Learning in Spoken Tutoring Dialogues,” *Journal of Natural Language Engineering*, vol. 12, no. 2, 2006, pp. 161-176.

[14] M. Rotaru and D. Litman. “Exploiting Discourse Structure for Spoken Dialogue Performance Analysis,” *Proceedings of EMNLP*, Sydney, Australia, 2006.

[15] G. Jackson, N. Person, and A. Graesser. “Adaptive Tutorial Dialogue in AutoTutor,” *Proceedings Workshop on Dialog-based Intelligent Tutoring Systems at ITS*, Maceio, Alagoas, Brazil, 2004.

[16] C. Shih and G. Kochanski, "Prosody and Prosodic Models," *7th International Conference on Spoken Language Processing*, Denver, Colorado, 2002.

[17] E. Shriberg, et al., “Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?” *Language and Speech*, vol. 41, no. 3-4, 1998, pp. 439-487.

[18] D. Surendran and G. Levow, “Dialog Act Tagging with Support Vector Machines and Hidden Markov Models,” *Proceedings of Interspeech*, Pittsburgh, PA, September, 2006.

[19] R. Fernandez and R. W. Picard, “Dialog Act Classification from Prosodic Features Using Support Vector Machines,” *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France, 2002.

[20] D. J. Litman and R. J. Passonneau, “Combining Multiple Knowledge Sources for Discourse Segmentation,” *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, MIT, Cambridge, MA, 1995, pp. 108-115.

[21] J. Hirschberg and C. Nakatani. “A Prosodic Analysis of Discourse Segments in Direction Giving Monologues,” *Proceedings of the 34th annual meeting on Association for Computational Linguistics* Santa Cruz, CA, 1996, pp. 286–293.

[22] M. Swerts and M. Ostendorf. “Prosodic and Lexical Indications of Discourse Structure in Human Machine Interactions,” *Speech Communication*, vol. 22, no. 1, 1997, pp. 25–41.

[23] J. Hirschberg and C. H. Nakatani, “Using Machine Learning to Identify Intonational Segments,” J. Chu Carroll & N. Green (Eds.), *Applying Machine Learning to Discourse Processing. Papers from the 1998 AAAI Spring Symposium.* Technical Report SS9801, pp. 52–59, Menlo Park, CA: AAAI Press.

[24] D. Jurafsky and J. H. Martin, “Speech and Language Processing: An Introduction to Natural Language Processing,” *Computational Linguistics and Speech Recognition*. Prentice Hall, New Jersey.

[25] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339-373.

[26] S. Wermter, and M. Löchel, "Learning Dialog Act Processing", Technical report, University of Hamburg, 1996.

[27] K. Ries, "HMM and Neural Network Based Speech Act Detection," *International Conference on Acoustics and Signal Processing (ICASSP '99)*, Phoenix, Arizona, USA, 1999.

[28] T. Fukada, D. Koll, A. Waibel, and K. Tanigaki, "Probabilistic Dialogue Act Extraction for Concept Based Multilingual Translation Systems," In Robert H. Mannell and Jordi Robert-Ribes, editors, *Proceedings of the International Conference on Spoken Language Processing*, volume 6, pp. 2771–2774, Sydney, December. Australian Speech Science and Technology Association.

[29] M. Louwerse and S. Crossley, "Dialog act classification using n-gram algorithms," *Proceedings of the 19th International Florida Artificial Intelligence Research Society*, Orlando, Florida, 2006.

[30] D. Surendran and G. Levow, "Dialog Act Tagging with Support Vector Machines and Hidden Markov Models," *Proceedings of Interspeech 2006* ICSLP, Pittsburgh PA, 2006.

[31] J. Edlund, J., and M. Heldner, "/nailon/ - Software for Online Analysis of Prosody," *Proceedings of Interspeech 2006* ICSLP, Pittsburgh PA, 2006.

[32] A. Tsymbal, M. Pechenizkiy, and P. Cunningham, "Diversity in Ensemble Feature Selection," Technical report, Trinity College Dublin, 2003.

[33] D. Opitz, "Feature Selection for Ensembles," *Proceedings of 16th National Conference on Artificial Intelligence*, 1999, pp. 379-384, AAAI Press.

[34] M. M. Louwerse, E. G. Bard, M. Steedman, X. Hu, and A. C. Graesser, "Tracking Multimodal Communication in Humans and Agents," *Technical report*, Institute for Intelligent Systems, University of Memphis, Memphis, TN, 2004.

[35] P. Boersma and D. Weenink, Praat: doing phonetics by computer (Version 4.6.01) [Computer program]. Retrieved May 16, 2007, from http://www.praat.org/

[36] A. Anderson, M. Bader, E. Bard, E. Boyle, G. M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson and R. Weinert, "The HCRC Map Task Corpus," *Language and Speech*, vol. 34, no. 4, 1991, pp. 351-366.

[37] C. E. Williams and K. N. Stevens, "Emotions and Speech: Some acoustical correlates," *JASA*, vol. 52, no. 4, 1972, pp. 1238-1250.

[38] R. Banse and K. R. Scherer, "Acoustic Profiles in Vocal Emotion Expression," *J. Personality and Social Psychology*, vol. 70, no. 3, 1996, pp. 614–636.

[39] S. Mozziconacci, "The Expression of Emotion Considered in the Framework of an Intonational Model," *Proceedings of ISCA Workshop of. Speech and Emotion*, 2000, pp. 45-52.

[40] R.W. Picard, "Affective Computing," MIT Press, Cambridge, 1997.

[41] M. E. Hoque, M. Yeasin, M. M. Louwerse, “Robust Recognition of Emotion from Speech,” *6th International Conference on Intelligent Virtual Agents*, Marina Del Rey, CA, August 2006.

[42] S. Kettebekov, M. Yeasin, and R. Sharma, "Prosody-based Audio Visual Co-analysis for Co-verbal Gesture Recognition," *IEEE Transaction on Multimedia*, vol. 7, no. 2, 2005, pp. 234-242.

[43] E. Brill, “A Simple Rule-based Part of Speech Tagger,” *Proceedings of the Third Annual Conference on Applied Natural Language Processing*, ACL, Morristown, NJ, USA, 1992.

[44] M. A. Hall, “Correlation-based Feature Subset Selection for Machine Learning,” Hamilton, New Zealand, 1998.

[45] I. H. Witten and E. Frank, “Data Mining: Practical Machine Learning Tools and Techniques,” 2 ed., San Francisco: Morgan Kaufmann, 2005.

[46] H. Liu, R. Setiono, “A Probabilistic Approach to Feature Selection - A Filter Solution,” *13th International Conference on Machine Learning*, 1996, pp. 319-327.

[47] J. Kittler, “On Combining Classifiers,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, March 1998.

[48] E. Bauer and R. Kohavi, “An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants”, *Machine Learning*,” vol. 36, no. 1- 2, 1999, pp. 105-139.

[49] D. Wolpert. “Stacked Generalization”. *Neural Networks*, vol. 5, no. 2, 1992, pp. 241–259.

[50] M. E. Hoque, M. S. Sorower, M. Yeasin, M. M. Louwerse, “What Speech Tells us about Discourse: The Role of Prosodic and Discourse Features in Dialogue Act Classification,” *IEEE International Joint Conference on Neural Networks (IJCNN)*, Orlando, Florida, 2007.

[51] J. Platt, “Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines,” Microsoft Research Technical Report MSR-TR-98-14, 1998.

[52] G. W. Wallace and S. Lawrence, “Efficient SVM Regression Training with SMO,” *Machine Learning*, vol. 46, no. 1-3, 2002, pp. 271-290.

[53] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, no. 2, 1996, pp. 123-140.

[54] A.K. Seewald, “How to Make Stacking Better and Faster While Also Taking Care of an Unknown Weakness”, *Nineteenth International Conference on Machine Learning*, 2002, pp. 554-561.

[55] A. Kapoor, H. Ahn, R.W. Picard, "Mixture of Gaussian Processes for Combining Multiple Modalities," *Proceedings of Multiple Classifier Systems*, Eds. N. C. Oza, R. Polikar, J. Kittler, and F. Roli, 6th International Workshop, MCS 2005, Seaside, CA, June 2005, pp. 86-96.

[56] R. Kompe, “Prosody in Speech Understanding Systems”, Springer-Verlag New York, Inc. Secaucus, NJ, USA, 1997.

[57] C. W. Wightman and M. Ostendorf, “Automatic Labeling of Prosodic Patterns,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, 1994, pp. 469–481.