DIALOGUE ACT DETECTION FROM HUMAN-HUMAN SPOKEN CONVERSATIONS

A DISSERTATION

Submitted in partial fulfilment of the
requirements for the award of the Degree of

Master of Technology

in

Computer Science and Engineering

(University of Calicut)

by

NITHIN T R

KMCT College of Engineering
Kallanthode, NIT P.O, Calicut

Department of Computer Science and Engineering

July 2013

KMCT College of Engineering
Kallanthode, NIT P.O, Calicut

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the thesis report entitled **Dialogue Act Detection from Human-Human Spoken Conversations** is the bonafide record of the Masters Research Project done by **NITHIN.T.R**(Register No: **CTALCCS012**) of fourth semester Computer Science & Engineering, KMCT College of Engineering, Calicut, towards the partial fulfillment of the requirement for the award of the Degree of Master of Technology by the University of Calicut.

Project Guide: **Head of the Department:**

Contents

List of Figures

List of Tables

Acknowledgements

I take this opportunity to acknowledge all the people who have helped me in every stages of the Masters Research Project.

It is the matter of great pleasure and satisfaction for me to express my sincere thanks and profound gratitude to **Dr. P. Janardhanan**, Principal KMCT College of Engineering, Calicut. I also express my sincere gratitude to **Mr. Pratap. G. Nair**, Dean of the Department of Computer Science & Information Technology, KMCT College of Engineering, Calicut.

I express my thanks to **Mr. Adarsh T K**, Head of the Department, Department of Computer Science & Engineering, KMCT College of Engineering, Calicut. I am extremely grateful to course coordinator and my internal guide **Ms. Swagatha**, Assistant Professor, Department of Computer Science, KMCT College of Engineering, Calicut, for her valuable suggestion and guidance.

Above all, I thank the almighty for enabling me to be what I am.

Nithin T R

Abstract

Understanding the human-human spoken conversation requires understanding not only of the semantic meaning of an utterance, but also of the intended meaning behind that utterance. Dialogue acts gives this intended meaning behind an utterance. This thesis presents an effective approach to improve the accuracy of dialogue act recognition from audio signal and transcription by combining prosodic and discourse features. The prosodic feature used is fundamental frequency (F0) of speech audio which correlates to pitch, and some key words corresponding to each of the dialogue acts are introduced and the number of keywords in an utterance is used as the discourse feature. This requires the segmentation of human-human dialogs into turns and utterances and the classification of each segment according to a DA tag. Utterance segmentation is done by detecting pauses between each utterance in a turn.

Chapter 1

Introduction

Whenever people speak, they express intentions for something. Everyone doesn't just talk to each other to exercise vocal cords, but rather they express an intention or meaning through their speech. Dialogue Acts gives the intended meaning of an utterance.The intention behind an utterance may be different from the structured sequence of words that the utterance contains.Language is not merely a collection of words to be processed and acted upon according to set rules. Quite apart from all the syntactic, grammatical and semantic difficulties encountered when trying to work out the structure of language, there are many other extraneous influences that govern any sentence; such as intonation, the character of the speaker, the occasion, the motives of the speaker, the use of jokes, sarcasm, metaphor, idioms and hyperbole (when words may not be taken at face value), etc. The understanding of language necessarily entails some understanding not only of the world, but also of the conversational 'context' of an utterance. This point is made by Reichman (1985), who says, "in addition to our knowledge of sentential structure, we have a knowledge of other standard formats (i.e. contexts) in which information is conveyed." Socio-linguists have put forward the idea that all speech can be seen as a variety of 'social action', such as greeting, promise or declaration, etc.

Dialogue acts are the underlying actions we perform when we speak [1]. Some examples are: INFORM, COMMAND, PROMISE, REFUSE, etc. Recognising the dialogue act that is being performed in the production of an utterance is important because it is the dialogue act that to some extent tells us what the speaker intends us to do with the propositional content of what he says. The identification of the dialogue act that is intended by the production of an utterance is vital then as it provides appropriateness constraints for our responses. This means that after every utterance, conversational expectations are created (either implicitly or explicitly) which serve us in understanding later conversation, in producing a relevant and appropriate response, and, very importantly, in being able to identify when and where a conversation goes wrong.

Not only this, but if we cannot understand the function intended by the production of a certain utterance, then we will also be unable to form opinions about the position of a speaker with respect to the content of his utterance. So, recognising dialogue acts could be essential for ascribing the correct beliefs and goals to a participant, for gleaning background knowledge of that participant and thus for being able to build on the knowledge gained from the current conversation in order to facilitate future interactions with that speaker.

Dialogue act interpretation can be modeled as a supervised classification task, with dialogue act labels as hidden classes to be detected. Machine-learning classi- fiers are trained on a corpus in which each utterance is hand-labeled for dialogue acts. The features used in dialogue act interpretation derive from the conversational context and from the act's microgrammar lexical, collocation, and prosodic features characteristic of the act [4], for example, used three kinds of features:

1. Words and Collocations: Please or would you is a good cue for a RE- QUEST, are you for YES-NO-QUESTIONs.

2. Prosody: Rising pitch is a good cue for a YES-NO-QUESTION. Loudness or stress can help distinguish the yeah that is an AGREEMENT from the yeah that is a BACKCHANNEL.

3. Conversational Structure: A yeah following a proposal is probably an AGREEMENT; a yeah after an INFORM is likely a BACKCHANNEL.

In this study first two features, i.e, words and collocations and prosody are used. A key task in human-human conversation is utterance boundary segmentation, the task of separating out utterances from each other. This is an important task since many computational dialogue models are based on extracting an utterance as a primitive unit. The segmentation problem is difficult because a single utterance may be spread over several turns , or a single turn may include several utterances.

Segmentation algorithms use boundary cues such as:

1. cue words: Cue words like well, and, so, that tend to occur at beginnings and ends of utterances.

2. N-gram word or POS sequences: Specific word or POS sequences that often indicate boundaries. N-gram grammars can be trained on a training set labeled with special utterance-boundary tags.

3. prosody: Utterance-final prosodic features like boundary tones, phrase-final lengthening and pause duration.

4. gaze: In face-to-face dialogue, gaze is an important cue. A related task in human-human dialog is diarization: assigning each utterance to the talker who produced it; this can be quite hard in multi-speaker meetings.

   In this study prosodic features are used as the boundary cues for utterance segmentation.

1.1 Background

Language is the “verbalisation of thought” and is neither straightforward to process nor to break down into manageable pieces. Humans themselves spend decades and many thousands of conversations to fully acquire all the intricacies of any given language. It may be that any system for computer language acquisition would require the same range of data input (although obviously computers are capable of much faster processing than human beings) for a similar learning process to take place.

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 give me a pen?" does not necessarily inquire whether person is capable of giving or not, but rather indirectly asks for the pen. 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. 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

When used in dialogue, as opposed to monologue, speech acts tend to exhibit adjacency pairs. That is, one particular type of speech act commonly follows another, such as questions are typically followed by answers. Speech acts are thus extended in dialogue to model the conversational functions of an utterance. To avoid confusion between the speech acts as used in monologue, the term dialogue acts is used to refer to this extended version used in dialogue.[4].

The term dialogue, as used in this thesis, refers to not only the spoken sentences but also the exchange of messages between two parties. The two parties may be either human-human or human-computer.The study is limited to to only two parties because that is how customer support is usually conducted. However, the methods we describe may be useful in dialogue with more than two participants. The messages that are exchanged during the dialogue are not necessarily sentence- based in the traditional sense; they instead contain one or more utterances, which are sometimes called non-sentential units.

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.

1.1.1 The Roles of the Speaker and Hearer

A speaker does not form his utterances using the only possible set of words for the 'correct' communication of his ideas, but packages what he says in a way he believes the hearer is most likely to understand in the context of the discourse situation.

If the speaker includes too much detail in his conversation then it becomes boring for the hearer, or the hearer might become overloaded by too much information and so be unable to process it to make a 'correct' interpretation; too little information on the other hand, will lead to ambiguity. Speech is thus constantly balanced between too much and too little information. A speaker is always vying for a hearer's attention and so must try to convey his message as simply as possible. Minimal specification is often the best strategy for speakers to follow.

This is often the way that children behave in conversation because they tend to believe that others (especially adults) are already aware of all the background information necessary to decode their message. (In fact this belief in very young children extends to all behaviour -they are incapable of deceit because of the assumption that the other person has complete knowledge of all that they themselves know.) It is interesting to note that minimal specification is often enough, and is easily expanded at need when extra information is required. This is negotiated between the participants in a conversation at the time the need for it occurs. If it becomes apparent that a hearer is unable to understand all that is said, the speaker can easily switch from a strategy of under-specification to that of over-specification (for example when a hearer's background information is inadequate to follow the references being made by the speaker, as in the case of an outsider joining a closely knit group of friends).Likewise, it is expected that the hearer will try to make sense of what he is hearing and cooperate in the process of communication.

For example, even in a simple referring phrase “In some ways, she’s very like Indira Gandhi” it is not always certain that the hearer will have interpreted the phrase correctly, even if the hearer correctly identifies ‘Indira Gandhi’ as ‘the former Indian prime minister’. In Fregean terms, he will have, because he has correctly associated the phrase with the individual. But that surely is not enough. The hearer must also have an exact copy of the background knowledge of the speaker concerning Indira Gandhi in order to be able to interpret what this means correctly. Even if the hearer infers the same attributes as the speaker intends, say Indira Gandhi’s standing up for her beliefs in a particularly aggressive and uncompromising manner, because the hearer believes that this is what the mutual friend under discussion is like, the hearer may still make the wrong assumption about what the speaker means. The hearer may think such traits admirable, whereas in actual fact the speaker is criticising their friend for them. Can the hearer really be said to have understood the speaker correctly? If we restrict our view of interpretation to just the correct ascription of sense and reference, then the answer would be yes. But plainly the hearer has failed to understand the speaker’s intention in producing the utterance.

1.2 Motivation

Dialogue acts are useful as they provide some semantic information about an utterance. If a system reliably tagged utterances in instant messaging with such dialogue acts, downstream tasks could be aided with this information. For example, a dialogue system needs to know if it was just asked a question or ordered to do something. However, messages received via instant messaging may contain more than one utterance, as demonstrated with the example ok, I will do it,. If a textual message contain more than one utterance, that message must be segmented into its utterances before classification.Dialogue models aim to account for the properties observed in dialogue, of which dialogue acts play an important part. We use the term dialogue modelling in this thesis to refer to the process of designing a dialogue model, whereas dialogue management serves to control and restrict the interaction.

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 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. Dialogue acts have also proven to be helpful as a unit of analysis in multimodal communication. 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 for time as a unit of analysis in multimodal communication.

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 . 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/sec2?", or "Can you repeat the question?" or "gravity equals 9.8 m/sec2, 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 that longer turns or statements (explanations, instructions) positively correlate with learning. Dialogue acts such as questions and feedback also known to maximize learning when used in appropriate context by the tutor.

Most companies that provide customer support are continuously trying to balance expensive human support costs, such as call centres, with cheaper methods such as FAQ pages on the World Wide Web and Interactive Voice Response (IVR) telephone systems. From a customer's perspective, the advantages of human support are clear: communication is easy, leading to questions being answered efficiently and problems resolved quickly. However, human support is expensive for many companies, sometimes prohibitively so. The lower cost of providing online resources, such as web sites, is attractive, but such resources require customers to search for their specific problem, which can be a frustrating process and inadequate when questions are not listed or not easily locatable. The desire to reduce costs whilst offering satisfactory support is fuelling significant commercial activity and research in high-quality, auto- mated support services.

1.3 Goals

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. 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. 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.

Empirical studies have demonstrated that discourse features provide satisfactory accuracy in classification of dialogue acts[16]. However, the successful performance of the discourse model in real-time environment is contingent upon the 100 percentage success rate of the speech recognition engines. Studies show that even the best speech recognition system can have up to 30 percentage 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. 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 keywords that are effective to identify some ambiguous dialogue act classes.

1.4 Methodological Outline

The proposed approach consists of five main components , namely, i) Segment the conversation automatically into utterances using pauses, ii) Manually checking to verify the automated segment of dialogues and then label as dialogue acts them using human experts, iii)Feature selection from speech and text data, iv) Train SVM to extract prosodic features v)Combine prosodic and discourse features,and detect dialogue acts.

1.5 Thesis Overview

The remainder of the thesis is organized as follows. Chapter 2 provides details about the literature survey.Chapter 3 defines the problems of the existing method.Chapter 4 represents the big picture of the proposed solution, with a detail description of prosodic and discourse features.Chapter 5 explains how the task was conducted and employed to collect data. Chapter 6 presents the experimental results.Chapter 7 shows the evaluation of results. Chapter 8 discuss the conclusions and future research direction.

Chapter 2

Literature Survey

Because of the importance of dialogue act classification within dialogue systems, it has been an active area of research for some time.Traditionally, the problem of identifying the different DA segments within an utterance has been approached in a separate fashion: first, DA boundary segmentation within an utterance was addressed with generative or discriminative approaches then, DA labels were assigned to such boundaries based on multi-classification [3]. Most of the attempts to detect dialogue acts include some basic steps such as dialogue act labeling using a corpus, extract different features,tagging dialogue acts as a multi-way classification, and the use of machine learning algorithms for classification. Dialogue act classification is performed using a probabilistic formulation, which helps to use a principled approach for combining multiple knowledge sources (using the laws of probability), as well as the ability to derive model parameters automatically from a corpus, using statistical inference techniques.

2.1 The Dialogue Act Labeling Task

In earlier work of A. Stolcke, [3] the domain chosen to model was the Switchboard corpus of human-human conversational telephone speech, distributed by the Linguistic Data Consortium. In the Switchboard speech corpus each conversation involved two randomly selected strangers who had been charged with talking informally about one of several, self-selected general interest topics. Traing of the statistical models on this corpus, human hand-coding of DAs for each utterance is combined, together with a variety of automatic and semiautomatic tools. The data consisted of a substantial portion of the Switchboard waveforms and corresponding transcripts, totaling 1,155 conversations.

2.1.1 Utterance Segmentation

Before hand-labeling each utterance in the corpus with a DA, it is needed to choose an utterance segmentation, as the raw Switchboard data is not segmented in a linguistically consistent way. To expedite the DA labeling task and remain consistent with other Switchboard-based research efforts, the work used a version of the corpus that had been hand-segmented into sentence-level units prior to our own work and independently of our DA labeling system . The units of this segmentation is referred as utterances. The relation between utterances and speaker turns is not one-to-one: a single turn can contain multiple utterances, and utterances can span more than one turn (e.g., in the case of backchanneling by the other speaker in mid-utterance). Each utterance unit was identified with one DA, and was annotated with a single DA label. The DA labeling system had special provisions for rare cases where utterances seemed to combine aspects of several DA types.

2.1.2 Tag Set

The previous work done by A. Stolcke, [3] used the modified DAMSL markup system, to make it more relevant to the corpus and task. Dialogue Act Markup in Several Layers (DAMSL) tag set, which was designed by the natural language processing community under the auspices of the Discourse Resource Initiative . DAMSL aims to provide a domain-independent framework for dialogue annotation, as reflected by the fact that our tag set can be mapped back to DAMSL categories. However, our labeling effort also showed that content-and task-related distinctions will always play an important role in effective DA labeling.

The resulting SWBD-DAMSL tag set was multidimensional; approximately 50 basic tags (e.g., QUESTION, STATEMENT) could each be combined with diacritics indicating orthogonal information, for example, about whether or not the dialogue function of the utterance was related to Task-Management and Communication-Management. Approximately 220 of the many possible unique combinations of these codes were used by the coders. To obtain a system with somewhat higher interlabeler agreement, as well as enough data per class for statistical modeling purposes, a less fine-grained tag set was devised. This tag set distinguishes 42 mutually exclusive utterance types and was used for the experiments. While some of the original infrequent classes were collapsed, the resulting DA type distribution is still highly skewed. This occurs largely because there was no basis for subdividing the dominant DA categories according to task-independent and reliable criteria.

The tag set incorporates both traditional sociolinguistic and discourse-theoretic notions, such as rhetorical relations and adjacency-pairs, as well as some more form-based labels. Furthermore, the tag set is structured so as to allow labelers to annotate a Switchboard conversation from transcripts alone (i.e., without listening) in about 30 minutes. Without these constraints the DA labels might have included some finer distinctions, but we felt that this drawback was balanced by the ability to cover a large amount of data.

2.1.3 Major Dialogue Act Types

The more frequent DA types used in earlier works are briefly characterized below.

1. Statements and Opinions

The most common types of utterances were STATEMENTS and OPINIONS. This split distinguishes "descriptive, narrative, or personal" statements (STATEMENT) from "other-directed opinion statements" (OPINION). The distinction was designed to capture the different kinds of responses we saw to opinions (which are often countered or disagreed with via further opinions) and to statements (which more often elicit continuers or backchannels).

2. Questions

Questions were of several types. The YES-NO-QUESTION label includes only utterances having both the pragmatic force of a yes-no-question and the syntactic markings of a yes-no-question (i.e., subject-inversion or sentence-final tags). DECLARATIVE-QUESTIONS are utterances that function pragmatically as questions but do not have "question form." By this we mean that declarative questions normally have no wh-word as the argument of the verb (except in "echo-question" format), and have "declarative" word order in which the subject precedes the verb.

3. Backchannels

A backchannel is a short utterance that plays discourse-structuring roles, e.g., indicating that the speaker should go on talking. These are usually referred to in the conversation analysis literature as "continuers" and have been studied extensively. Recognition of backchannels is useful because of their discourse-structuring role (knowing that the hearer expects the speaker to go on talking tells us something about the course of the narrative) and because they seem to occur at certain kinds of syntactic boundaries; detecting a backchannel may thus help in predicting utterance boundaries and surrounding lexical material.

4. Turn Exits and Abandoned Utterances

Abandoned utterances are those that the speaker breaks off without finishing, and are followed by a restart. Turn exits resemble abandoned utterances in that they are often syntactically broken off, but they are used mainly as a way of passing speakership to the other speaker. Turn exits tend to be single words, often so or or.

5. Answers and Agreements

YES-ANSWERS include yes, yeah, yep, uh-huh, and other variations on yes, when they are acting as an answer to a YES-NO-QUESTION or DECLARATIVE- QUESTION. Similarly, we also coded NO-ANSWERS. Detecting ANSWERS can help tell us that the previous utterance was a YES-NO-QUESTION. Answers are also semantically significant since they are likely to contain new information. AGREEMENT/ACCEPT, REJECT, and MAYBE/ACCEPT-PART all mark the degree to which a speaker accepts some previous proposal, plan, opinion, or statement. The most common of these are the AGREEMENT/ACCEPTS. These are very often yes or yeah, so they look a lot like ANSWERS. But where answers follow questions, agreements often follow opinions or proposals, so distinguishing these can be important for the discourse.

2.2 Feature Set Used

The following set of features either alone or in combination were used in the previous studies. The random variables used in dialogue modeling are
Sequence of DA labels (U).
Evidence,i.e.,complete speech signal (E).
Prosodic evidence (F).
Acoustic evidence, i.e., spectral features used in ASR (A).
Sequence of words (W).
Speakers labels (T).

2.2.1 Prosody

Prosody has been introduced in dialogue act classification to segment speech.[9]. Often pitch range, pause patterns, speaking rate, energy patterns, utterance duration, and patterns of the pitch contour provide useful clues about utterance segmentation.[16]. Prosodic features uses the evidence given by the acoustic features F capturing various aspects of pitch, duration, energy, etc., of the speech signal.

2.2.2 Transcribed words

This refers to the true (hand-transcribed) words spoken in a conversation. Use of discourse feauters such as POS, N-gram has been used in the previous works[6]. Recent works involves use of some keywords as discourse feauters for tagging dialogue acts, which were evolved from the field of emotion recognition [10] [11]. Feautures such as POS and

N-grams are focused on the 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 can also considered a crucial factor. Patterns of dialogue such as a question is normally followed by a reply, whereas, a properly executed instruction or explanation yields an acknowledgement, are extremely helpful to disambiguate intentions even though they may contain similar lexical information.[17]

2.2.3 Recognized words

This involves considering multiple alternative recognized word sequences.[3] For fully automatic DA classification, the above approach is only a partial solution, since it is not yet able to recognize words in spontaneous speech with perfect accuracy. A standard approach is to use the 1-best hypothesis from the speech recognizer in place of the true word transcripts. While conceptually simple and convenient, this method will not make optimal use of all the information in the recognizer, which in fact maintains multiple hypotheses as well as their relative plausibilities.

2.3 Multi-way Classification Problem

Once DA segments have been identified, tagging them according to their DA tag becomes a multi-way classification problem. In [5], combinations of word n-grams and prosodic features were deployed in a semi-supervised learning setting to assign a unique DA label to an utterance, assuming that the latter contained a single dialogue act.

2.4 Machine learning algorithms

Previous studies have used various machine learning algorithms to correlate prosodic and discourse features to various dialogue acts.[16] [17] Examples are the Markov Model, Hidden Markov Model,[15] Neural Networks, Self-Organizing Map Kohonen Networks, Support Vector Machine, Transformation-Based Learning, word-N-gram modelling, Polygram language model, Decision Tree, [3] Bayesian Networks, and Conditional Random Fields[6].

Early work on automatic dialogue act classification modelled discourse structure with hidden Markov model or SVMs [15] experimenting with lexical and prosodic features, and applying the dialogue act model as a constraint to aid in automatic speech recognition. Traditionally, the problem of identifying the different DA segments within an utterance has been approached in a separate fashion: first, DA boundary segmentation within an utterance was addressed with generative or discriminative approaches then, DA labels were assigned to such boundaries based on multi-classification [5] [8]. In [5], combinations of word n-grams and prosodic features were deployed in a semi-supervised learning setting to assign a unique DA label to an utterance, assuming that the latter contained a single DA.

2.4.1 Hidden Markov Model

The goal is to perform DA classification and other tasks using a probabilistic formulation, giving us a principled approach for combining multiple knowledge sources (using the laws of probability), as well as the ability to derive model parameters automatically from a corpus, using statistical inference techniques.[3] Given all available evidence E about a conversation, the goal is to find the DA sequence U that has the highest posterior probability P (U|E) given that evidence. Applying Bayes' Rule and find U* using the equation 2.1.

$$U^* = \underset{U}{\operatorname{argmax}} P(U|E)$$

$$= \underset{U}{\mathrm{argmax}} \frac{P(U).P(E|U)}{P(E)}$$

$$= \underset{U}{\mathrm{argmax}} P(U).P(E|U) \tag{2.1}$$

Here P (U) represents the prior probability of a DA sequence, and P (E|U) is the likelihood of U given the evidence. The likelihood is usually much more straightforward to model than the posterior itself. This has to do with the fact that our models are generative or causal in nature, i.e., they describe how the evidence is produced by the underlying DA sequence U .Estimating P (U) requires building a probabilistic discourse grammar, i.e., a statistical model of DA sequences. A computationally convenient type of discourse grammar is an n-gram model based on DA tags, as it allows efficient decoding in the HMM framework.The statistical discourse grammar models the prior probabilities P (U) of DA sequences. In the case of conversations for which the identities of the speakers are known (as in Switchboard), the discourse grammar should also model turn-taking behavior. A straightforward approach is to model sequences of pairs $(U_i; T_i)$ where U_i is the DA label and T_i represents the speaker.

2.4.2 Conditional Random Fields

As a learning algorithm, first-order linear-chain CRFs, a category of probabilistic learners frequently used for labeling and segmenting structured data[6] CRFs are undirected graphical models used to specify the conditional probability of assigning output labels given a set of input observations. A conditional probability distribution is defined over label sequences given a particular observation sequence (of e.g. DA surfaces), rather than a joint distribution over both label and observation sequences. CRFs simultaneously segment and assign labels to the tokens of an unsegmented, unlabelled input.

2.4.3 Support Vector Machines

A support vector machine (SVM) implements an approximation to the structural risk minimization principle in which both the empirical error and a bound related to the generalization ability of the classifier are minimized [17]. The SVM fits a hyperplane that achieves maximum margin between two classes, and its decision boundary is determined by the discriminant f(x) which is defined as

$$f(x) = \sum_i y_i \lambda_i K(x, x_i) + b \tag{2.2}$$

where x_i and $y_i \in \{-1, 1\}$ are the input-output pairs, K(x,y) $= \phi(x).\phi(y)$ is a kernel function which computes inner products, and $\phi(x)$ is a transformation from the input space to a higher dimensional space. In the linearly separable case $\phi(x) = x$. An SVM is genereralizable to non-linearly separable cases by first applying the mapping $\phi(.)$ to increase dimensionality and then applying a linear classsifier in the higher dimensional space.The parameters of this model are the values λ_i, non-negative constraints that determine the contribution of each data point to the decision surface, and b, an overall bias term.The data points for which $\lambda_i \neq 0$ are the only ones that contribute to the discriminant and are known as support vectors. Fitting an SVM consists of solving the optimization

$$maxF(\Lambda) = \Lambda.1 - 1/2\Lambda.D\Lambda \tag{2.3}$$

subject to the conditions $\Lambda.y = 0, \Lambda \leq C1, \Lambda \geq 0$. where $\Lambda = [\lambda_1...\Lambda_l]'$. and D is a symmetric matrix with elements $D_{i,j} = y_i y_j K(x_i, x_j)$ and C is a non-negative constant that bounds each Λ_i, and which is related to the width of the margin between the classes.Having solved Λ from the equation 2.3, the bias term can be found.

$$b = 1/2 \sum_i \lambda_i y_i (K(x_-, x_i + K(x_+, x_i)) \tag{2.4}$$

where x_- and x_+ are two correctly classified support vectors from classes -1 and $+1$ respectively.

Chapter 3

Problem Definition

The problem in accurate interpretation of speaker is that, a speaker has some meaning in mind when he produces an utterance; so, unless the speaker is completely irrational, the utterance can be said to have a single correct interpretation in the mind of the speaker, even when the meaning of the utterance is ambiguous in the context. Very rarely, as mentioned previously, a speaker will use knowledge shared with one of his hearers in order to communicate different meanings to different members of his audience, but this is not the normal way we communicate. So the question is, how does the hearer know he has identified the speaker's one correct interpretation of the utterance?

Standard techniques from statistical language modeling have been applied to the dialog act tagging task. One of the most common approaches uses n-grams to model the probabilities of DA sequences. The model proposed by uses bigrams and trigrams conditioned by the preceding DAs to predict the upcoming DAs, and a tagging accuracy of about 40 percentage was reported.

The best results in the previous works were obtained with large amounts of training data which are quite expensive to produce. All of these studies also have in common the fact that they try to detect one DA per utterance.The need of a simple way to do an automatic DA annotation using a small amount of training data and still get reasonable results will be make the task easier.

A recently proposed approach uses a discriminative approach, namely Conditional

Random Fields (CRFs),[6] to simultaneously segment an utterance into its DA boundaries and label such segments according to a DA tag using lexical features POS, N-gram , and CoNLL. In the previous approach syntactical structure of an utterance, the sequences or repetition of certain parts of speech were used to provide useful clues about the intentions of an utterance.The pattern adjacent pairs of utterances were also considered such as a question is normally followed by a reply, whereas, a properly executed instruction or explanation yields an acknowledgement.

It produced some ambiguous results of the dialogue acts opinion vs. non-opinion statements and agreements vs. acknowledgements, occurs while classifying the dialogue acts. This work uses discourse feautures, i.e, some keywords to detect the dialogue act such as acknowledgement, opinion, agreement, and expression.The SVM is used to label segments according to a DA tag where HMM, CRF were used in the previous works. Moreover, it takes a step beyond the previous work by including the use of prosodic features based on frequency variation to do utterance segmentation more effectively.

Chapter 4

Proposed Approach

The proposed approah uses prosodic and disourse features.It uses conversations from Buckeye corpus [18]for experiment.It uses Support Vector Machines for classification.14 dialogue act tags are used in the proposed approach.

The proposed approach consists of five main steps, namely,

i) Segment the conversation automatically into utterances using pauses.

ii) Manually checking to verify the automated segment of dialogues and then label as dialogue acts them using human experts.

iii)Feature selection from speech and text data.

iv) Train SVM to extract prosodic features.

v)Combine prosodic and discourse features,and detect dialogue acts.

In the proposed method , audio clip of human-human conversations and its transcriptions are used as input.The dialogue act recognition from both the input are performed using prosodic and discourse features respectively.After extracting prosodic and discourse features they are combined, giving more weightage to the discourse features.The flow chart of the dialogue act detection system is given in figure 4.1.

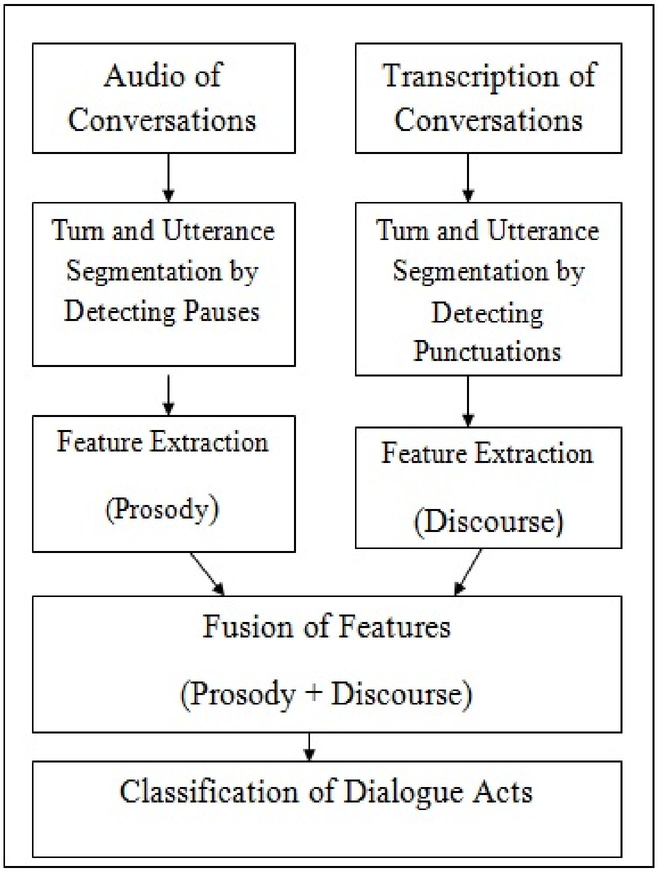

Audio of Conversations
Transcription of Conversations
Turn and Utterance Segmentation by Detecting Pauses
Turn and Utterance Segmentation by Detecting Punctuations
Feature Extraction (Prosody)
Feature Extraction (Discourse)
Fusion of Features (Prosody + Discourse)
Classification of Dialogue Acts

Figure 4.1: Dialogue Act Detection System

The proposed method adopts 14 DA tagsets from DAMSL [12] and ADAMACH taxonomy,[7] where utterances have been manually transcribed and annotated according to the DAs. The tag set used are shown in Table 4.1. Subsequent subsections briefly discuss each module of the proposed speech act classification system.

Table 4.1: Dialogue Acts and Description

Dialogue Act	**Description**
OPENING	greetings
CLOSING	farewells
THANKING	thanking and responding to thanks
YES-ANSWER	Affirmative answer to a question
NO-ANSWER	negative answer to a question
REQUEST	a question by the speaker
ACKNOWLEDGEMENT	demonstrated via continuer or assessment
WH-QUESTION	question starting with who, what, where, how, which,when, whom, why, ?
YN-QUESTION	question starting with did, do, does, are, was, were, is
OPINION	speaker's opinions
AGREEMENT	speaker's response to previous proposal
EXPRESSION	speaker's expression
STATEMENT	a claim made by the speaker
FILLER	gap fillers

4.1 Turn and Utterance Segmentation

The pauses in spoken words were used as the feature to detect the beginning and end of an utterance. For each turn from its starting point to end, checking significant amount of silence by counting continuous low frequencies.If the count value exceeds 20000,then it is determined that there is a pause.The utterance are automatically segmented using pause.Each utterances are then saved.These utterances are then retrieved to extract the features.

4.2 Feature Extraction

The basic algorithm for dialogue act segmentation and classification is very similar to sequence classification problem. Most of the dialogue act detection system uses the feature-based statistical classification with combination of several features using appropriate machine learning methods. In this paper the fundamental frequency variation for each utterance is studied, and using the measures of mean, variance, standard deviation of the fundamental frequency, which correlates to pitch, is used as the prosodic features. A small database of the keywords is formed and by spotting these keywords in each utterance the corresponding dialogue acts are classified. By combining the results from discourse features and prosodic features better results are obtained.The feature set used are shown in the table 4.2.

Table 4.2: The Prosody and discourse features extracted

Feature Type	**Feature**
Prosody	min, max, mean, variance, standard deviation of pitch
Discourse	keywords

4.2.1 Prosodic Features

Prosody contains speech related information, which is not entirely predictable at word or sentence level, by analyzing phoneme sequences . Speech features like pitch, energy, pauses, rhythm, formant, and intensity are called prosodic features. These features are independent of words and cannot be deduced from lexical channels. Prosody, therefore, provides valuable information about various dialogue acts that are difficult to disambiguate with only text [6]. 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. By 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.

Pitch is fundamental frequency of speech signal.[13] The pitch signal depends on the tension of the vocal folds and the sub glottal air pressure when speech is generated. The pitch signal is produced due to the vibration of the vocal folds. By analyzing the characteristics of pitch, this study considers subjective assessment of pitch frequency, statistical analysis of pitch mean, variance, and standard deviation, minimum value and maximum value. In this paper 'Cepstral method' of pitch extraction has been implemented. The analog signal is converted to .wav digital format by sampling with a suitable rate and quantized. The digital signal is then hamming windowed to convert it into a suitable frame size. The signal is converted into frequency domain by using Fast Fourier Transform. The absolute values of the signal are considered and then the logarithm of the signal is obtained. The signal is then transformed into Cepstral domain by taking its IFFT.[14]

4.2.2 Discourse Features

Discourse is one of the four systems of language, the others being vocabulary, grammar and phonology. Discourse has various definitions but one way of thinking about it is as any piece of extended language, written or spoken, that has unity and meaning and purpose. One possible way of understanding 'extended' is as language that is more than one sentence. Example, Something as short as two phrases in a conversation or as long as an entire extended essay are both examples of discourse and both show various features of discourse. In linguistics, a discourse particle is a lexeme or particle which has no direct semantic meaning in the context of a sentence, having rather a pragmatic function, it serves to indicate the speaker's attitude, or to structure their interactions with other participants in a conversation. Discourse particles are primarily a feature of spoken language; in written language they indicate an informal or jocular tone.

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. Discourse provides context information often not available through prosodic channels. Syntactical structure of an utterance, the sequences or repetition of certain keywords could provide useful clues about the intentions of an utterance. A detailed study about use of such keywords is conducted and a database of keywords words for each dialogue acts is formed. The presence of such keyword in an utterance is also considered as a crucial factor. In total, 64 keywords were assigned. We checked the occurrence of keywords over 14 dialogue acts in manually-labeled dialogues and found that all keywords were found in the conversations. These keywords are spotted in each utterance using mean of their ASCII value.By comparing ASCII values of each words in an utterance presence of a keyword is identified. There set of keywords used are summarized in the Table 4.3. The use of discourse features and prosodic features can reduce the ambiguity between opinions vs. statement, agreement vs. statement etc.

Table 4.3: Dialogue Acts and Keywords

Dialogue Act	**Keywords**
OPENING	hi, hello, welcome, good morning/evening/afernoon
CLOSING	see you, bye, good night
THANKING	thanks, thank you
YES-ANSWER	yes,hmm
NO-ANSWER	no, not, n't
REQUEST	may, can, could, shall, will, should, would, please
ACKNOWLEDGE	great, okay, sorry
WH-QUESTION	who, what, where, how, which,when, whom, why, ?
YN-QUESTION	did, do, does, are, was, were, is
OPINION	in my view, to my mind, to be honest, I think as far as I'm concerned, It seems, I would argue, I beleive
AGREEMENT	of course, agree, right
EXPRESSION	Is it, so
STATEMENT	this, that
FILLER	you know, shhh, hmmm, uh

4.3 Algorithm

The algorithm of the modules are as follows:

1. Read Audio File
2. Finding length for initial silence removal.
3. Detecting less amplitude, i.e silence and do turn segmentation
4. For each turn from its starting point to end, checking significant amount of silence to detect a pause for utterance segmentation.

5. Check significant amount of silence by counting continuous low frequencies.If the count value exceeds 20000,then it is determined that there is a pause.

6. Compute a power spectral estimate of each turn for the analysis Using melscale, based on pitch perception.

7. Windowing analysis region with a hamming window.

8. Compute square magnitude

9. Compute fast Fourier transform

10. Compute mean,variance , standard deviation, minimum ,maximum value of the signal.

11. Train SVM to group utterances according to prosodic features

12. Obtain the mean value of trained data.

13. Read Transcription file in .txt format.

14. Convert all the alphabets into small letters

15. Turn segmentation is done by identifying :

16. Utterance segmentation is done by identifying . or ?

17. Read the text file containing keywords

18. Check the presence of the keyword in an utterance by comparing its ASCII values and extract it.

19. Combining both prosdic and discourse feaures, more weightage is given to discourse features.

20. Detecting dialogue acts category using the features.

Chapter 5

Task Seutp

For one trying to investigate features of spoken language, there are really only two paths to follow; one must either use an available spoken dialogue corpus, or collect and transcribe one's own data. The former option is obviously preferable for time- and labour-saving reasons. However, there are many obstacles to be overcome if one is to use speech corpora for research into spoken language phenomena. In this next section, I will consider the problems attendant on these various options, and indicate at each stage the grounds for my choices. I shall first discuss the available spoken language corpora, then the collection of my own data for transcription.

5.1 Speech Corpora

At the start of this research, I had hoped to use existing spoken dialogue corpora for my investigations into speech act use and dependencies. However, this proved to be less easy than I thought. To begin with, the numbers of freely (or at least reasonably freely) available spoken language corpora are relatively few. I detail here the main relevant (English) corpora that were found, and the reasons why they did not fulfil my requirements.

5.1.1 Linguistic Data Consortium

The United States Defence Department's Advanced Research Projects Agency (DARPA) made a policy change in 1986 in an effort to centralise resources for its speech research programs. The success of this data sharing led to rapid progress in the areas of speech recognition, message understanding, document retrieval, and machine translation. The Linguistic Data Consortium (LDC) was founded on the back of this success in 1992, and has provided a useful forum and resource for large-scale development and widespread sharing of data for both academic and industrial researchers in emerging linguistic technologies. The main spoken language corpora available from the LDC that are directly relevant to my research:

SWITCHBOARD: The SWITCHBOARD corpus consists of about 2400 telephone conversations of 6-minute duration. There were 543 different speakers (302 male and 241 female) from all over the United States, who were strangers to each other. The callers would dial into an automated telephone exchange, and a computer-driven operator system would prompt the caller for recorded information (such as selecting a topic about which the participants would speak) and select an appropriate 'callee'. The speakers could then talk to each other until they finished their conversation about the topic and ended the call. There were about 70 different topics from which to choose; the only constraints on the participants were that they would only talk to the same person, and use the same topic, once.

To obtain the standard corpus is difficult, as it is not available free of charge. This reflects the fact that such corpora are extremely expensive to collect in terms of man-hours; huge quantities of data are required to satisfy the demands of speech processing computer programs, in order to build robust lexicons and grammars. Not only is obtaining the data in the first place costly, but once it has been amassed, there are the additional costs of transcription, documentation, maintenance and distribution.

5.1.2 Collecting Recordings

The audio files of conversations from BBC learning English website are downloaded and converted into suitable format.The transcription of the conversations are prepared manually.

5.1.3 The Buckeye Corpus

For the standard some conversations from, Buckeye Corpus,[18] which is freely available is also used. Dialog act detection is the task of identifying the function or goal of a given utterance: thus, it provides complementary information to the identification of domain concepts in the utterance, and a domain independent dialog act scheme can be applied. For training and testing purpose some conversations from the Buckeye corpus [7] of spontaneous American English speech, a 307,000-word corpus containing the speech of 40 talkers from central Ohio, USA is used. For the current study, 5 conversations were randomly taken from the corpus totalling 45 minutes. The 5 conversations had 10 participants in total. The gender distribution of the participants is about 60 percentage female and 40 percentage male.

5.2 Transcription

The transcription process itself throws up many representational difficulties.There is significant loss of information in the process of transcribing a conversation into some kind of orthographic representation. We lose certain features such as voice quality, whether the speaker is interested, amused, bored, etc. (i.e. attitude of expression), volume (in a heated debate for instance), emotion, intonation (although this can be represented perhaps in a very crude way by punctuation), speed, pauses, hesitations, stutters, mispronunciations, as well as elision, assimilation, omission of sounds, or lengthening of syllables, etc. We also lose all visual and physical contextual clues, such as the speaker's facial expression, as I have already mentioned.

The process of transcription is by its very nature based on the speaker, which ignores what goes on with the listener (for example when signalling a wish to speak). But the act of listening may well be vitally linked to planning the hearer's next contribution as speaker. This is one of the reasons why multi-modal analysis is so much in vogue these days, so that researchers can take into account and have access to the visual context as well as just the audio reproduction.

Perhaps, as I mentioned earlier, there is an argument for studying telephone conversations in the analysis of purely spoken interaction, as this medium naturally eschews visual signals. So getting a transcription in standard format is also difficult.In this study We use a freely available sample transcription from the SWITCHBOARD corpus.By using this as a standard some transcriptions of conversations are also generated manually.Standard good quality audio of spoken conversations and its transcriptions are freely available in the BBC learning English website andIn this research work I also used some of them.The transcriptions are loaded in the text format.

5.3 Speech/Dialogue Act Annotation Schemes

There are an increasing number of competing speech act labelling and classification schemes in the computational linguistic community. This fact reflects the enormous interest that has been generated in recent years for analysing language at a functional level, with the idea of aiding in the design of human-computer dialogue systems. Often in the literature relating to dialogue management these are called dialogue acts rather than speech acts because the medium of interaction can be either written or spoken.

There have been a variety of attempts to come up with a definitive set of dialogue act labels. The main problems are that each different scheme was influenced by the theory behind it, and also by its application and use (i.e. what the developers wanted to use each scheme for). The results seem to be that each scheme differs from the rest just enough so that a direct mapping from one scheme to another is impossible. A comparison of the different dialogue act schemes shows the expected similarities between those that are used for equivalent or comparable domains; surprisingly, sometimes schemes that have totally different functional backgrounds show a considerable overlap. There is no way to tell however, whether this is because the design of one scheme has had an influence on the design of another. Trying to map all the representations onto each other, or amalgamate the various acts left unaccounted for in some schemes, into one overarching, generalised scheme causes problems at a theoretical level, because phenomena that should be considered at different levels of abstraction are conflated. In this research we used the dialogue act annotation scheme from DAMSL,TRAINS and DIT++. A Total of 14 dialogue acts were extracted from these schemes.

5.3.1 DAMSL and (SWITCHBOARD) SWBD-DAMSL

The DAMSL annotation scheme was influenced by the design of VERBMOBIL, and aimed at producing a generic, standard tag-set, from which specific dialogue act schemes could be developed for task-specific domains (Core 1998). This work is also related to work in plan recognition in that the development of the annotation schemes was strongly influenced by researchers in dialogue systems such as TRAINS (the simulated TRAINS dialogues were used to test the DAMSL scheme for instance). Allen, for example, was one of the two authors of the manual for annotation using the DAMSL scheme.[12]

Although the DAMSL scheme has stemmed originally from an intention-based background, it also borrows heavily from work in structure-based methods of analysis. It contains theories which categorise dialogue acts according to whether they are initiative, forward-looking, or responsive, backward-looking.

Forward-Looking-Function

An utterance has a forward-looking function if its production has an effect on the subsequent dialogue. The different types of tags for this category of dialogue act are sown in table 5.1.

Table 5.1: Forward-Looking-Function

Dialogue Act	**Meaning**
STATEMENT	A claim made by the speaker
ASSERT	A claim intended to be believed by hearer
REASSERT	A claim repeated by the speaker
OTHER	A comment made by the speaker
OPEN-OPTION	A weak suggestion or listing of options
ACTION-DIRECTIVE	An actual command
INFO-REQUEST	Request for information
OFFER	Speaker offers to do something
COMMIT	Speaker commits to doing something
CONVENTIONAL	Conventional formulations
OPENING	Greetings
CLOSING	Farewells
EXPLICIT-PERFORMATIVE	Dialogue act named by the verb
EXCLAMATION	An exclamation (of surprise, etc.)
OTHER-FORWARD-FUNCTION	Any other forward-looking function

Backward-Looking-Function

An utterance has a backward-looking function if it refers back to a previous utterance in a dialogue. The different types of tags for this category of dialogue act are given in table 5.2.

Three main groups of dialog acts had been identified in the previous works.

1. Core Acts: which represent the fundamental action performed in the dialogue, requesting and providing information, or executing a task. These include initiatives (often called forward-looking acts) and responses(backward-looking acts);

2. Conventional/Discourse management Acts: which maintain dialog cohesion and delimit specific phases, such as opening, continuation, closing, and apologizing.

3. Feedback/Grounding Acts:used to elicit and provide feedback in order to establish or restore a common ground in the conversation.

Our taxonomy follows the same threefold partion, summarized in the table 5.3.

Table 5.2: Backward-Looking-Function

Dialogue Act	**Meaning**
AGREEMENT	A speaker's response to a previous proposal
ACCEPT	Accepting the proposal
ACCEPT-PART	Accepting some part of the proposal
MAYBE	Neither accepting nor rejecting the proposal
REJECT-PART	Rejecting some part of the proposal
REJECT	Rejecting the proposal
HOLD	Putting off response, usually via sub-dialogue
UNDERSTANDING	Whether speaker understood previous utterance
SIGNAL NON UNDERSTANDING	Speaker did not understand
SIGNAL-UNDERSTANDING	Speaker did understand
ACKNOWLEDGE	Demonstrated via back-channel or assessment
REPEAT-REPHRASE	Demonstrated via repetition or reformulation
COMPLETION	Demonstrated via collaborative completion
ANSWER	Answering a question
INFORMATION-RELATION	How the content relates to previous content

Table 5.3: Dialogue Act Taxonomy

Groups	**Dialogue Acts**
Core Acts	statement,request,wh-question, yn-question, opinion, expression
Conventional/Discourse management Acts	opening,closing, thanking
Feedback/Grounding Acts	acknowledgement,filler, yes-answer, no-answer, agreement

5.4 Implementation

The algorithm is implemented using MATLAB.Support Vector Machines(SVM) are used for classification of features. SVM Structure is ceated using MATLAB command SVM-Struct. Support Vector Machines are used to train using prosodic features. A support vector machine (SVM) implements an approximation to the structural risk minimization principle in which both the empirical error and a bound related to the generalization ability of the classifier are minimized. The SVM fits a hyperplane that achieves maximum margin between two classes, Support Vector Machine (SVM) was also taken into consideration for its robust performance in speech act classification based on previous studies [15] [17]. 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 had been practical due to the small dataset. For larger dataset Sequential Minimal Optimization (SMO) can be used as a fast method to train SVMs.The experiment is conducted using an audio file in wave format and its transcription in text format. The algorithms were implemented using MATLAB. Using SVM, the mean, variance, standard deviation, maximum value, minimum value etc., of the prosodic feature, i.e., the fundamental frequency and the discourse feature the keywords are combined.SVM is trained in a two way step. In the first step the classification using prosodic features are performed, and then the discourse features are also incorporated. The testing is also done using the casual conversations.

5.5 Software Requirements

Matlab Version 7.7 (R2009b) 7.9.0.529 64-bit(win-64) is the tool used for implementation.

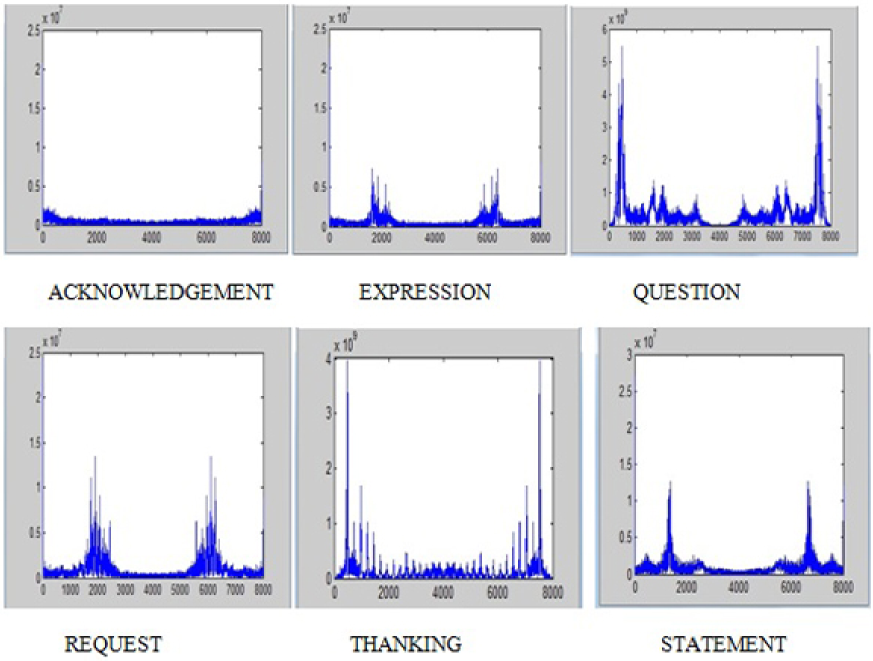
5.6 System Requirements

Operating System : Windows (Windows XP, Windows Server 2003, or Windows Vista), Linux (Red Hat Enterprise Linux v.4 and above, Fedora Core 4 and above, or Debian 4.0 and above). Processors : Intel Pentium (Pentium 4 and above), Intel Celeron, Intel Xeon, Intel Core, AMD Athlon 64, AMD Opteron, or AMD Sempron. Disk Space : 510MB. RAM : 1GB.

Chapter 6

Experiments and Results

The experiment was conducted in three steps. In the first step the turn and utterance segmentation is done using prosodic features, then the pitch variations of each utterances are studied.The pitch variations of 6 different dialogue acts are shown in the figure 6.1.

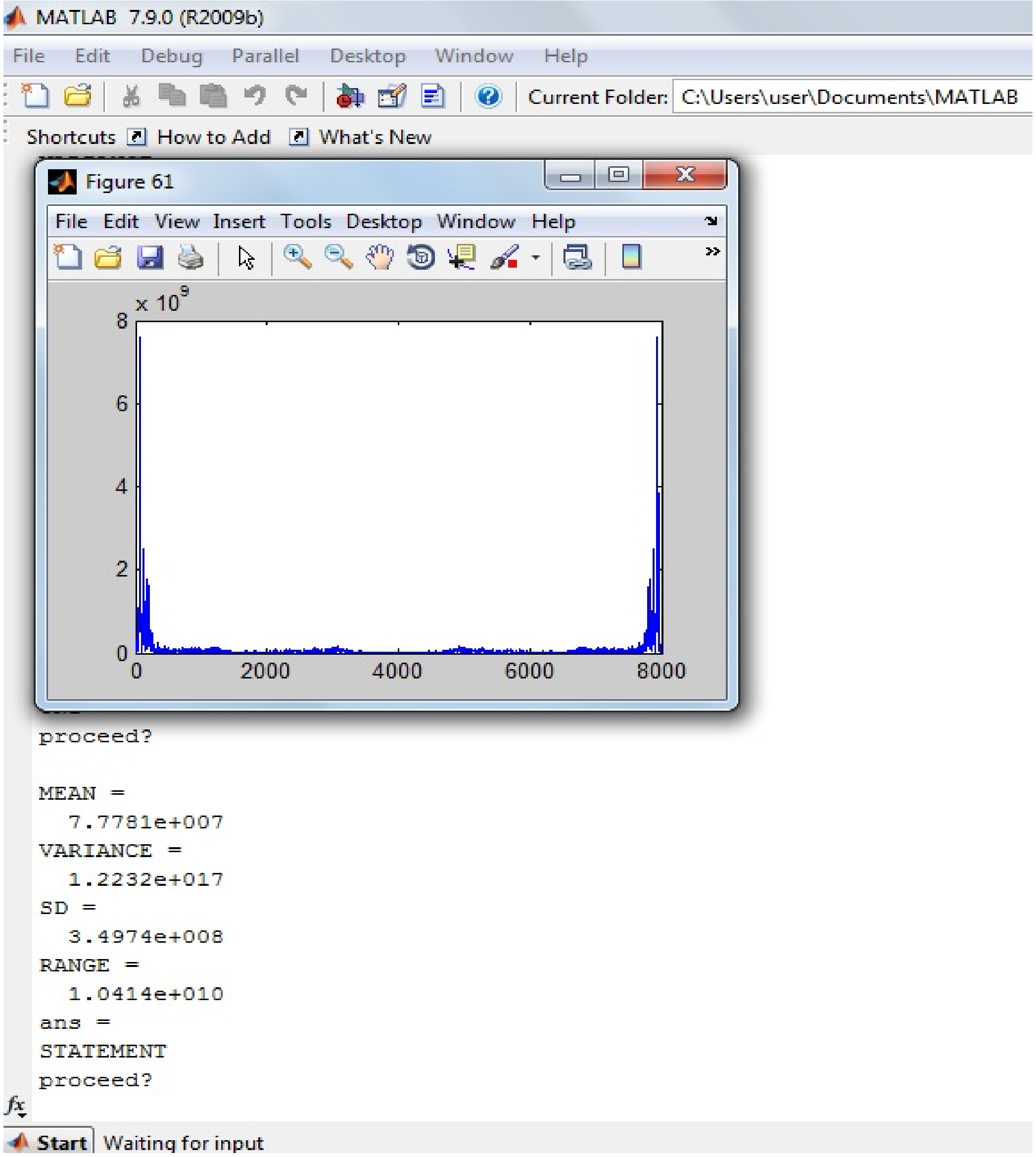


ACKNOWLEDGEMENT EXPRESSION QUESTION

REQUEST THANKING STATEMENT

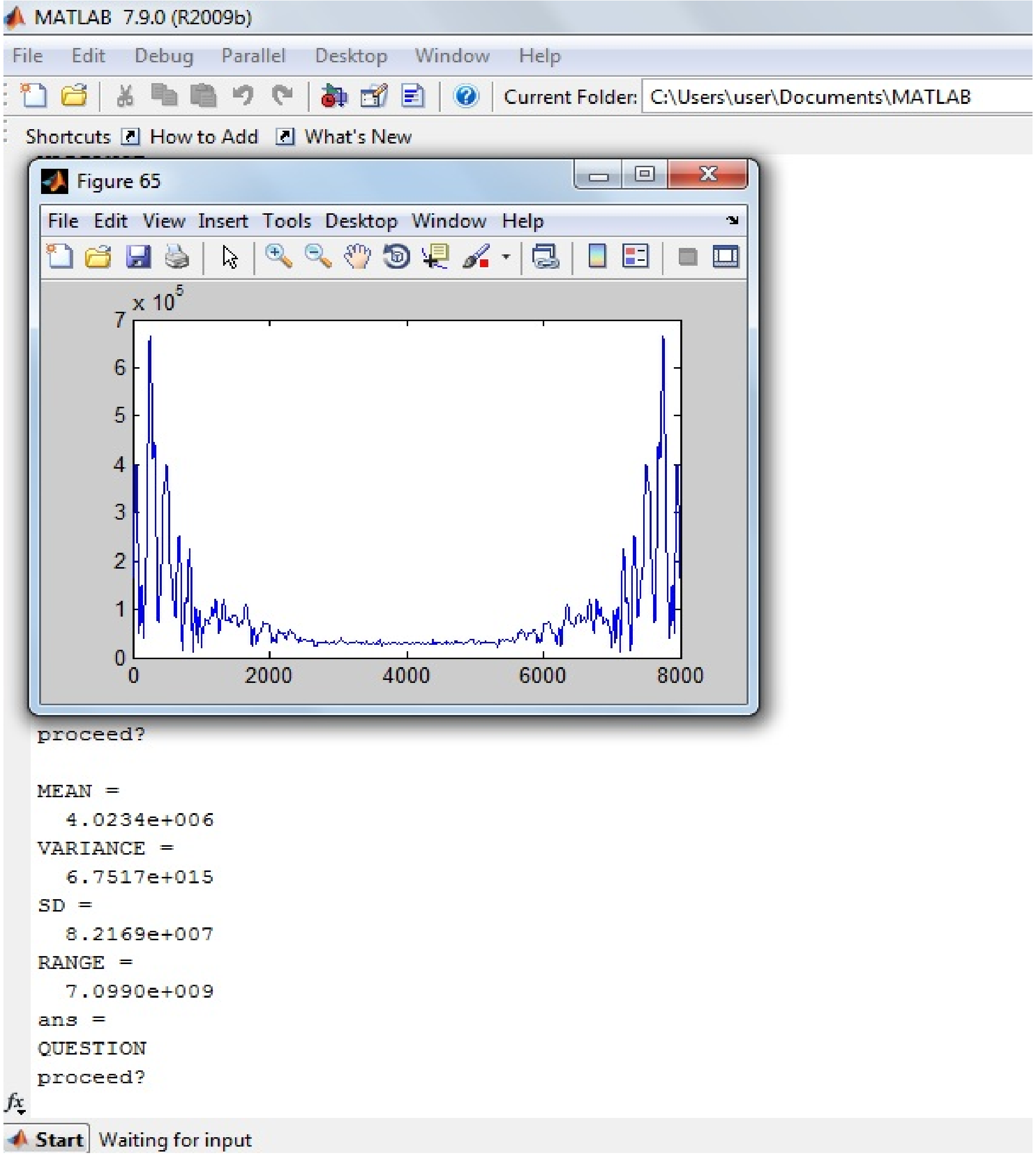
Figure 6.1: Pitch variations of some dialogue acts

Dialogue act classification using only the prosodic features are performed.An audio file in the .wav format is given as the input. Six categories of dialogue acts are successfully identified. According to the intonation pattern of frequency six dialogue acts such as request, question, thanking, statement, acknowledgement, expression are identified.The results obtained using matlab is shown in figures 6.2, 6.3 and 6.4.


MATLAB 7.9.0 (R2009b)
File Edit Debug Parallel Desktop Window Help
Current Folder: C:\Users\user\Documents\MATLAB
Shortcuts How to Add What's New
Figure 61
File Edit View Insert Tools Desktop Window Help
x 10^9
8
6
4
2
0
0
2000
4000
6000
8000
Start
Waiting for input

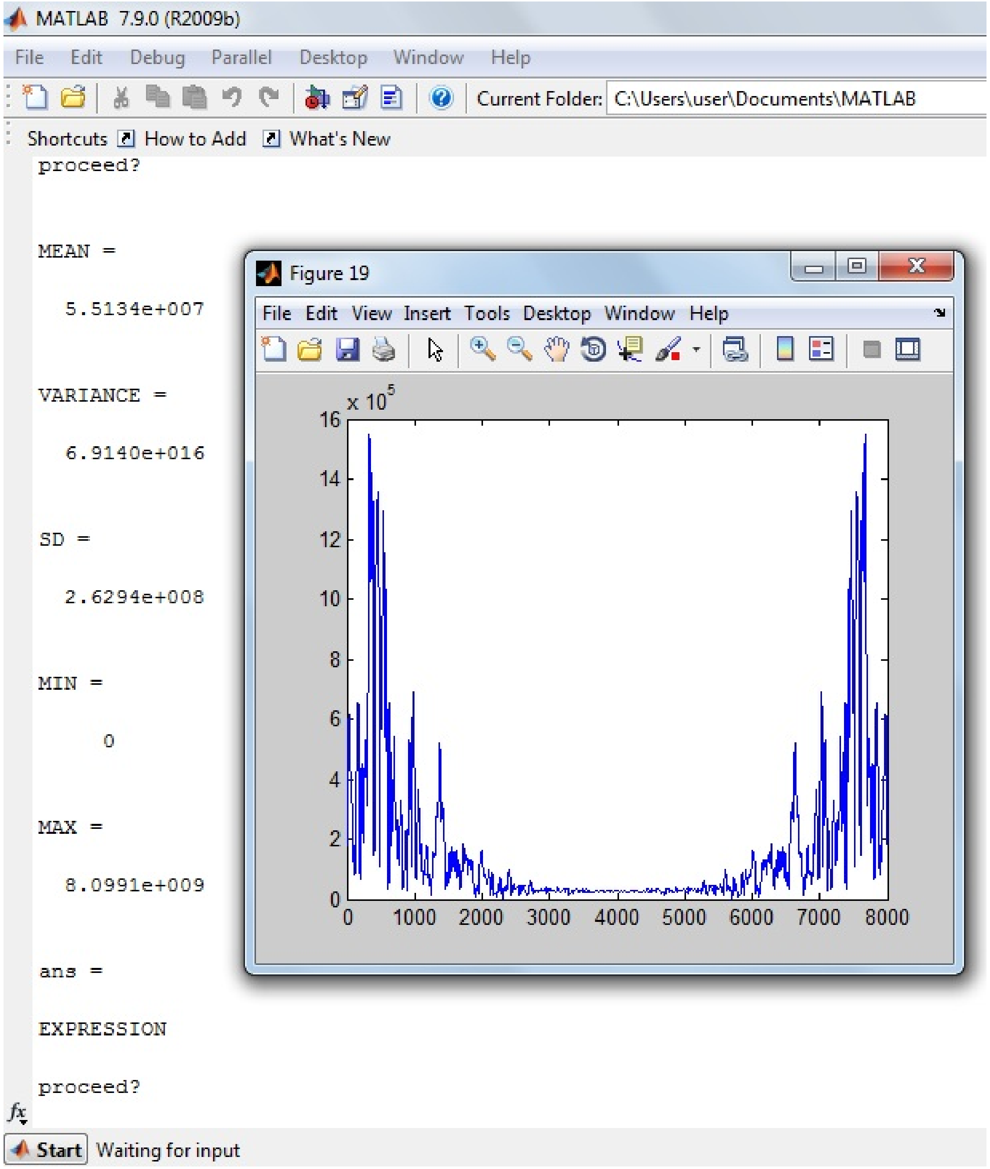
```
proceed?

MEAN =
  7.7781e+007
VARIANCE =
  1.2232e+017
SD =
  3.4974e+008
RANGE =
  1.0414e+010
ans =
STATEMENT
proceed?
```

Figure 6.2: DA detection using only prosodic features-STATEMENT


MATLAB 7.9.0 (R2009b)
File Edit Debug Parallel Desktop Window Help
Current Folder: C:\Users\user\Documents\MATLAB
Shortcuts How to Add What's New
Figure 65
File Edit View Insert Tools Desktop Window Help
x 10^5
7
6
5
4
3
2
1
0
0
2000
4000
6000
8000
Start
Waiting for input

```
proceed?

MEAN =
  4.0234e+006
VARIANCE =
  6.7517e+015
SD =
  8.2169e+007
RANGE =
  7.0990e+009
ans =
QUESTION
proceed?
```

Figure 6.3: DA detection using only prosodic features-QUESTION

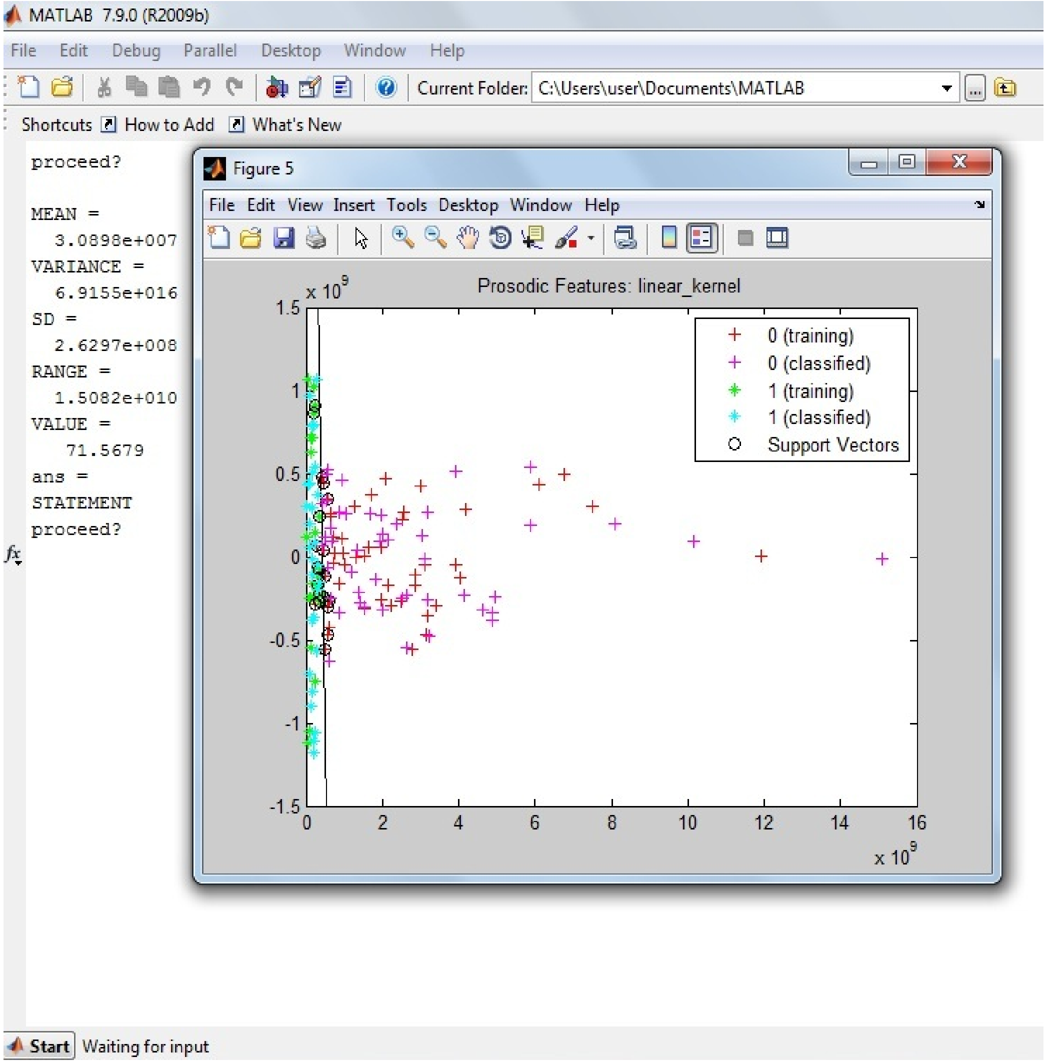


MATLAB 7.9.0 (R2009b)
File Edit Debug Parallel Desktop Window Help
Current Folder: C:\Users\user\Documents\MATLAB
Shortcuts How to Add What's New
Figure 19
File Edit View Insert Tools Desktop Window Help
x 10^5
16
14
12
10
8
6
4
2
0
0
1000
2000
3000
4000
5000
6000
7000
8000
Start Waiting for input

```
proceed?


MEAN =

   5.5134e+007


VARIANCE =

   6.9140e+016


SD =

   2.6294e+008


MIN =

     0


MAX =

   8.0991e+009


ans =

EXPRESSION

proceed?
```

Figure 6.4: DA detection using only prosodic features-EXPRESSION

The SVM is trained using the prosodic features.The results after training svm is shown in figure 6.5

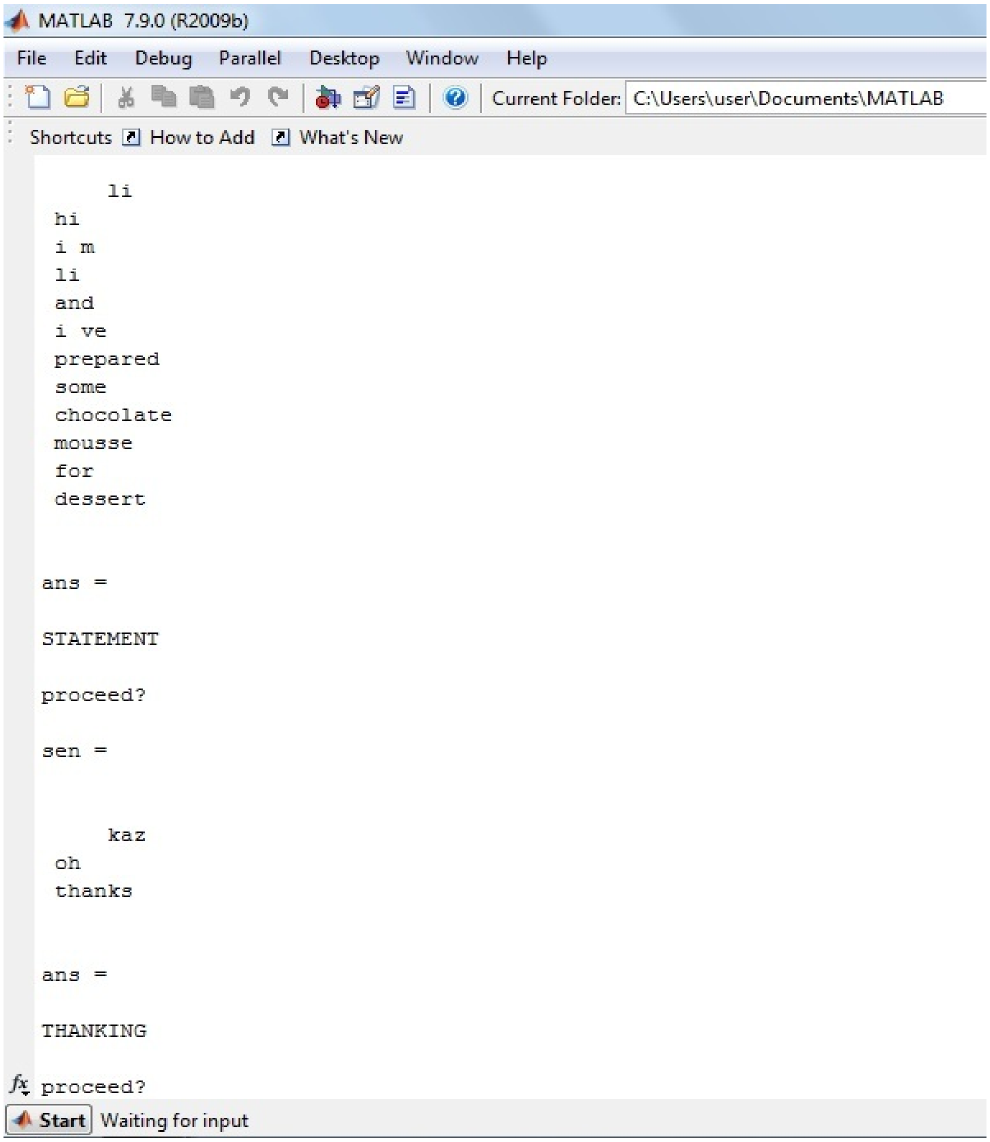


Current Folder: C:\Users\user\Documents\MATLAB

```
proceed?

MEAN =
  3.0898e+007
VARIANCE =
  6.9155e+016
SD =
  2.6297e+008
RANGE =
  1.5082e+010
VALUE =
   71.5679
ans =
STATEMENT
proceed?
```

Figure 6.5: Traing of SVM using Prosodic features

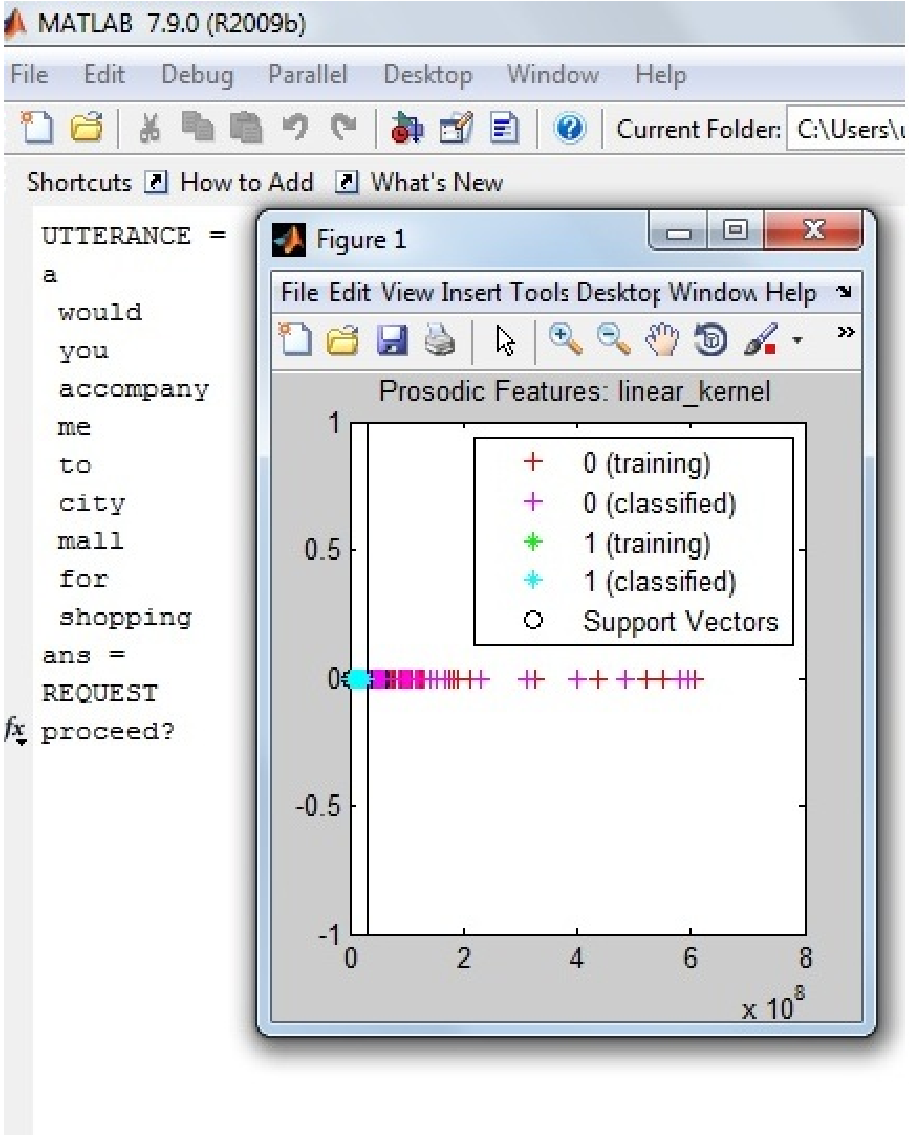
In the second step dialgue act classification is done using discourse features is performed.The transcription of human-human dialogue is stored in the .txt file is given as the input.The text file containing all the keywords are also given as the input.Keywords from the keyword file are extracted and spotted in the utterances of the input file, and detected the corresponding dialogue acts.All the 14 classes of dialogue acts are identified.The results obtained using matlab is shown in figure 6.6.


MATLAB 7.9.0 (R2009b)
File Edit Debug Parallel Desktop Window Help
Current Folder: C:\Users\user\Documents\MATLAB
Shortcuts How to Add What's New
Start Waiting for input

```
     li
 hi
 i m
 li
 and
 i ve
 prepared
 some
 chocolate
 mousse
 for
 dessert


ans =

STATEMENT

proceed?

sen =


     kaz
 oh
 thanks


ans =

THANKING

proceed?
```

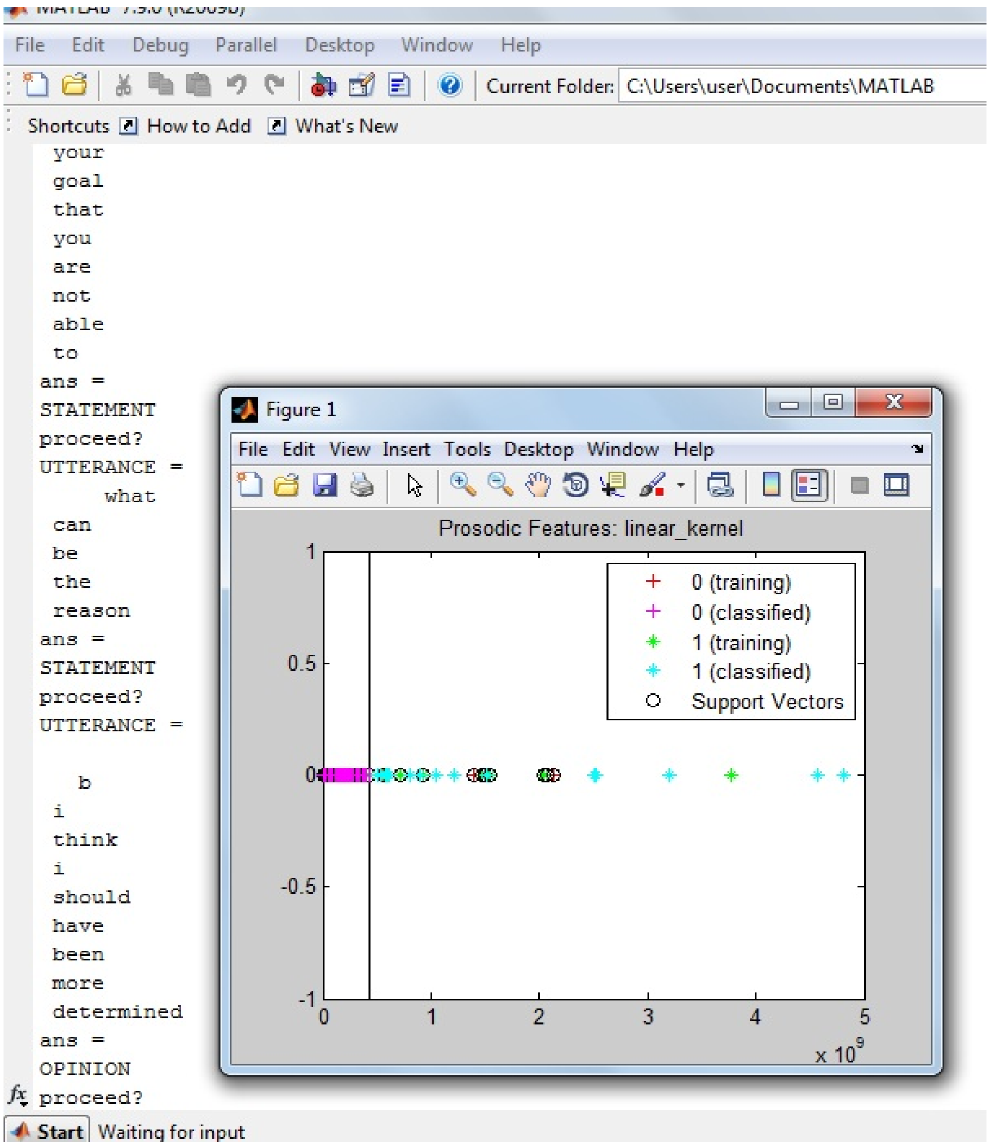
Figure 6.6: Dialogue Act Classification using keywords

In the third step dialogue act classification is performed by combining prosdic and discourse features.It could improve the results.The results obtained using matlab are shown in figures 6.7, 6.8.


MATLAB 7.9.0 (R2009b)
File Edit Debug Parallel Desktop Window Help
Current Folder: C:\Users\
Shortcuts How to Add What's New
Figure 1
File Edit View Insert Tools Desktop Window Help
Prosodic Features: linear_kernel
0 (training)
0 (classified)
1 (training)
1 (classified)
Support Vectors
1
0.5
0
-0.5
-1
0
2
4
6
8
x 10^8

```
UTTERANCE =
a
 would
 you
 accompany
 me
 to
 city
 mall
 for
 shopping
ans =
REQUEST
proceed?
```

Figure 6.7: Dialogue Act Classification using prosodic and discourse features-i


MATLAB 7.9.0 (R2009b)
File Edit Debug Parallel Desktop Window Help
Current Folder: C:\Users\user\Documents\MATLAB
Shortcuts How to Add What's New
Figure 1
File Edit View Insert Tools Desktop Window Help
Prosodic Features: linear_kernel
1
0.5
0
-0.5
-1
0
1
2
3
4
5
x 10^9
+ 0 (training)
+ 0 (classified)
* 1 (training)
* 1 (classified)
○ Support Vectors
Start
Waiting for input

```
  your
  goal
  that
  you
  are
  not
  able
  to
ans =
STATEMENT
proceed?
UTTERANCE =
      what
  can
  be
  the
  reason
ans =
STATEMENT
proceed?
UTTERANCE =

    b
  i
  think
  i
  should
  have
  been
  more
  determined
ans =
OPINION
proceed?
```

Figure 6.8: Dialogue Act Classification using prosodic and discourse features-ii

Chapter 7

Result Evaluation

The precision is taken as the performance measure.The definition of precision is precision = categories found and correct/total categories found. The results using only the lexical features, i.e., keywords produced some ambiguous results for the dialogue acts such as acknowledgement, expression, question, request, thanking and statement. By adding the prosodic features, i.e., variation of pitch, those ambiguities are successfully resolved. The combination could produce better results. The performance measures are summarized in the table 7.1. The accuracy is measured as the percentage of correctly labelled dialogue acts. The results were analyzed by comparing with the manually detected dialogue acts.It could successfully reduce the ambiguity of the confused dialogue act pairs.The comparison with baseline features are shown in table 7.2.

Table 7.1: Performance Measures

Features	**Precision**
Prosody	0.42
Discourse	0.71
Prosody + Discourse	0.78

Table 7.2: Comparison with baseline features

Feature set used	**Precision in percentage**
N-gram + POS + CoNLL	70.20
Discourse	71.00
Prosody + Discourse	78.40

Chapter 8

Conclusions and Future Works

This thesis presents an approach to recognize dialogue acts in utterances. It 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.

A hybrid method combining the prosodic and the keyword spotting features is proposed to the automatic detection of the dialogue acts of user for the affective user interfaces. Finding utterance boundaries in dialogue is a critical step. The work focused on the confusion pairs of dialogue acts like opinion vs. non-opinion statements and agreements vs. acknowledgements, occurs while classifying as dialogue acts.The standard information retrieval metrics precision is used to evaluate the results. The use of key word spotting feature were proved to reduce this ambiguity. The experiments were conducted using Buckeye corpus.

There are several techniques worth investigating to improve the dialogue recognition accuracy reported in this thesis. The methods used to detect dialogue acts presented here do not take into account sentential structure and word ordering in the transcript. The sentences such as

a. Ramu has been to India

b. Has Ramu been to India

be treated equally with the keyword approach. Without the punctuation as is often the case with informal typed dialogue, the keyword approach will not differentiate the sentences, whereas if we look at the ordering of even the first two words we can see that " Ramu has been to India" is likely to be a statement, whereas "Has Ramu been to India" would be a question.This difference is identified using the prosodic feature here. It would be interesting to research other types of features, such as phrase structure or even looking at the order of the first few words and the parts of speech of words in an utterance to determine its dialogue act from the transcript itself.

Although other studies have attempted to automatically tag utterances with dialogue acts it is difficult to fairly compare results because the corpora used were significantly different, the domains were different transcribed spoken dialogue versus typed dialogue), and the dialogue acts were also different ranging from a set of 9 to 42 . It may be possible to use a standard set of dialogue acts for a particular domain, but inventing a set that could be used for all domains seems unlikely. This is primarily due to different labelling requirements in various applications. A superset of dialogue acts that covers all domains would necessarily comprise of a large number of tags at least the 42 identified by Stolcke et al. But some of those tags not being appropriate for other domains.

Another clear task for future work is gathering a larger task-based on instant messaging corpus. The current study is done using Buckeye corpus. Although this was sufficient for the purposes of the study, to further this research, a larger corpus will help to avoid problems with scarce data, particularly if more granular dialogue acts are to be used. It would also be interesting to evaluate the models presented here in real-life situations with many different users and agents. Doing so will provide more data with which to evaluate the models' usefulness in assisting customer support agents and verifying the observations made in the present study.In future this work can be extended to repeat the experiments using standard corpus like SWITCHBOARD.

Bibliography

[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] 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, 2000.

[4] 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.

[5] U. Guz, S. Cuendet, D. Hakkani-Tur, and G. Tur, “Multi-view semi-supervised learning for dialog act segmentation of speech,” IEEE TASLP, vol. 18, no. 2, 2010.

[6] Silvia Quarteroni, Alexi V. Ivanov, Giuseppe Riccardi, “Simultaneous Dialogue Act Segmentation and Classification from Human-Human Spoken Conversations,”978-1-4577-0539-7/2011 IEEE.

[7] M. Dinarelli, S. Quarteroni, S. Tonelli, A. Moschitti, and G. Riccardi, “Annotating spoken dialogs: from speech segments to dialog acts and frame semantics,”in Proc. SRSL, 2009.

[8] Sophie Rosset and Lori Lamel, “Automatic Detection of Dialog Acts Based on Multi-level Information,”Spoken Language Processing Group, LIMSI-CNRS, B.P. 133, 91403 Orsay cedex, France.

[9] Gina-Anne Levow , “Prosodic Cues to Discourse Segment Boundaries in Human-Computer Dialogue”, University of Chicago, levow@cs.uchicago.edu

[10] Cheongjae Lee and Gary Geunbae Lee, “Emotion Recognition for Affective User Interfaces using Natural Language Dialogs”,Department of Computer Science and Engineering ,Pohang University of Science and Technology, South Korea.

[11] Ze-Jing Chuang and Chung-Hsien Wu, “ Emotion Recognition using Acoustic Features and Textual Content”, Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, ROC.2004 IEEE.

[12] M. G. Core and J. F. Allen, “Coding dialogs with the DAMSL annotation scheme,” in Proc. AAAI Fall Symposium on Communicative Actions in Humans and Machines, 1997.

[13] Ben Gold and Nelson Morgan,“ Speech and Audio Signal Processing”, John Wiley and Sons, New York, 2000. ISBN 0471351547.

[14] Daniel P.W. Ellis, “An introduction to signal processing for speech”, LabROSA, Columbia University, New York, October 28, 2008.

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

[16] Mohammed E. Hoque, Mohammad S. Sorower, Mohammed Yeasin, Max M. Louwerse, “What Speech Tells Us About Discourse: The Role of Prosodic and Discourse Features in Dialogue Act Classification”, IJCNN 2007.

[17] 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.

[18] Mark A. Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling ,William Raymond, “The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability,” Speech Communication 45 (2005) 89-95,ELSEVIER.

Publications out of the Thesis Work

Journal Publications

1. "Dialogue Act Detection from Human-Human Spoken Conversations". International Journal of Computer Applications 67(5):24-27, April 2013. Published by Foundation of Computer Science, New York, USA ,
http://research.ijcaonline.org/volume67/number5/pxc3886688.pdf

2. "Dialogue Act Recognition from Audio and Transcription of Human-Human Conversations" International Journal of computer Trends and Technology (IJCTT) , June 2013 issue volume number 4 issue 6 . ISSN: 2231-2803,
http://www.ijcttjournal.org/volume-4/issue-6/IJCTT-V4I6P181.pdf