# Dialogue Act Recognition from Audio and Transcription of Human-Human Conversations

Nithin Ramachandran

*CSE Department, KMCT College of Engineering, University of Calicut*

*Kerala, India*

***Abstract*— Understanding the human - human conversations requires both understanding the semantic meaning of an utterance, and the intended meaning behind that utterance. Dialogue acts gives this intended meaning behind an utterance. This paper 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 the classification of each segment according to a DA tag. Turn segmentation is done using silence removal algorithms and pitch separation.**

***Keywords*— *Utterance, Dialogue Act, Prosodic Features, Discourse Features, Turn segmentation***

## I. 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. There are 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 paper, the focus is mainly on illocutionary acts which are also referred as dialogue act. The intentions or meanings are conveyed through various ways, such as, by making assertions, declarations, questions, expressions, etc.

This characterization provides a representation of conversational function and is especially useful in systems that require an automatic interpretation of dialog act to facilitate a meaningful response or reaction. Dialog act tagging has been successfully integrated in speech recognition, speech understanding, text-to-speech synthesis and speech translation systems. Dialogue acts can be detected from 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 lexical-related features using a set of keywords. The acoustic data is used to capture the intonation patterns rather than the semantic meaning of the utterance.

## II. RELATED WORK

Because of the importance of dialogue act classification within dialogue systems, it has been an active area of research for some time. Early work on automatic dialogue act classification modelled discourse structure with hidden Markov model [2] or SVMs [3] 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 [1, 2]; then, DA labels were assigned to such boundaries based on multi-classification [3, 4]. In [3], 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. A recently proposed alternative approach uses a discriminative approach, namely Conditional Random Fields (CRFs), [4] 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. Lots of study has been carried out to investigate prosodic indicators to detect dialogue acts in speech [5]. The features that are commonly considered include Fundamental frequency F0, duration, intensity, spectral variation and wavelet based features. Of the approaches noted above, in this paper linear feature extraction techniques and their extraction algorithms are explained. These features are then used to classify dialogue acts.This paper uses combination of lexical and prosodic features and SVM classifier for simultaneously segment an utterance into its DA boundaries and label such segments according to a DA tag. Moreover, it takes a step beyond the previous work by including the use of a set of keywords based on lexical features. It could successfully reduce the ambiguity of opinion vs. non-opinion statements

and agreements vs. acknowledgements, occurs while classifying the dialogue acts.

## III. 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 [5]. 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 lexical features and prosodic features better results are obtained. The variation of pitch for each dialogue acts and frequently used keywords in an utterance for each dialogue acts are investigated and those are extracted as feature set.

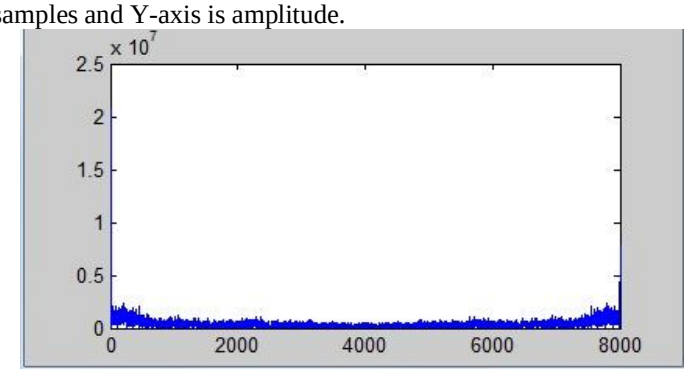
### *A. Prosodic Features*

Prosody contains speech related information, which is not entirely predictable at word or sentence level, by analyzing phoneme sequences [5]. 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. 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.

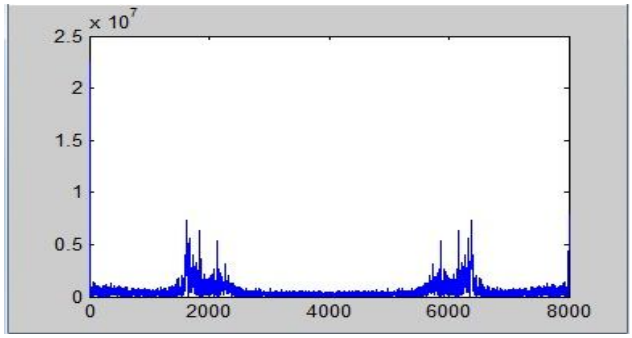
The analog signal is converted to .wav digital format by sampling with a suitable rate and quantized. The digital signal is then hamming windowed to convert it into a suitable frame size.

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.

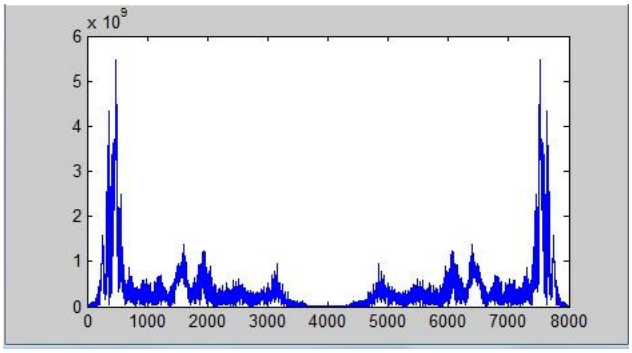
The range of the frequency variance for each utterance in corpus is studied. The frequency variation of dialogue acts such as acknowledgement, expression, question, request, thanking, statement are shown in figure 1. X-axis is number of samples and Y-axis is amplitude.



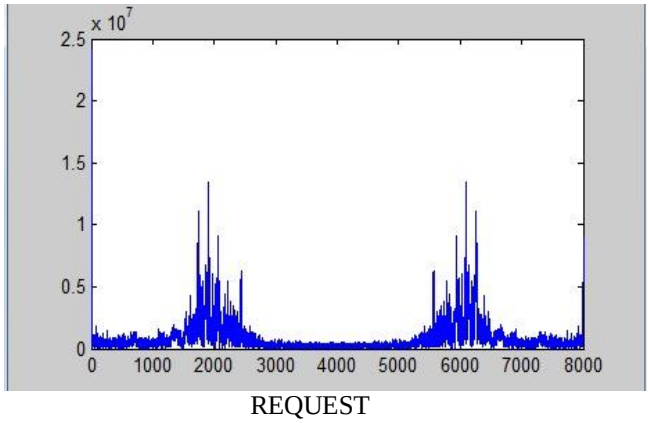

ACKNOWLEDGEMENT



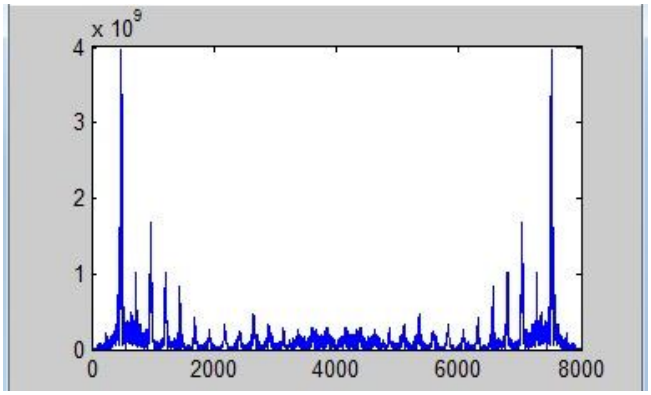

EXPRESSION



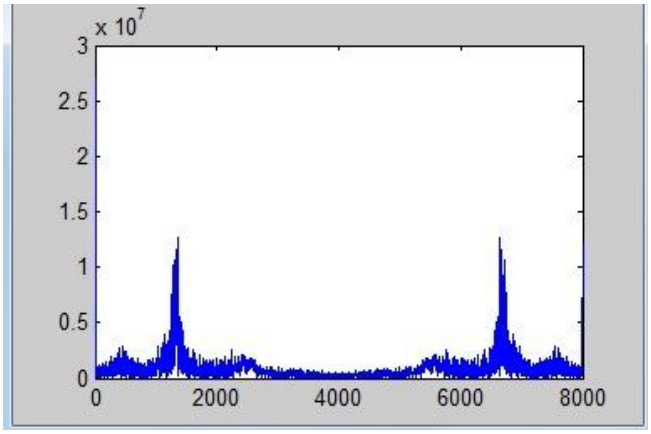

QUESTION




REQUEST




THANKING




STATEMENT

Fig. 1 The frequency variation of different utterances and corresponding dialogue acts.

### *B. Lexical features*

Lexical 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. Lexical features such as keywords and word identity 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. These patterns of dialogue are extremely helpful to disambiguate intentions even though they may contain similar lexical information. 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. These keywords are spotted in each utterance using mean of their ASCII value. The feature set used are summarized in the Table I.

TABLE I
SET OF KEYWORDS SPOTTED IN DIALOGUE ACTS

| Dialogue Act | Keywords |
|---|---|
| OPENING | hi, hello, welcome, good morning, good afternoon, good evening |
| CLOSING | see you, bye, good night |
| THANKING | thanks, thank you |
| YES-ANSWER | yes |
| NO-ANSWER | no |
| REQUEST | would, can, please, may, let us |
| ACKNOWLEDGE MENT | great, okay |
| WH-QUESTION | ?, how, what, which, when, who, whom, where, why |
| YN-QUESTION | do ,does, did |
| OPINION | I, I am, In my view, To my mind, To be honest, As far as I'm concerned, I think, It seems, I would argue, I believe, I think |
| AGREEMENT | Of Course, Hmmm, I Agree, You are right. |
| EXPRESSION | Is it, so |
| STATEMENT | this, that |
| FILLER | you know |

## IV. TASK SETUP

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 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. The proposed method adopts 14 DA tagsets from DAMSL [8] and ADAMACH taxonomy, [9] where utterances have been manually transcribed and annotated according to the DAs. The DA tag set and labels used are shown in Table II.

TABLE II
DIALOGUE ACTS AND LABELS

| Dialogue Act | Label |
|---|---|
| OPENING | [OPEN] |
| CLOSING | [CLOSE] |
| THANKING | [TH] |
| YES-ANSWER | [YA] |
| NO-ANSWER | [NA] |
| REQUEST | [RQT] |
| ACKNOWLEDGEMENT | [ACK] |
| WH-QUESTION | [WHQ] |
| YN-QUESTION | [YNQ] |
| OPINION | [OP] |
| AGREEMENT | [AG] |
| EXPRESSION | [EX] |
| STATEMENT | [ST] |
| FILLER | [FL] |

## V. EXPERIMENTS AND RESULTS

Support Vector Machines are used for classification. 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 keywords are performed, and then the prosodic features are also incorporated. The testing is also done using the casual conversations. The experiment is conducted using an audio file in wave format and its transcription in text format. The algorithms were implemented using MATLAB.

The training is done using dialogues from the Buckeye corpus and the testing is done using the dialogues from Buckeye corpus and the casual chats from BBC learning English website. 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 obtained results using MATLAB containing transcript of conversations annotated with dialog acts are shown in figure 2.

**Li: Hi**
**ans = OPEN**
**I'm Li and I've prepared some chocolate mousse for dessert.**
**ans = ST**
**Kaz: Oh thanks.**
**ans = TH**
**I love chocolate mousse.**
**ans = ST**
**Mmm it's delicious.**
**ans = EX**
**Li: Good!**
**ans = EX**
**Anyway what was I saying?**
**ans = WHQ**

Fig. 2 The utterances in conversations, annotated with dialog acts

The accuracy is measured with test set using turn level, utterance level and DA accuracy (*AccT, AccU, AccD*); these are respectively defined as the number of correctly segmented and labelled turns out of the total number of turns, the number of correct utterance-level predictions out of the total number of utterance and 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 classification accuracy of dialogue acts given different feature sets are shown in Table III. Better results were obtained by using combination of prosodic and lexical features.

Table III
CLASSIFICATION ACCURACY OF DIALOGUE ACT GIVEN DIFFERENT FEATURE SET

| Feature sets Used | AccT | AccU | AccD |
|---|---|---|---|
| N-gram, POS, CoNLL (BASELINE) | 58.8% | 70.9% | 65.9% |
| Prosodic Features + Keywords | 74.4% | 71.2% | 78.4% |

The comparison of turn level, utterance level and dialogue act detection accuracy with baseline features is shown in the chart in figure 3.

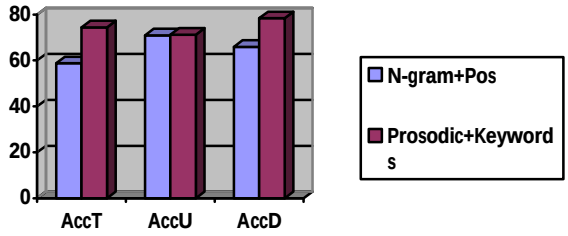



Fig.3 The comparison of turn level, utterance level and dialogue act detection accuracy with baseline features.

## VI. CONCLUSIONS

This paper, presents an approach to recognize dialogue acts in utterances. A hybrid method combining the prosodic and the keyword spotting features is proposed for the automatic recognition of the dialogue acts of user for the affective user interfaces. The work focused on the confusion pairs of dialogue acts like opinion vs. non-opinion statements and agreements vs. acknowledgements, occurs while classifying dialogue acts using only lexical features such as POS and N-grams. Instead of those, the use of key word spotting feature were proved to reduce this ambiguity. The experiments were conducted using Buckeye corpus and the algorithms were implemented using MATLAB. In future this work can be extended to repeat the experiments using standard corpus like SWITCHBOARD, which is a larger corpus.

## REFERENCES

[1] N.K. Gupta and S. Bangalore, "Segmenting Spoken Language Utterances in to Claused for Semantic Classification," *in Proc. ASRU*. Citeseer, 2003, pp. 525-530.
[2] Y. Liu, A. Stolcke, E. Shriberg, and M. Harper, "Using conditional random fields for sentence boundary detection in speech" in Proc. ACL, 2005.
[3] U. Guz, S. Cuendet, D Hakkani-Tur, and G. Tur, "Multi-view semi supervised learning for dialogue act segmentation of speech," IEEE TASLP, vol. 18, no.2, 2010.
[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.