Analysis of Speech Parameters as Indicators of Engagement in Conversation

by

Christy Elias

**A Dissertation submitted to the
University of Dublin, Trinity College
in partial fulfillment of the requirements
for the Degree of Master of Science (by Research) in Computer Science**

University of Dublin, Trinity College

November 2020

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Christy Elias

November 19, 2020

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this dissertation upon request.

Christy Elias

November 19, 2020

Summary

This dissertation aims to analyse the engagement of interlocutors in casual conversations to find the speech parameters that are indicators of engagement. Engagement is often described as the quality of the interaction, the internal state of interest in the conversation. Although the definitions of engagement vary across the researchers, engagement is seen as social behaviour of humans that makes the conversation effective and achieve its goal (e.g. complete a transaction/ carry out a casual conversation). In this dissertation, the engagement is analysed in the scope of casual conversation where the goal is to take part in the conversation itself.

To understand how conversational and social behaviours are studied by researchers, a study of related works was performed. Social behaviours, physiological and psychological states, conversational behaviours were analysed in the past. Different modalities have been used in analysing conversational behaviours (audio, video, touch). In this dissertation, conversations are analysed using the audio with the assumption that a change in a speakers engagement may be reflected in their voice and can be determined automatically by analysing the speech parameters such as Prosodic (F0, VUV), Mel-Frequency Cepstral Coefficients (MFCCs), and voice quality parameters.

A corpus of casual conversations, Table Talk Corpus, with recordings of people in casual conversations was used to analyse the speech parameters and how they can be used as indicators of engagement. The annotations of engagement as a positive change in engagement (high-engagement) and negative change in engagement (low-engagement) from the norm were used, and conversational segments where no change in engagement was found were marked as neutral. A voting average of these annotations was derived for each utterance from the original annotations to make them more consistent. Machine learning techniques were used to validate the hypothesis and to gain insight. Supervised machine learning algorithms were trained using extracted features (acoustic features) of engagement labelled utterances. Support Vector Machines (SVM) and Random Forests were utilised to see

how the performance of the classifiers was affected by the use of the speech parameters. The classifiers were used along with cos sensitive meta classifiers to deal with the class imbalance in the data. Random Forests were used as they are useful in understanding internal measures of performance, such as the average impurity decrease, that can be used for calculating the relative importance of features in classification.

The results indicate that the speech parameters used are strong indicators of engagement in conversations. The results were statistically significant and produced better results than the baseline classifier (ZeroR). The results of the classifications are described in terms of accuracy, precision, recall and F measure. The results of the classifications are shown in tables of performance measures, histograms showing the importance of individual features in predicting engagement. The combination of the speech parameters was found to achieve better results in the classifications.

Based on the results of the classification, the conclusions of the dissertation are presented. The difficulty in collecting conversations where the participants have low engagement in laboratory settings is difficult; this points to the need for developing a larger corpus of casual conversations with engagement annotations. The prosodic, cepstral and voice quality features have been shown to be good indicators of engagement in conversations. Alternative approaches using other modalities are suggested. The work described in this dissertation is a step towards understanding the conversational engagement in casual conversations. The insights gained from this work can be useful in making computers socially intelligent.

Acknowledgments

I would like to thank my supervisor, Prof. Nick Campbell, for the guidance and support provided in completing this dissertation. I am also grateful to the past and present members of Speech Communication Lab at Trinity College Dublin, ADAPT Centre as well as my friends and family for their support.

Christy Elias

University of Dublin, Trinity College
November 2020

Analysis of Speech Parameters as Indicators of Engagement in Conversation

Christy Elias

Master of Science (by Research) in Computer Science
University of Dublin, Trinity College, 2020

Supervisor: Prof. Nick Campbell

Abstract

In human conversations, the words being said may not always contain the whole of the information intended to be conveyed; instead, such information may be conveyed through the way those words are being said. Thus understanding the information beyond the words is vital in understanding the behaviours of the speaker in a conversation. The acoustic-prosodic features in speech are excellent sources of information which can indicate the internal states of a speaker. Humans have the capability of doing this processing of extra-linguistic information to understand their conversational partner — computers, on the other hand, lack this ability. The analysis of speech signal can reveal a lot of the information which may not be captured by analysing the linguistic content alone. In this study, the information encapsulated by the features of the speech signal is analysed to model conversational engagement of the participants. Casual conversations are analysed as part of this study to identify features of the speech signal, which can be used as cues for detecting engagement in conversations. Voice quality features (including glottal parameters) are mainly analysed in this study to model engagement in conversation. The proposed methods attempt to understand the conversational engagement of participants in casual conversations.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Speech is the primary form of human communication. In a typical speech communication scenario, the speaker conveys an idea or concept through linguistic information in a vocalised form, through words, non-verbal utterances, conversational behaviours and other signals such as gestures and body movements. The conversational behaviour is supplemented with nonverbal signals. These nonverbal signals encapsulated in speech can communicate more than what is contained in words being vocalised. The internal state of the speaker, affective states, sentiments, emotions etc. are discovered by an attentive conversational partner naturally. This extra information or signals other than the linguistics content is referred to as paralinguistic information[1].

The basic form of human communication is face to face conversations, where much of the information is conveyed through different modalities (audio, visual, haptic). With the advent of electronic communication systems, humans communicate with each other using electronic means (Telephones, Video conferencing, Internet Messaging, Emails). In Human-Human conversations, an interlocutor can understand the linguistics information and the paralinguistic information with minimal effort. In the case of communicating with a machine or a robot, the comprehension of these paralinguistic pieces of information is still a challenge for machines. As the need to interact with the non-human counterparts in conversational situations such as in automated call centre scenarios, automated banking calls, and interactive voice response systems it is useful to understand the human interlocutors' conversational behaviours and internal states to make the interactions natural and effective.

In communication, how something is being said is equally important as what is being said. The way in which something is said can indicate many conversational aspects of the speakers. Humans have an innate abil-

ity to understand the social signals and conversational behaviours of their counterparts, and it makes humans socially intelligent. Computers lack this ability to interpret social signals and conversational behaviours. By giving computers, social-intelligence can make interactions with them more fruitful and natural.

So to make interactions with a computer effective and natural, it is essential to give the machine the capability to understand the conversational behaviours and social signals. Engagement is a vital conversational or social behaviour in human-human or human-computer interactions as it makes the conversation effective and meets the purpose of the interaction which may accomplish something (make an online purchase, provide help to customers, or make a banking transaction) or not (to carry on a casual conversation).

In this research, the conversational Engagement is considered as "the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction" [2]. In this analysis, we investigate how features of the speech signal can be used as indicators of Engagement in conversations. In order to understand the Engagement of the participant as expressed by the characteristics of their speech and how it elicits Engagement in other participants, it is necessary to analyse the speech signal.

The definitions of Engagement varies between researchers, but they all are concerned with the analysis of the quality of the interactions [2]. Sidner et al. .[3] defines Engagement as the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake. Rather than being involved, being engaged represents how one actively controls and participates in an interaction. Gatica-Perez et al. [4] considers the degree of Engagement displayed as an expression of an internal state of "interest" of a person resulting from the attraction towards the interlocutor, interest in the theme of the conversation or the social rapport.

1.1 Motivation

The primary motivation behind this research on conversational Engagement is the interest in conversational behaviours of humans. How humans understand their conversational partners' behaviours or intentions from the paralinguistic information is in itself an exciting capability. This ability adds to the social intelligence of human beings. In the case of computers

that take part in conversations with humans, this social intelligence or the capacity to understand the behaviour or intentions expressed through the paralinguistic phenomena or social signals is insufficient or absent.

Conversational Engagement of the speaker has been analysed in many contexts and using many methods. Speech parameters have been used as indicators of many human behaviours, emotion, affect and internal states. This motivates to use these parameters for classification of Engagement in casual conversation. Casual Conversations are used in this research work to focus on the Engagement in conversation rather than the Engagement elicited by a goal/task as in the case of task-oriented conversations.

So the central theme of this dissertation is around the subject of Engagement in casual conversations and the use of speech parameters as indicators of Engagement. The aim is to use the speech parameters to create a model that can classify Engagement as High, Low and Neutral in Casual Conversations.

1.2 Research Questions

The main questions to be answered in the context of this research are as follows. Can conversational behaviours of interlocutors be collected from speech and Engagement be detected from speech features? Is Engagement reflected in the way a person talks? Do the features in speech indicate something about his/her Engagement in the conversation? Is it possible to generalise the variations in speech features as cues for Engagement? Can conversational Engagement of a human interlocutor be modelled based on speech features? Can a social robot detect the Engagement of a human interlocutor and direct its speech based on a model built on speech features? All these can be summarised into one research question,

"Is it possible to generalise the patterns in speech features as indicators for measuring the Engagement of a participant in a conversation?"

There are many ways Engagement can be measured. Analysis of visual cues such as the nods, shakes, gestures, gaze, hand and body movements of both the speaker and the listener can give indications of their Engagement in a conversation. Another way is by looking at the features of the speech signal, backchannels, and linguistics content etc. The Engagement of a speaker is affected by different factors. The internal emotional state of the speaker, the statements made by the other party involved in the conversations, interest in the topic of the conversation, previous level of Engagement. In this research, the acoustic cues are used as a measure of

Engagement. The Engagement measured can then be used to improve the quality of the interaction.

The central hypothesis of this research is that the conversational behaviours, such as Engagement of the participants in a conversation, are reflected in the features encapsulated in the speech signal. Thus the goal becomes understanding the correlation of the speech parameters and Engagement of the participant in a conversation and develop models to recognise it using computational methods.

By measuring Engagement in a conversation, the internal state of interest of a person can be understood, and this could help in determining the smoothness of the interaction and can indicate the work performance and relationship[5]. It is necessary to estimate the Engagement from speech features which are observable representations of Engagement to understand the internal state of interest of a person. A typical application of an engagement detection system could be in automated call centres where a decision to transfer a call to a human operator can be made by understanding the engagement level of the customer. In general, understanding the engagement state of a person can help the system to respond in a more suitable way to the situation.

The rest of the dissertation is organised as follows, chapter 2 describes related works and the knowledge on which the dissertation builds on. Chapter 3 gives an overview of the corpus and the annotation of Engagement used. The classification of Engagement using the speech parameters and the feature extraction methods used are explained in chapter 4. Chapter 5 describes the voting average of engagement annotation and the random forest method used for classification and discusses the results. The use of additional voice quality features for the classification and how they affect the classification performance is described in chapter 6. The conclusions of the dissertation are detailed in chapter 7 and discuss the purpose, results, and limitations.

Chapter 2

Related Works

There has been an increased interest by researchers in the measurement of Engagement in various contexts. Bohus and Horvitz have experimented with measuring multiparty Engagement in open word dialogues using an avatar-based system[6] where they created policies to understand Engagement using visual and speech inputs. The Engagement state captures whether an agent is engaged in interaction and is modelled as a deterministic variable with two possible values: engaged and not-engaged. They updated the engagement state based on the joint actions of the system and the agent. In this research, a similar approach in modelling the Engagement is used where states of Engagement are considered as high/low/neutral.

Bonin et al. analysed the behaviour of conversational participants in a group context [7]. The group involvement and individual Engagement were analysed in four people casual conversational data, Table Talk Corpus. They had annotated the conversation for individual and group Engagement; they found a significant amount of agreement between the annotators even though there were no predetermined timing of chunks, the changes of perceived Engagement were synchronised. The annotations from this study are used for the engagement classification experiments in this dissertation.

Voigt et al. studied how raw bodily movements correlated with prosodic phenomena distributed over a phrase, which are indicative of speaker engagement [8]. A collection of video monologues were analysed using acoustic and movement measurements for each phrase in the data. They found that phrases where speakers are engaged, they use higher pitch and intensity as well as higher variance in their pitch and intensity. In this work, the correlation between speaker movements and prosodic indicators of Engagement was analysed, whereas this dissertation intends to analyse how the prosodic and acoustic features can be used to detect Engagement. The data in their work was composed of monologues, but the scope of this

dissertation is to analyse Engagement in conversations.

Engagement in collaborative conversations between Robots and humans were studied in [3], where an architecture to understand how Engagement occurs was developed. They used a robot as conversational agent that could process spoken utterances and visual information to make conversational decisions and engage the user in physical activities. In this study, the conversational Engagement of the user was not measured; instead, the robots ability to elicit Engagement using visual and conversational cues were analysed.

Yu et al. analysed Engagement in telephonic conversations using affective information in users' speech[9]. They use low-level acoustic information extracted from each utterance to find users' affective states using a support vector machine (SVM) and then use those as inputs for a coupled hidden Markov model (CHMM) to estimate Engagement. Prosodic and Spectral features extracted from the speech signal were used as features for the SVM-based classifier for affective states and the CHMM was trained with those affective state information. In our method, we make use of the acoustic/prosodic information, including voice quality parameters directly to analyse Engagement.

Hsiao et al. follows the CHMM model proposed by Yu et al. .[9] to measure engagement [5]. Instead of extracting features for utterances each slice of the CHMM model was based on a 30-second sliding window. They combined features from turn-taking with acoustic features to recognise Engagement.

Gustafson and Neiberg showed how prosodic cues, including change in syllabicity, pitch slope and loudness in non-lexical response tokens in Swedish could be used to detect engagement [10] while this study is only analysing the prosodic cues in non-lexical tokens for Engagement of the listener our work analyses lexical and non-lexical utterances for Engagement detection for the participants.

Rich et al. identified four types of connection events, such as directed gaze, mutual facial gaze and adjacency pairs and backchannels and found that these events, occurring at some minimum frequency, are the process mechanism for maintaining engagement[11]. The mean and maximum delays between these events and the number of failed events were calculated. These statistics were compared with the current time window and the whole interaction to recognise the Engagement, the failure rates were used to recognise disengagement.

Lei et al. used turn-taking features based on speaker activity to auto-

matically derive measures of group-level involvement for finding relevant segments in conversation helpful in the summarisation of meeting data[12]. The speech parameters or non-lexical features were not analysed in their work. Here the measuring of involvement was not the primary focus but to use it as a feature in finding useful measures for summarisation.

In a study of quantifying engagement Norris et al. analysed a system for rating involvement in live human avatar interaction games [13]. The Engaging features of the game design were found to be associated with higher involvement ratings. This work mainly focused on the design of the game and how they affected the users' Engagement using a statistical analysis of involvement ratings. The interactions of the users with the system were not analysed using audio or visual parameters.

Szafir and Mutlu analysed student attention drops, using Engagement measured from electroencephalography (EEG) and found that using verbal and nonverbal cues from a robot tutor helped in recapturing the attention of the student [14]. The EEG data, and the engagement levels before and after the cues from robot tutor were tested using analysis of variance. The results showed that a higher level of Engagement was achieved when the instructions to recapture attention was given early by the robot tutor than when they were delayed. This research shows that the measure of Engagement can be used in applications for tutoring, where Engagement can be elicited using a robotic interaction partner.

Conversational Engagement has also been measured in human-robot social interactions using dialogue activities by Duplessis and Devillers in [15]. The Engagement of the human participant was given a score based on the proportion of dialogue activities that have been a successful dialogue in a given interaction, weighted by the length of the body of that activity. The model proposed in their work shows that the engagement score is higher for interlocutors that coordinate their contributions with that of the robot partner to interact at the right time.

Much attention has also been given to the study of visual correlates of Engagement. Hernandez et al. [16] used visual features including head pose, five facial points, head size, and head position to measure Engagement in television viewers. Richardson and Dale [17] focused on understanding the relationship between speakers' and listeners' eye movements and discourse comprehension based on experimental studies. Schmidt et al. [18] also carried out related experiments using hand and leg movement as cues.

Engagement behaviour in children using vocal cues in non-verbal vocalisation was analysed by Gupta et al. [19], where they achieved the highest

accuracy of 62.9% in classifying Engagement. Their results show the role of speech in defining the engagement interactions involving children. They used a 2 class as well as a 3 class approach to engagement classification.

Glas and Pelachaud showed that a user's preference for a physical object is positively correlated with the user's Engagement during the discussion of it with a virtual agent [20]. They also found that by detecting the level of Engagement during an interaction can also reflect a user's preference towards an object. This correlation was used to develop an agent model that can enhance the user's Engagement by personalising the topic of the interaction.

Vels and Jokinen analysed techniques for object recognition to automatically detect human behaviours in video conversations [21]. They found that by detecting the body movements of the conversational participants and analysing their conversational styles, the Engagement in the communicative activity can be understood. They also found that the movement trends and irregularities in their behaviour can be used to study the synchrony and adaptation between participants.

The voice quality features have not been used previously in the analysis of Engagement. They have been used to predict other social behaviours in conversations. Charfuelan et al. have used voice quality features to predict the dominance of participants in meetings from their voice [22]. They used prosody and voice quality features and found that the most dominant speaker uses a louder than average voice quality and the least dominant speaker uses a softer than average voice quality. In this dissertation, we intend to use voice quality features in the classification of Engagement in casual conversations.

This chapter provided an overview of the related research performed for understanding and modelling engagement in conversations and the various tasks involved in achieving that. This chapter also discussed the various modalities and features used for recognising Engagement and related social behaviours of interlocutors in conversations. The concept of Engagement, defined by various researchers, has also been discussed.

Chapter 3

Table Talk Corpus

In this research, the speech parameters are analysed in the context of casual conversations to use them as indicators of engagement. The corpus used in the study should contain casual conversations, where there is no goal for the conversational participants other than to take part in the conversations.

In this chapter, the corpus used in this research, Table Talk Corpus is described. The engagement annotation, the annotation scheme and the inter-annotator agreement measures are described.

3.1 The TableTalk Corpus

TableTalk is a corpus [23] of free/casual conversation between four/five(on the third day) participants of various linguistic and cultural background. The recordings were performed at the ATR Research Labs, Japan. The participants were all living in Japan at the time and shared some knowledge about the culture in Japan. The language of the conversation was English. Only one of the participants was a native speaker, while the others were Finnish, Belgian and Japanese. The Japanese speaker was a paid participant, and the others were researchers. The participants were not familiar with each other except one speaker who knew all the others.

In order to keep the conversations as casual/free as possible, the participants were not given prior instructions about the topics or activities so as not to restrict the participants and keep the conversation as natural as possible. The participants were informed that the conversation is being recorded.

The corpus contains recordings performed over three days. The total duration of the corpus is 210 minutes. The video was recorded using a 360-degree camera to capture the frontal face while the participants were

Figure 3.1: A Screenshot from the video recording from the TableTalk Corpus

seated around a table. The audio was captured using a centre mounted microphone, in order not to encumber speakers in any way.

The conversation of the first day was used for this study.

3.2 Annotation of Engagement

In an experiment conducted by Bonin et al. [7] the first 34 minutes of the day 1 of TableTalk corpus was annotated for engagement. There was no pre-determined chunking or segmentation of the videos, and annotators were not time-constrained.

Annotator	Number of High Engagement Segments	Number of Low Engagement Segments	Number of Neutral Engagement Segments	Duration of Segments
1	40	26	66	1 second to 48 seconds
2	38	24	42	3 seconds to 19:23 Minutes
3	13	12	0	200 milliseconds to 13 seconds
4	104	87	191	8 seconds to 2:22 minutes
5	22	19	20	3 seconds to 2:03 minutes

Table 3.1: Annotation of Engagement on the Table Talk Corpus

There were five annotators who were psychology students trained to evaluate human behaviour. They were not given any strict definition of engagement. The annotators were asked to rate the data freely in time.

The engagement annotations were discreetly marked with '+' (plus) and '-' (minus) for increase or decrease in engagement. The stable periods of

engagement were left without any annotation. They are also annotated for both group and individual engagement. In this dissertation, the classes are addressed as High-engagement (increase in engagement) Low-engagement (decrease in engagement) and Neutral classes.

We reformatted the annotation data as a continuous-time sequence by filling these unannotated segments with' 0' as a label for neutral segments.

An inter-annotator agreement test was performed to find if the similarities in the annotations were merely accidental or not. It resulted in k=0.9 (annotators with Cohens Kappa greater than 0.5 are considered to have a good agreement [24]). The Cronbach's Alpha test was also conducted to verify if the annotations were consistent and reliable. The group reliability test result was 0.68 (the Alpha coefficient varies between 0 and 1, where the higher value represents high internal consistency).

The labels of engagement from these annotations are used as the target labels in the classifications performed in the following chapters.

Chapter 4

Engagement Classification Using Acoustic Features and Engagement Annotations

This section details the methods used for modelling engagement in casual conversations using speech parameters. By using speech parameters (acoustic/prosodic) for classification of Engagement, an understanding of which features are indicators of Engagement in conversations can be made. The process of understanding how speech parameters can be used as indicators of Engagement involves the preprocessing of the speech data in the corpus, extraction of speech parameters and machine learning to create models that can be used in the classification Engagement.

In order to investigate whether Engagement could be automatically classified by using the speech parameters, an automatic engagement classifier using machine learning approach was used. In this task, a Support Vector Machine (libSVM [25], 10-fold cross-validation approach) was used as the supervised classifier to model the conversational Engagement from the feature matrix and to classify the instances into any one of the classes(High/Low/Neutral). The figure 4.1 shows the steps involved in the classification of Engagement. The extracted features constitutes the features matrix and each row represents a feature vector which contains the engagement label. In this dissertation, the focus is on the classification of speech frames as one of the three engagement levels (High/Low/Neutral). The segments in the corpus which were annotated as 0 (no change in Engagement from the normal) were used as Neutral instances. As the speech parameters used in this experiment are found to be useful recognition of emotions, affect and other physiological and psychological states of

a speaker they were expected to be useful features in the classification of Engagement, and the results validate our expectations.

This chapter is organized as follows, the preprocessing and extraction of features is explained in 4.1, then the classification algorithm and tools used are explained in 4.2. The results of the classification are explained in 4.3.

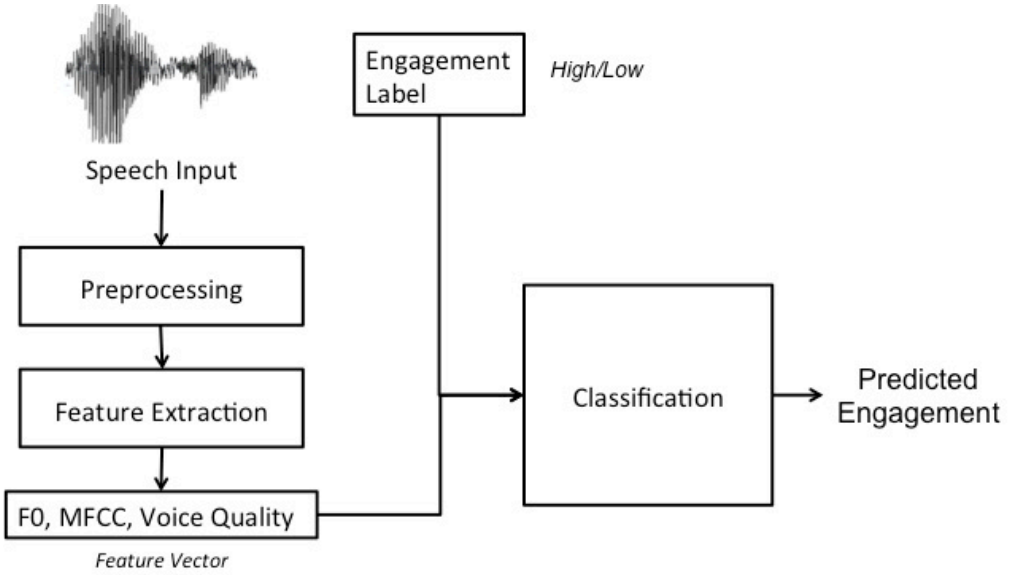

Figure 4.1: Steps in the Classification of Engagement

4.1 Feature Extraction

The audio recording was down-sampled to 16kHz for feature extraction in this work. The 34 minutes and 24 seconds of data form the day-one recordings from the Table Talk Corpus, which contained the annotations for Engagement was used for the feature extraction. The features extracted from the speech signal consisted of the prosodic parameter (F0), Mel-frequency Cepstral Coefficients (MFCCs) and Voice Quality (glottal parameters).

The analysis was performed using 25 ms frame duration and 5 ms frameshift. F0 was estimated using the RAPT algorithm [26] and 12-dimensional MFCC coefficients (with log energy) were computed with the SPTK toolkit (http://sp-tk.sourceforge.net/). The voice quality (glottal parameters) consisted of the open quotient (OQ), return quotient (RQ), and speech quotient (SQ). They were estimated using the method described in [27]. Open quotient (OQ) is the ratio of the duration of the open phase of the glottal cycle (when the glottal folds are open) by the pitch period. Speed quotient (SQ) represents the asymmetry of the glottal pulse. The return quotient (RQ) is related to the abruptness of the transition between

the open phase and the closed phase, which is proportional to the spectral tilt. These voice quality parameters have not been used for the classification of Engagement in a casual conversational context. These features were used to make feature vectors to model engagement using the SVM.

Different combination of features was used to make the feature matrices (F0, VQ, MFCCs, VQ+MFCCs, F0+VQ+MFCCs). Each of these feature matrices was used for classifying Engagement to understand how the combination of these features affected the performance of the classification.

4.2 Classification of Engagement

Support vector machines (SVMs) are a useful classification technique. SVMs have been used extensively in speech emotion recognition [28] [29] [30]. The input for the SVM is a set feature vector extracted using the feature extraction methods described in 4.1 along with the high/low/neutral engagement annotations. The class labels and the attributes(speech parameters) constitute the feature set. Since there are three target values (class labels), the problem at hand is a multi-class classification problem. The goal of the SVM is to produce a model, which is based on the training data, that can predict the class labels using the attributes separately and in combinations.

SVM is a supervised learning model i.e, it uses labelled data. For each feature vector of attribute-label pairs $(x_i, y_i), i = 1, ..., l$ where $x_i \in R^n$ and $y \in \{1, -1\}$, the Support vector machine is supposed to find the solution for the optimization problem which has been discussed in[31] and [32].

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \\ \text{subject to} \quad & y_i\big(w^T \phi(x_i) + b\big) \geq 1 - \xi_i, \\ & \xi_i \geq 1 \end{aligned} \tag{4.1}$$

As in the above equation, feature vectors x_i are mapped into a higher dimensional space by the function ϕ [33]. The SVM finds a hyperplane that separates the vectors into corresponding target classes. There may be many hyperplanes that separate the data, but the objective of the SVM is to find the optimal separating hyperplane that maximizes the margin of the training data. The term $C > 0$ in equation 4.1 is the penalty parameter for the kernel function.

In this classification experiment we are using a Radial Basis Function as the SVM kernel function which has the form, $K\big(x_i, x_j\big) = exp\big(-\gamma \|x_i -$

$x_j\|^2), \gamma > 0$. where γ is the kernel parameter. The RBF kernel is an appropriate choice as this kernel maps the feature vectors into the higher dimensional space non-linearly so that it can handle a non-linear separation problem by transforming the data and find the hyperplane that can separate the data based on the labels/output parameters specified.

The splitting of data into train, validation and test sets are not feasible in this case as the number of samples in the dataset is low, as it could drastically reduce the number of samples to learn the model. So a cross-validation method is used for evaluating the performance of the classifier. In this approach (k-fold cross-validation) the training set is split into K smaller sets. A model is trained using the k-1 of the folds as training data and the resulting model used with the remaining part of the data to compute performance measures. The performance measure for the k-fold cross-validation is the average of the values for the k iterations. In this classification experiment, we are using a 10 fold cross-validation (stratified cross-validation is used across all the classification experiments in this dissertation so that instances from all three classes are in each fold).

The Support Vector Machine implementation in Weka[34] using the LIBSVM package [25] was used as the classifier as it supports multi-class classification. A c-SVC type SVM with radial basis function kernel was used, where the C is the penalty parameter of the error function. The default parameters were used for C, epsilon and gamma for the classification.

As the instances in each class were not balanced, a meta classifier that makes the SVM classifier cost-sensitive was used. The Low-Engagement and Neutral-Engagement classes had fewer instances than the High-Engagement class. So the instances were reweighed to give a higher cost to misclassifying the Low-Engagement and Neutral-Engagement classes that had less number of instances. The cost-sensitive meta classifier was supposed to minimize the expected misclassification cost[35]. The use of such a meta-classifier doesn't require the change in the underlying classifier.

4.3 Classification Results

The results of the classification using different speech parameter combinations have been given in the table 4.1.

Pitch related features (F0), MFCC, and voice quality parameters are widely used in emotion recognition in acted and spontaneous speech data and have produced state of the art results [36], so the classification of Engagement with these features was expected to achieve competent per-

formance. The classification results were better than the ZeroR baseline classifier accuracy of 50% and were statistically significant with $p<0.05$.

Feature Set	Accuracy(%)	Precision	Recall	F-Measure
F0	70.92	0.51	0.71	0.59
VQ	72.23	0.73	0.72	0.62
F0 + VQ + MFCCs	71.62	0.54	0.72	0.64
MFCCs	73.34	0.73	0.73	0.66
VQ + MFCCs	73.99	0.72	0.74	0.71

Table 4.1: Results of classification using original engagement annotation

The performance measures accuracy, precision, recall, and F-measure were calculated to evaluate the performance of the different features in the classification task. All the features and their combination performed better than the baseline classifier performance. Voice Quality parameters and MFCCs resulted in accuracies of 72.23% and 73.34% when these features were used separately. The combination of these features gave a slightly better performance of 73.99%. The F0 performed with the lowest accuracy and F-Measure, and when F0 was combined with the Voice quality and MFCC parameters, the accuracy and other performance measures were lower. This performance decrease while using the F0 parameter, could be the result of silences in the samples, which needs to be further investigated.

The combination of voice quality features with the MFCCs improved the precision of 0.51 in the F0 feature set to 0.73 while which was higher in MFCCs alone (0.73). VQ+MFCCs improved the F-Measure and 0.59 in F0 to 0.71. The Recall values also show a similar trend in the classification, where the combination of features improved the recall than when the features were used separately.

Although the combination of these features didn't make a tremendous difference in the classification accuracy, the combination of features improved the classification performance significantly. The results of the classification achieved better performance from the methods in [9]. The results automatic classification of Engagement using the speech features shows that these features are good indicators of Engagement in casual conversations.

In this chapter, the use of features extracted from the speech samples for the classification of Engagement in conversational data was studied. The SVM classification was explored in order to understand whether these features were useful in modelling engagement in conversations. The results of the classification show that the speech parameters can be used as indica-

tors of Engagement in conversations and the combination of MFCCs with Voice Quality parameters improve the overall performance of the classifier.

Chapter 5

Engagement Classification Using Voting Average of Annotations

In this chapter, the process of deriving a voting average of engagement annotations from the original annotations used in chapter 4 is used as the target labels for classification. The method for deriving the voting average of annotations, the classification of engagement using the derived target labels and the feature matrix similar to the one used in previous chapter, the random forest classification algorithm are described here.

5.1 Voting Average of Engagement Annotations

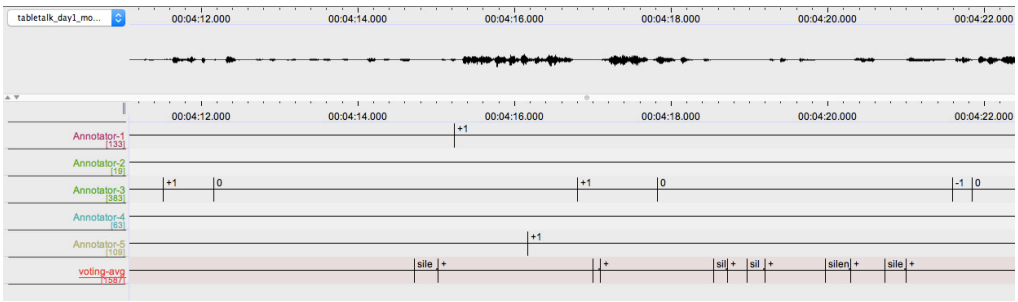

Figure 5.1: Voting Average of Engagement Annotations

A new annotation was derived from the original annotations using a voting average. Although the original annotations were in agreement mostly, there was slight disagreement in the case of some segments, i.e. segments annotated as engaged in one annotator was not annotated in another or

they were annotated with a different label, or the length of the annotation was different. To avoid inconsistency among annotations, a segment was annotated with a label of high/low/neutral Engagement if the majority of annotations agreed on it as high/low/neutral. Figure 5.1 shows the original annotations, and the last tier shows the derived annotation.

The audio samples were segmented into silence and speech. As the low level, acoustic features were of interest the silence segments were not used for the classification experiments. The speech segments were further segmented into utterances. The annotations of Engagement as high/low/neutral was based on the voting average of annotation from the original annotations. The majority label of Engagement from the original annotation was used as the label in the new annotation to avoid any ambiguity in the labels of speech segments.

5.2 Feature Extraction

In this experiment, the features were extracted in a similar manner, as described in section 4.1. The features were extracted from the utterances. The frame-level acoustic features were extracted from frames of 25 ms duration and 5 ms overlap.

The feature matrix for each set of features and the derived annotation of engagement labels were used for the classification. F0, MFCCs, and voice quality parameters were extracted and were used for classification.

5.3 Classification of Engagement

Random forests are an ensemble learning method for classification and regression, and they were first introduced by Ho [37]. Random forests algorithm fit several decision tree classifiers on sub-samples of the dataset and employs averaging the predictive accuracy and control over-fitting. Each tree-structured predictors are created from a random sample of features, thus the name random forests. Random forests are a combination of tree predictors in which each tree depends on the values of a random vector sampled independently and with the same distribution of all trees in the forest [38].

The final outcome of the random forest classification is taken as the average of the predictions of the trees [39]:

$$RandomForestPredictions = \frac{1}{K}\sum_{K-1}^{K} K^{th} tree\, response \qquad (5.1)$$

where the index K runs over the individual trees in the forest.

Random forests algorithm builds a number of prediction trees, each tree votes to make the random forest prediction. If the majority of trees in the random forest predicts an instance as a particular class, the random forest predicts it as that class. The number of features used to create each tree is limited so as to reduce the correlation between trees in the forest, which reduces the error rate of the random forest. The optimal number of features selected randomly for each tree in the forest is given by $log_2 M + 1$, where M is the number of features in the original dataset.

As the number of trees in the random forest increases, the generalisation error for random forest converges. The generalisation error of the random forest depends on the correlation between the individual trees and their strength. As the trees are created with randomly selected features from the dataset to split each node, the error rates of the random forest are favourable and robust against noise.

The internal measures such as the "Gini Impurity" or information gain/entropy can be used as an indicator of feature importance. This measure, Gini Impurity, indicate how often a particular feature was selected for a split. As each tree is constructed, the calculation of how much the error function drops for a variable at each split point can be made. The drops in error can be averaged across all the trees, and an estimate of the importance of each feature can be made. The importance of a variable is higher if the error drop was higher when that feature was used in the split.

A random forest learning algorithm implemented in Weka [34] was used for the classification of engagement. A cost-sensitive meta classifier was used to deal with the imbalance of classes in the data. The random forest classifier with bagging and 100 iterations was used as the base classifier. Attribute importance based on average impurity decrease (and the number of nodes using that attribute) was also measured to understand the relative importance of the speech parameters in the feature matrix.

The classification experiment was performed using different combinations of the speech features together with the engagement labels. Seven feature sets were used: F0, MFCC, VQ, F0+VQ, MFCC+F0, MFCC+VQ, MFCC+F0+VQ. A 10-fold cross-validation approach was performed to as-

sess the performance of the classifier.

The standard performance measures such as accuracy, precision, recall and F-measure were calculated to evaluate the performance of the classifier. These measures are useful for us in identifying the best combination of speech parameters for modelling engagement in casual conversations.

5.4 Classification Results

The results of the classifications using the various feature sets are shown in table 5.1 - 5.7. When F0 was used alone for classification, the accuracy has decreased from the experiment in 4. But the Precision and Recall for Low and Neutral Engagement instances have improved. The accuracy of the classification was improved by 4.15% from the previous experiment when VQ features were used alone. The Precision and Recall of Low and Neutral instances were significantly improved from the previous experiment.

MFCCs achieved the best performance when used alone as in the case of the previous classification. This was expected as MFCCs are found to be a robust set of features and has been used widely in similar tasks. The combination of features actually improved the classification performance in terms of all the measurement metrics. When all the features were used (F0, MFCCs & VQ), the accuracy improved from 71.62% to 88.836%. A similar improvement in the precision, recall and F-measure is also achieved.

Results of classification for each feature set is described in the following confusion matrices with corresponding precision and recall values.

	High Engagement	Low Engagement	Neutral Engagement	Precision
High Engagement	49798	17234	11490	0.63419
Low Engagement	5127	15322	3711	0.63419
Neutral Engagement	1776	836	4526	0.63407
Recall	0.87826	0.45885	0.22943	

Table 5.1: Confusion Matrix: Classification using F0 as feature

The overall accuracy of the classification while using F0 as the feature was 63.418% and the F-Measure was 0.53.

	High Engagement	Low Engagement	Neutral Engagement	Precision
High Engagement	59975	8553	9994	0.7638
Low Engagement	2873	18453	2834	0.76378
Neutral Engagement	708	977	5453	0.76394
Recall	0.94366	0.65944	0.29829	

Table 5.2: Confusion Matrix: Classification using Voice Quality (VQ) as feature

The overall accuracy of the classification while using VQ as the feature was 76.38% and the F-Measure was 0.53.

	High Engagement	Low Engagement	Neutral Engagement	Precision
High Engagement	67075	7298	4149	0.85422
Low Engagement	2234	20639	1287	0.85426
Neutral Engagement	397	642	6099	0.85444
Recall	0.96226	0.72217	52.874	

Table 5.3: Confusion Matrix: Classification using MFCCs as features

As shown in table 5.3, the overall accuracy of the classification while using MFCCs as the feature has improved to 85.424% and F-Measure of 0.78 is a significant improvement. Also, the recall values on Low Engagement and Neutral Engagement instances have also improved.

	High Engagement	Low Engagement	Neutral Engagement	Precision
High Engagement	62299	10879	5344	0.7934
Low Engagement	2234	20639	1287	0.79346
Neutral Engagement	397	642	6099	0.79364
Recall	0.94518	0.61859	0.43854	

Table 5.4: Confusion Matrix: Classification using F0+VQ as features

The overall accuracy of classification was 79.343% while using F0+VQ as the features and the F-Measure was 0.78.

	High Engagement	Low Engagement	Neutral Engagement	Precision
High Engagement	67734	5140	5648	0.86261
Low Engagement	1762	20852	1556	0.86308
Neutral Engagement	523	458	6157	0.86257
Recall	0.9675	0.78836	0.46082	

Table 5.5: Confusion Matrix: Classification using MFCC+F0 as features

The table 5.5 shows there has been slight increase the the classifiers performance as features are combined to classify Engagement. The overall accuracy of the classifier was 86.271% and the F-Measure was 0.7075.

	High Engagement	Low Engagement	Neutral Engagement	Precision
High Engagement	69664	4603	4253	0.88719
Low Engagement	1564	21434	1162	0.88717
Neutral Engagement	376	430	6332	0.88708
Recall	0.97291	0.80978	0.53903	

Table 5.6: Confusion Matrix: Classification using MFCC+VQ as features

While using MFCC+VQ as features for classification the overall accuracy was significantly improved to 88.718% and the F-Measure was 0.81 which is an improvement to the previous classifications.

	High Engagement	Low Engagement	Neutral Engagement	Precision
High Engagement	69751	5391	3378	0.88832
Low Engagement	1791	21461	908	0.88836
Neutral Engagement	248	546	6344	0.88876
Recall	0.97162	0.78331	0.5968	

Table 5.7: Confusion Matrix: Classification using MFCC+F0+VQ as features

As the table 5.7 shows the overall performance of the classifier improved when all the features were combined to create the feature set. The overall accuracy here was 88.836%, and the F Measure was 0.824.

The performance of the random forest classifier was better than the ZeroR baseline classifier accuracy of 50% and is statistically significant with p-value$<$0.05.

The improvement in the accuracy of the combination of features from the previous classification can be attributed to the removal of silence segments from the data and the refinement of the annotation of segments using the voting average. As the features were extracted from the utterances and avoided the silence segments, the F-measure of the classification while using the combination of features has improved from 0.64 to 0.824.

The improvement in accuracy can also be seen in the Voice Quality features (72.23% to 76.38%) and MFCCs (73.34% to 85.424%).

5.4.1 Importance of Features

The feature importance graph in figure 5.2 shows the importance of the contribution of each feature in the set towards correctly predicting the engagement labels. This is useful information in gaining insight into what features are indicative of engagement. Log energy coefficient of the MFCCs has high relative importance among the features. Most of the MFCCs have high relative importance, as shown in the graph. This was also seen in the performance metrics of the classification as the MFCCs performed with higher accuracy in the classification. F0 also has been shown to have high relative importance in the classification of engagement, voice quality feature RQ also has been one of the highly significant features in classifying engagement.

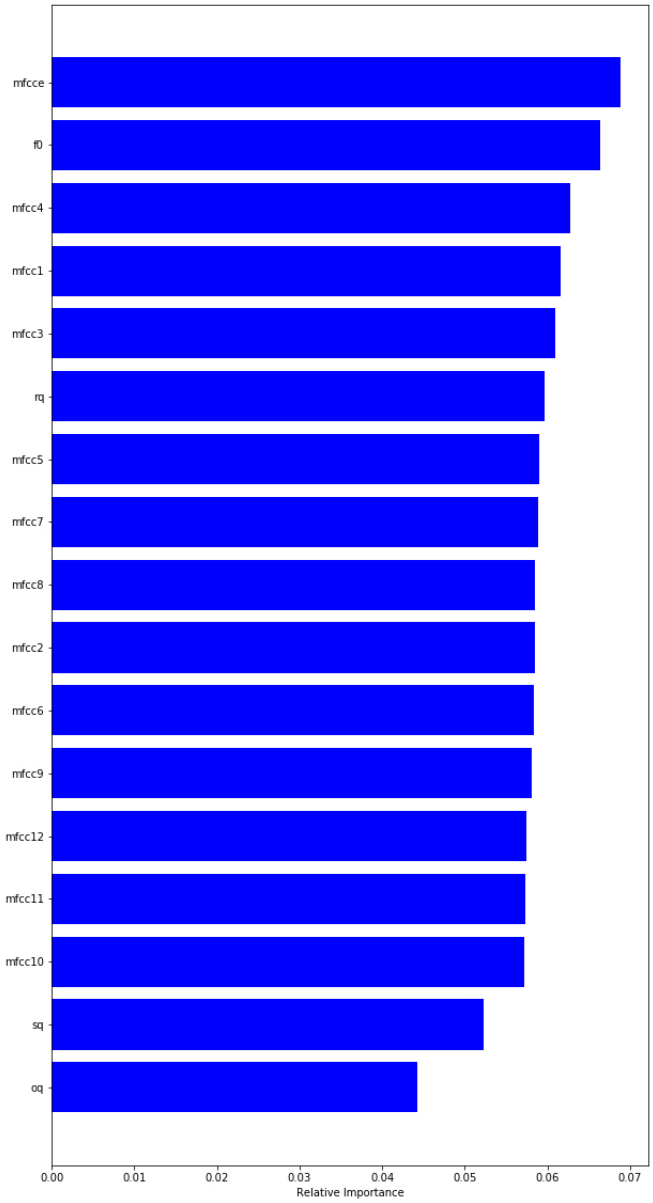

Figure 5.2: Relative Importance of Features for Engagement Classification

Although these relative importance measures are indicative of the usefulness of the features in classification of engagement, the impurity based subject to specific issues, feature importance measure based on the impurity reduction has a bias towards variables that have more categories [40]. Another issue is that when there are multiple features that are correlated, the model has no preference for either of those features, as they both may be equally important. If any of those features are used for the split, the impurity that any subsequent features can reduce is already being reduced by the first feature. Thus the first feature may appear to be a more meaningful indicator than the others. In the case of the engagement classification, the voice quality features are also important indicators of engagement as

the improvement in the accuracy and other metrics have been evident in the result of the classification. The real importance of these features in the classification of engagement may be very similar. The importance of these features may be interpreted from the relative importance measures as well as from the results of the classifications.

Chapter 6

Engagement Classification Using Additional Voice Quality Features

From the previous experiments, it was found that the combination of Prosodic (F0), Spectral and Voice Quality features extracted from speech samples can be used to recognize engagement. The voice quality features are not widely used for engagement recognition tasks. We expect that adding more information about voice quality can improve the performance of the engagement classification task.

6.1 Feature Extraction

The algorithms for speech feature extraction provided in the COVAREP [41] tool-kit was used for extracting features. Features were extracted with a 25 ms window with a 5 ms overlap. The following features were extracted from the speech samples,

Prosodic: Fundamental frequency (F0) and voicing (VUV)

Voice Quality: Conventional Glottal Parameters such as Normalized amplitude quotient (NAQ), Quasi open quotient (QOQ), the difference in amplitude of the first two harmonics of the differentiated glottal source spectrum (H1H2), parabolic spectral parameter (PSP) and wave-let based parameters like maxima dispersion quotient (MDQ), spectral tilt/slope of wavelet responses (peakSlope), and shape parameter of the Liljencrants-Fant model of the glottal pulse dynamics (Rd) also the posterior probability of the creaky voice detection is added as a feature.

Spectral: Mel cepstral coefficients (MCEP0-24), where the 0th coeffi-

cient is energy-related.

These features have been used in emotion detection [42], psychological disorder analysis [43], predicting respondent reactions in dyadic negotiation [44] etc.

The COVAREP feature set is selected in this classification experiment as it has been used previously in psychological state detection[45], emotion detection from speech [46], speaker attitude analysis [47] etc.

6.2 Classification of Engagement

The Random Forest classifier, as specified in section 5.3, was trained using the feature set and engagement labels. A 10 fold cross-validation was performed to evaluate the performance of the classification. The feature VUV (voiced/unvoiced) was used to select only the voiced speech frames for the classification of engagement as most conversational behaviour research examines voiced speech and have found it to contain more information useful for such classifications.

An improvement of the classification accuracy is expected as a result of additional voice quality features in the feature set. The feature importance of each feature for the classification of engagement is also analysed.

6.3 Classification Results

The table 6.1 shows the confusion matrix for the classification of engagement using the covarep feature set.

	High Engagement	Low Engagement	Neutral Engagement	Precision
High Engagement	72593	4744	1185	0.92449
Low Engagement	2355	21019	786	0.86999
Neutral Engagement	562	482	6495	0.86152
Recall	0.96137	0.80088	0.76719	

Table 6.1: Confusion Matrix: Classification using COVAREP features

The Number of Low Engagement and Neutral instances is less compared to the number of High Engagement segments. The lower precision and

recall of the Low Engagement and Neutral classes can be attributed to this unbalance in the data. This unbalance exists as it is challenging to get low engagement in data collected in laboratory settings.

The accuracy of the classification has significantly improved from all the previous experiments. The overall accuracy of the classification of engagement with the COVAREP feature set was 90.824%. We estimated that the use of additional voice quality features would improve the performance of the classifier and the results show that the new feature set with the additional voice quality features have significantly improved the performance of the classifier. Precision and Recall values have improved from the previous classifications, which shows that use of additional voice quality features has improved the classifier's performance by correctly classifying the high, low and neutral engagement instances.

The F-measure, a description of the balance between the learned model's exactness and completeness, has also been increased from 0.824 to 0.8627 than in the previous classification experiment.

6.3.1 Importance of Features

The chart shows the importance of features in the classification in figure 6.1. As in the previous classification experiment, Mel-Frequency Cepstral Coefficient 0 has the highest relative importance among the features used for classification. The probability measure of the speech segment to be creaky is also an essential measure of engagement.

It is interesting to note that the feature VUV has the lowest relative importance in the features used. This can be the result that the information about the voicing itself has not contributed to the performance of the classifier, but it is to be noted that the voiced instances have information that reflects the engagement and other behavioural traits of the human interlocutor.

As in the previous classification experiments, the MFCC features have high relative importance in the features used. F0 and Voice Quality features have moderate importance in the classification of engagement. Among the voice quality parameters creak, peakSlope, are the most significant features. MDQ and H1H2 have moderate relative importance while the other Voice quality parameters have relatively low importance among the features used for classification.

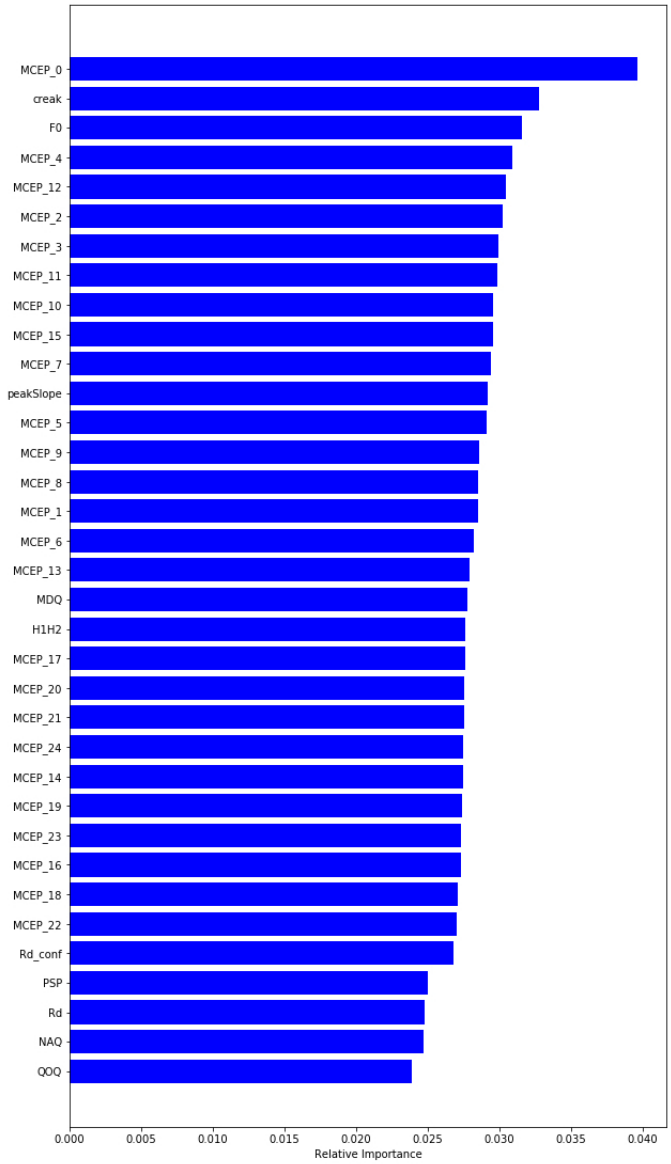

Figure 6.1: Relative Importance of Features for Engagement Classification(Additional Voice Quality Features)

As the correlation of these speech parameters may be influencing the relative importance of each feature in the classification when one feature is used for a split in the forest other features that are correlated with it might not have the chance to reduce the impurity at nodes. This effect is minimised by the random selection features for splitting in the random forest learning but might still affect the feature importance. So it is better to take into account the improvement of the performance measures of the classification as well to gain insight on the usefulness of these features.

In this chapter, the classification of engagement using speech parameters (with additional voice quality features) and engagement labels using

a random forest classifier (with meta classifier for cost sensitivity to deal with class imbalance) was performed. The use of more parameters of voice quality has contributed to the improvement of classification accuracy and performance measures. The relative importance of features in classification based on the mean impurity decrease was also measured.

Chapter 7

Conclusions

This dissertation aimed to analyse the speech parameters from conversational speech data and classify Engagement in casual conversations. The goal was to use speech parameters as indicators of Engagement in casual conversational data. These goals were achieved and were verified by various classification experiments using speech parameters and the engagement labels. The results provide a measure of how well these features indicate the Engagement of the interlocutors in casual conversations. The accuracy measures of various classification experiments show that these speech parameters are good indicators of Engagement of interlocutors in casual conversations.

From the experiments it was noted that the positively engaged(high engagement) segments were the majority in the corpus, this happens as it is challenging to find segments of negative Engagement and neutral Engagement (no increase or decrease) in data collected in laboratory settings. This was a limitation of this dissertation, and future research could be oriented towards the negative/low engagement and neutral Engagement of an interlocutor and how it is reflected in the speech parameters.

The research carried out as part of this dissertation gives insight into the use of speech parameters (prosodic, spectral, voice quality) as indicators of Engagement in casual conversations. A Three class classification of Engagement (High/Low/Neutral) was performed in this research, analysing the use of these speech parameters to recognise varying levels of Engagement could be an exciting step. The results of these classification experiments provide a measure of how well each feature indicates Engagement and several accuracy measures for the classification process. The F-measure for the classification of Engagement has been improved when additional voice quality features were added to the feature set. The MFCCs are good indicators of Engagement as they have shown consistent

improvement in the performance of classification tasks. The use of the random forest as the classification algorithm was the right decision as they can generalise well when the number of features is extensive. As the corpus used for this study was relatively smaller, the evaluation was performed using k-fold cross-validation(10 fold) as splitting the data into train, validation and test sets were not feasible as it would reduce the number of instances for the model to train. The research question "Is it possible to generalise the patterns in speech features as indicators for measuring the Engagement of a participant in a conversation?" has been answered as the classifiers modelled using speech features were able to classify Engagement in the conversation with significant levels of accuracy. The use of a combination of various speech features such as voice quality and MFCCs with the widely used pitch based features (F0) has significantly improved the overall accuracy as well as the recall and precision of Neutral and Low Engagement instances even though they were at a lesser quantity in the corpus used.

The relative importance of the speech parameters used for the classification gives insight on which parameters are contributing more towards the classification of Engagement. This information can be used to test these parameters further to improve classification performance. The development of a corpus for engagement analysis is required to further investigate the use of speech parameters as indicators of Engagement. Analysing Engagement at tri-level is the approach followed in this dissertation; multiple levels of Engagement can also be analysed, given annotation of the data into varying levels of Engagement. A multi-modal approach where audio, visual and other social signal are used as indicators of Engagement is a potential alternative approach. Conversational Engagement is a social behaviour of humans; the ability to understand it makes us socially intelligent. In human-computer interaction scenarios, this ability to understand the Engagement of the human partner can make the computer more social and make the interactions more natural.

Bibliography

[1] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[2] I. Poggi, *Mind, hands, face and body. A goal and belief view of multimodal communication*. Weidler, 2007.

[3] C. L. Sidner and M. Dzikovska, "Human-robot interaction: Engagement between humans and robots for hosting activities," in *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, p. 123, IEEE Computer Society, 2002.

[4] D. Gatica-Perez, "Modeling interest in face-to-face conversations from multimodal nonverbal behavior," *Multimodal Signal Processing*, pp. 309–323, 2009.

[5] J. C.-y. Hsiao, W.-r. Jih, and J. Y.-j. Hsu, "Recognizing continuous social engagement level in dyadic conversation by using turn-taking and speech emotion patterns," in *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[6] D. Bohus and E. Horvitz, "Models for multiparty engagement in open-world dialog," in *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '09, (Stroudsburg, PA, USA), pp. 225–234, Association for Computational Linguistics, 2009.

[7] F. Bonin, R. Bock, and N. Campbell, "How do we react to context? annotation of individual and group engagement in a video corpus," in *2012 International Confernece on Social Computing (SocialCom)*, pp. 899–903, IEEE, 2012.

[8] R. Voigt, R. J. Podesva, and D. Jurafsky, "Speaker movement correlates with prosodic indicators of engagement," in *Speech Prosody*, vol. 7, 2014.

[9] C. Yu, P. M. Aoki, and A. Woodruff, "Detecting user engagement in everyday conversations," *arXiv preprint cs/0410027*, 2004.

[10] J. Gustafson and D. Neiberg, "Prosodic cues to engagement in non-lexical response tokens in swedish," in *DiSS-LPSS Joint Workshop 2010*, 2010.

[11] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human-robot interaction," in *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*, pp. 375–382, IEEE, 2010.

[12] C. Lai, J. Carletta, and S. Renals, "Detecting summarization hot spots in meetings using group level involvement and turn-taking features," in *Proc. Interspeech 2013, Lyon, France*, 2013.

[13] A. E. Norris, H. Weger, C. Bullinger, and A. Bowers, "Quantifying engagement: measuring player involvement in human–avatar interactions," *Computers in human behavior*, vol. 34, pp. 1–11, 2014.

[14] D. Szafir and B. Mutlu, "Pay attention!: designing adaptive agents that monitor and improve user engagement," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 11–20, ACM, 2012.

[15] G. Dubuisson Duplessis and L. Devillers, "Towards the Consideration of Dialogue Activities in Engagement Measures for Human-Robot Social Interaction," in *International Conference on Intelligent Robots and Systems*, Designing & Evaluating Social Robots for Public Settings Workshop, (Hambourg, Germany), pp. 19–24, Sept. 2015.

[16] J. Hernandez, Z. Liu, G. Hulten, D. DeBarr, K. Krum, and Z. Zhang, "Measuring the engagement level of tv viewers," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pp. 1–7, IEEE, 2013.

[17] D. C. Richardson and R. Dale, "Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension," *Cognitive science*, vol. 29, no. 6, pp. 1045–1060, 2005.

[18] R. C. Schmidt, C. Carello, and M. T. Turvey, "Phase transitions and critical fluctuations in the visual coordination of rhythmic movements

between people.," *Journal of experimental psychology: human perception and performance*, vol. 16, no. 2, p. 227, 1990.

[19] R. Gupta, C.-C. Lee, D. Bone, A. Rozga, S. Lee, and S. Narayanan, "Acoustical analysis of engagement behavior in children," in *Third Workshop on Child, Computer and Interaction*, 2012.

[20] N. Glas and C. Pelachaud, "User engagement and preferences in information-giving chat with virtual agents," in *Workshop on Engagement in Social Intelligent Virtual Agents (ESIVA)*, pp. 33–40, 2015.

[21] M. Vels and K. Jokinen, "Recognition of human body movements for studying engagement in conversational video files," in *Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication, August 6-8, 2014, Tartu, Estonia*, pp. 97–105, Linköping University Electronic Press, 2015.

[22] M. Charfuelan, M. Schröder, and I. Steiner, "Prosody and voice quality of vocal social signals: the case of dominance in scenario meetings," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[23] N. Campbell, "An audio-visual approach to measuring discourse synchrony in multimodal conversation data," in *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*, pp. 2159–2162, ISCA, 2009.

[24] K. L. Gwet, *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

[25] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[26] D. TALKIN, "A robust algorithm for pitch tracking (rapt)," *Speech Coding and Synthesis*, pp. 495–518, 1995.

[27] J. P. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Towards an improved modeling of the glottal source in statistical parametric speech synthesis," in *6th ISCA Workshop on Speech Synthesis, Bonn, Germany*, 2007.

[28] Y. Chavhan, M. Dhore, and P. Yesaware, "Speech emotion recognition using support vector machine," *International Journal of Computer Applications*, vol. 1, no. 20, pp. 6–9, 2010.

[29] T. Seehapoch and S. Wongthanavasu, "Speech emotion recognition using support vector machines," in *Knowledge and Smart Technology (KST), 2013 5th International Conference on*, pp. 86–91, IEEE, 2013.

[30] P. Shen, Z. Changjun, and X. Chen, "Automatic speech emotion recognition using support vector machine," in *Electronic and Mechanical Engineering and Information Technology (EMEIT), 2011 International Conference on*, vol. 2, pp. 621–625, IEEE, 2011.

[31] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152, ACM, 1992.

[32] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[33] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," tech. rep., Tech. rep., Department of Computer Science, National Taiwan University., 2003.

[34] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, pp. 10–18, Nov. 2009.

[35] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[36] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pp. 474–477, IEEE, 2005.

[37] T. K. Ho, "Random decision forests," in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, vol. 1, pp. 278–282, IEEE, 1995.

[38] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[39] T. Hill and P. Lewicki, *STATISTICS Methods and Applications*. StatSoft, Tulsa, USA, 2007.

[40] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, p. 25, Jan 2007.

[41] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep—a collaborative voice analysis repository for speech technologies," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 960–964, IEEE, 2014.

[42] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, (New York, NY, USA), pp. 3–10, ACM, 2016.

[43] S. Scherer, G. Stratou, G. Lucas, M. Mahmoud, J. Boberg, J. Gratch, L.-P. Morency, *et al.*, "Automatic audiovisual behavior descriptors for psychological disorder analysis," *Image and Vision Computing*, vol. 32, no. 10, pp. 648–658, 2014.

[44] S. Park, S. Scherer, J. Gratch, P. J. Carnevale, and L.-P. Morency, "I can already guess your answer: Predicting respondent reactions during dyadic negotiation," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 86–96, 2015.

[45] N. Cummins, B. Vlasenko, H. Sagha, and B. Schuller, "Enhancing speech-based depression detection through gender dependent vowel-level formant features," in *Conference on Artificial Intelligence in Medicine in Europe*, pp. 209–214, Springer, 2017.

[46] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition.," in *Interspeech*, pp. 3603–3607, 2016.

[47] D.-K. Mac, T.-L. Nguyen, A. Michaud, and D.-D. Tran, "Influences of speaker attitudes on glottalized tones: A study of two vietnamese sentence-final particles," in *ICPhS XVIII (18th International Congress of Phonetic Sciences)*, 2015.