DEVELOPMENT OF AN AUTOMATIC ATTITUDE RECOGNITION SYSTEM: A MULTIMODAL ANALYSIS OF VIDEO BLOGS

Noor Alhusna Madzlan

Thesis submitted for the Degree of Doctor in Philosophy
School of Linguistics, Speech & Communication Sciences

University of Dublin
Trinity College

Under the Supervision and Direction of
Asst. Prof. Breffni O'Rourke and Prof. Nick Campbell

October 2017

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work. Wherever there is published or unpublished work included, it is duly acknowledged in the text.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Noor Alhusna Madzlan

Summary

Communicative content in human communication involves expressivity of socio-affective states. Research in Linguistics, Social Signal Processing and Affective Computing in particular, highlights the importance of affect, emotion and attitudes as sources of information for communicative content. Attitudes, considered as socio-affective states of speakers, are conveyed through a multitude of signals during communication. Understanding the expression of attitudes of speakers is essential for establishing successful communication. Taking the empirical approach to studying attitude expressions, the main objective of this research is to contribute to the development of an automatic attitude classification system through a fusion of multimodal signals expressed by speakers in video blogs. The present study describes a new communicative genre of self-expression through social media: video blogging, which provides opportunities for interlocutors to disseminate information through a myriad of multimodal characteristics. This study describes main features of this novel communication medium and focuses attention to its possible exploitation as a rich source of information for human communication. The dissertation describes manual annotation of attitude expressions from the vlog corpus, multimodal feature analysis and processes for development of an automatic attitude annotation system. An ontology of attitude annotation scheme for speech in video blogs is elaborated and five attitude labels are derived. Prosodic and visual feature extraction procedures are explained in detail. Discussion on processes of developing an automatic attitude classification model includes analysis of automatic prediction of attitude labels using prosodic and visual features through machine-learning methods. This study also elaborates detailed analysis of individual feature contributions and their predictive power to the classification task.

Acknowledgement

All praises, above and beyond, to Allah the Most Gracious and Most Merciful, for giving me the will and strength to complete this doctoral dissertation. Greatest appreciation to my sponsors in Malaysia, the Ministry of Higher Education and Sultan Idris Education University (UPSI). Deepest gratitude goes to my academic supervisors in Trinity College Dublin, Assistant Professor Dr. Breffni O'Rourke and Professor Dr. Nick Campbell for believing in me. I may have not completed this research if not for their constant support and guidance. For that, I am greatly indebted. Special thanks also to Dr. Loredana Sundberg Cerrato from ADAPT Centre, Trinity College Dublin for the helpful insight and encouragement.

Deepest love and gratitude to my beloved husband, Nik Saiful Nizan for believing in me. I also wish to convey my appreciation to my beloved parents, Madzlan Aziz and Sharifah Ahmad for their unconditional love and support. In extension, I thank my supportive siblings, Noor Alhuda and family, Noor Al Iman and family, Noor Alyaqeen and family, Ahmad Munzir and family and Ahmad Basyeer for their endless advice and encouragement. You have all kept my spirit alive and I am sincerely thankful.

Special gratitude also to my incredibly supportive colleagues Francesca Bonin, Frank Han, Emer Gilmartin, Yuyun Huang, Justine Reverdy, Christy Elias, Sucheta Ghosh and others from the Speech Communications Laboratory and SCSS, TCD. Last but not least, huge thanks to my loyal friends and support system in Dublin, Nur Adyanie, Azim, Azita Iliya, Julinawati, Hasbullah, Ika Rastika, Zurina, Asnul Hadi and family, Siti Norhayati, Raihana Zahirah, Marlini and family and others. Thank you for keeping my sanity intact and being with me in this long winding journey. Through all the laughter and pain, I truly value our friendship and memories here in Ireland. Appreciation also goes to friends and family in Malaysia. Lastly, thank you to those who have contributed to this research.

Related Publications

[1] Noor Alhusna Madzlan, Jingguang Han, Francesca Bonin and Nick Campbell. Towards automatic recognition of attitudes: Prosodic analysis of video blogs. *Speech Prosody*. Page 91-94, 2014

[2] Noor Alhusna Madzlan, Jingguang Han, Francesca Bonin and Nick Campbell. Automatic recognition of attitudes in video blogs - Prosodic and visual feature analysis. *Fifteenth Annual Conference of the International Speech Communication Association, INTERSPEECH*. Page 1826 - 1830, 2014

[3] Noor Alhusna Madzlan, Yuyun Huang and Nick Campbell. Automatic classification and prediction of attitudes: Audio-visual analysis of video blogs. *Speech and Computer: 17th International Conference, SPECOM*. Springer. Volume 9319, Page 96 - 104, 2015

[4] Noor Alhusna Madzlan, Justine Reverdy, Francesca Bonin, Loredana Sundberg Cerrato and Nick Campbell. Annotation and Multimodal Perception of Attitudes: A Study on Video Blogs. *Third European Symposium of Multimodal Communication (MMSYM)*. Page 50-54, 2015

Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Meaning	Page
SVM	Support Vector Machine	9
RF	Random Forest	9
VLOG	Video Blog	10
AES	Affective-Epistemic States	18
F0	Fundamental Frequency	19
HNR	Harmonicity-to-Noise ratio	19
B1	Bandwidth of the First Formant	25
H1	Amplitude of the First Harmonic	26
A1	Amplitude of the First Formant	26
H1-H2	First harmonic relative to the Second Harmonic	26
H1-A3	First harmonic relative to the Third Formant	26
NAQ	Normalised Amplitude Quotient	26
ASM	Active Shape Model	28
FAC	Facial Action Coding system	29
AAM	Active Appearance Model	29
DCT	Discreet Cosine Transform	30
HMM	Hidden Markov Model	30
ANOVA	Analysis of Variance	31
HCI	Human Computer Interaction	32
wMEI	Weighted Motion Energy Images	33
GP	Gaussian Process	36
GMM	Gaussian Mixture Model	40
LDA	Linear Discriminant Analysis	41
Vlogger	Video Blogger	47
Hz	Hertz	80
dB	Decibel	80
SEC	Second	80
FMEAN	Mean value of the fundamental frequency	80
FMIN	Minimum value of the fundamental frequency	80
FMAX	Maximum value of the fundamental frequency	80
FPCT	Percentage of shape of the pitch contour	80
FVCD	Percentage of the vibration of the vocal folds	80
PMEAN	Mean value of power/ intensity	80
PMIN	Minimum value of the power/ intensity	80
PMAX	Maximum value of the power/ intensity	80
PPCT	Intensity movement	80
DN	Duration	80
PCA	Principal Component Analysis	90
PC	Principal Components	90
PC1	First Principal Component	97
PC2	Second Principal Component	97
PC6	Sixth Principal Component	97
ASR	Automatic Speech Recognition	100
RBF	Radial Basis Function	114

Chapter 1

Introduction

This dissertation highlights affective expressions as a source of information content in human communication. The study explores the dynamics of attitudinal expressions conveyed through several types of non-verbal signals. Attitudinal expressions between communicators must be understood and displayed appropriately to ensure successful message transfer during the communication process. This research further extends understanding of attitudinal aspects of expression, not only through the exploration of human perception of attitudes, but also through the development of automatic attitude recognition which will be of great utility for automatic understanding of user perception and information retrieval.

This chapter briefly discusses human communication, its different signalling modalities for communicating content of various types, as well as the role of attitudinal states in the communicative setting. Further discussion about the meaning of attitudes and their relation, or non-relation, to affect and emotions is also explored. Several definitions and concepts are introduced to give a clearer overall view of the methodology involved in the study. The final part of this chapter explains the research motivation and objectives.

1.1 Human Communication

Research in domains including social psychology, social signal processing and affective computing has increasingly focused on understanding the dynamics of human communication. Humans communicate with each other in a number of different modalities and the interpretation of their utterances may vary from a person to the other. Generally, humans communicate with each other to share information. Human communication involves a series of processes and can take place between two or more interlocutors. This sharing of information is transferred through signals.

Allwood [1] elaborates on the purpose of human communication, which is dissemination of information. Allwood mentions several types of information involved in the communication process:

1. physiological states (fatigue and hunger)
2. character, identity, personality (being timid, aggressive)
3. affective-epistemic attitudes (showing joy, friendliness, surprise)
4. factual content giving information about beliefs, assumptions about facts
5. communication management (feedback, turn-taking)

Among the information streams used in communication, one interesting type of information is the expression of affective-epistemic attitudes. This type of information is useful for understanding affective states and emotions of the people involved in the communicative setting. Emotions, affect and attitudes seem to be similar concepts but there are clear distinctions between them. Damasio [2] distinguishes between emotion and feelings, stating that feelings refer to the inner, cognitive experience of an emotion while emotion is the observable response of these feelings. The frequently cited five basic emotions are sadness, anger, fear, surprise, disgust and happiness [3]. There are certainly other emotional ways of representing emotional states than the simply categorical, including valence based representations [4] but such concepts are not sufficient to describe the complexities of people's affective states [5]. A person could express a mixture of sometimes contradictory feelings simultaneously. For instance, when a person finds that her friend is leaving the country to pursue her dream job, she expresses first the feeling of happiness and shifts her expression to sadness. Hence, human emotion, although easily shown, are often complex and confusing to understand or interpret.

Affect, on the other hand, is considered a general term for the inward feelings of the human experience. This concept is a broader representation of feelings where emotion contributes to a large part of the overall definition. While Zanna and Rempel [6] refer to affect as "any thoughts that are infused with strong, weak or no emotion at all", Shouse [7] claims that affect has no relation with feelings and emotion. The term affect is sometimes used interchangeably with emotion. It is believed that there are differences between the two concepts. Emotion refers to the display of feelings while affect is a non-conscious experience. Affect is a pre-personal experience of the speaker that unconsciously affects the consequent feelings and actions of the speaker. Hence, affect is viewed as a broad, abstract concept of the humanly cognitive experience, while emotion refers to the inner feelings of the person. Figure 1.1 summarizes the relationship between affect and emotion.

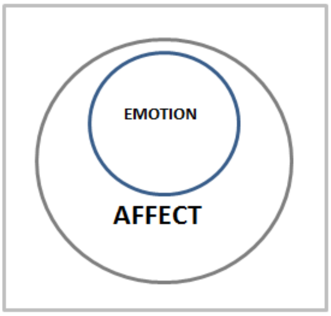


Figure 1.1: Relationship between Affect and Emotion

Affect is a general and broader representation of a person's feelings while emotion is the displayed response of affect, and thus is incorporated into the affective state of a person.

Recent empirical studies, particularly in the field of Affective Computing, use affect to refer to affective-epistemic states of humans when interpreted by machines. This use of affect is a broad term as a point of reference to studies on emotion, attitude and behavioural states of humans displayed in the communicative setting. This study supports the concept of attitude as representation of speakers' affective-epistemic characteristics. Attitude is viewed differently from concepts of affect and emotion. The following section details attitude as a concept for study.

1.2 Attitude as a Communicative Function

The previous section explains the differences between affect and emotion. The concept of attitude is related to affect and emotion, but attitude has distinctive traits. Zanna and Rempel [6] describe attitude as states that may be expressed as strong emotions or may be identified solely from the way an individual behaves with an object. Research on the psychology of attitude finds limited agreement on precise definitions and characteristics of attitude. Psychologists relate attitude with beliefs, opinions, habits and values [8]. Oskamp [8] explains attitude as having three main components, as shown in Table 1.1:

Components	**Description**
Cognitive	Ideas and beliefs of the agent towards the object
Affective (emotional)	Feelings and emotions the agent has towards the object
Behavioural	The agent's actions towards the object

Table 1.1: Atittude Components

Fishbein and Ajzen [9] suggest that attitude is mainly associated with the affective component. This is believed to be true as these components are independent and separate entities but are still interrelated. The affective component of attitude is an object of interest to several fields of research including psychology, economics, and marketing. However, some researchers use the term attitude loosely and often interchangeably with affect or emotion. In this study, these concepts, although similar, are treated independently. Attitude is of particular interest for this study as it represents observable traits of speaker's cognitive experience. Attitude refers to actions, outer representation of feelings, while emotion refers to inner feelings that are difficult to evaluate. Attitude is believed to be a pragmatic concept for the understanding of affect and emotions [10]. This interpretation of emotional states involves the crucial aspect of intended, voluntary and controlled actions. Auberge [10] explains communicative functions in Figure 1.2.

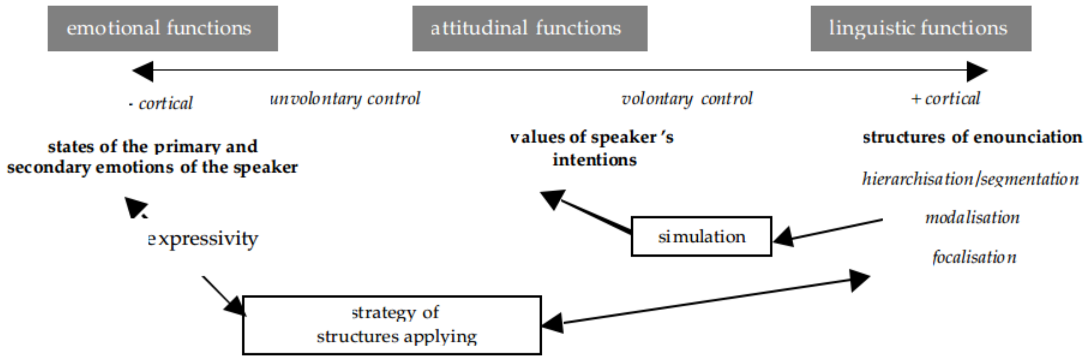


Figure 1.2: Attitude as a Communicative Function

Figure 1.2 shows attitudinal functions as a component for communication. The scale indicates emotional functions to be at the far left of the control scale while linguistic functions are on the far right. Attitudinal function is in the middle of the control scale. This means that attitudes, unlike emotions, are not vague and involuntarily expressed. Rather values of speaker intention are expressed in a conscious, controlled and voluntary manner.

The definition of attitude may or may not have a strong emotional component but always has an evaluative component. In fact, attitude and emotion can be seen as a continuum, a degree of emotional involvement in attitudes. Emotion, conversely, need not have a strong attitudinal - i.e. evaluative - component. This is because emotions are not necessarily directed towards any particular content, that is, you can be happy or sad (for example) without being able to identify why. Additionally, emotions can be internal and less susceptible to conscious control. Whereas attitudes are held more consciously, often deliberately expressed (in language), and subject to control in the sense of reflective evaluation.

Emotion and attitude are conceptually distinct, but interact with one another in reality. Attitudes always have an element of evaluation. Oskamp [8] has identified three components of attitude; cognitive, affective/emotional and behavioural. In this thesis I am concerned with the overt expression of attitude, which I take to be evidence of the first two, internal, components, i.e., the cognitive and affective/emotional states of the speaker. Under one understanding of the term attitude, attitudes can be entirely internal, comprising the cognitive and affective states, but never overtly expressed. For the purposes of this research, the focus is on attitudes as overtly displayed. Therefore, when I use the term attitude, it will frequently refer to the overt expression of attitude.

1.3 Multimodal Communication

Attitude is expressed in the content of communication through several different channels or modalities. Multimodal communication is defined as "co-activation, sharing and co-construction of information simultaneously and subsequently through several modes of perception and production" [11]. This concept adds to the concepts of communication with the sharing of information as its main function. However, multimodal communication involves several modes of information sharing using simultaneous sensory channels. These include sight, hearing, touch, smell and taste. Multimodal communication involves two main processes; firstly multimodal integration (perceiving information based on several modalities) and multimodal distribution (producing information using multimodalities). This is similar to the traditional communicative process, where both communicative agents perceive information and produce feedback. The difference is that there is emphasis on simultaneous sensory modalities.

Multimodal communication has become an important research area as humans perceive and produce communicative expressions through several modalities. A combination of speech, visuals and gesture facilitates communicators to meet their communicative goal of information sharing. As mentioned in Section 1.1, the contents of information include affective-epistemic attitudes. The simultaneous use of multimodal signals facilitates expression of attitudes. Due to the complexities of affective attitudes, the use of more than one sensory modality is beneficial to better understand and interpret attitudes. For example, a person can perceive a friend's expression of friendliness through rising tone of voice, display of a smile and the wave of a hand. Simultaneous use of speech, facial expressions and gestures facilitates the communicators' message transmission.

1.3.1 Multimodalities as Signals in Communication

Numerous studies suggest the relation between display of affective states and non-verbal gestures [12] [13] [14] [15]. Vinciarelli and Valente [16] refer to non-verbal communication as the transmission of a message through non-verbal behavioural cues, such as facial expressions, vocalisations, gesture and posture. Non-verbal communication relays speakers' inner feelings, whether intentionally or unintentionally. For instance, a mother outwardly expresses approval when her child voluntarily makes the bed by establishing eye contact with the child, praising her with a rising tone and giving a bright smile. In another scenario, the mother tries not to show disapproval when the child throws a tantrum by speaking with a level or falling tone and controlling her facial expressions from showing contempt. This example shows how people communicate feelings and intentions through the dynamics of several and simultaneous non-verbal signals.

With the emergence of the body of knowledge of Social Signal Processing, the role of this area is to firstly provide physical quantification and synthesis of non-verbal signals through which the affective behaviour of humans are expressed, and, secondly, to implement non-verbal signals in conversational agents [12]. This area is especially interesting as multimodal traits of humans are quantified through development of recognisers and synthesisers. This study aims to evaluate and measure communicative contents, in particular attitudinal representation of speakers. The outcome can be helpful in facilitating better understanding of relationships in human-human and human-computer interaction.

1.4 Multimodal Affective Systems

One method of quantifying attitudinal states of speakers through multimodal signals is by applying machine-learning methods. Research in the fields of affective computing, speech technology and human-computer interaction employs machine-learning techniques. There are numerous studies conducted in emotion recognition based on analysis of sentiment [17] [18], prosody [19], facial features [15], gestures [20] and posture [21]. Multimodality (using combination of signals) is also studied in great detail in emotion recognition. However, when a distinct separation of terminology between emotions and attitudes is made, there is little research conducted on automatic attitude recognition. Although there are some notable studies on prosodic attitudes [22] [23][24], research in automatic classification of attitudes using multimodal signals is still scarce.

In this century, there is growing interest in social media as a rich source of human expression. People from different areas of study, including sociolinguistics, psychology, affective computing and human-computer interaction conduct studies using data from social media. Social psychologists, for example, investigate people's communicative purposes in online blogs [25] [26] [27]. Researchers are also interestered in the social activity of micro-blogging on Facebook [28] [29]. Other empirical studies involve research in online search engines [30] [31], Twitter [32] [33] and YouTube [34] [35] [36]. Through conceptual and empirical research, it is evident that social media offers a rich source of social discovery and scientific application. Similarly, open access to instances of human behaviour, in particular, attitude expressions, from social media users shows a dynamic social representation worthy of exploration and understanding.

1.4.1 Use of Machine Learning Techniques

Machine learning is used in the area of Artificial Intelligence to enable computers to learn without being explicitly programmed. The learning takes place through observations of new data, hence the outcome of this implementation is a system that enables automatic classification of the learned data. Essentially, there are several methods of machine learning – supervised, semi-supervised and unsupervised learning. Supervised learning involves the task of training a dataset that has a group of features and object labels. The machine is constructed to predict and identify the label of an object given the set of features. Unsupervised learning however functions without the identification of labels. Instead, labels are derived based on the training of the dataset given by the features provided.

Classification in supervised learning is used to predict categorical labels given a set of observations. Figure 1.3 [37] illustrates the work flow of building a classification model using supervised learning technique:

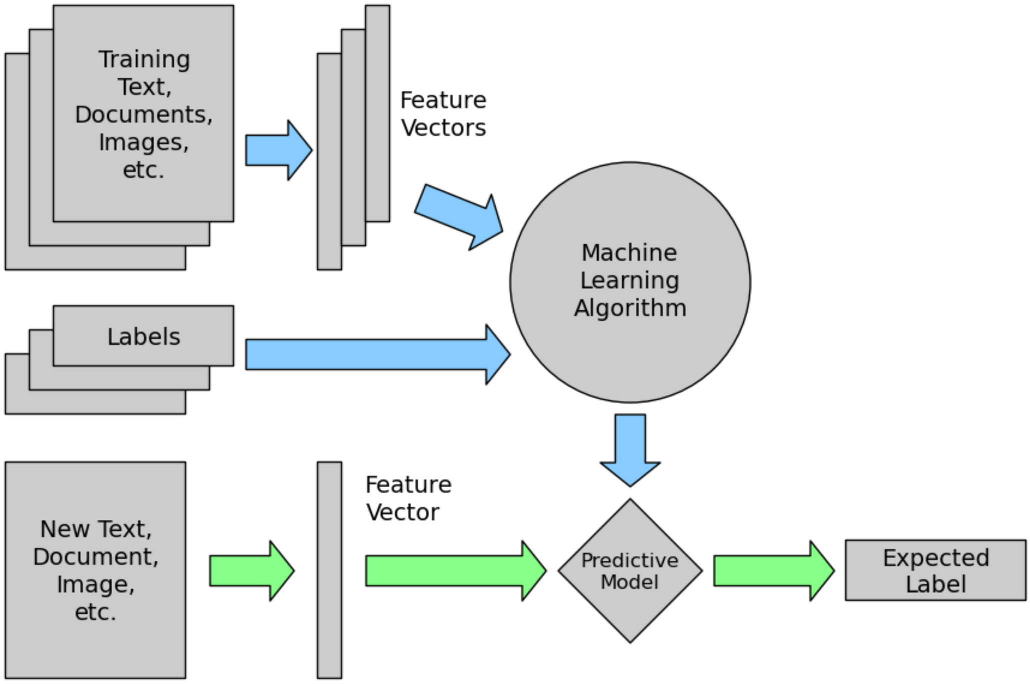


Figure 1.3: Flowchart for Supervised Machine Learning

Figure 1.3 generally describes the processes involved in developing a predictive model using a supervised machine learning method. Following this training process, a predictive model is derived and will produce expected labels.

There are several learning algorithms developed for machine learning classification tasks. One popular method of implementation is the use of statistical learning algorithms [38]. This study adopts this method by using Support Vector Machine (SVM) and Random Forest (RF) classifiers, as will be explained in Chapter 5.

1.5 Conceptual Definitions

This section addresses conceptual definitions used in this thesis. The list describes definitions for the following terminologies:

1. **Annotation**: Annotation is a metadata that is associated to another data by commenting and noting. In this study, annotation involves the act of assigning attitude labels on relevant parts of the video blogs. Annotation of attitude states creates another set of data comprising of only video segments that are annotated with an attitude label.

2. **Attitude Expression**: Attitude refers to pragmatic interpretations of emotion [10]. The expression of attitudes are socio-affective states of speakers expressed in voluntary and controlled settings.

3. **Machine-Learning**: Machine-learning is an area of study in Artificial Intelligence where automatic algorithms are constructed to allow learning and prediction of computers from the given data.

4. **Prosody**: Prosody refers to the suprasegmentals in the voice. The use of prosody in this study covers aspects of fundamental frequency and pitch contours, intensity of the voice, voice quality and duration of speech segments.

5. **Visual**: Visual refers to facial observations from videos used in this study. This reference to visual features includes movement of the eyes and other facial expressions of the speaker's face.

6. **Vlog**: Vlog is an abbreviated term for video blog. This abbreviation is used in this research to refer to YouTube videos of speakers where they share stories of daily life and events.

1.6 Motivation

There has been much work in the fields of Social Signal Processing and Affective Computing in developing recognisers for affective states of speakers through multimodal signals. Human beings can detect differences in affective expressions using data gathered through their eyes and ears. Machines can also do the same using information recorded by cameras and microphones. So by physically capturing and then analysing affective states of speakers, researchers are able to understand human behaviour through concrete displays of affective expressions. Treating attitude as dissimilar to affect and emotion, the development of an attitude recognition system through the means of multimodal signals is an interesting step forward towards understanding human behaviour. Apart from that, development of automatic recognisers has two main purposes:

First Automatic information retrieval and categorisation. Recognition systems are used to index and retrieve information from metadata or other forms of media. Collection of information obtained from these systems enables other systems to implement this source of information for affective modelling.

Second Applications for human-machine interaction. This system could be applied to artificial communicative agents such as social robots for technological advancement. Communicative content can be interpreted by an automatic system and used to inform the behaviour to robots or avatars. This is useful to then develop socially intelligent robots that understand and respond appropriately to humans.

This study aims to contribute to the first purpose of developing automatic recognition systems. Communicative content, in particular, affective information is useful to be embedded into systems for information retrieval and categorisation. This creates a source for people to retrieve information on attitude expressions during speech.

Attitude expressions are especially prevalent in social media where people find comfort and openness in expressing themselves better than in face-to-face communication. One advantage of social media is the ease of public access. This allows researchers to investigate human behaviours, particularly affective expressions from a dynamic source. YouTube, for example, allows public access for people to share videos about their daily life. This rich source of naturally-occurring communication is especially interesting to further understand human behaviour. Treating these videos as representations of people's attitude expressions makes the collection of this corpus interesting for development of an attitude recognition system that could be applicable to other areas of study.

1.7 Statement of Problem

To the best of my knowledge, little research has so far been conducted in the area of automatic attitude recognition, particularly in exploiting social media corpus such as YouTube video blogs to understand multimodal expressions of humans. This study introduces the concept of attitudes as an evaluative representation of human social experience through spontaneous talk in vlog speech. Following that, internalising attitude states in a computer system for artificial intelligence also contributes to this study's major work by applying multimodal signals into the recognition system.

1.8 Research Objectives

The main objective of this research is to develop an automatic attitude recognition system for information retrieval and categorisation. To achieve this, the following objectives are defined:

Objective 1 Collection of a corpus that represents attitude expressions. Expressions of attitudes are observed and investigated from social media. The use of social media is essential as they provide dynamic and rich multimodal signals that include speech, facial expressions and gestures to indicate speakers' attitude expressions. This study annotates and segments five attitude expressions from vlogs in YouTube.

Objective 2 Investigation of the use of multimodal features as contributors to developing a reliable attitude recognition system. Relevant prosodic and visual features are extracted and analysed. Subsequently, specific prosodic and visual features are selected and examined to understand their contribution towards the classification model.

Objective 3 Development of a reliable automatic attitude recognition system. This research highlights the development of an attitude recognition system that can predict different attitude states of speakers. The development of this system creates a concrete understanding of attitude expressions through multimodal signals, namely prosodic and visual features. Supervised machine learning is conducted using Support Vector Machines (SVM) and Random Forests (RF).

1.9 Detailed Outline

This thesis is divided into six main chapters which are listed below:

Chapter 1 Discussion of general concepts of human communication and multimodalities in affective expression of attitudes. This chapter also outlines research motivation, statement of problems and three main research objectives for the present study. Conceptual definitions used in this thesis are also listed in this chapter.

Chapter 2 Elaboration of discussion and criticisms of theories of attitudes, multimodalities and use of social media for understanding human communication. This chapter also outlines some of the recent works in the study of automatic recognition through multimodalities using machine learning techniques.

Chapter 3 Discussion of a collection of the vlog corpus. Vlogs are collected from an online video sharing website, YouTube. An ontology of attitude annotation scheme is introduced in this chapter. The processes of collection, annotation and segmentation of attitude states of speakers are described in detail.

Chapter 4 Discussion of the processes involved in developing an automatic attitude classification system. This process involves steps for multimodal feature extraction, from prosodic and visual means. The second part of the chapter discusses processes of feature selection. This section highlights prominent prosodic and visual features that provide greatest contribution to the classification system in recognising different attitudinal states of speakers.

Chapter 5 Supervised machine-learning techniques and results on building a reliable attitude classification system. A total of three experiments are reported:

a) development of an attitude classification system through prosodic modality

b) use of prosodic and visual modalities to develop an automatic attitude classification model

c) improved use of multimodal feature sets for improving the classifier

Chapter 6 Statement of general conclusions, research limitations, suggestions for improvement and future direction of this study.

Chapter 2

State of the Art

This dissertation focuses on developing an attitude recognition system for the purpose of classification and retrieval of human attitudes from video recordings. Prior to more elaborate discussion, this chapter presents the state-of-the-art from related literature addressing the notion of attitude, its relevance to multimodal and novel forms of media, and recent applications in recognition interfaces. The present work adopts some of the conceptual frameworks and methods used from past literature to address the research goals. This chapter begins with a description of related literature concerning attitude definitions and their relation to communicative content.

2.1 Communicative Content

As we have noted, Allwood [1] sees the exchange of information between communicators. We will now explore the five types of communicative information that he identifies (repeated here for convenience):

1. Physiological states
2. Character, Identity and Personality
3. Affective-Epistemic Attitudes
4. Factual Contents
5. Communication / Feedback Management

Physiological states are expressed in the communicative setting, and these include fatigue and hunger. The character of a person is also communicated in speech, examples include the aggressiveness and openness of an individual. Affective-epistemic attitudes refer to the affective or emotive state of the speaker: for instance friendliness, joy and surprise. Examples of factual content include information about beliefs, theories and assumptions, while communication management consists of turn-taking, sequences of feedback and topic change [1]. Another aspect of communicative content that is relevant to the present work in this thesis is Affective-Epistemic Attitudes. This content is essential for the communicative process during information transfer. The next section describes relevant literature relating to attitudes.

2.1.1 Attitude - Definition

Communicative content during communication involves several types of information. One crucial source of information, as mentioned by Allwood [1] is the affective-epistemic attitudinal states of communicators. To better understand the communicative process as well as the modes of interaction, it is essential to first understand the role that attitudes have [1]. In the literature, attitudes are interchangeably related to affect and emotions. However, this thesis identifies attitudes as a distinctive term and this is described through the definition and components of attitudes.

Malhotra [39] states distinctions between affect and attitude, which agrees with Ajzen that attitude refers to summary evaluations of an object or behaviour [9] [40]. Bargh and Chartrand explain attitude as being judgments that are outcomes of spontaneous and unconscious effort [41]. They claim that attitude can be automatically activated without any prior goals of judgment [41]. This perspective, however, is not supported by Auberge [10] who views attitude as driven by intention, voluntary and controlled action. Auberge describes a continuum (see Figure 1.2) of controlled expression of affect where emotion is regarded as an inward and involuntary cognitive experience. Attitude is pragmatic representations of emotions, voluntary and controlled expressions of the cognitive state while linguistic function is the linguistic means of affective expression. Hence, attitude stems from a pragmatic level of conscious and controlled action, and not from a spontaneous and unconscious effort.

The definition of attitude relates to three dimensions: cognitive, affective and behavioural [8]. All three components constitute the definition of attitude, where attitude is built on the assumption towards an object and the reaction of the communicator through affective and behavioural judgments of the said assumption. Agarwal and Malhotra [42] conducted a study on cognitive and affective components of attitude. The study involved public opinion of use of sneakers among selected undergraduate students. A self-administered questionnaire was distributed to participants to understand their choice of sneakers as well as their mood states prior to the test. Results from the study proposed a model for a multidimensional method of assessing opinions and affective states of participants to better understand participant attitude. Their work provides insight to the depiction of multiple dimensions that define attitude, namely cognitive, affective and behavioural components and their correspondence with each other.

It is inevitable to regard cognitive and behavioural components as important in defining attitude. However, assumptions, beliefs and opinions: features of cognitive and behavioural attitudinal components, involve personal states of the being. Affective attitudes are highly related to reactions triggered by the communicative process between communicators. They are components existing in both intrapersonal and interpersonal settings. To better understand the concept of affective attitudes, the next section briefs on the related literature of affective attitudes.

2.1.2 Affective Attitudes

Yan et al. [43] conducted a study on perception of attitudes among Chinese subjects. They regard attitudes as social affects which are expressed outwardly in controlled settings. They define attitudes with elements of beliefs and opinions (whether intended or not) within social settings. The study discusses results of a perception test for the recognition of the selected attitudes. In their classification of social affect, there are three main groups comprising attitudes, social parameters and social context. The corpus contains recordings of a female Chinese speaker speaking 399 utterances using 19 attitudes. These attitudes are shown in Table 2.1.

Attitudes			
Declaration	Question	Admiration	Irony
Confidence	Irritation	Doubt	Contempt
Disappointment	Resignation	Obviousness	Neutral Surprise
Positive Surprise	Negative Surprise	Politeness	Authority
Seduction	Intimacy	Infant Directed Speech	-

Table 2.1: 19 affective attitudes

The perception test was given to 30 native Mandarin Chinese listeners who were asked to recognise different attitudes in the corpus. The results show that all participants were able to recognise all the proposed attitudes except for Confidence. They extended the study to a clustering analysis of attitudes. Declaration, was mostly confused with Confidence and Politeness. Resignation and Disappointment were also confused between the two. The strength of this study is the validation of attitude states through a perception test. Finding 19 attitudes to be quite elaborate, they conducted cluster analysis to identify which attitudes were distinctive or confused the most. In addition, the definition of attitudes as social affects as explained in this study is relevant to this research.

Allwood et al. [44] conducted a similar study on attitude perception through audio-visual modalities. In this work, they attributed attitudes as affective-epistemic states (AES). The definition of AES is defined from their previous work [45], which refers to "internal states that simultaneously involve cognition, perception and emotion". This includes affective epistemic attitudes such as boredom and surprise. Similarly, Allwood et al. conducted a perception test to validate these AES using multimodal stimuli. The study involved 12 Swedish participants who were presented with recordings from the NOMCO First Encounter Corpus [46] which contained gestures annotated according to an adapted version of the MUMIN annotation scheme. Participants were shown a 2-minute long clip of the corpus and they were asked to choose any words that describe both affective-epistemic and behavioural states. Results from semantic analysis led to seven types of AES: happiness, interest, nervousness, confidence, disinterest, thoughtfulness and understanding. The audio-visual modality shows most attributions for nervousness, interest and thoughtfulness. Interesting findings from this study suggest that the expression of an AES may be conflicting or complementing according to different modalities. For example, happiness is expressed best through the audio modality but not vividly shown in video modality. Hence, multimodal expressions of AES are more complex to perceive [44].

Treating attitudes as a generic terminology, Henrichsen and Allwood [47] used a standard 10-attitude set for prosodic analysis in the NOMCO corpus [46]. They conducted prosodic analysis on a selection of the 10 attitudes for the preparation of machine-learning methods. This work, however, focuses attention on attitude annotation and subsequent development of a standard attitude annotation scheme. Annotators of the corpus chose attitude labels and 439 attitude events were labelled. Further analysis resulted in a reduced number of attitude labels that show rich representation of the corpus. The ten attitudes are: Interested, Friendly, Casual, Bored, Thoughtful, Confident, Amused, Enthusiastic, Uninterested and Impatient. Using these attitude labels as predictors, machine-learning was conducted using prosodic features, in particular fundamental frequency (f0), Intensity and Harmonicity-to-Noise Ratio (HNR). Results show that prosodic signals provide faster and better information in predicting different attitudes. The strength of this work is its development of a simple and standard attitude annotation scheme that is useful for researchers to exploit in machine-learning and empirical research.

2.1.3 Prosodic Attitudes

Recent studies in social signal processing build upon the definition of attitudes as a source of information in the communicative process. The use of speech, in particular, prosody as a signal for social actions has close congruity with attitudes [48][49] [50] [22].

Morlec et al. [48] highlight prosodic attitudes with reference to the expression of what people feel (attitudes) using intonation (part of prosody). They suggest that attitudes have strong correspondence to the prosodic conditions of the person. This work makes clear distinctions from cognitive and lexicon-based approaches to measuring attitudes. Their study introduced six attitudes expressed in French from the inter-perceptual-center group (IPCG) [51] melodic curve corpus. The corpus consisted of 322 utterances for each of the six attitudes – Assertion, Question, Exclamation, Incredulous Question, Suspicious Irony and Evidence. They conducted a perception study among 20 participants to validate the six attitudes using training and testing sentence module. Results suggest that there exist confusions between Incredulous, Question and Suspicious Irony despite clear prosodic distinctions. Apart from that, production of synthetic sentences using statistical evaluations performed slightly lower than natural production of utterances. This initial study on prosodic attitudes makes a compelling case for bringing a standard definition of attitudes from prosodic perspectives.

Rilliard et al. [50] conducted a study concerning perception of prosodic attitudes through audio-visual modalities. Extending work on the six prosodic attitudes developed by Morlec et al. [48], they included audio-visual recording of the six attitudes from two French speakers. They were asked to speak three sentences with a 5-syllable length for each of the six attitudes. They further conducted a perception test that presented different modalities to 32 French listeners. Results from the perception test show that listeners were able to recognise different attitudes better by the first speaker than the second speaker. Audio-visual modality also proved most helpful for listeners to identify these prosodic attitudes, particularly Obviousness and Suspicious Irony. The interesting part of this study is that they conducted a cluster analysis to understand confusions between these 6 attitudes, despite attaining good recognition rates for each attitude. From analysis, they found that Doubt-Incredulity and Surprise-Exclamation were confused in the audio modality, while Question and Doubt-Incredulity were confused when presented in video stimuli. For the audio-video stimuli, video helps in distinguishing Exclamation from Doubt-Incredulity. This work is helpful in providing clear distinctions between the six prosodic attitudes by conducting a perception study and cluster analysis through different modes of stimuli.

A similar study was conducted on prosodic attitudes in cross-cultural settings. Mac et al. [22] define prosodic attitudes as a person's speech relaying opinions about the interlocutor. They conducted a cross-cultural perceptual study on audio-visual attitudes from Vietnamese participants. A total of 20 Vietnamese and 20 French listeners were involved with the listening experiment. This test required participants to listen and watch recordings in audio, video and audio-visual modalities of 16 Vietnamese attitudinal expressions produced by one native speaker from Hanoi. Results suggested mixed performance of native and non-native speakers of Vietnamese. Both participants performed best in perceiving different attitudes through audio-visual means. French participants performed best in understanding Authority and Irritation attitude classes through the audio modality while native Vietnamese participants were able to indicate Declaration, Obviousness, Authority and Colloquial attitudes successfully through audio modality. Through clustering analysis, French participants performed better than the native participants in perceiving Admiration attitude.

This research agrees with the definition of prosodic attitudes as socio-affective states of speakers. Attitudes are affective states expressed in social settings. However, the essence of prosodic attitudes is their close connection to prosody, particularly when it involves linguistic (phonological) differences between different languages and cultures. These attitudes are derived based on melodic curves of the utterances. In creating an attitude recognition where multimodal signals including prosody is applied, this study focuses attention to treating attitudes as socio-affective states in simpler terminology. Despite differences in the aspects of terminology, this research supports the works of Henrichsen and Allwood [47] in its simpler and standard version of attitude definition within a larger realm of signaling, which does not merely involve speech and prosody, but also other aspects of multimodalities.

Recognition of attitudinal states is not always universally perceived and understood [43] [22]. Several factors influence people's perception of attitudinal states such as age and gender. Research in emotion recognition (considered equivalent to attitude recognition) shows the influence of age and gender in recognising different emotion states. Orgeta and Phillips [52] conducted analysis on age effects for emotion recognition. Forty undergraduate participants aged between 17 to 37 and 40 community dwellers above 61 years old participated in the study. Participants were presented with images indicating six basic emotions and based on the facial stimuli, they were required to attribute the image to the emotion. Results found that young participants were able to identify sadness better than the older group. A notable observation from this study is that age contributes to differences in emotion recognition although results from both groups do not show a vast difference. Young people show slightly higher ability to identify emotion faster and more accurately than older people.

This hypothesis is supported by Di Domenico et al.[53]. They conducted a similar study on positive and negative emotion recognition among two groups of participants with varying age range. Forty young and forty older participants were required to recognise different emotional states based on 10 videos. They had to indicate positive and negative emotions when watching the videos. Results show that the older group recognised the happy or positive expression better than the younger group. Hence, this study supports the notion that difference in age contributes to the recognition of emotions.

Hall and Matsumoto [54] conducted a study on gender variations in perceiving emotions. They conducted two related studies on gender variation between 69 male and 27 female participants, who were required to view facial expressions for 10 seconds and rate the indicated emotion based on seven types of emotions. Findings show that women provided varying emotions compared to the male participants. This indicates that women are more successful in identifying different types of emotions. The second study involved a larger group of participants, consisting 126 male and 237 female students. They were presented with videotapes from the JACBART corpus [55], where a 1 second emotion expression was included in the speaker's neutral face. Analysis found similar results from the first study, where women performed better than men in recognising different emotions through non-verbal signals. This work is helpful to better understand gender differences in perceiving affective states of speakers when presented with non-verbal conditions. To provide a concise overview of attitude research, Table 2.2 summarises past literature outlining different attitude categories.

No.	Author(s)	Title	Attitudes
1	Lu, Yan, Aubergé, Véronique and Rilliard, Albert (2012) [43]	Do you hear my attitude? Prosodic perception of social affects in Mandarin	Declaration, Question, Admiration, Confidence, Irritation, Doubt, Contempt, Disappointment, Resignation, Obviousness, Neutral Surprise, Positive Surprise, Negative Surprise, Politeness, Authority, Seduction, Intimacy, Infant Directed Speech
2	Allwood, Jens, Lanzini, Stefano and Ahlsén, Elisabeth (2014) [44]	Contributions of different modalities to the attribution of affective-epistemic states	Happiness, Interest, Nervousness, Confidence, Disinterest, Thoughtfulness and Understanding
3	Morlec, Yann, Bailly, Gérard and Aubergé, Veronique (1997) [48]	Generating the prosody of attitudes	Assertion, Question, Exclamation, Incredulous Question, Suspicious Irony and Evidence
4	Rilliard, Albert and Martin, Jean-Claude and Aubergé, Véronique and Shochi, Takaaki (2008) [50]	Perception of French audio-visual prosodic attitudes	Assertion, Question, Exclamation, Incredulous Question, Suspicious Irony and Evidence
5	Mac, Dang-Khoa, Aubergé, Véronique, Rilliard, Albert and Castelli, Eric (2010) [22]	Cross-cultural perception of Vietnamese Audio-Visual prosodic attitudes	Declaration, Interrogation, Exclamation of Neutral Surprise, Exclamation of Positive Surprise, Exclamation of Negative Surprise, Obviousness, Doubt-Incredulity, Authority, Irritation, Sarcastic Irony, Scorn, Politeness, Admiration, Infant-Directed Speech, Seduction, Colloquial
6	Henrichsen, Peter Juel and Allwood, Jens (2012) [47]	Predicting the attitude flow in dialogue based on multi-modal speech cues	Amused, Enthusiastic, Casual, Friendly, Impatient, Confident, Interested, Bored, Uninterested, Thoughtful
7	De Moraes, João Antônio, Rilliard, Albert, De Oliveira Mota, Bruno Alberto and Shochi, Takaaki (2010) [23]	Multimodal perception and production of attitudinal meaning in Brazilian Portuguese	Arrogance, Authority, Contempt, Irritation, Politeness, Seduction, Doubt, Irony, Incredulity, Obviousness, Surprise

Table 2.2: Attitudes in Past Literature

2.2 Multimodal Expression

People communicate through dynamic means of multimodal signals [56] [57] [58]. Multimodal expression puts emphasis on the fusion of different sensory channels in expressing different types of communicative content. Multimodal signals are used in both perception and production of communicative content. Allwood [56] suggests that communicators utilise speech and different bodily gestures as modes of production and make use of hearing and vision modalities for perception of communicative content. He further observes different modalities used by communicators according to different types of information. These types of information are listed in Table 2.3:

Type of Information	**Modality**
Emotions and Attitudes	Prosody, Gesture
Illustrations	Iconic or Conventional Gesture
Feedback	Content, Gesture

Table 2.3: Modalities used to indicate types of information

Table 2.3 illustrates relations between modalities used for different purposes of information sharing. Attitudes are best expressed using prosodic and facial gestures or head movements. Sharing an illustrated type of information, iconic gestures such as hand pointing makes the delivery livelier and engaging [56]. Interactive communication management involves the use of verbal (i.e., lexical and syntactic), prosodic and gestural modes to seek feedback from the interlocutor. The combination of verbal and non-verbal signals is central to face-to-face communication [56].

The current study adopts Allwood's notion of multimodal expression particularly in sharing affective or attitudinal information. Attitudes are best perceived through prosodic and visual signals. The following sections address related literature pertaining to prosodic, visual and facial signals used when expressing affective states.

2.2.1 Prosodic Signals

Prosodic information in speech is integral for delivery of different communicative contents [56] [59] [60]. Nooteboom [61] refers to prosody as the melody of speech that is derived from non-segmental phonemes within an utterance. As Laver [62] and Roach [63] mention, prosody, also referred to as a part of suprasegmentals, contains four main components: pitch, loudness, duration and articulatory quality.

Pitch refers to the perceptual concept of melodic movement of the voice [62]. The acoustic and measurable equivalent is fundamental frequency [62] [47]. Changes in pitch give indications of changes to speaker's different attitudinal states. Roach [60] develops a model for attitudinal change according to changes in speaker's tone of the voice. The fall tone indicates an attitudinal state of finality and certainty while the rise tone indicates a questioning state in the speaker. A fall-rise tone indicates uncertainty or doubt while rise-fall indicates the state of surprise or being impressed. Henrichsen and Allwood [47] conducted a study on attitude labels and applied fundamental frequency as one of the parameters for machine learning applications. Results found fundamental frequency, a measurement unit for pitch, is essential for understanding the role of attitudes in the communicative setting.

Laukka et al. [64] conducted a study on the relationship between vocal expressions and emotion. By means of vocal expressions including vocal intensity or loudness of the speaker's voice, emotive states of speakers were identified. Actors expressed seven different emotions and 30 listeners involved in this study were required to recognise the emotions. These emotional dimensions were activation, valence, potency and intensity. Analysis of vocal cues particularly in vocal intensity of the voice showed that listeners were able to recognise positive valence when the intensity of the voice is low. The study shows that activation, potency and emotion intensity are indicated through high intensity rate.

Hanson and Chuang [65] conducted a study on glottal characteristics of male speakers. This work provides knowledge on the influence of voice quality in recognising speaker characteristics. They analysed the voice quality of 21 adult male speakers and compared results with previous work on female glottal characteristics. They introduced a theoretical measurement of voice quality through speech waveforms and spectrum. Their voice quality measurements involved measuring first-formant bandwidth (B1), amplitude of the first harmonic relative to the second harmonic (H1-H2), amplitude of the first harmonic (H1) relative to the first-formant prominence in the spectral domain (A1) and the first harmonic relative to the third-formant spectral peak (H1-A3). Male speakers were required to speak one sentence in a normal tone. The sentences were repeated using three different vowels and participants needed to repeat the same sentence three times using these three vowels. Results showed that there are significant differences in glottal characteristics in H1-H2 and H1-A3. What this shows is that changes in spectral tilt correlate with perceived voice quality and contribute to gender variations. This finding based on changes in spectral tilts and open quotients is relevant for acoustic measurement of speaker's voice quality.

Campbell and Mokthari [66] propose voice quality as a prosodic dimension by measuring breathiness through Normalised Amplitude Quotient (NAQ). They measured glottal phonation of one female Japanese speaker, uttering over 13,000 utterances. Speaking styles (polite, friendly and casual) and speech acts (exclamations, giving information, requesting information, muttering and requesting repeats) of the female speaker with groups of interlocutors (child, family, friends, others, self) were annotated by labellers. Results showed that voice quality shows significant differences according to the speech acts and speaking styles. This work contributes to the belief that voice quality features should be regarded as a prosodic characteristic.

Crystal [59] and Roach [63] support an integrated analysis paradigm combining different prosodic characteristics such as pitch, intensity, voice quality and other vocal features when measuring speaker's overall vocal activity. This is helpful for applications of prosodic characteristics in measuring attitudinal states of speakers. Table 2.4 summarizes relevant literature on applications of prosodic parameters (fundamental frequency, power, / energy, voice quality, speech rate, tempo) in recognition systems:

No.	Author(s)	Title	Prosodic Features
1	Dellaert, F, Polzin T, Waibel, A (1996) [67]	Recognizing emotion in Speech	Pitch, Rhythm
2	Nwe, TL, Wei, FS (2001) [68]	Speech based emotion classification	Mel-Frequency speech power coefficients (Power / Energy)
3	Gobl, C, Ni Chasaide A (2002) [19]	The Role of Voice Quality in communicating emotion, mood and attitude	Voice Quality
4	Yanushevskaya, I, Gobl, C, Ni Chasaide A (2008) [69]	Voice Quality and Loudness in affect perception	Voice Quality, Loudness
5	Mozziconacci, S (2001) [70]	Emotion and attitude conveyed in speech by means of prosody	Pitch Range, Speech Rate
6	Blanc, JM, Dominey, PF (2003) [49]	Identification of prosodic attitudes by a temporal recurrent network	Fundamental Frequency (f0)
7	Hirschberg, J, Litman, D, Swerts, M (2004) [71]	Prosodic and other cues to speech recognition failures	Fundamental frequency (f0), Energy values (RMS), Length of Pause, Tempo, Silence
8	Busso, 2004	Mood Affect Attitude	Pitch, Intensity, Speech and Non-Speech Ratio

Table 2.4: Applying Prosody in Recognition Systems

2.2.2 Visual and Facial Signals

Non-verbal behaviour consisting of expressions of visual activity, influences content delivery during the communicative process [72] [73]. Ekman [72], one of the pioneers of emotive facial expression research suggests the notion of the face as an indication of a person's emotional state. Scherer [73] further states various functions of non-verbal features. Emotions (used interchangeably with attitudes), categorised under the pragmatic function of non-verbal signals, are indicated determinately by the face, and assisted with hand gestures, posture, gaze patterns and other bodily movements. In point of fact, Luettin and Thacker [74] state that not only does visual information facilitate acoustic signals for speech production, it accommodates information in noisy settings, a situation that is difficult to capture by merely using acoustic signals which are determinately more sensitive to noise.

Much research in facial recognition highlights individual facial contours to recognise different communicative contents [75] [76] [77] [74]. Sadro et al. [77] conducted a study on the role of eyebrows in face detection. A collection of still images of celebrities was shown to 18 participants of the study. These images were manipulated to show eyes only or eyebrows only. Participants were required to identify the celebrities based on name or other form of identification. The results showed that the percentage of recognition without eyes or without eyebrows significantly dropped. However, recognition error was most notable for the no-eyebrow stimuli. This work supports its hypothesis that eyebrows play a major role in detecting faces. Although the work of Sadro et.al [77] focuses mainly on face identification, the present study (highlighting attitudinal states of speakers) aims to examine the role of eyebrows and other features that contribute most to attitude identification.

Luettin and Thacker [74] postulate a model for speechreading using visual information. They believe that the lip contour, particularly the inner lips, provides most visual information. To test their theory, they test their algorithm for lip tracking using the Active Shape Model (ASM), developed by Cootes et al. [78]. Samples for training data consisted of a subset of the Tulips 1 database [79] containing 96 images of 12 speakers. From the sample data, they analysed the performance of their algorithm from three perspectives, as shown in Table 2.5:

Category	Position of Lip Contour
Good	One quarter of the lip thickness deviation
Adequate	One quarter and half the lip thickness deviation
Miss	Out of lip contour

Table 2.5: Categories for Lip Tracking using ASM

Findings from this work show that the Good category for lip tracking achieved slightly 98%. This serves as a reliable model for tracking of lips, particularly for the purpose of speechreading. The use of the ASM approach provides insight for automatic tracking of individual (in this case, the lips) visual features for recognition systems [80].

Shifting from studies that focus on individual facial features, research in facial recognition develops models for face recognition paying particular attention to detecting multiple facial features. Ekman and Friesen [81] postulated the Facial Action Coding system (FAC) for the purpose of classifying facial activity. This system provides a grammar for describing facial expressions of emotional meaning [82]. However, Ekman and Friesen developed this model not only to cater for emotion detection based on facial movements, but as a comprehensive model that can be implemented for any study on facial behaviour [81]. Numerous studies applied the FAC model in quantitative studies of facial behaviour. For instance, Bartlett et al. [82] extended the study on FAC by developing a fully automated facial recognition system in spontaneous settings. They selected two datasets, the first dataset from the DFAT-504 database [83] which consists of videos of 100 students displaying 27 different facial expressions, with initial neutral faces. The second dataset involved videos of 24 participants displaying facial actions annotated by FAC experts. Analysis revealed a 91% agreement between the system (trained with Support Vector machine) and the human FAC labels. The strength of this work is evident in a successful agreement rate achieved through machine learning methods.

Another notable model for multiple visual feature detection is an extension of the Active Shape Model (ASM), called Active Appearance Model (AAM). As previously mentioned, ASM highlights visual feature tracking using the statistical model of shape in the facial regions [78] [80]. AAM, an extended work of Cootes et al. [84] is a model-based approach which postulates a robust and fast matching of not only the shape but texture or appearance of an object to a new image. Cootes et al. [85] conducted a study that compares performance of ASM and AAM for facial detection. The experiment involved two sets of data, the first dataset includes 400 face images with 133 points, while the second data consists of 72 MR brain images with 133 marked points. Results from this work suggest that AAM worked best in the face data due to minimal texture errors while ASM performed best in localised regions of the facial contour, as is evident in brain images from the second dataset. The AAM is robust because it combines a larger capture region which subsequently tackles not only the shape, but the texture of grey-scale images. The ASM is useful for image interpretation because it concentrates on the localised region or shape of the object, by which the model points provide most helpful information [85].

Neti et al [86] conducted a study on audio-visual recognition of speechreading. They conducted visual analysis by implementing the AAM approach on over 4000 images from the IBM ViaVoice database. A comparison of visual tracking approaches was conducted between AAM and Discreet Cosine Transform (DCT), using Hidden Markov Model (HMM) for the speechreading recognition task. Results show that DCT performed better in recognition performance in comparison to AAM. Neti et al. state two justifications for the poor performance of the AAM which are errors in modelling and tracking. Modelling errors refer to poor tracked sequences, which impedes accurate tracking of the object image. Insufficient labels of the training data is also a contributing factor to AAM's poor performance. This finding sheds light on the factors that influence the success or failure of visual tracking using AAM.

Research in visual recognition of interlocutor's behavioural states involves methods and approaches for facial tracking. Identification of behavioural states, particularly attitudinal states through facial movements is difficult to achieve [57] [82]. Nevertheless, facial signals complement other modalities in better understanding expressions of attitudes. The current research in this thesis makes use of facial information to recognise different attitude states of speakers in the vlog corpus.

2.2.3 Fusion of Multimodal Expressions

Research demonstrates that a fusion of several modalities increases the recognition of behavioural activity [56] [57] [58] [23] [87] [88] [89]. Use of multimodal sensory streams is prevalent during the communication process. Allwood [56] proposes two modes of production: speech and bodily gestures. He supports the notion that speech is the main mode of communication while bodily movements act as additional enforcers to the delivery of communicative content. Multiple modalities can either co-occur or happen in isolation during the process of human communication [90].

Several researches apply multimodal channels in perceptual studies of attitudes. Shochi et al. [91] conducted a study on audio-visual recognition of Japanese attitudes. Twelve Japanese attitudes were identified: Declaration, Interrogation, Admiration, Irritation, Exclamation of Surprise, Sincerity-Politeness, Doubt-Incredulity, Simple-Politeness, Evidence, Authority, Arrogance and Kyoshuku (ashamedness and embarrassment). Two male Japanese speakers produced a Japanese sentence with 12 types of attitudes recorded in audio, video, and in combined audio plus video. Results showed that an individual modality (audio or video) may be sufficient to recognise some attitudes without the complementary aid of the other modality. However, the researchers also observed that most attitudes were recognised best with the combination of audio and video. These findings therefore agree with Allwood's claim that other modalities are enforcers to a predominant modality.

De Moraes et al. [23] conducted a similar perceptual study on attitudes among Brazilian Portuguese speakers. Thirty participants speaking Brazilian Portuguese participated in the study. Multimodal stimuli consisting of audio, video and audio-video were presented to listeners. They were required to recognise two types of attitudes: social (Arrogance, Authority, Contempt, Irritation, Politeness, Seduction) and propositional attitudes (Doubt, Irony, Incredulity, Obviousness, Surprise). After conducting analysis of variance (ANOVA), both types of attitudes were distinguished based on the modality presented. Social attitudes were least recognised through the audio modality while propositional attitudes were best identified through a fusion of audio and visual modalities. This work contributes to the notion of multimodalities and their role as facilitators of attitudinal expression and behaviour.

The study of multimodalities is not only prevalent in human-human communication, but there is growing interest in exploring multimodal sensory streams in Human Computer Interaction (HCI) [87]. This body of knowledge explores the applications of multimodality into recognition systems through various approaches and techniques. Turk [89] explains that the objective of creating interfaces based on multimodal information is to produce systems that are robust, flexible and adaptable. Vo and Waibel [87] presented a recognition model aided by multimodal sensories. They conducted multimodal recognition analyses on numerous communicative contents, which include speech recognition, lip-reading, word-spotting, eye-tracking, gesture recognition and handwriting recognition. By creating a fusion of multimodal recognition systems, they found that recognition rates significantly improved. This work is prominent for introducing multimodal aspects that occur during communication and using this information for developing interfaces for human-computer interaction. Kessous et al. [92] conducted a study on multimodal automatic recognition of affect and emotion by using Bayesian Networks, a probabilistic method. Ten participants, from five different language backgrounds were involved in the recording. Both overall body and close-up face expressions were recorded when participants acted out eight different emotions. They were also encouraged to use gestures when expressing each emotion. Face, body and speech features were extracted and selected for the classification task. A Bayesian network was used as the classifier to train these multimodal features to automatically recognise the emotions. The classifier's performance is summarised in Table 2.6:

Modality	**Classification rate [%]**
Face	48.3
Body	67.1
Speech	57.1
Multimodal	78.3

Table 2.6: Classification rate per modality from Kessous' work [92]

Table 2.6 clearly shows the highest recognition rate of emotions is achieved through multimodal information. This finding from a machine-learning approach agrees with literature that multimodal information is helpful for better understanding of communicative contents and interlocutor's behaviour.

2.2.4 Multimodalities in Vlogs

Video blogs (vlogs) refer to a new media that exists for the purpose of personal online publishing [93]. Nardi et.al [25] describe vlogging as one activity of blogging that is user-generated and focuses on personal content or themes, discussing about current events, opinions on issues and about daily life in general. A prominent social platform for vlogging activity is inherent in YouTube, a website created for the purpose of video sharing and publishing. Trier [94] states that YouTube has established a reputation as the most popular video sharing website compared to other video sites. Vlogging requires minimal expert skills, which makes it easy for beginners to produce [95]. Deh [95] states that it is attractive and simple for anyone to produce, with only a little equipment and an interesting story to deliver to the audience. The production of vlogs is portable and efficient for speaker's expressions [96]. Vlogs have unique characteristics where expressions are portrayed through multimodal signals [97] [98] [99] [100]. Recent research explores this unique spontaneous genre of communicative content [101] [102] [103] [104].

Biel and Gatica-perez [99] conducted a study on personality impressions though audio-visual signals of vlogs. A collection of over 2000 vlogs from over 400 vloggers were used for this study. For annotation of vlogger personality impressions (based on the Big 5 Personality traits [105]) , researchers in this study used Mechanical Turk, a pieceworks service hosted by Amazon, to allow paid annotators to recognise these impressions. Audio features including pitch, speaking energy and voice rate were extracted using PRAAT [106]. Visual cues comprised aspects of looking time, length of looking segments, number of looking turns, proximity to camera and vertical framing. The visual cues were extracted using a normalised Weighted Motion Energy Images (wMEI) [107]. Multimodal cues were a combination of speech / non-speech and looking / not looking segments. Results from automatic recognition using 10-fold cross validation for training the Support Vector Machine (SVM) showed multimodal signals significantly improved prediction for Conscientiousness. The automatic recognition task however performed below baseline for traits of Agreeableness and Openness to Experience [99]. This current work on vlogs and its methodologies used for constructing an automatic recognition of personality traits is of particular interest for understanding the relationship between multimodal signals and this novel genre of social expression.

Morency et al. [17] conducted a study on multimodal sentiment analysis in conversational vlogs. Sentiment analysis refers to the study of automatic analyses of private states, such as opinions, emotions, and beliefs [17]. Sentiment analysis is an interesting approach to evaluate different attitudes of speakers, as attitudes themselves consist of affective states, opinions, beliefs and assumptions of people towards an object, as mentioned by Ajzen [40]. Morency et al. [17] selected 47 vlogs based on three criteria: diversity (age, gender, topics etc), multimodal (facial expressions, body postures, intonation, choice of words), ambient noise (presence of real-world noise). All videos were normalised to 30 seconds, the first introductory remarks (with animations of titles) were removed. With accordance to the experiment's objective, this is possibly the best way to normalise the length of data. However, introductory remarks may contain emotive or attitudinal components, such as friendliness ([47]) which should not be discarded. Annotation of the sentiments expressed by the vloggers was conducted by three annotators, labelling three sentiments: positive, neutral and negative. Percentile rankings on multimodal cues were analysed. These cues were polarized words, smile, look away, pauses and pitch [17]. Further analysis involved automatic classification using Hidden Markov Models (HMM)s. Textual features were extracted using polarity analysis, OKAO Vision software for visual features, and OpenEAR for speech features. A leave-one-out approach was used for training and testing and analysis found that the tri-modal classification, consisting of text, visual and audio cues performed best. This finding highlights the advantage of having multimodal cues as signals of speaker sentiments.

Rosas et al. [18] explored aspects of sentiment analysis, similar to Morency et. al [17], but this study used Spanish vlogs. A total of 105 videos were collected from YouTube consisting of different age, gender and topics (movie suggestions, political opinions, video games etc). Annotation of sentiments (positive, negative and neutral) was conducted by two annotators, and 47 videos were labelled positive, 54 videos were negative and 4 videos were labelled as neutral. Automatic sentiment classification was conducted using a Support Vector Machine (SVM) and a 10-fold cross validation was run on the dataset. Setting 51% as the baseline, the modality that performed best in the classification task was the combination of textual, speech and visual modalities, with an accuracy of 75%. This work is in agreement with previous literature [17] [99] that multimodalities present in vlogs are indicators of behavioural activity, such as sentiments and personality.

Sanchez-Cortes et al. [100] conducted a study on inferring mood through multimodal cues taken from vlogs. A total of 264 vlogs were annotated according to 11 moods through crowdsourcing and manual speech transcriptions. These moods included Happiness, Excitement, Sadness, Relax, Boredom, Disappointment, Surprise, Nervousness, Stress, Anger and Overall Mood. Audio cues were extracted using PRAAT to measure pitch, energy and speaking rate. Visual cues were extracted using weighted motion energy images (wMEI) measuring entropy, mean, median and vertical and horizontal center of mass. Multimodal cues were also extracted by measuring looking / non-looking and speech / non-speech segmentation. This analysis is similar to the study conducted by Biel and Gatica-Perez [99] where they also analysed multimodal cues using the same method. Classification was conducted using two classifiers, a Support Vector machine (SVM) and a Random Forest (RF). Results show recognition of moods was best obtained using multimodal information. For some moods, this phenomenon was not necessarily important as most moods could be detected through the audio information alone. This finding is related to the present study in that mood labels, as presented in vlogs, are similar to that of attitude labels in vlogs presented in this thesis. Similar machine learning methods for automatic classification are used in this study.

The above literature presents different methods of analysis and classification using vlogs. These studies use vlogs to study personality, sentiments and moods through machine-learning techniques. The present work also applies machine-learning methods for analysing vlogs.

2.3 Affective Recognition Systems

Recent empirical research in behavioral studies has explored the application of psychology to computational interfaces. Picard [108] pioneered affective computing as a field of study that refers to the work of understanding hidden affects (includes emotion, attitude and assumptions) of a person and translating them to computerised interfaces. A primary part of this recent field of study is enabling computers to recognise human affective states by using annotated real-world data to train on automated recognition interfaces [108]. Picard also states that elaborate research in this field is dedicated to studies of enabling computers not to merely recognise affective states of humans but to "have" emotions in order to make intelligent decisions [108]. This theory recognises the need for interfaces to be able to provide sufficient information for artificial agents to operate with decision making abilities [109].

Recent research in affective computing focuses on developing affective recognition and synthesis interfaces through verbal and non-verbal signals [110] [111] [109]. These multimodal signals applied to recognition of affect are applied in Kapoor and Picard's work [111]. Kapoor and Picard [111] postulated a framework for automatic recognition of affect based on multimodal signals for the purpose of embedding the component into a computerised learning companion. This work identified different sets of modalities, namely facial features, head gestures and pose. Eight children (monitored by their teachers) were asked to play a game called Fripples Place and their reactions during the task were recorded. Teachers were asked to indicate high, low and medium levels of interest and a fourth state 'taking a break' defined as fidgeting; 78% agreement was achieved for teacher perception of the affective states. A total of 50 samples contained all instances of multimodal signals. Classification was conducted using Gaussian Process (GP) and SVM classification methods. Results show that GP performed better than SVM for recognising affect through multimodal channels. Analysis of this work introduces a mixture of Gaussian Processes approach that contributes to the performance of the classification task in recognising affective attributes in learning environments [111].

Affective computing generally revolves around processes of data annotation, affect (includes emotion and attitude) classification, and feature extraction and selection through one or several modalities. Machine learning techniques are often applied to develop automatic affective interfaces. The following sections elaborate on processes involved in building such interfaces within the realm of affective computing.

2.3.1 Annotation

Development of recognition systems involves the process of data annotation. Annotation of data is conducted primarily for the purpose of data indexing and retrieval [108]. Although this process is the most elaborate and time-consuming, it is crucial for the classification and recognition task [112]. As mentioned by Pedregosa et al. [37], developing a recognition task using machine-learning techniques involves a series of processes and tasks, (see Figure 1.3 in Chapter 1). The process of obtaining labels for the machine to learn and train from requires real-world data to be transcribed and annotated. This annotation task can either be manually or automatically retrieved. Manual annotation by expert and non-expert annotators may be laborious but it ensures accuracy and controlled flexibility of annotation which is useful as preceding examples of fully automated annotation systems [17].

Schultze-Berndt [113] states that documentation does not merely involve recoding of raw data, but a certain amount of the raw data should be processed in order to meet the objective of a specific research project. Schultze-Berndt further postulates that derivation of annotation that represents the interpretation of raw data is a challenging task. Annotation from the raw data reduces information from the original data [113]. However, Schultze-Berndt also believes that annotation enriches the data because crucial information combines together different aspects of the raw data that could be possibly beneficial for the research goal. Schultze-Berndt claims that linguistic annotation produces or converts the data into machine-readable formats. This annotation may consist of multi-tier and inter-tier levels of annotation that produce a detailed and comprehensive annotation scheme. However, annotators need to determine the amount of time and detailed information that is necessary for the goal of the research. Schultze-Berndt advises that careful attention should be paid to the time allocated for this initial process of annotation because over time, other annotators can build up more annotations from the initial work, provided that the raw data and initial annotation is still available.

Riek et al. [112] claims that one of the primary issues in Affective Computing is data annotation, as processes of data collection, conversion and segmentation are time-consuming and costly. Recent work on annotation aims to overcome this issue by introducing crowdsourcing [114] (known also as Human Computation [112]). As stated by Riek et al., crowdsourcing refers to data labelling conducted by multiple non-experts. Several studies apply this method of annotation as it is time and cost efficient [103] [112]. Riek et al [112] present a study on generating tags for detecting social context through crowdsourcing. Thirty-three people (invited by word-of-mouth and Facebook) were involved with the experiment. Participants were asked to solve the game Guess What?, which Riek et al. developed for the purpose of crowdsourcing. Participants watched 39 videos containing different social scenes (birthdays, sporting events, concerts etc) taken from YouTube. They were required to indicate their answer based on a four fixed-choice question. Results from the game indicated a 70% inter-annotator agreement, which showed high agreement between all participants in indicating different social contexts.

With regards to multimodal annotation, Allwood et al. [115] created a standard coding scheme, MUMIN, to annotate multimodal video clips in Swedish, Finnish and Danish. The data involved short clips from movies and broadcast interviews. A comprehensive annotation process was conducted to determine annotation labels for each modality and between multiple modalities. For example, facial displays were given coding labels and tags, as indicated in Table 2.7[115]:

Type of Facial Display	Value	Tag
Eyebrows	Frowning Raising Other	Frown Raise Other
Eyes	Exaggerated Opening Closing-Both Closing-One Closing-Repeated Other	X-Open Close-BE Close-E Close-R Other
Gaze	Towards Interlocutor Up Down Sideways Other	Interlocutor Up Down Side Other

Table 2.7: Example of MUMIN coding labels

The example above illustrates the coding tags listed for the annotator's reference. Annotators consisted of participants involved in a workshop and were divided into groups of 2 or 3. Each annotator in the group was given the same video clip with the same coding tool (tools involved were ANVIL [116], MultiTool [117] and NITE [118]). Annotators were required to first practice annotation together as a group, and then individually annotate the videos. Finally they worked together and compared their annotations within the group. Inter-annotator agreement measurement was analysed if there were discrepancies on the annotations. The MUMIN coding scheme is a comprehensive scheme that takes into account multimodal features on communicative and social functions of interaction.

Henrichsen and Allwood [47] also conducted annotation of attitudes and developed the standard A10 attitude annotation scheme. One annotator conducted annotation of the NOMCO speech corpus while another annotator checked the annotation. Although given sufficient flexibility and freedom to annotate any attitude label of the audio-visual corpus, exclusions were made to attitude labels that were inconsistent and sparse. Results from the annotation process identified 10 attitude labels that best represented the data based on accumulated duration and number of instances. Through manual annotation of attitudes, it provided necessary information for the extraction of attitude tags for machine-learning.

Annotation is essential for generating data tags. Manual annotation requires knowledge from expert and non-expert annotators. To obtain reliability of the data labels, it is crucial to conduct validation analysis. There are several methods to achieve this, including perception tests or running inter-annotator agreement measurements. Perception tests typically involve non-expert participants identifying people's behavioural activity, particularly attitudinal states of speakers [50] [22] [48]. Reliability and validity of annotated data is also measured through a statistical method of inter-rater agreement [119]. Artstein and Poesio [119] describe two main statistical methods for measuring rater agreement. The first method of measurement is Cohen's Kappa [120], which measures agreement between two coders. However, annotation from two coders are reliable only for small-scale analysis [119]. Hence, when multiple raters are involved with the validation test, the method used for agreement measurement is Fleiss Kappa [121]. This measurement is advantageous as it allows for interpretation of agreement of judgments between multiple raters [119].

2.3.2 Machine Learning Techniques

Machine learning refers to an area of computational study that involves applications of algorithms to data classification and prediction [122]. Development of automated affective systems typically involves the application of machine-learning techniques. There are several machine-learning algorithms that are applied to automated data classification and learning tasks. Examples of these algorithms are Bayesian Networks [92], Hidden Markov Model (HMM) [17], Gaussian Mixture Model (GMM) [123], Support Vector (SVM) [124], and Random Forest (RF) [99] and others.

Szekely et al. [123] conducted a study on automatic voice style detection using a mixture of fuzzy SVMs and GMMs. This study aimed to automatically detect different voice styles of people based on the voice quality. A total of over 3000 utterances were annotated and segmented from four open source audio-books. These classifiers were applied to measure the confidence thresholds of each voice style by normalising separate hyperplanes using the sigmoid function (using FSVM) and confidence level in GMM was measured based on the normal approximation of the parameter's sampling distribution [125]. Results pointed to the notion that, with the application of these classification algorithms, different voice styles were successfully detected and distinguished from the entire corpus.

Bartlett et al. [126] developed an automatic face detection system using FAC through machine-learning techniques. They presented comparisons of different classification algorithms to produce the best detection rate of facial expressions. A dataset called DFAT-504 [83] was used for this study. This dataset consists of 23 different emotional expressions acted out by 100 university students. Six basic emotions were labelled and 313 sequences were derived. A classification task was then conducted to identify recognition rates of these basic emotions. Comparisons between performances of SVM, AdaBoost and (Linear Discriminant Analysis) LDA classifier were analysed. The comparison of classifier performance between SVM and Adaboost resulted in Adaboost showing faster recognition rates compared to linear SVM. When combining SVM (non-linear) with selected Gabor features of AdaBoost, results show that this combination of classifiers performed better than SVM and Adaboost alone. The performance of SVM was then compared to LDA, and results show that SVM performed significantly better than LDA. Bartlett et al. concluded that the combination of AdaBoost and SVMs displayed better performance in recognising six basic emotions [126].

Jokinen et al. [127] conducted a study to develop a recognition interface for detecting communicative functions of gesture. In order to find the best classification algorithm, experiments were conducted using several classification algorithms. Classifiers in WEKA, such as SVM, Decision Trees and Naive Bayes were tested and they found the SVM classifier performed best in the classification task. Lin et al. [124] highlight the advantage of using SVM compared to other classification tools, such as Naive Based and Neural Networks, in that SVM is primarily simple and provides high precision. LibSVM is a package for the SVM classification tool developed by Chang et al [128]. Chang et al [128] explain that the LibSVM package is typically used for two main purposes:

1. to develop a model based on training of the dataset

2. to apply the model to a testing data for information prediction

This extension of the SVM classifier suits the purpose of automatic recognition of communicative contents [129] because it provides faster and accurate prediction of the categories. With an optimised SVM algorithm, LibSVM allows for multi-class classification to be measured in a highly accurate manner.

2.4 Conceptual Underpinning

This thesis reports on the development of an automatic attitude recognition system based on multimodal cues expressed in vlogs. Previous literature explains attitudes as a source of information for better understanding of human expressivity. This expressivity is inherent in a new form of open access and multimodal media, commonly called vlogs. The present work acknowledges the importance of embedding attitudinal expressions into machine-readable interfaces for the purpose of information retrieval and classification. Hence, this thesis adopts conceptual definitions and methods from previous literature to address the research objectives, stated in Chapter 1. Figure 2.1 summarizes related literature that is relevant to the present study:

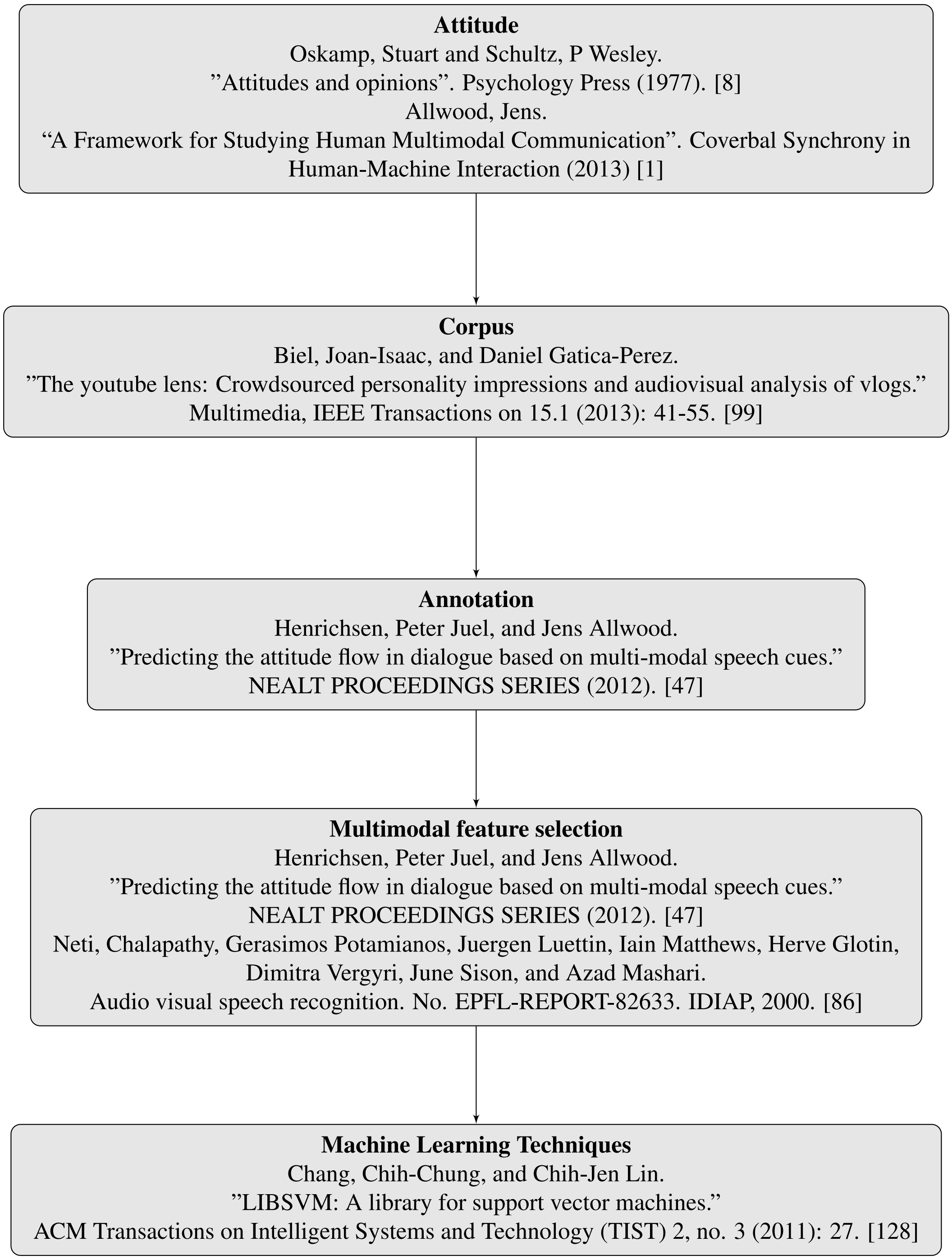


Figure 2.1: Conceptual Framework

Chapter 3

Vlog Annotation Scheme

3.1 Introduction

Utilising supervised machine-learning techniques for the purpose of developing an automatic recognition system requires data tags derived from specific labels. For this purpose, this chapter elaborates on attitude labels collected from a novel dataset. Expressions of attitude and affect are notable in different speech settings. This behavioural expression is manifested in both broadcasted and spontaneous corpora. In speech technology, numerous corpora are electronically stored and available to the public for research purposes. To name a few, the AMI [130] and D-ANS [131] corpora are some examples of public corpora. These corpora contain recordings of multimodal conversations between two or more interlocutors recorded in spontaneous settings. This research introduces a novel corpus containing speech from single speakers talking in their video blogs. Video blogs are a unique speech genre that contain single speakers talking to their "imaginary" audience in a semi-spontaneous speech setting. An elaboration of the characteristics of video blogs is described in the following section.

This chapter describes a novel corpus which illustrates multimodal expressions of speakers extracted from video blogs on YouTube. This chapter also describes criteria for speaker and video selection necessary for developing the video blog corpus. Additionally, this chapter explains techniques for annotating and segmenting attitude labels from the video blogs. The final section of this chapter discusses the validation process of attitude selection for this research.

3.2 The Vlog Corpus

3.2.1 Why Vlogs?

The purpose of this research is to develop an attitude recognition system through multi-modal features. Data selection is crucial for understanding attitudinal expressions through multimodal settings. This research specifically selects data from video blogs as they contain necessary requirements for multimodal feature selection. Video blogs contain both speech and visual expressions of a single speaker. This nature of video content is essential for the purpose of exploring multimodalities in speech.

Besides that, the focus of this research is on understanding speaker expression of attitudes and not the speaker's interaction with other interlocutors. Other readily available corpora typically contain two or more interlocutors interacting with each other. Research conducted using these corpora focuses less on speaker expression and more on the interaction between interlocutors. Selection of a dataset that focuses on a person's visible expression is crucial for identifying a person's attitude states. A single speaker's expressions are more visible in video blogs than in corpora that contain recorded interactions of multiple speakers. The AMI meeting corpus, [130] for instance, contains recordings involving multiple speakers and these videos do not fully capture the entire face of the speakers as cameras are set in positions where only half of the speaker's face is seen when speakers converse with other interlocutors. The position of the cameras (in corners or less visible parts of the room) might be strategically set so as not to cause intrusion to participants during recording for spontaneous talk. This type of corpus, however, is not suitable for capturing a speaker's expressions particularly when displaying attitudes.

Attitudes expressed by a single speaker bring forth valuable information for this research purpose compared to having multiple interlocutors. On that note, having multiple speakers during interaction carries as much valuable information in expression of attitudes, but the choice of focusing on single speaker vlogs in this study eases the technicalities of video preparation for analysis. This includes minimising noise (back-channels, overlapping talk etc) that could happen in interaction involving multiple speakers.

There are indeed other speech corpora that contain single speakers, for example news reports and political speech. This type of speech genre is often scripted and behavioural expressions are controlled to suit the genre of serious talk. Speech in video blogs on the other hand, is different from broadcasted speech. This type of speech is interesting because it involves speakers' attitudinal expressions in spontaneous settings, and is not fixed to serious and scripted speech. Yet due to its nature as recorded speech, the preparation of speech and controlled expressions are similar to broadcasted speech. The following section discusses in greater detail about the characteristics of video blogs for the development of an automatic attitude recognition system.

3.2.2 Vlog Characteristics

Video blogs contain unique characteristics which portray dynamic expression of speaker behaviour. The term video blog originated from the term "weblog". When someone creates a blog on the World Wide Web, they are able to write, save, upload, store and update entries. These entries are called "posts". These posts, typically arranged chronologically, are opinions and discussions about any topic decided by the writer (commonly called "blogger"). A video blog is an innovative blog genre that possesses similar characteristics to a conventional blog. Video blogs (in short, "vlogs") are defined as online personal diaries, or personal narratives pre-recorded in the form of a video [132]. People who create these vlogs are called video bloggers (known as "vlogger"). Typically, vlogs are videos recorded from home. The background settings of these videos are most commonly set in the vloggers' living room or bedroom. This is to create a sense of familiarity, proximity and some level of attachment between the speaker and the audience. Video contents involve topic-related thoughts or opinions and description of daily activities, happenings and events. These videos are shared in a large online social community, such as via YouTube. Vlogs are produced with minimal professional editing or production. They are easy to create and produce. Virtually anyone can produce their own vlogs on their own. A step-by-step guide to creating a vlog is illustrated in Appendix C.

Vlogs hold a specific set of characteristics. Some of the common characteristics [132] of vlogs are listed below:

1. single person
2. monologue
3. talking while facing the camera
4. video frame captures the head and shoulders
5. short length - typically from 2 to 5 minutes
6. stored in reverse chronological order
7. asynchronous communication - delayed feedback
8. semi-spontaneous speech

Figure 3.1 illustrates common characteristics one notices when watching a typical video blog:

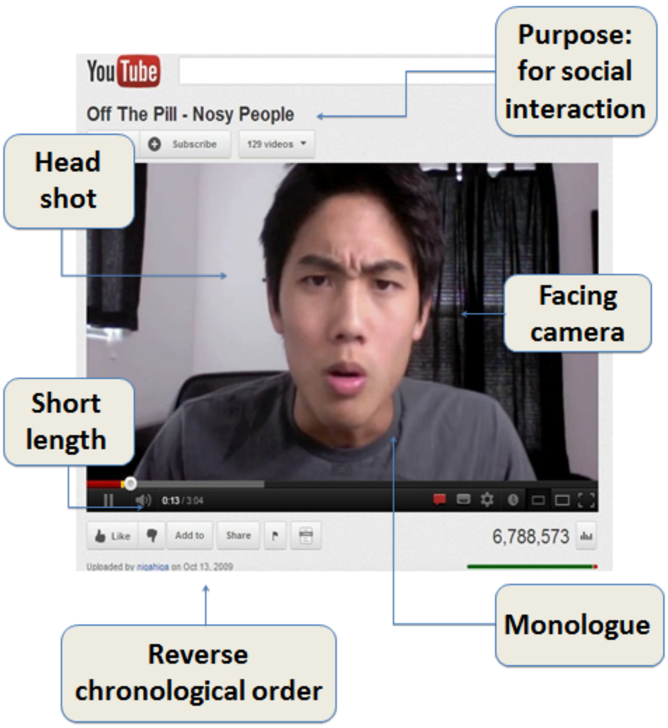


Figure 3.1: Characteristics of Video Blogs

As seen in Figure 3.1, a typical vlog displays a single person talking facing the camera. Eye contact is established with the camera to send a signal that the speaker is addressing and connecting with their "imaginary" audience. Generally, videos capture the head and shoulders of the speaker, however some videos may capture their entire body. Vlogs are different from other speech data mainly because of the length of the videos. Speakers keep their speech concise with a duration of 2 to 5 minutes. However, some videos can span up to more than 10 minutes, depending on the topic of presentation and the speaker's decision to include worthy clips that are necessary. The settings on YouTube channels allow for videos to be stored in reverse chronological order. This means that the newer videos are shown first whilst the older videos are shown in the latter parts of the channel.

Another interesting characteristic of vlogs is their asynchronous type of communication. Vlogs are similar to blogs where feedback from audience is not instantaneous. Vloggers on YouTube attain audience feedback through two methods. The first method of feedback is by dropping written comments in the comment feature made available below each video. The second method of getting feedback is by posting a video response on the audience's channel. Vloggers are made aware of the video response when the audience makes use of the tagging feature. When posting videos, YouTube users indicate keyword tags that may associate them with the vlogs that they intend to respond to. These two methods of feedback are considered delayed feedback as vloggers have the freedom to respond to comments or completely ignore them. Hence, this type of communication is somewhat one-sided. This is not considered a major problem because the main purpose of vlogging is for self-expression, and less likely for gaining immediate feedback from the audience.

A distinct characteristic of video blogs is the semi-spontaneity in speech. The narrative of a vlog is typically scripted beforehand [133] [134] but the format of delivery is in a spontaneous manner. However, some characteristics of this genre is similar to broadcasted speech as a significant amount of time is allocated for speech preparation. Time is dedicated to planning what to say and how to deliver speech effectively. Vloggers typically prepare scripts detailing introductory speech, main content and concluding remarks [133] [134]. This process is similar to broadcasted speech such as political speech and news reports.

As mentioned, characteristics of vlogs are a combination of aspects in broadcasted speech and spontaneous speech. Such spontaneity is apparent as there exists disfluencies in speech. Dufour et.al [135] define this as filled pauses, repetitions, repairs and false starts in speech. This element of prepared but spontaneous speech creates an interesting genre of new social media. The following is an example of transcribed video blog speech[136]:

```
"Alright y'all know me. I'm all about love and positivity
and spreading joy throughout the world."
"Ain't nobody wanna be your friend anyways"
```

The speech is in a highly informal register, using casual, colloquial and dialect expressions like yall, cause, and aint nobody, and taboo expletives like What the fuck are you talking about?.

Grammatical disfluencies including hesitations, repetitions and repairs are also present in video blog speech [137]:

```
"I had to..I felt like..I cou..I was staring at my brain."
"Don't you..don't you da..sit down"
```

It is apparent from the examples above that vlog speech contains aspects of colloquialism, expletives and disfluencies characteristics that make it similar to spontaneous speech. This genre can therefore be categorised as semi-spontaneous.

3.2.3 Ethical Considerations

Social media research undertakes long winding discussion about the ethical issues encompassing participation and content obtained from online public sites. Some may disagree with the free use of online participants and their contents without prior consent. The British Psychological Association, for instance, addresses the ethical framework where researchers are bound to obtain informed consent from participants before proceeding with data collection. If this is difficult to accomplish, researchers are advised to keep the anonymity of participants [138] [139].

However, multiple research appears to support the dynamic use of content that is publicly available to the general online community. Neuhaus and Webmoor [140] argue that informed consent is not feasible in circumstances where sources of data are posted on a public domain that is easily accessed by the online community. This proposition is supported by the British Association of Applied Linguistics (as cited in[141]) where informed consent is not necessary nor required for data that is publicly archived and published in open-access sites.

Hence, general consensus from the research community reaches common understanding that informed consent is needed from online participants if the data is retrieved in private sites, while data that is easily and publicly accessible does not require prior consent [141].

In such case of YouTube content, where videos are publicly accessible (unless account users change viewer settings to private posts), it is stated in the site's Terms and Condition that videos published on YouTube may be redistributed to the Internet, and may be viewed by the general online community. In the case of reproducing materials such as images and texts with considerable transformations, and acquire only short lengths from the original video's duration, these materials are allowed to be used and re-produced under The Fair Use policy. Fair use, as termed by YouTube, refers to permission for reusing copyright-protected material for non-profit conditions without prior consent from the copyright owner [142].

This study utilises and reproduces information available from public videos on YouTube as a part of data collection and analysis. Videos used in this research work are publicly accessible and are not used for the purpose of gaining financial profit. Therefore, as stated by the Fair Use policy, prior informed consent from each individual involved with the study is not required.

3.2.4 Speaker selection

For ease of analysis, videos are selected with certain characteristics to create a homogeneous set of data. Characteristics of speaker selection are limited to four provisional criteria (conservative dataset). These characteristics include:

1. Native speaker of American English
2. Male speaker
3. Aged between 18-28 years

The data was chosen to include several individual speakers, while controlling for variables associated with factors such as dialect, gender and culture. So in order to eliminate as much variation as possible in such variables - especially accent, voice pitch and quality, and speech style, including pragmatic phenomena such as politeness - the vloggers chosen for analysis were all male speakers of American English aged between 18 and 28.

Collection of data vlogs was conducted over a span of ten weeks. Speakers were selected according to the categories listed on YouTube. YouTube lists different categories for ease of searching and viewing of videos such as Sports, Music, Comedy, Education etc. The videos were searched under “Comedy” and “Most Popular” videos recommended by YouTube. When watching the vlog, other suggestions were given by YouTube on other related videos having similar comedic genre. Another method of widening the search for suitable speakers was by going to the channels that the first vlogger is subscribed to. These are mainly the vloggers’ acquaintances that share similar interest and have collaborations with each other. These channels are displayed in the vlogger’s subscription bar on his channel page. For example, Niga Higa’s channel is located based on YouTube’s recommendation at the “Most Popular” page. Then, in Niga Higa’s page, the side bar displays Kev Jumba’s channel, a channel that Niga Higa has subscribed to. Kev Jumba’s page is then explored to find suitable vlogs. The process continues until 10 suitable vloggers are identified for this research purpose. Figure 3.1 shows still video images of the vloggers under study:

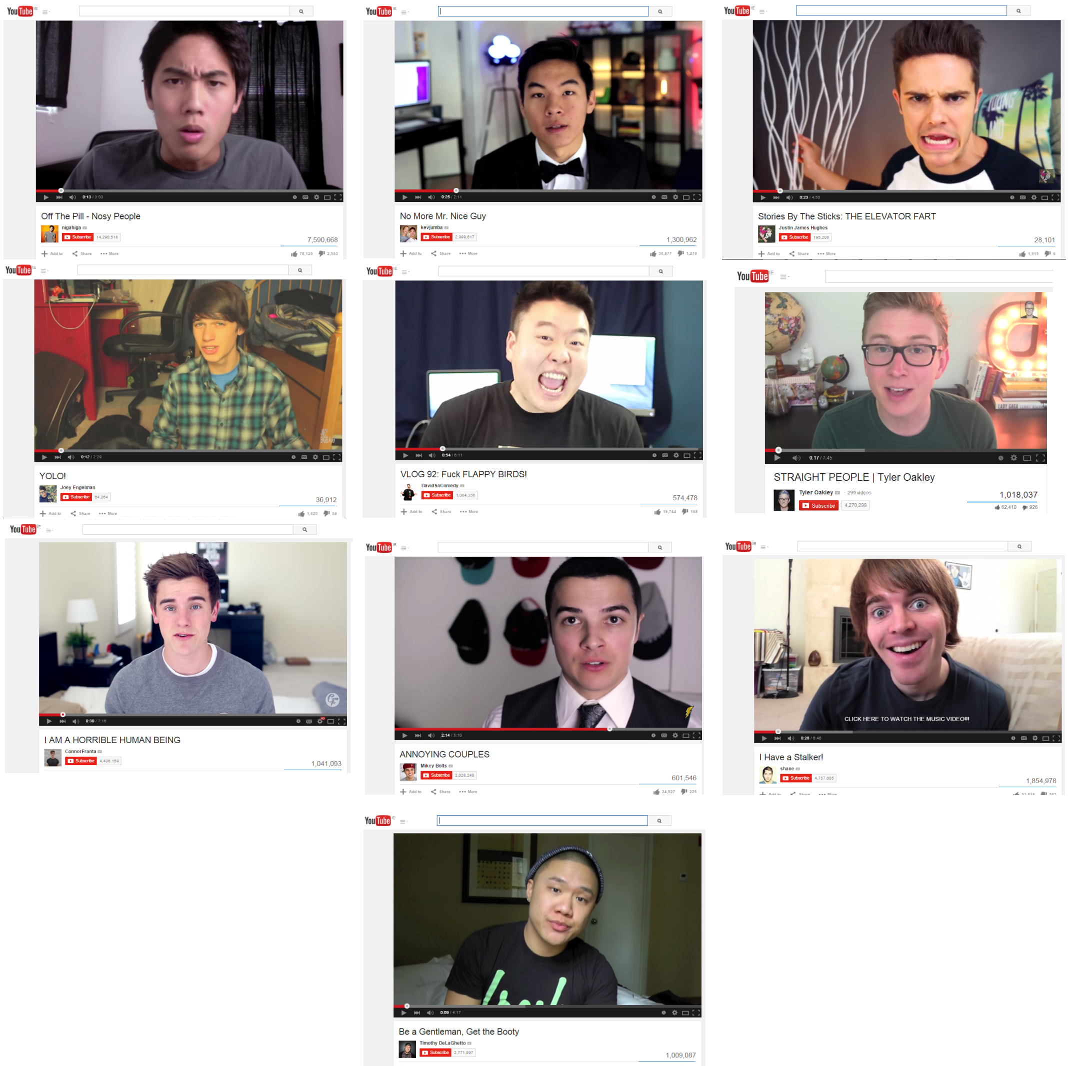


Figure 3.2: Related Video Bloggers

3.2.5 Video Selection

Video blogs from YouTube are carefully selected based on speaker criteria mentioned in the previous section. In order to ensure the quality and quantity of videos, the following list outlines criteria for video selection:

1. number of views
2. topic of talk
3. limited background noise
4. sufficient lighting / clear view of the face
5. position of camera

Vlogs taken from each speaker's channels are selected according to their large number of views. This is seen at the view counts on each video. View counts for all videos in this study range from 100,000 to 5,000,000 views. This factor is crucial as high numbers of viewership is an indicator of the social attention given to the these vloggers [102]. On average each speaker is represented by 25 videos. Table 3.1 presents information on a subset of the videos under study. The full list of video information is listed in Appendix A.

Vlogger	Channel URL	Video Title	Video Length	Index No.
Niga Higa	https://www.youtube.com/user/nigahiga	Off the Pill - 2009	03:00	V001
Niga Higa	https://www.youtube.com/user/nigahiga	Off the Pill - Arrogant People	02:06	V002
KevJumba	https://www.youtube.com/user/kevjumba	Asians Aren't Short!	03:16	V027
KevJumba	https://www.youtube.com/user/kevjumba	Asians Just Aren't Cool Enough?	02:52	V028
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	69 Things I Hate(d) About High School	03:28	V060
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	10 RANDOM FACTS about ME	04:19	V135
Joey Engelman	https://www.youtube.com/user/HashbrownLIVE	Anything but the Laughter!	01:46	V081
Joey Engelman	https://www.youtube.com/user/uncuthashbrown	BOOM, Things Happen	02:28	V082
Tyler Oakley	https://www.youtube.com/user/tyleroakley	Why 2011 Was Fricking Amazing	03:16	V104
Tyler Oakley	https://www.youtube.com/user/tyleroakley	I'm Gonna Kill Him	04:11	V105
David So	https://www.youtube.com/user/DavidSoComedy	VLOG 92: Fuck FLAPPY BIRDS!	6:11	V115
David So	https://www.youtube.com/user/DavidSoComedy	VLOG 91: Common Manners!	5:00	V116
Connor Franta	https://www.youtube.com/user/ConnorFranta	My First Time	5:09	V129
Connor Franta	https://www.youtube.com/user/ConnorFranta	Things Girls Should Know About Guys	3:56	V130
Shane Dawson	https://www.youtube.com/user/shane	A Message to Haters	9:16	V138
Shane Dawson	https://www.youtube.com/user/shane	ALL ABOUT THAT BASS!	5:51	V139
Mikey Bolts	https://www.youtube.com/user/michaelbalalis	Amanda Bynes: A Symbol of Hope	3:25	V142
Mikey Bolts	https://www.youtube.com/user/michaelbalalis	ANNOYING COUPLES	3:10	V143
Timothy DeLaGhetto	https://www.youtube.com/user/TimothyDeLaGhetto2	Annoying People I Hate	5:45	V144
Timothy DeLaGhetto	https://www.youtube.com/user/TimothyDeLaGhetto2	Annoying People I Hate #2	5:59	V145

Table 3.1: Subset list of Video Information

Another factor in deciding which videos should be selected is the topic of the vlog. Topics that are interesting such as rants about life events are selected because they convey several attitudinal responses of the speakers. For example, Niga Higa talks about arrogant people in his vlog [143]. He begins his vlog by greeting his audience. This expresses his "friendly" state. He then continues to rant about how he once met an arrogant person. He shows "impatience" with the person's arrogant demeanour. In the last part of his vlog, he advises people to shy away from being arrogant. This example shows that the choice of topic in the vlog is important to identify different attitude expressions.

To ensure the quality of each video, a number of additional factors are taken under consideration during the selection phase. Videos with high audio and visual quality are mostly preferred and used for this study to achieve an accurate feature analysis of speakers' prosodic and facial characteristics. In order to attain high quality videos, videos with loud background noise, such as music and environmental sounds, or with echoes are discarded. Apart from that, videos with insufficient lighting, whether recorded in darkly lit rooms or outside the home with poor lighting which may impede the visuals of the speaker's face are also deleted from the list. Videos are also disregarded when they do not capture clear visuals of the speaker's face due to awkwardly positioned cameras, whether too high or too low from the speaker's upper body and facial regions.

For convenience, Table 3.2 provides a summary of the video content for each selected speaker. As seen below, the list consists of the name of each speaker's YouTube channel, the channel's URL, number of videos selected for each speaker as well as the minimum and maximum length of videos per speaker.

Vlogger	**Channel URL**	**No. of Videos**	**Min. Length of Video**	**Max. Length of Video**
Niga Higa	https://www.youtube.com/user/nigahiga	34	02:06	07:48
KevJumba	https://www.youtube.com/user/kevjumba	32	01:52	05:01
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	43	01:42	09:49
Joey Engelman	https://www.youtube.com/user/HashbrownLIVE	21	01:00	04:26
Tyler Oakley	https://www.youtube.com/user/tyleroakley	20	01:28	09:59
David So	https://www.youtube.com/user/DavidSoComedy	22	02:49	06:20
Connor Franta	https://www.youtube.com/user/ConnorFranta	17	2:10	08:58
Shane Dawson	https://www.youtube.com/user/shane	29	2:45	11:46
Mikey Bolts	https://www.youtube.com/user/michaelbalalis	18	2:00	05:47
Timothy DeLaGhetto	https://www.youtube.com/user/TimothyDeLaGhetto2	14	4:08	09:14
	TOTAL	**250**		

Table 3.2: Summary of Videos per Speaker

3.2.6 Video Preparation

Following video selection, preparing videos in the required format is essential for subsequent processing of data. Videos are downloaded from YouTube using a free Add-On Extension tool [1] on Mozilla Firefox Web browser. Using the same downloader, audio and video files are downloaded in different formats; MP3 for audio files and MP4 for video files. Separate audio and video formats are necessary because the files are processed using different software to extract multimodal features for machine-learning. The processing of speech files is conducted using an online freeware conversion tool [2] as indicated in Figure 3.3:

[1] Easy YouTube Video Downloader Express 8.01
[2] http://audio.online-convert.com/convert-to-wav

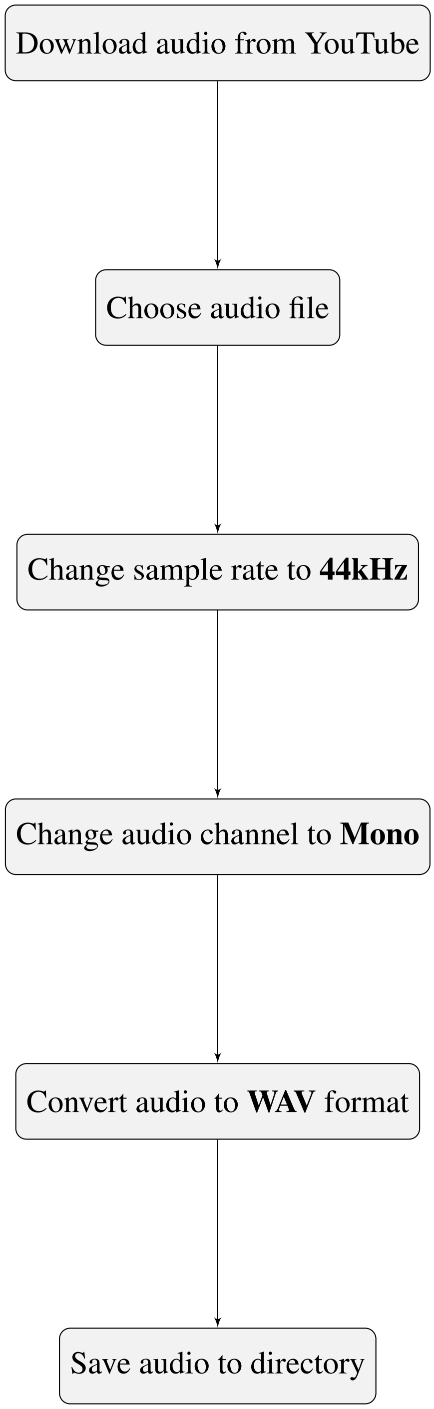


Figure 3.3: Details of Audio File Format

The purpose of having this process of conversion is necessary because the online YouTube downloader does not offer an option for downloading videos directly to WAV files. Audio files can only be downloaded into MP3 format. The software used for annotating attitude labels is WaveSurfer and this software can only read WAV files. Besides that, in order for the audio to work in the software, the preset setting only allows the audio channel to be changed from Stereo to Mono channel. This is probably set to achieve a standard sound signal. It also serves the purpose of speech reinforcement and intelligibility [144]. Sampling rate is also standardised to 44kHz as a popular standard for the file's audibility. Figure 3.4 outlines the processes involved in retrieving videos selected for this study:

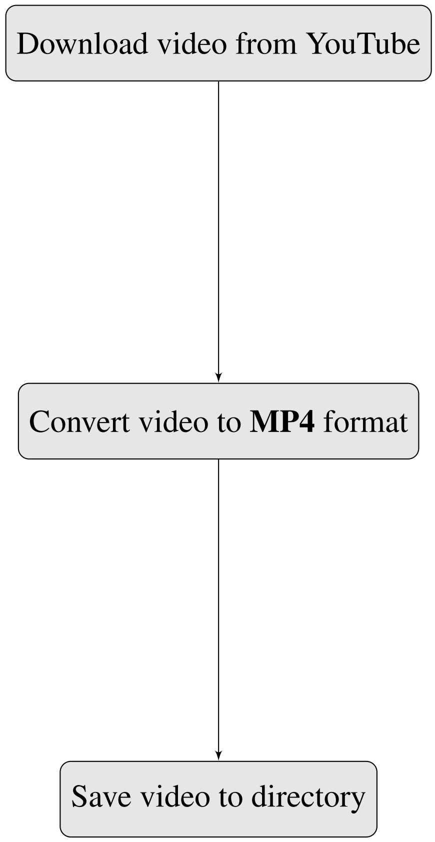


Figure 3.4: Details of Video File Format

Figure 3.4 illustrates the process of retrieving and preparing video files in the required format. This process utilises Mozilla Firefox as the main browser for video retrieval. When using this tool, the MP4 format is selected for the video files. It is crucial to select this format as the tool for visual feature extraction (which will be detailed in the following chapter) only supports MP4 or AVI videos. The extension for video download gives the option of downloading videos in MP4 format and not AVI. This leads to the decision to choose MP4 format for the videos. After the videos are converted, videos are indexed numerically and saved in the directory.

3.3 Annotation and Segmentation

3.3.1 Attitude Annotation Scheme

The next stage of corpus collection is the annotation of attitude in vloggers' speech. In order to identify suitable attitude states of the speakers, an annotation scheme is adopted for this process. Given the scarce literature on the modelling of attitudes in computer systems, the researcher selected Henrichsen and Allwood's [47] annotation scheme, as it clearly constitutes a simple attitude annotation scheme that has become standard in the field and can be used for computational modelling. Their annotation scheme, the A10, lists 10 attitude labels. The complete list is stated in Table 3.3:

No.	Attitudes
1	Amused
2	Bored
3	Casual
4	Confident
5	Enthusiastic
6	Friendly
7	Impatient
8	Interested
9	Thoughtful
10	Uninterested

Table 3.3: Standard A10-based Annotation Scheme

From the A10 list of attitudes, this study selected four attitudes and included one additional attitude that are most salient in these vloggers' expressions. The selected attitudes are Amusement, Enthusiasm, Friendliness and Impatience, while Frustration is included as an additional attitude. A detailed description of each attitude is listed in Table 3.4:

Attitude	Description
Amusement	speaker laughs, chuckles
Enthusiasm	speaker appears excited, laughs, high voice activity
Friendliness	speaker seems friendly, greets the audience politely
Frustration	speaker seems upset, sighs, low voice
Impatience	speaker appears disconcerted, shouts, loud voice

Table 3.4: Five Attitude Labels

The reason for selecting a subset of the attitude states is due to initial observation and careful analysis from annotators on attitude states of vloggers. The annotators, who themselves have experience in research using the vlogging genre, found that not all 10 attitudes indicated by Henrichsen and Allwood [47] are likely to occur in this vlog dataset, or that some attitude states are inconsistently and insufficiently represented in the vlogs. It was felt that ten attitudes were likely to be too many for volunteer raters to manage comfortably, and that the task would be more feasible with a reduced set. In order to reduce the number of states presented to raters, annotators scrutinised the videos and determined that six in particular were not much represented in the videos.

For the case of the attitudes "Confident" or "Interested", they can be difficult to observe as they may appear concurrently with other states, such as "Friendliness". Additionally, the annotators found infrequent occurrences of the attitude state "Bored" and "Uninterested" from the speakers in vlogs. This is possibly due to the fact that the intention of vloggers' speech is mainly to express attitudes that seem more "Engaged" with their audience. Due to this reason, attitudes that show low arousal such as "Bored" and "Uninterested" (indicative in the Valence-Arousal Circumplex Chart in [145]) are not frequently present or expressed in the vlog dataset. Given the characteristics of vlogs where they merely consist of one person present in the video, it is highly unlikely that a vlogger would express boredom or disinterest when speaking. In order to validate the choice of attitudes in the vlog dataset, a perception test is conducted to investigate reliability of the five attitudes chosen in the vlog dataset.

Details of the perception test is elaborated in Section 3.4.

After selecting attitudes that are most interestingly observed in the vlog corpus, as indicated in Table 3.4, the attitudes and their descriptors are used as guidelines for the next stage of annotation and segmentation. This procedure along with detailed description of the freeware tools used in this process is elaborated in the following section.

3.3.2 Annotation Procedure

Attitudes are annotated by indicating the start and end times in each video. A total of 250 videos are selected based on the criteria mentioned in Section 3.1. Annotation is conducted by one main annotator and checked by another expert annotator. The main annotator's task is to perceive and transcribe videos with suitable attitude labels. The expert annotator's task is to annotate a small number of videos. The overlap between these two annotators' transcriptions is measured using Cohen's Kappa [120]. An inter-annotator agreement measure is conducted to validate the annotation of all five attitude states. The result obtained a kappa value of 0.75 which is interpreted as "substantial agreement" between annotators [120]. This result of agreement is also the same in the video segmentation and labelling process. The annotation procedure involves annotators' pre-selection of attitudes by watching the entire videos on YouTube. The annotator then records the start and end times of each identified attitude. Following that, annotation was conducted using a Freeware speech analysis tool called Wavesurfer [146]. This tool allows for annotators to annotate relevant attitudes that they have observed in each video. A visual guide to the annotation process using Wavesurfer is illustrated in Appendix D whilst Figure 3.5 illustrates the annotation procedure:

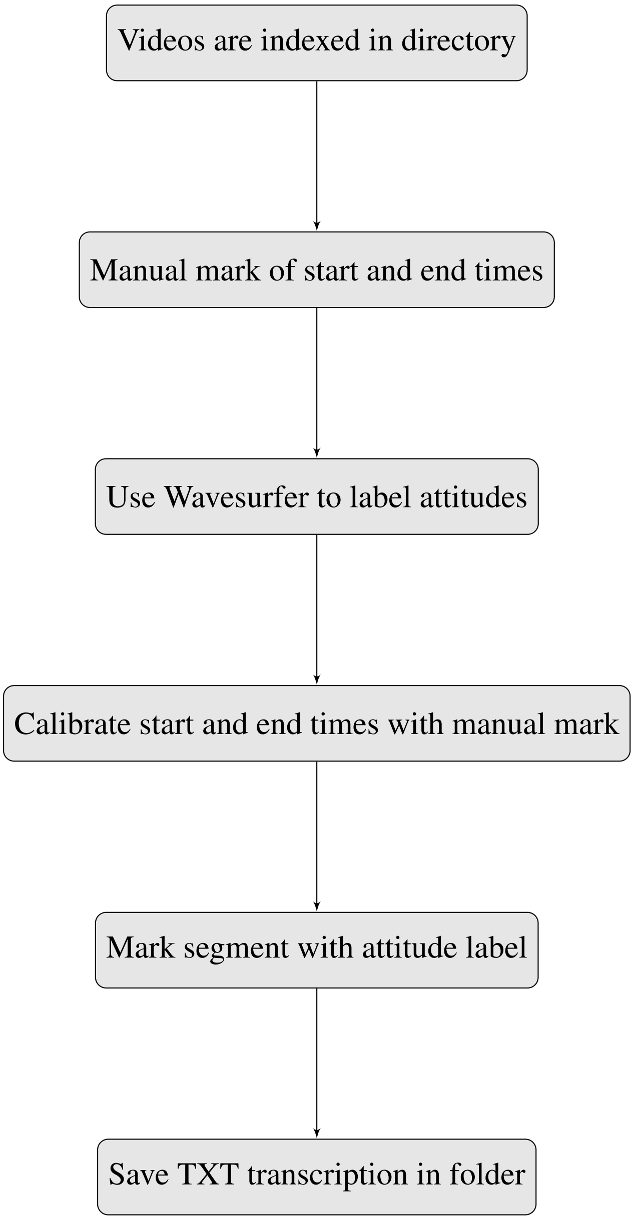


Figure 3.5: Process of Annotation

Figure 3.5 illustrates the stages of annotating the vlog dataset. As previously mentioned, annotation is conducted using Wavesurfer, a speech technology freeware tool for speech analysis. Annotation with just the speech modality is not the ideal method of annotating attitudes as this vlog corpus contains both visual and audio modalities and the visual component provides helpful information. Several software tools were tested, such as ELAN [147] to accommodate both visual and speech components during annotation. ELAN was only able to recognise PRAAT [106] audio scripts for audio annotation, hence it was a challenge to conduct annotation using this software. Using Wavesurfer speech analysis tool is preferred for this annotation task as it is easier to automatically extract annotated segments and subsequently extract speech features, while PRAAT requires customised scripts to complete this task. Manual alignment of audio and video segments, although elaborate and repetitive, is observed to work best to achieve precise attitude segments. Several attempts at automating the alignment of the audio and video segments using Wavesurfer scripts were done. However, the scripts returned errors, where misalignment between 2-3 seconds of the start time occurred. This led the researcher to conduct manual alignment of the audio and video segments. Further details of the audio and video segmentation process is reported in the following section.

3.3.3 Attitude Segmentation

Following the annotation process, annotated attitudes are transformed into attitude segments. The purpose of segmentation is to mark the annotated attitudes which merely involves small portions of the overall video. A total of 513 attitude segments are collected for this experiment. The total duration of the segmented files are 793 seconds with a mean duration of 1.55 seconds. The reason for using only a small part of the video is that the focus on attitudes is of prior importance. Most parts of these videos contain information that is unnecessary for the research. Informations such as musical opening sketch, speech with overlapping music background, and acted clips that are edited are excluded. Only segments that express attitudes of the speakers in these videos are used. In fact, most of the attitudes appear in a short length of time, only expressed in brief seconds. The total number of segments for each attitude category is listed in Table 3.5:

Attitude	No. of Segments
Amusement	100
Enthusiasm	107
Friendliness	101
Frustration	103
Impatience	102
TOTAL	**513**

Table 3.5: Segments by Attitude Category

For a deeper understanding of the dataset based on the segmentation process across speakers, Table 3.6 shows the total number of segments for each speaker and attitude state:

Attitude / Speaker	Amusement	Enthusiastic	Friendliness	Frustration	Impatience	Total per speaker
NihaHiga	25	11	28	8	26	98
KevJumba	6	4	17	3	8	38
Justin James Hughes	24	25	8	25	14	96
Joey Engelman	3	3	9	1	2	18
Tyler Oakley	10	11	8	5	2	36
Connor Franta	8	15	3	12	5	43
Mikey Bolts	3	2	11	5	3	24
Shane Dawson	21	22	12	27	9	91
David So	-	8	3	13	20	44
Timothy DeLaGhetto	-	6	2	4	13	25
TOTAL	**100**	**107**	**101**	**103**	**102**	**513**

Table 3.6: Segments across Speaker

With reference to Table 3.6, the total number of attitude segments varies across speaker. The highest number of segments is 98 as expressed by NigaHiga while the lowest number of segments totals to 18, showed by Joey Engelman. It is observed that these numbers have a considerably large difference. This imbalance of segments across speaker may be because some speakers may upload fewer videos that include attitude expressions. In addition, different speakers show differing amounts of video uploads, which are highly dependent on how active they are on their channels. This factor thus influences the varying number of relevant videos selected for this study.

Audio Segmenting

Stored TXT files containing start and end times as well as the attitude transcription are automatically extracted using a TCL/TK script. Running this script creates output sound files which involves only the annotated portion. Output files have short sound files with varying lengths depending on the marked start and end times of each attitude label. The average length of each sound clip is between 1 to 4 seconds. It is crucial to note that output folders named after the five attitude states are created prior to running the script. After running the script, sound files are automatically directed and stored in five output directories named with the five attitudes; Amusement, Enthusiastic, Friendliness, Frustration and Impatience.

Video Segmenting

With reference to the start and end times marked during the annotation stage, video files are segmented using a trimming tool in a video editor. The tool that is used for this purpose is Windows Live Movie Maker. Instructions for using the tool is described in Appendix E. Alignment of the start and end times is conducted based on the exact times marked in the audio segments. These times are found in the output folders assigned by the TCL/TK script used in the audio segmenting phase. Segmented video files are automatically saved in WMV format. The files are then converted to MP4 format in order to meet the requirements for the software used in the visual feature extraction stage. Format conversion is performed using an online freeware conversion tool [3].

After this stage, video and audio files are fully annotated and segmented according to the five attitude states. A complete list of all 513 video segments with the assigned attitude labels is shown in Appendix B. The following is a sample of the index of the attitude segments indicating the start and end times for each attitude label, as described in Table 3.7:

[3]http://video.online-convert.com/convert-to-mp4

No.	Speaker	Video URL	Video	Start Time	End Time	Attitude
1	Niga Higa	https://www.youtube.com/watch?v=7sz5cI51enE	V002	0:58	0:59	Amusement
2	KevJumba	https://www.youtube.com/watch?v=aIrbxxsLgTk	V027	1:14	1:15	Amusement
3	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V061	2:21	2:22	Amusement
4	Joey Engelman	https://www.youtube.com/watch?v=sU2e4Xeuqpw	V081	1:15	1:16	Amusement
5	Tyler Oakley	https://www.youtube.com/watch?v=v1nQuJUpkrU	V104	2:40	2:42	Amusement
6	Connor Franta	https://www.youtube.com/watch?v=DwdDJx9mhto	V132	1:35	1:36	Amusement
7	Mikey Bolts	https://www.youtube.com/watch?v=3L3qZLmYyc0	V136	2:02	2:04	Amusement
8	Shane Dawson	https://www.youtube.com/watch?v=3hOCo2Rg8JE	V138	2:10	2:13	Amusement
9	Connor Franta	https://www.youtube.com/watch?v=nJPNGSR7LDo	V244	1:15	1:18	Amusement
10	KevJumba	https://www.youtube.com/watch?v=9pG2HmQiB0U	V052	2:48	2:49	Amusement
11	Niga Higa	https://www.youtube.com/watch?v=gErOFu61v-A	V001	0:10	0:11	Enthusiasm
12	KevJumba	https://www.youtube.com/watch?v=clXERkuQmXM	V037	0:52	0:53	Enthusiasm
13	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V064	1:08	1:09	Enthusiasm
14	Joey Engelman	https://www.youtube.com/watch?v=_Bm9XzFrs0I	V099	0:54	0:55	Enthusiasm
15	Tyler Oakley	https://www.youtube.com/watch?v=v1nQuJUpkrU	V104	0:01	0:03	Enthusiasm
16	David So	https://www.youtube.com/watch?v=-D18CZYbrGc	V115	0:09	0:11	Enthusiasm
17	Connor Franta	https://www.youtube.com/watch?v=V02wBggJiCs	V130	2:00	2:01	Enthusiasm
18	Shane Dawson	https://www.youtube.com/watch?v=u1o5-H0bYbw	V139	5:10	5:12	Enthusiasm
19	Mikey Bolts	https://www.youtube.com/watch?v=xR2L4loWPa8	V142	0:07	0:08	Enthusiasm
20	Timothy DeLaGhetto	https://www.youtube.com/watch?v=pY7n5mpG5mU	V145	2:15	2:17	Enthusiasm
21	Niga Higa	https://www.youtube.com/watch?v=7sz5cI51enE	V002	0:00	0:01	Friendliness
22	KevJumba	https://www.youtube.com/watch?v=aIrbxxsLgTk	V027	0:12	0:13	Friendliness
23	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V080	2:47	2:48	Friendliness
24	Joey Engelman	https://www.youtube.com/watch?v=kM99wsh-4aM	V085	0:00	0:01	Friendliness
25	Shane Dawson	https://www.youtube.com/watch?v=F0lD9OHaWLk	V154	0:03	0:05	Friendliness
26	Tyler Oakley	https://www.youtube.com/watch?v=KbnZLFz6c4I	V158	0:00	0:01	Friendliness
27	Mikey Bolts	https://www.youtube.com/watch?v=AQ4Mk_8kNeQ	V160	3:16	3:18	Friendliness
28	Connor Franta	https://www.youtube.com/watch?v=gwsX9e8uYU4	V221	0:16	0:18	Friendliness
29	Timothy DeLaGhetto	https://www.youtube.com/watch?v=W8tvdiRPAA8	V242	0:00	0:03	Friendliness
30	David So	https://www.youtube.com/watch?v=XgVvvGCM37k	V247	0:12	0:14	Friendliness
31	Niga Higa	https://www.youtube.com/watch?v=NTnG4OFaFYc	V023	0:06	0:07	Frustration
32	KevJumba	https://www.youtube.com/watch?v=1qk23jdUT1g	V032	1:11	1:12	Frustration
33	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V080	0:34	0:37	Frustration
34	Joey Engelman	https://www.youtube.com/watch?v=HNnPnA-whEg	V100	0:41	0:42	Frustration
35	Tyler Oakley	https://www.youtube.com/user/tyleroakley	V111	1:18	1:19	Frustration
36	David So	https://www.youtube.com/watch?v=r3ufPb6AP6g	V117	0:05	0:06	Frustration
37	Connor Franta	https://www.youtube.com/watch?v=-rlwrZ_Zg4E	V133	1:36	1:38	Frustration
38	Shane Dawson	https://www.youtube.com/watch?v=3hOCo2Rg8JE	V138	8:17	8:18	Frustration
39	Timothy DeLaGhetto	https://www.youtube.com/watch?v=pY7n5mpG5mU	V145	4:47	4:49	Frustration
40	Mikey Bolts	https://www.youtube.com/watch?v=K2FdwLwl37o	V166	0:04	0:05	Frustration
41	Niga Higa	https://www.youtube.com/watch?v=K1aLtgEjzPk	V026	1:57	1:58	Impatience
42	KevJumba	https://www.youtube.com/watch?v=Hk-VrU0FoKw	V030	0:44	0:46	Impatience
43	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V075	4:24	4:26	Impatience
44	Joey Engelman	https://www.youtube.com/watch?v=sU2e4Xeuqpw	V081	0:38	0:39	Impatience
45	David So	https://www.youtube.com/watch?v=_f5XgiK1mNc	V124	0:36	0:37	Impatience
46	Mikey Bolts	https://www.youtube.com/watch?v=ltRl19Uw_fE	V143	0:24	0:26	Impatience
47	Timothy DeLaGhetto	https://www.youtube.com/watch?v=8FNb1abOhF0	V144	1:01	1:04	Impatience
48	Connor Franta	https://www.youtube.com/watch?v=ALspMqxZ9dM	V178	0:12	0:13	Impatience
49	Shane Dawson	https://www.youtube.com/watch?v=tasZxBMuKWM	V185	3:14	3:16	Impatience
50	Tyler Oakley	https://www.youtube.com/watch?v=v_4vcKn_ASU	V195	0:19	0:20	Impatience

Table 3.7: Subset List of Attitude Segments

Table 3.7 shows a subset of the full list of attitude segments that is ready for the next phase of experimentation. After data preparation is complete, the next stage of analysis involves feature extraction using prosodic and visual features. Further details of multimodal feature extraction and selection are elaborated in Chapter 4.

3.4 Validity of Attitudes

3.4.1 Motivation

As noted in 3.3.1, the set of attitudes was initially reduced from 10 to 5, including the new Frustration state. This subset (named N5) comprises five attitude categories, which are Amusement, Enthusiasm, Friendliness, Frustration and Impatience. After annotators observed the videos several times, they found that the Frustration attitude appears frequently in the videos. In fact, this attitude state also appears alongside other attitudes, such as Impatience, within the same video. That further motivated the researcher to include Frustration as an additional attitude as it is prevalent and meaningful in the dataset, although it is not included in the A10 annotation scheme.

Hence, selection of the attitudes contained in the N5 attitude annotation scheme is thought to be the most observable for the representation of attitudes in this vlog dataset. For the purpose of clarity, Figure 3.6 illustrates the attitude categories from the A10 standard Attitude Annotation Scheme and the subset attitudes adapted to develop the N5 attitude categories. In order to validate the choice of attitudes in the vlog dataset, a perception test is conducted to investigate reliability of the five attitudes chosen in the vlog dataset. Details of the perception test is elaborated in Section 3.4.

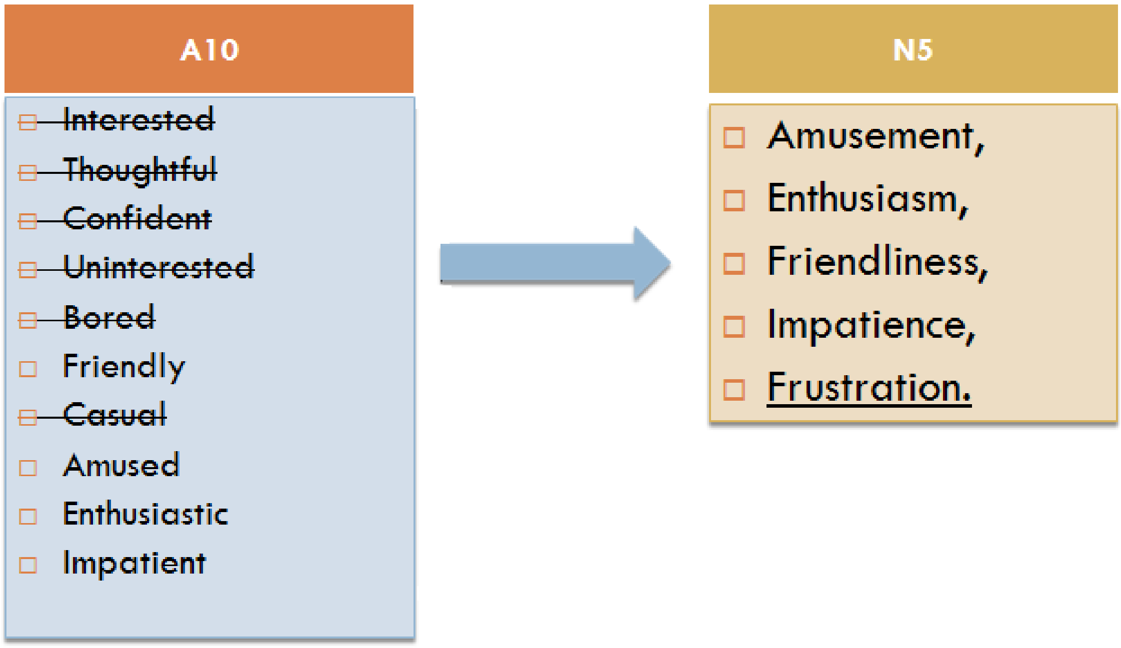


Figure 3.6: Description of the A10 and N5 attitude categories

As indicated in Figure 3.6, four attitude categories from the A10 attitude annotation scheme were selected and one additional attitude was included. As mentioned in Section 3.3.2, the selection was conducted by two annotators. Although the inter-annotator agreement between two expert annotators was 0.75 kappa, there is a need for the selection of attitudes to be verified by a larger public opinion. The objective of this small-scale study is to test the validity of the expert annotators' selection of attitude categories in the N5 attitude annotation scheme within a larger public perception. To achieve this validity of attitudes, a perception test is designed in the form of an online survey. Further details, experimental setup and procedures are elaborated in the following sections.

3.4.2 Experimental Setup

The perception test was developed to obtain perceptual understanding of attitudes from a set of anonymous participants. They consisted of 20 post-graduate students from the FASTNET and CNGL research groups from the School of Computer Science and Statistics, Trinity College Dublin, Republic of Ireland. Participants are not obliged to participate in the online survey since participation is completely voluntary. Participants are required to register and login to the online survey webpage. They are advised to use nicknames which do not identify them to their actual identity. This is necessary to keep the participants' anonymity and avoid biased experimental conclusions. Participant identities are kept strictly anonymous and no information of the participants is distributed to third parties. A confirmation letter of Ethical Approval from the Research Ethics Committee, School of Linguistics, Speech and Communication Sciences, Trinity College Dublin, Republic of Ireland is appended in Appendix F. Participant information is indicated in the registration page of the online survey, as illustrated in Figure 3.7:

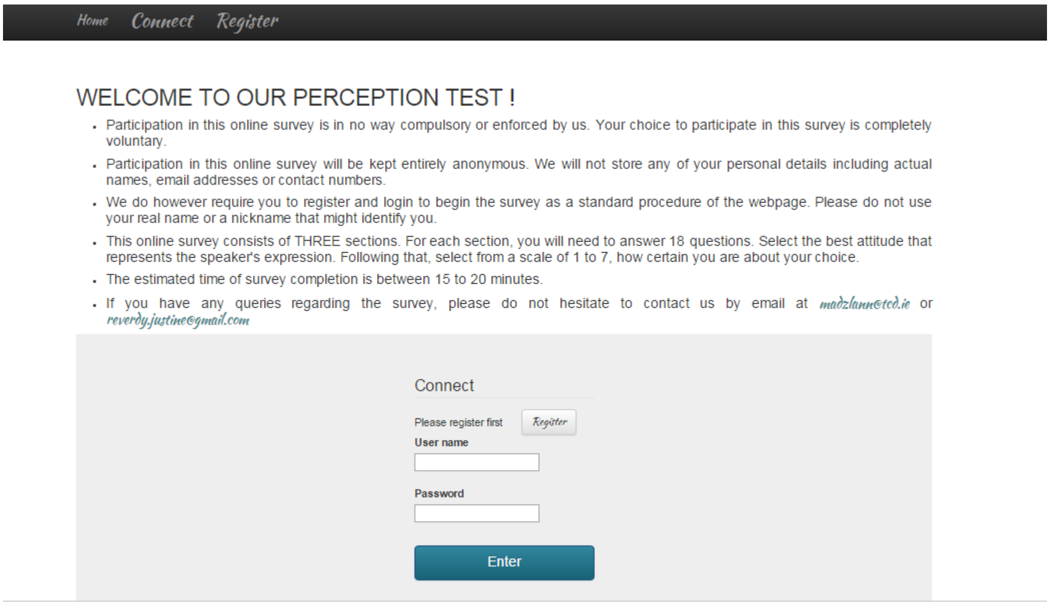


Figure 3.7: Participant Information

The link to the webpage of the online survey is circulated within the internal mail of the FASTNET and CNGL research groups [4]. The online survey is developed using PHP5 with an MVC architecture associated with a MySQL database. Participants are required to complete 3 sections containing several modalities, with 18 questions for each section. They are presented with the Audio only stimuli for Section A, Video only stimuli for Section B and Audio-Video Stimuli for Section C. Figure 3.8 shows examples of the three sections in the survey.

[4]http://tcd-fastnet.com/perceptionExpBroad/

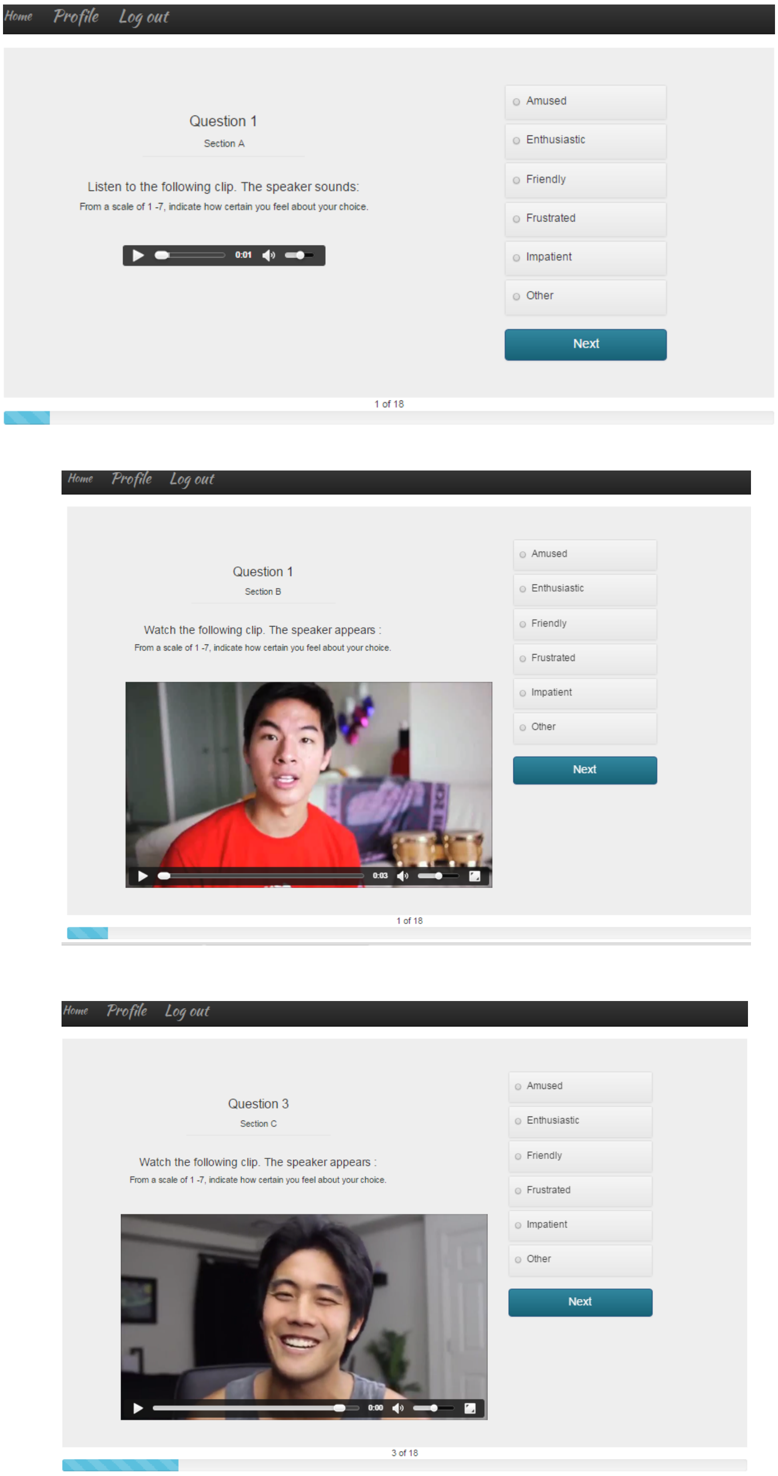


Figure 3.8: Sections in Online Survey

Participants are given 15 to 20 minutes to complete all sections. For each question, participants are given N5 attitude choices; Amusement, Enthusiasm, Friendliness, Frustration, Impatience and one additional "Other" drop-down menu showing the remaining 6 attitudes from the A10 attitude annotation scheme; Interested, Uninterested, Thoughtful, Casual, Bored and Confident. The reason for this decision to include the "Other" attitude choice is to provide more choices, rather than indicating definitive options to the participants. Figure 3.9 provides an illustration of the attitude choices with the addition of the "Other" attitudes.

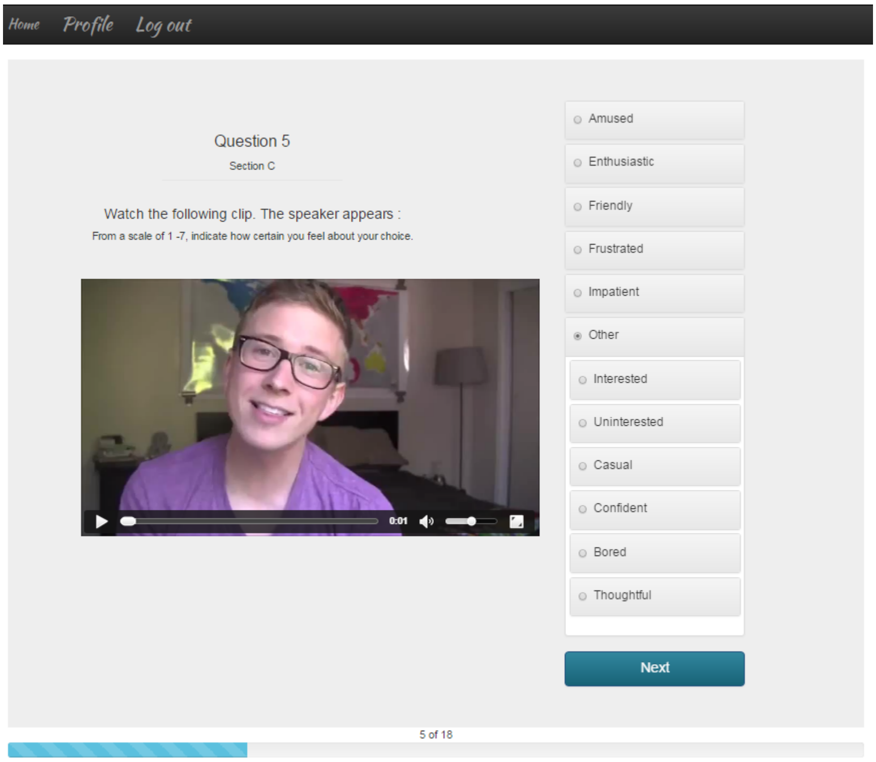


Figure 3.9: Example of Attitude Choices in Survey

Participants are required to select the attitude choice that they perceive as the best attitude expressed by the speakers. During the video selection phase, different speakers were selected for all sections so that participants are not presented with the same speaker expressing the same attitude. This is on the basis of minimising biased interpretations or providing clue indications for the participants. Additionally, the analysis included certainty level of participants in their attitude choices. Participants are prompted to express their certainty based on their choice of attitudes using a 7-point Likert Scale ranging from 1 to 7 (Unsure to Certain). To gain a better understanding of this procedure, Figure 3.10 illustrates the certainty scale shown from the survey:

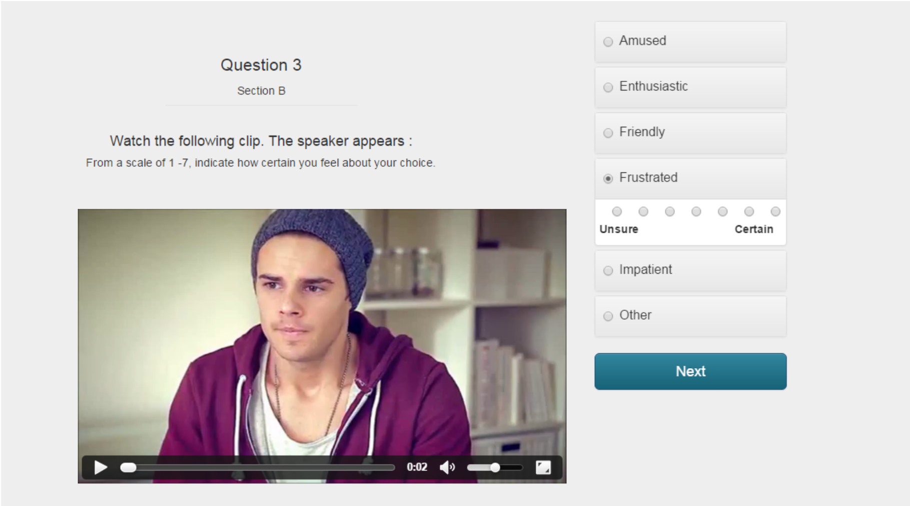


Figure 3.10: Example of the 7-point Likert Scale of Certainty

Figure 3.10 above illustrates the 7-point Likert Scale indicating how certain participants are with their choice. The far left of the scale indicates "Unsure" while the far right of the scale indicates "Certain". Participants are required to select the certainty level for each attitude choice or the system would not allow the participant to continue further with the survey. This process is then repeated until the participant completes all the sections in the survey.

3.4.3 Results

Inter-annotator Agreement

Inter-annotator agreement was conducted to investigate the level of agreement in attitude perception among the participants. Findings report 100% agreement among all 20 participants and the expert annotator for 37% of the stimuli [148]. This indicates that 37% of the attitudes were perceived by all the participants. Then, further analysis of agreement was conducted using a multiple-rater inter-annotator agreement measurement using the Fleiss Kappa measurement [121], which is typically used for measuring agreement between more than two raters. Results obtained a kappa score of 0.27 which is interpreted as a "fair agreement" between all 20 raters and the expert annotator. This low agreement is not unexpected when having many raters in the experiment as they hold mixed beliefs, perceptions and viewpoints of affective and attitudinal states. Several factors such as difference in age, gender and cultural backgrounds may contribute to such low agreement of affective states of speakers.

Further analysis investigates the performance of a subset of the 20 participants. Agreement was measured between three annotators and the expert annotator using Fleiss Kappa [121] and achieved a kappa score of 0.47, interpreted as a “moderate agreement”. Further investigation was conducted to measure agreement between one single participant with the highest agreement with the expert annotator. Analysis was conducted using weighted Cohen’s Kappa Inter-Annotator measurement and results found a kappa score of 0.73 which suggests “substantial agreement” between raters. This result is promising as one participant achieved high agreement with the perception of the expert annotator in identifying the N5 attitude labels. Volkova et.al [149] conducted a study on Emotional Perception of Fairy Tales and they performed an inter-annotator agreement (IAA) task involving 6 annotators. They achieved 0.34 on average kappa score with the task of perceiving fifteen emotions. For the word list task, the average kappa score was 0.45. Seeing that the work by Volkova et.al [149] is similar to attitude perception, we can deduce that the agreement rates between their work and with this current work achieved similar range of agreement, which is between fair to moderate agreement. Hence, it may be considered difficult to obtain a high level of agreement for human annotators in perceiving attitudes and emotions.

Attitude Selection

To validate the attitude categories in the N5 Attitude Annotation Scheme, analysis was conducted on the attitude selections of the participants in the online survey. Results on the frequency of occurrence of the attitude choices among all participants are indicated in Figure 3.11:

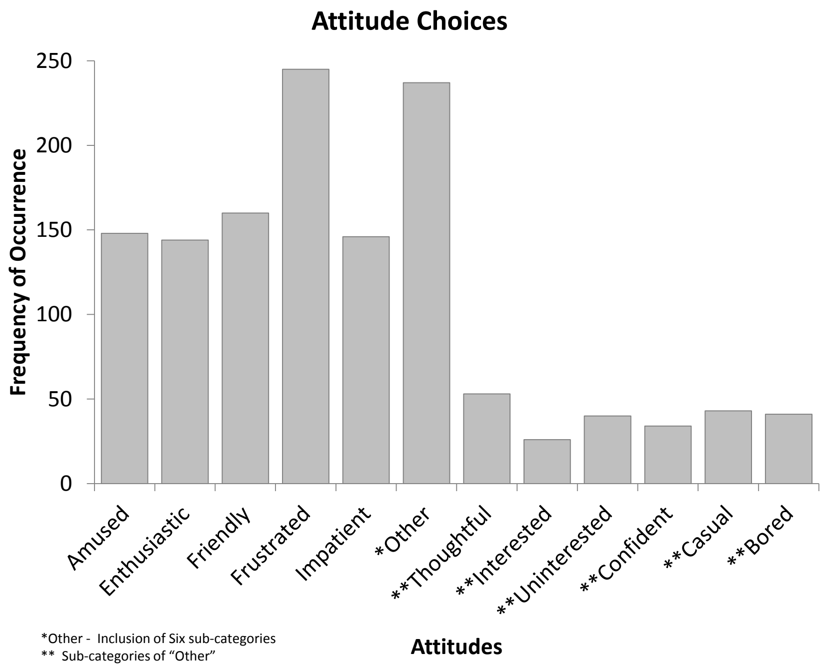


Figure 3.11: Frequency of Occurrence for Attitude Selection

From the Figure 3.11 above, it is evident that the attitude "Frustration" obtained highest attitude selection among the participants with a frequency percentage of 22.69% recognition. The lowest attitude selection among the participants is "Enthusiasm" with 13.3%. A complete list of percentages per frequency of occurrence is stated in Table 3.8:

Attitude	Percentage of Selection
Amusement	13.70%
Enthusiasm	13.33%
Friendliness	14.81%
Frustration	**22.69%**
Impatience	13.52%
Other	21.94%

Table 3.8: Percentages of Selection for Each Attitude

Table 3.8 highlights the percentage of attitude selection among the participants involved in the perception study. It is evident that the attitude "Frustration" is the most frequent attitude choice. This result is interesting as this attitude is not included in the original A10 attitude categories in the attitude annotation scheme, but it is included as an addition to the N5 subset attitude categories. This finding supports the hypothesis that the attitude categories, with the inclusion of "Frustration", in the N5 attitude annotation scheme are sufficient to annotate the vlog dataset.

To gain more understanding of the results obtained for the “Other” category, Figure 3.11 indicates participants’ selection of the “Other” attitudes. It is observed that the “Other” attitude that was most frequently selected is Thoughtful. This finding is interesting as participants might perceive this attitude as a state which occurs concurrently with other more notable attitudes. For example, the attitude Friendly might be perceived as Thoughtful too. This finding is not surprising given the different levels of attitude perception among participants.

Certainty Level

This perception study aims at understanding participants’ certainty level in selecting attitude choices presented through the stimuli. Investigating participants’ level of certainty is important to understand which attitude is expressed clearly and which of the attitudes that the participants have reservations about. Results of the certainty level of participants for all attitudes are indicated in Figure 3.12.

Based on results of the 7-point Likert Scale shown in Figure 3.12, it is found that “Friendliness” is the attitude that participants are most certain in selecting. With reference to the first graph on the second row in Figure 3.12, where no. 7 of the 7-point Likert scale showed highest participant selection. This result implies that participants are most certain when selecting the “Friendliness” attitude. This is not surprising as people could perceive the attitude friendliness almost instantly by merely watching the speaker smile and greeting the audience. This is evident in the N5 annotation scheme where “Friendliness” is described as the expression of the vlogger when greeting his viewers and smiles to the camera. Due to the rather obvious expression of vloggers’ friendly state, participants show high certainty when selecting this attitude.

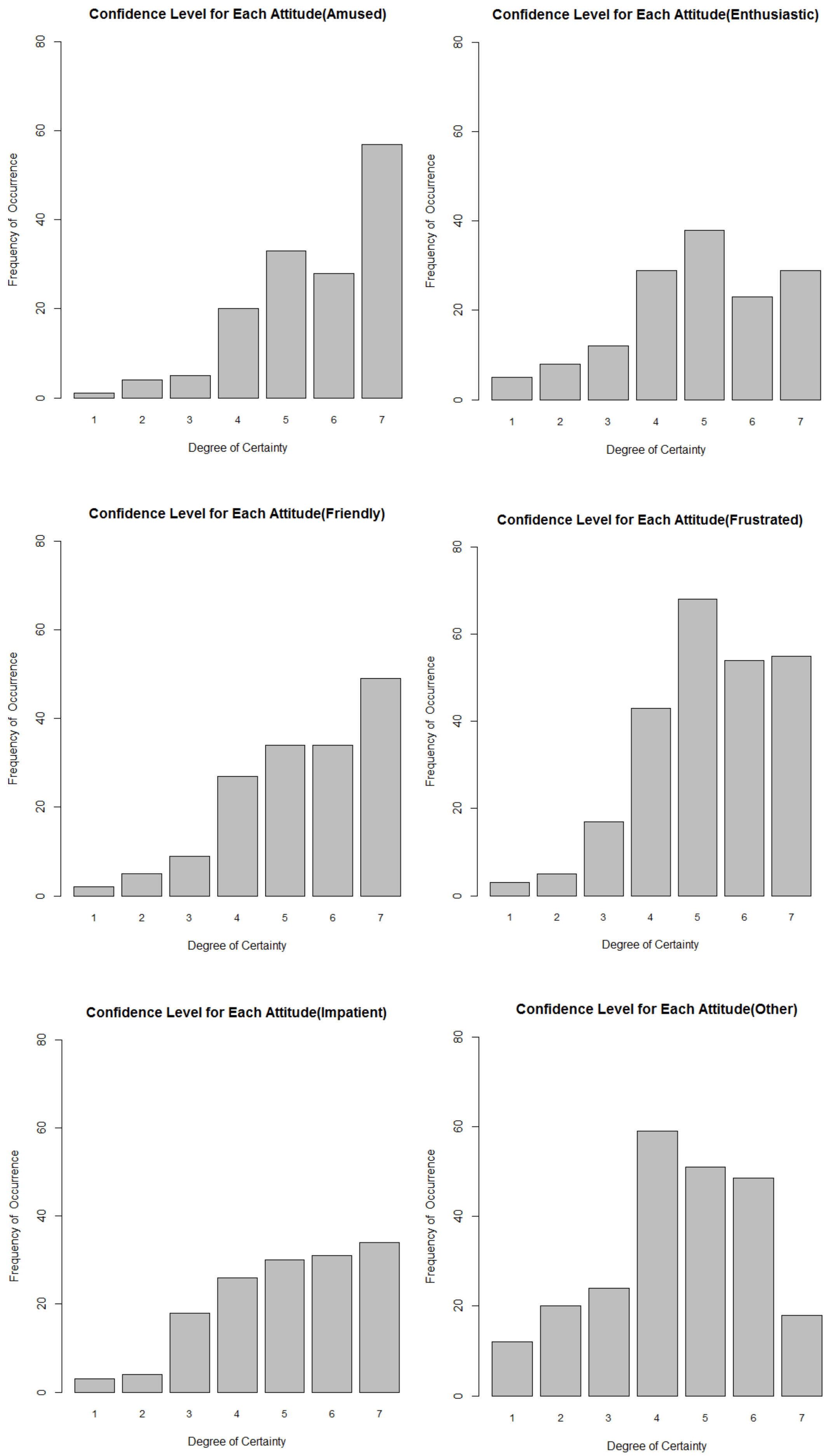


Figure 3.12: Barplots for Each Attitude’s Certainty Rate

Another observation from this measure of certainty is that the participants showed greatest uncertainty when selecting the “Other” attitude, as indicated in the last graph from Figure 3.12. It is evident that the scales indicating uncertainty (4) is highest for the “Other” attitude, while the scale that showed most certainty (7) is lowest for the “Other” attitudes in comparison to the other states. This can indicate that participants showed least confidence when selecting attitudes listed under the “Other” state. This is an interesting finding as the “Other” attitude category is not included in the N5 attitude annotation scheme. Instead the “Other” attitudes are attitudes from the A10 Attitude annotation scheme by [47] that are not included in the N5 scheme.

For a deeper analysis of the “Other” category, Figure 3.13 presents a breakdown of the certainty levels of participants when selecting the remaining 6 attitude states included in the “Other” menu:

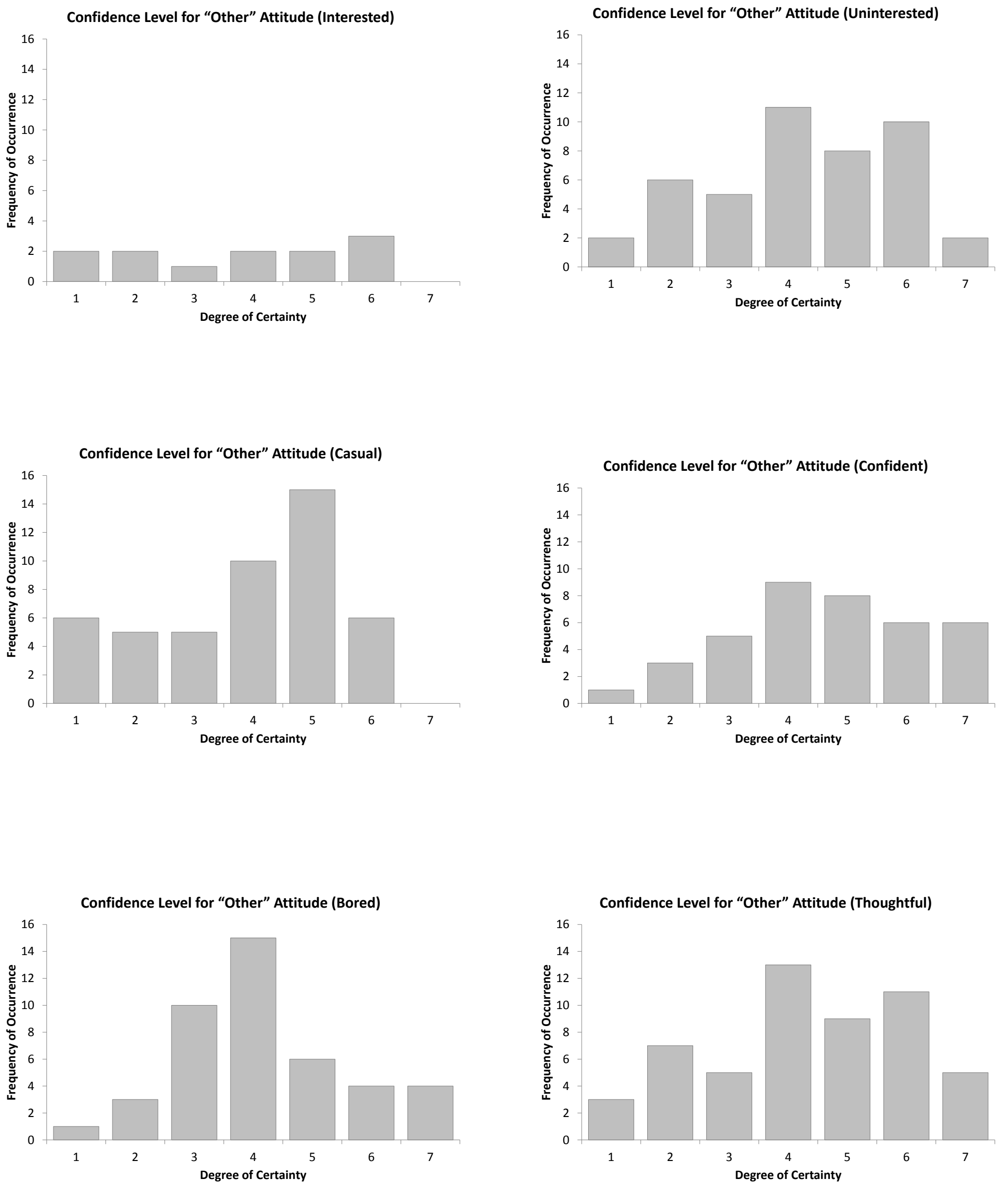


Figure 3.13: Barplots for Each “Other” Attitude’s Certainty Rate

It is observed, from Figure 3.13, that participants showed less confidence with their choice for states under the “Other” category. This is notable in their choice from the Likert Scale indicating uncertainty, which is between 1-4, with 4 being the most frequent rate selected (as shown in Figure 3.12). This low confidence among participants is further supported by a limited selection of the most certain scale (marked as 7). It is apparent that participants did not select certainty level 7 for Interested and Casual states. This may be an indication that they show reservation and uncertainty in identifying these states from the videos presented. Thus, the fact that the level of certainty of the participants in choosing the attitudes from the “Other” list is low, can be an indication that the attitudes from the“Other” category are not be so representative of the vlog genre.

3.4.4 Discussion

Based on the findings of the perception study of attitudes, there are several key observations worthy of further discussion. The main objective of this study is to test the validity of the attitude categories from the N5 attitude annotation scheme adapted from a standard A10 attitude annotation scheme [47]. Findings from the inter-annotator reliability measurement suggest that there is fair agreement between public perception and the expert annotators in perceiving vloggers’ attitude expressions. This is not a surprising result because different people have differing worldviews and perceptions, particularly in determining people’s affects, emotions and attitudes. Schuller [150] states that anomaly in obtaining reliability in the annotation of affect is expected because of the equivocal nature of affect data. Treating attitude data as a part of affect data, this present work, agrees with Schuller’s statement that total agreement in annotating attitudes is difficult to achieve.

Differences in age, for example may be a contributing factor to the low agreement between participants. Participants consist of post-graduate Computer Science students and they have a relatively mixed age group. People in their 40s may perceive Frustration, for instance, differently from people in their 20s. Recent work by Di Orgeta and Philips [52] studied age-related differences in recognising emotions. They found that positive and negative emotional facial expressions were perceived more intensely among the older adult group compared to the younger group. In relation to this study, this finding is interesting as age differences between participants may contribute to the difference in perceiving attitudes. The study on age differences as a factor of attitude perception is reserved for future work.

Gender variation is also a factor for differences in attitude perception. Studies suggest that there exists gender differences in understanding non-verbal cues [151]. Hall and Matsumoto [54] for example, conducted a study on gender differences in perceiving emotions, and they found that women showed highest recognition of different emotions. Attitudes, considered to display outward emotions, may be perceived differently according to different genders. With relation to this perception test, different genders may perceive mixed result of attitude perceptions. Hence, future analysis of gender role in perceiving attitudes will be conducted.

Besides that, cultural differences also contribute to mixed perceptions of attitudes of participants involved with the study. Unfortunately, information on participants' cultural and ethnic identities is not available, hence there is a constraint in investigating whether participants' cultural backgrounds have an influence to the results. It is however mentioned in previous studies that cultural variations may affect people's perception of attitudes and/or emotions. Elfenbein and Ambady [152] studied cultural specificity for the judgment of emotion expressions. They propose the need for gaining cultural knowledge prior to giving judgments of people's emotion expressions. An example as stated in their study, shows the emotion of "Contempt" obtained the lowest accuracy rate among subjects across cultures, hence concluding that emotion recognition is culture specific. Although their study highlights recognition of emotions, and not attitudes, observations drawn from their study about cultural differences can bring similar effects to the perception of attitudes.

Due to the anonymity of the participants, where information on age, race and gender are not available, the researcher was not able to conduct further analysis on these factors. These factors such as age, gender and cultural background were not requested so as to protect the anonymity of participants and also to avoid any form of bias in analysing the data. Indeed these are crucial factors for this study as they can provide more understanding of the difference levels of human perception with regards to identifying attitudes. However, if the experimental setup allowed for such information to be obtained, this could give further insights and discussion on the relation between these influences and attitude perception. Although this setup is considered adequate for this study, a possible method may be to setup the online survey to allow the retrieval of the participants' personal details in order to investigate such factors at a deeper level of analysis.

A relevant observation from results of the perception study is the attitude selection among the participants. It is found that "Frustration" is the attitude that is most frequently selected by participants. This is a good indicator of the attitude's relevance as an attitude that needed inclusion in the annotation scheme for annotating the vlog dataset. Note that "Frustration" was not included in the original attitude categories in the A10 attitude annotation scheme. This finding can be interpreted to support the inclusion of this attitude category in the N5 annotation scheme as a representative of attitudinal expression in vlogger speech.

In relation to attitude choices, the construct of this test is able to obtain certainty levels of participants in their attitude choice. Participants are required to rate their level of certainty about their choice of attitudes. Based on findings, "Friendliness" showed greatest certainty rate. This might be the case as friendliness may be indicated by vloggers quite clearly. Vloggers' greeting gesture when addressing the viewers may be seen as an indication of the speaker's friendly state. Hence, participants showed certainty that the speaker expresses friendliness in their speech. In contrast to that, the "Other" category showed highest reservation and doubt among the participants when choosing this category. This result is interesting as the attitudes in the "Other" category may not be as representative in the vlog data as compared to the five attitudes. This may indicate the possibility that attitudes in the Other category may not be sufficient to justify their inclusion in the N5 attitude annotation scheme.

It was decided to use a forced-choice format for the attitude-perception experiment, since this yields more usable data. Open-question responses require a further step of interpretation and post-hoc categorisation (e.g., where participants use different words for similar attitudes) and this introduces a step of subjective interpretation prior to quantification. It can be argued that the forced-choice format involves researcher interpretation at the experimental design stage, since the researcher must choose the categories to include. However, these categories are based on prior research in the field and are therefore not wholly subjective.

It could be argued that the pre-selection and placement of the remaining A10 attitudes in the “Other” sub-menu might lead to biased results as they are somewhat hidden from the participants. Hence, this could give the assumption that participants show lesser frequency to select these attitudes. However, this is not the case as presented in Table 3.8, the “Other” category was more frequently selected compared to the rest. In addition to that, findings from the certainty level of participants when selecting the attitudes showed that it is obvious that the confidence level for the “Other” category was low. This could imply that the attitudes from the “Other” category were not representative of the videos presented. Hence, this finding may, to some degree, support the pre-selection of the N5 attitudes.

Given the limitation of the experimental design, an alternative method to conducting this experiment is by possibly presenting participants with initial open-ended questions so they are allowed to provide a wider selection of attitude states. Then, answers may then be clustered into several categories or sub-categories to discover patterns of attitude states. Next, an error matrix or cluster analysis can then help to investigate consistent attitudes that appear as a result of the human annotation exercise.

3.5 Conclusion

This section outlines a summary of the third chapter of this dissertation. Content from this chapter includes the introduction of a novel corpus of vlogs. Definition, examples, characteristics and advantages of vlogs as a novel corpus for attitude recognition are elaborated in the first part of the chapter. Complete indexes of vlogs that are collected for this study are outlined in the dissertation’s appendix section. This chapter also includes explanation of the data collection process which gives emphasis on speaker and video selection.

This chapter also introduces the derivation of the N5 attitude annotation scheme, an annotation scheme that is developed from a standard A10 attitude annotation scheme by Henrichsen and Allwood [47]. Explanation of the attitude categories as well as their descriptors is elaborated. The annotation scheme is developed to derive a standard guideline for the annotation of attitude states most representative in the vlog genre. This chapter also describes the processes involved in the annotation stage, which consist of labelling and segmenting attitude states using free software tools. A step-by-step guide to using these softwares for annotation is attached in the appendix.

To validate the choice of attitude classes in the annotation scheme, an elaboration of a perception test is discussed in the final part of this chapter. A clear explanation of the perception test is stated in detail. This includes the motivation of the study, experimental setup, results and discussions. The study explains that the choice of the attitudes in the N5 annotation scheme may show sufficient representation of the vlog dataset.

Overall, this chapter details methods of acquiring data tags of attitudes through data collection, annotation and segmentation. These data tags are essential as they represent labels for training the automated attitude interface using supervised machine-learning techniques. The next chapter presents the next stage of developing this interface, which is extraction and selection of features using prosodic and visual signals.

Chapter 4

Multimodal Feature Contribution

4.1 Introduction

Past research including Social Signal Processing focuses attention to the different methodologies and approaches used for processing social signals. Non-verbal signals, in particular, provide useful information in describing attitudinal expressions of speakers during the communicative process. One of the contributions of this study is to describe the methods used for multimodal signal processing through the extraction and selection of prominent features. As mentioned in the previous chapter, after collection of data tags, the next process of devising a recognition interface is the processing of signals which involves feature extraction and selection. The following sections elaborate on the methods of prosodic and visual feature extraction.

This study also addresses the use of non-verbal features to identify different attitude states of vloggers in their speech. These non-verbal components involve prosodic and visual information of speakers when displaying different attitude states. Several studies analyse modelling of recognition systems through prosodic features. Additionally, utilising visual features in automatic interfaces are also conducted by past studies. This present work applies knowledge of the past literature and highlights difference in methodologies used to process non-verbal features for the classification system of attitude categories.

Apart from investigating methods of feature extraction, there is a need to conduct deeper feature analysis of this non-verbal information. Given the task of developing an attitude classification system where feature selection plays a crucial role, this chapter further describes individual features that contribute most to the classification task in distinguishing between different attitude states of vlog speech. This chapter takes an in depth look at the process of identifying multimodal feature vectors. This chapter further reports on analyses conducted to identify features that are most useful to the classification task.

4.2 Feature Extraction

In machine-learning techniques, the classifier is trained based on different types of features. Data processing is conducted to extract different features from the data segments. Note that this data processing phase involves processing of segments of the vlog data, not on the entire videos. When applying a supervised machine learning method for classification, a classifier is trained on different types of features. The data processing technique used in machine-learning is the identification of features or parameters that are considered most relevant for the classification task. This study highlights the use of multimodal signals, namely prosodic signals of the voice and facial features as parameters for the training of the machine to automatically classify attitude states. Sections 4.2.1 and 4.2.2 of this chapter discuss at length the procedures for multimodal feature extraction.

4.2.1 Prosodic Features

Human speech is a rich and dynamic source of information. Prosody, as a part of speech, is especially interesting as it gives communicative functions for intelligibility. Prosodic cues such as intonation, rhythm and stress can be correlated with expressions of affective information such as emotion, mood and attitude. They also serve as key indicators in facilitating semantic interpretation, meaning making and intelligibility during the communicative process. The function of intonation and stress is crucial for conveying one's attitude [63]. Roach [60] refers "attitudinal function" to mark the importance of intonation in interpreting the speaker's attitudinal states. Intonation, as defined by Crystal [59] refers to the interaction of prosodic characteristics. Prosody, according to Laver [62] and Roach [63], consists of four main components:

1. Pitch
2. Loudness (also known as Intensity)
3. Duration
4. Articulatory quality

The present work examines aspects of prosody that are meaningful for the classification task of the model. For this study, specific prosodic features are extracted and selected during the signal processing phase. The prosodic features are extracted using a TCL/TK script. By implementing this script, 14 prosodic features are automatically extracted from the audio files of each attitude category. The full list of the selected prosodic features is indicated in Table 4.1:

No.	**Prosodic feature**	**Unit**
1	Fundamental frequency (mean) [fmean]	Hertz [Hz]
2	Fundamental frequency (minimum) [fmin]	Hertz [Hz]
3	Fundamental frequency (maximum) [fmax]	Hertz [Hz]
4	Shape of the Pitch Contour [fpct]	Percentage Approximation [%]
5	Vibration of the Vocal Folds [fvcd]	Percentage of Voicing [%]
6	Intensity of the voice (mean) [pmean]	Decibel [dB]
7	Intensity of the voice (minimum) [pmin]	Decibel [dB]
8	Intensity of the voice (maximum) [pmax]	Decibel [dB]
9	Intensity movement (rising / falling) [ppct]	Percentage Approximation [%]
10	Voice Quality (harmonics1 - harmonics2) [h1-h2]	Decibel [dB]
11	Voice Quality (harmonics1 - formant3) [h1-a3]	Decibel [dB]
12	Voice Quality (harmonics1) [h1]	Decibel [dB]
13	Voice Quality (formant3) [a3]	Decibel [dB]
14	Length of utterance [sec] [dn]	Second [sec]

Table 4.1: Extracted Prosodic Features

Pitch, as measured through fundamental frequency (f0) is essential for understanding attitude expressions [47] [63] [153]. A person's attitudinal state can be detected by analysing the pitch of the voice. High pitch range is often associated with attitudes that are highly active. For example, the speaker adopts a high pitch range when expressing happiness or excitement. Contrary to that, speakers may adopt a low pitch range when bored or engaged in casual talk.

Measuring the mean, maximum and minimum values of fundamental frequency is essential to understand speaker's vocal pitch range. Not only that, inclusion of the pitch contour's shape measurement is also relevant as an acoustic parameter for communicative speech analysis as each speaker has varying degrees of pitch range. To find patterns of this variation, the shape of the pitch contour is measured according to the percentage of pitch movement. Pitch movement gives information on the falling and rising of pitch in the speaker's voice. Pitch contour is directly related to the tone of a person's voice and by using this measure, the relation between vocal activity and attitude expression can be better understood. As mentioned by Roach [60], speaker's slight changes to the tone; the rising and falling of the tones indicate differences in attitudinal meaning. Speaker's falling tone indicates certainty of the speaker's expression while rising tones may indicate uncertainty, impatience or friendliness [154]. In addition to that, vibration of the vocal folds is represented through the measurement of the percentage of voicing. Measuring the percentage of the vocal fold's vibration gives information on the extent of voiced characteristics within a segment. Due to this reason, measurements of mean, min, max of f0, pitch contour and voicing gives comprehensive information for pitch activity of the speaker's voice. Hence, measuring this acoustic parameter is relevant for the indication of vloggers' vocal activity in expressing attitudinal states.

Intensity (loudness) of the voice is related to how a speaker expresses his attitudinal state, whether in a loud or soft voice. Vocal intensity measurement is seen as a good feature to represent different attitude states of speakers. An increased value of intensity relatively means the person conveys high arousal of attitudes. For example, when a speaker expresses enthusiasm, intensity of the voice is measured with an increased decibel value while contrastively, intensity of the speaker's voice may be reduced when the speaker is expressing accounts of sadness or relaxed mode. Examples of past literature [64] [155] investigate the correlation between Intensity of the voice and speaker's emotional representation. Although emotions are different from attitudes, their relation to vocal intensity provides good knowledge of how vocal activity insinuates attitude expressions of speakers. Schroder et al. [155], for example, investigate emotional expressions and their correlate to vocal expressions. They found that increased intensity indicates the expression of active emotions and negative emotions show most prominent correlation with maximum intensity. This finding shows that intensity of a speaker's voice may indicate different levels of affective expressions.

Voice Quality is also a prosodic dimension worthy to be included in the parameters for attitude detection. Voice quality typically describes Breathiness, Creakiness, Harshness and Modular characteristics in a speaker's voice. Campbell and Mokhtari [66] found the significance of voice quality to be essential for understanding speaker style and speech act. Not only is voice quality useful for identifying speech acts, but is seen as a good acoustic parameter that is helpful as an identifier for attitude states. For instance, the breathy voice is likely to be associated with low arousal of emotion (also similarly referred to as attitudes by Roach [63]) while harsh and creaky voice signals a speaker's state of impatience or anger. Voice Quality can be represented through Harmonics in the speech waveforms and spectrum. Hanson and Chuang [65] conducted a study on glottal characteristics of male speakers and Figure 4.1 shows the acoustic measurements used in their study.

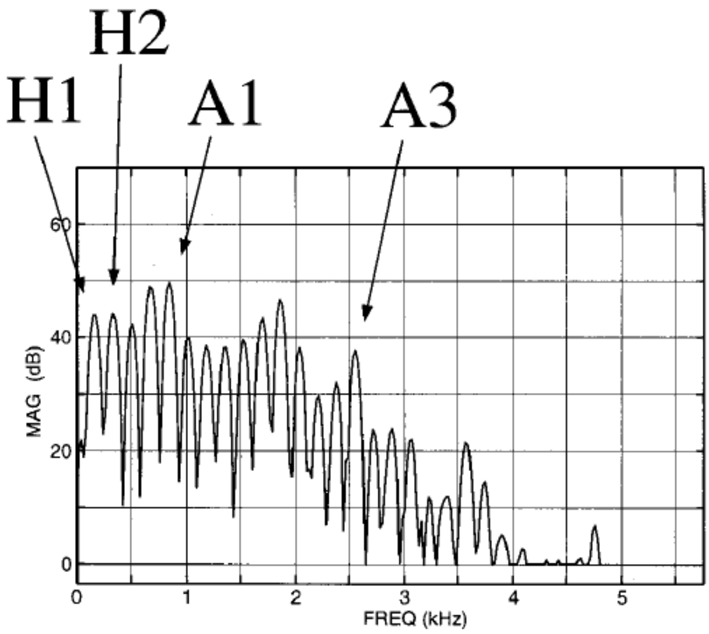


Figure 4.1: Acoustics measurements in the speech spectrum [65]

Figure 4.1 illustrates acoustic measurements for voice quality, where H1 refers to the first harmonic in the spectrum, H2 refers to the second harmonic, A1 refers to the first formant and A3 refers to the third formant. As described by Hillenbrand et al. [156], acoustic measurement of Breathiness is typically calculated through the decibel (dB) amplitude of the first harmonic (H1) relative to the second harmonic (H2), thus indicated as H1-H2. Hanson and

Chuang [65] refers to the source of spectral tilt as the amplitude of the first harmonic relative to that of the third-formant spectral peak, indicated as H1-A3. Shue et al. [157] conducted analysis on the dependency of voice quality with acoustic measurements and they found that H1-H2 and H1-A3 measures of the spectral tilt were largest for Breathy phonation and contrastively, showed smallest measurement for Pressed phonation. Seeing that the use of these measurements is identifiable to voice quality characteristic of the speakers' voice, the present work applies this approach (listed in Table 4.1) to measure voice quality for vloggers' different expressions of attitudes.

Duration or length of an utterance is also an essential acoustic parameter for communicative analysis. The duration of the utterance may signify differing attitudinal interpretations [47]. Speaker's long utterance may indicate more dynamic attitudinal expressions compared to a shorter utterance. Indication of friendliness may take a shorter amount of time to express while the expression of frustration may exhibit a longer duration.

Roach [63] and Crystal [59] promote the idea of integrating several acoustic elements such as pitch range, tone, loudness and rhythmicality and tempo since they are essential for understanding attitude expressions. As suggested by Crystal and Roach, there is a need to have a combination of varying speech elements to provide sufficient information for attitudinal meaning of speakers. To sum up, the use of the 14 prosodic parameters is likely to provide sufficient information for extraction of prosodic features for the training of a classifier that recognises attitude expressions of vloggers in the vlog speech genre.

4.2.2 Visual Features

Apart from prosodic features, extraction of visual information is considered an additional feature component for the purpose of training in the classification task of the interface. As suggested by Luettin and Thacker [74], visual information facilitates prosodic information for speech production. This section details methods and tools used for extracting visual information from vloggers' facial region.

Performing visual feature extraction of a single speaker, as is the case of vloggers in the vlog dataset, requires robust software tools. There are numerous interfaces for facial tracking. Some of the notable models used for facial tracking are Facial Action Coding system

(FAC) [81] and Active Shape Model (ASM) [78]. FAC is widely used for facial tracking as it provides a coding scheme of facial behaviour. This work is fascinating as it involves directly with affective behavioural expressions through facial activity. The downside of this model is that facial tracking is not fully automated (but recent work is dedicated to making this model fully automated [82]), making this quite an elaborate and time-consuming process. ASM is also one notable model for facial tracking which highlights statistical modelling of the shape of the facial landmarks. ASM is beneficial to automatically track localised regions of the face to give the regions a shape. This provides useful information for characterising facial behaviour of the object [85].

One of the models for computer vision and facial tracking is the Active Appearance Model (AAM) [84]. AAM, which is in fact an extended model of the ASM, offers a robust quick matching, not only by using the shape model from ASM but includes holistic information on the facial appearance feature of the object to an image. The present work adopts an algorithm of the AAM model which allows detection of 67 facial landmarks. The software tool used is Luxand FaceSDK [1], an open-source software developed as a cross platform face detection and recognition library. The sample application within the library that is used in this study is the LiveFacialFeatures application. This application provides tracking of the facial features in real-time using a web camera. The recent version supports recognition from 70 facial coordinates or landmarks, but this study used an older version which enables tracking on just 67 points of the facial region. The FaceSDK Face Detection application is initialised by firstly calling initialization from the data file path that is stored in Microsoft Windows. After clicking the application, the tool prompts to select a video stored in the specified path. Then, it proceeds to process the video. It then detects the face in it and returns the coordinates of the facial feature points. This application detects live human faces with a fast speed of slightly less than 1 second by which each frame is detected independently [2]. In cases where the tool failed to detect the face (depending on several factors such as the video resolution quality etc), it would return an empty output file. This suggests that the application failed and needs to be run again. This tool is used in several studies within the realm of visual recognition and face detection [158] [159]. Figure 4.2 illustrates the 67

[1] Luxand Face SDK, http://www.luxand.com

[2] https://www.luxand.com/facesdk/documentation/specifications.php

facial landmarks detected using FaceSDK :

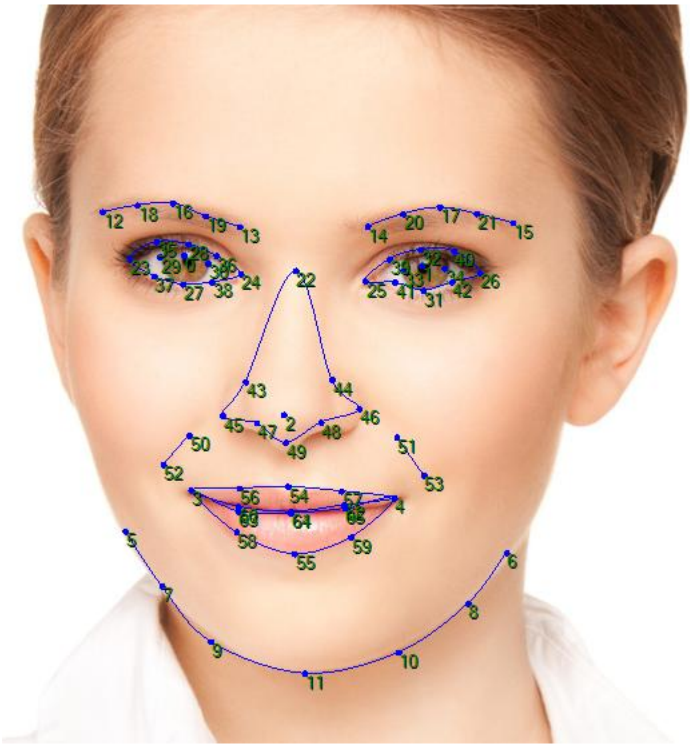


Figure 4.2: Facial Landmarks tracked by FaceSDK

Figure 4.2 illustrates the 67 visual dots on an interlocutor's facial region. The facial landmarks consist of key features such as the eyes, eyebrows, nose, lips, chin and nasolabial folds. For the purpose of data processing, all of the data points of the face are assigned with specific numbers ranging from V1 to V67 (all 67 data points). A complete list of all 67 facial landmarks is indicated in Appendix G. Some examples of the facial features are labelled in Table 4.2:

Visual Feature	**Code**	**Label**
V1	H	Head
V2	EL	Eye Left
V3	ER	Eye Right
V4	NT	Nose Tip
V5	MCL	Mouth Corner Left
V6	MCR	Mouth Corner Right
V7	FCUL	Face Contour Upper Left
V8	FCUR	Face Contour Upper Right
V9	FCL	Face Contour Left
V10	FCR	Face Contour Right
V11	CL	Chin Left
V12	CR	Chin Right
V13	CB	Chin Bottom
V14	EBCOL	Eye Brow Corner Outer Left
V15	EBCIL	Eye Brow Corner Inner Left
V16	EBCIR	Eye Brow Corner Inner Right
V17	EBCOR	Eye Brow Corner Outer Right
V18	EBML	Eye Brow Middle Left
V19	EBMR	Eye Brow Middle Right
V20	EBMLL	Eye Brow Middle Left Left

Table 4.2: Examples of Facial Landmark Labels

When running this software, each data point is assigned numbers, for instance, the Head position is indicated as “V1” in the output. This allows subsequent data processing to be conducted with clear indications of the data points. The procedure for visual feature extraction is indicated in Appendix H. A brief example of the processes involved in facial feature extraction with the FaceSDK software is indicated in Figure 4.3:

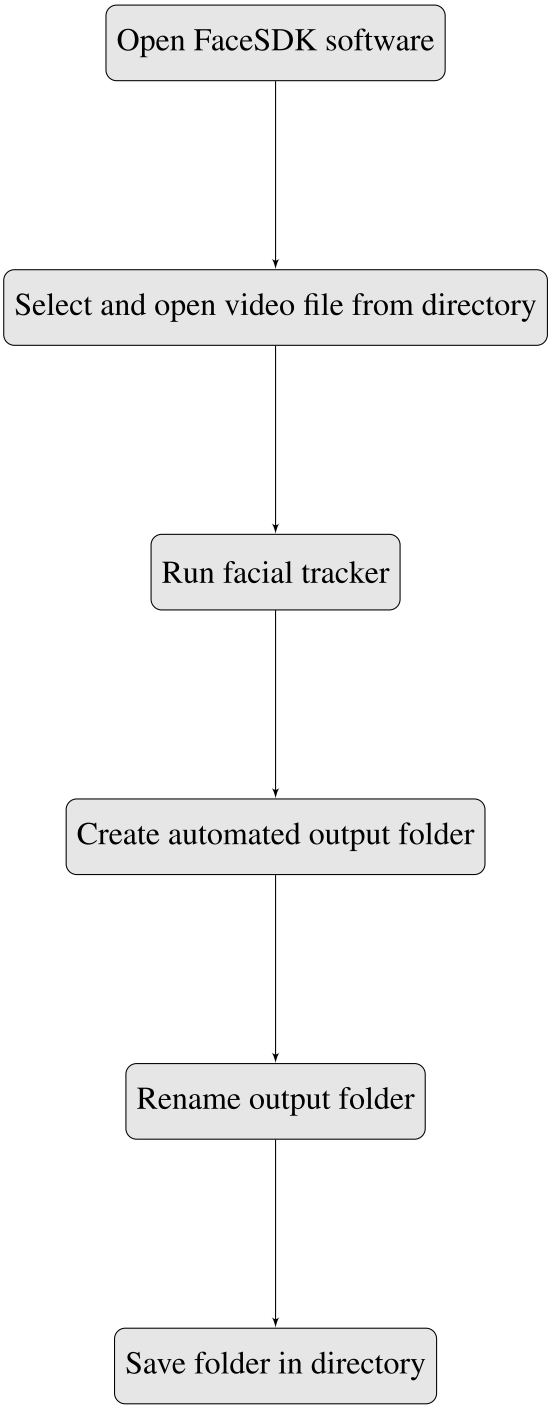


Figure 4.3: Process of Facial Feature Extraction

When running this software, video files of the speaker are input to the software. The software then automatically detects the facial movements and provides an output indicating the raw values of each facial landmark's movement. These values are then processed using several machine-learning techniques which will later be described in Chapter 5.

4.3 Feature Selection

After prosodic and visual features are extracted, the next stage of feature processing is the selection of relevant features. Feature selection is an important process in modeling recognition systems using machine learning techniques. This process ensures accuracy of the predictive model [160]. What this means is that the classifier would be able to perform better when the training is done using filtered feature vectors. For example, you have 100 prosodic features to train your predictor. The predictor would not be able to learn from the given features because of many features that are redundant or similar to each other. To solve this problem, we should conduct feature selection prior to training the classifier. By reducing the features and subsequently creating a subset of highly co-varied set of features, there is a higher chance for the classifier to produce a better accuracy rate in detecting different attitude states. This feature selection is best represented in Figure 4.4 below:

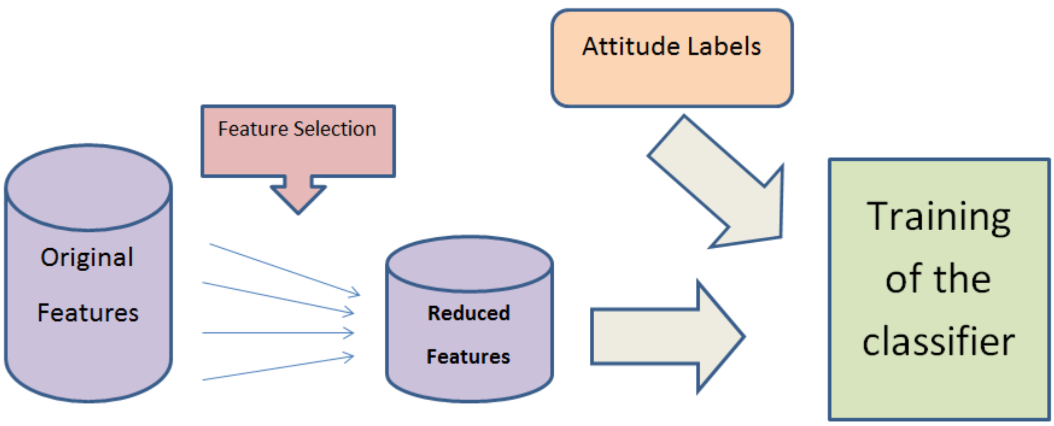


Figure 4.4: Feature Selection in Supervised Machine Learning

4.3.1 Statistical Methods for Feature Selection

As illustrated in Figure 4.4, the process of deriving“reduced features” is a crucial stage prior to training the classifier so that features that are most useful to the classification task could be identified. To derive these reduced features, feature reduction methods are typically used. The present study utilises statistical measurements to better understand the influence between features to the classification task. Statistical methods are measured using R [161], a software tool and programming language used for data analysis. R is a powerful tool that serves a number of functions; primarily for statistical analysis, data visualisation and predictive modeling. R is an open-source software that makes it easily downloaded and used, compared to other licensed softwares. The following statistical methods are computed in R to better understand relationships between features:

Pearson's Correlation Coefficient Correlations between features can be measured with Pearson's r [162]. The advantage of using this measurement is its ability to identify how strong the relationships are between variables. Measurement of the strength of the relationships is calculated based on the coefficient value. If the values are negative (for instance, -1.00), then the variables are negatively correlated. If the values are positive (maximum range of 1.00), then the variables are positively correlated. This measurement is suited for measuring the relationship between two variables. This study exploits this method by measuring correlates of prosodic features to obtain a subset of reduced prosodic features. The use of this measurement is described in the training process of classification in Chapter 5.

Welch Two Sample T-test T-tests are used to measure equal means between two populations. The Welch Two Sample T-Test [163], an extension of the Student T-Test [164], is used to compute reliability between two samples that have equal and unequal variances and sample size [165]. This measure gives better results compared to the standard Student's T-test because it addresses unequal variance between samples, whereas the Student's T-test provides low results for samples of unequal values. To interpret the significance level between two variables, we will need to look at the p-value. The typical cut-off p-value is less than 0.05 ($p<0.05$) but if more evidence is required to measure significance between variables, one could set a stricter cut-off p-value to smaller values like 0.005 ($p<0.005$) [166]. For this study, the Welch Two Sample T-Test is used to find co-variance between pairs of features.

If they are significantly different, then the p-value would be less than 0.005, whereas if they are significantly similar, the p-value would be more than the cut-off value. The use of this measurement will be explained later in the following section.

Analysis of Variance (ANOVA) ANOVA [167] is typically used in statistics to examine the amount of variability between multiple groups of samples. People would choose t-test measurements to compare between two samples, but if there are more than two samples, then this should be evaluated using ANOVA. The measurement of ANOVA aims to produce works of comparing variance not only between groups, but also within groups to decide whether the groups belong to one population or do they possess different characteristics. The present study adopts this measurement to evaluate different groups of features and statistically investigate differences and similarities between features.

Precision and Recall Precision is the total number of selected items that are relevant, while recall is the total number of relevant items that are selected. In simpler terms, high percentage of precision means that more relevant results are obtained compared to irrelevant results, whereas high recall value retains most of the relevant results. When combining these two measures, this means that the algorithm returns real positive cases (precision) that are correctly predicted positive (recall) [168]. In the classification task, precision and recall is typically used to measure correct association to data labels. This measure is not commonly used for feature selection, but is in fact used to measure the performance of the classifier. For this work, it is best to include the evaluation of the results in this section for describing prominent features observed from the classifier's performance. The standard measurement for values of precision and recall is the F-measure, which serves to better measure the skewed datasets. The f-score is obtained by calculating the average value of precision and recall. The application of precision and recall as well as the f-score in this study will be described in the following section.

Principal Component Analysis Another notable method of data dimension reduction is through the use of Principle Component Analysis (PCA). PCA refers to a multivariate technique that analyses a data table, and the outcome is described by several dependent variables [169]. The main purpose of using PCA is to extract important information through data dimension reduction. It then transforms this reduced dimension into new independent variables called principal components (PC) [169]. PCA is also used to prevent overfitting, which occurs when the sample data dimension is higher than the number of samples. By reducing the dimensionality of the dataset though PCA, selected dimensions that carry most information are preserved. This work finds this statistical approach relevant in finding relationships and variations between features, as will be better explained later.

4.4 Prominent Prosodic Features

After explaining general descriptions of statistical methods for feature selection, this section elaborates on specific features that contain most useful information, beginning with prosodic features. To better understand prosodic feature selection, boxplots, seen in Figure 4.5 provide a visual overview of the distributions of three primary modes of prosodic characteristics; Pitch, Intensity and Voice Quality. Details of analyses conducted using several statistical methods are described as follows.

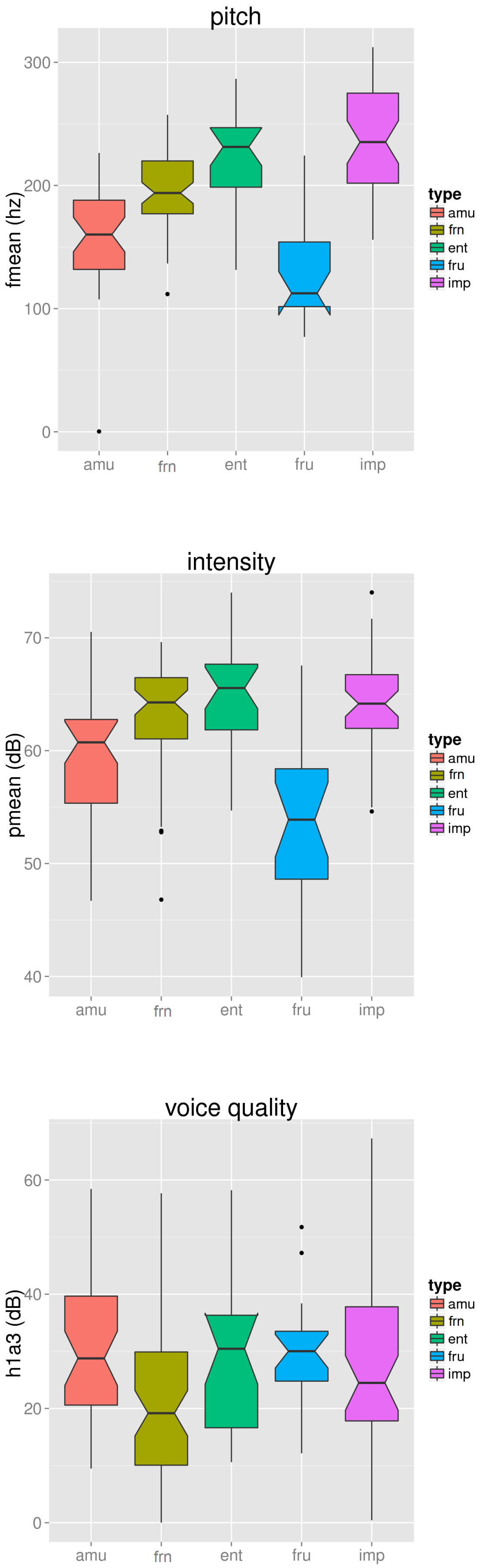


Figure 4.5: Average distribution of prosodic features

As illustrated in Figure 4.5, the y-axis represents measures for each prosodic feature, fmean (mean value of Pitch) is measured using Hertz (Hz), pmean (mean value of Intensity) is measured using Decibels (dB) and h1a3 (amplitude of the first harmonic relative to the third formant to represent Voice Quality) is measured using Decibels (dB). The x-axis shows five attitude classes, “amu” refers to Amusement, “frn” points to Friendliness, “ent” means Enthusiasm, “fru” refers to Frustration and “imp” relates to Impatience. Boxplot 1 describes the average mean value of Pitch (measured as f0) for each attitude state. The plot illustrates that “Impatience” showed highest pitch distribution while “Frustration” showed lowest distribution. To test the relationship between the attitude classes with the prosodic feature, a one-tailed t-test is conducted. A one-tailed t-test is used because of the assymetric distributions prevalent in this condition. Indeed ANOVA can and is used in other analyses of features, but in this particular analysis, Pairwise T-Test is conducted to find significance in pairs of features. The reason for analysis using t-test instead of ANOVA is for the purpose of comparing one feature against another by using Pairwise t-test. This is done to find patterns of features that are most prominent, and this information is necessary for the subsequent machine-learning tasks. It is observed that “Frustration” and “Amusement” differ significantly in terms of pitch from the other categories, showing a p-value of less than 0.005 ($p<0.005$) [170]. This means that the average pitch level of speakers in the vlog dataset is significantly lower when showing states of being frustrated and amused. Contrastively, “Impatience” and “Enthusiasm” show highest average pitch range.

Boxplot 2 illustrates the average distribution of Intensity of the speaker’s voice. Similar to Pitch, the one-tailed t-test is also conducted to measure the significant difference between each attitude state and Intensity level of the speaker’s voice. Results found that “Frustration” showed most significant difference, indicating a low value of Intensity with a p-value of less than 0.005 ($p<0.005$) [170]. This result points to the assumption that speakers express the state of being frustrated with a significantly low intensity (loudness). In contrast, the highest value of Intensity is predominant when speakers show the state of “Impatience”. Meanwhile, Boxplot 3 shows the average distribution of Voice Quality. Results from the t-test measure showed that “Friendliness” significantly differs from the other attitude categories with respect to voice quality, with a p-value of less than 0.005 ($p<0.005$) [170].

Analysis from this significance measure identifies three prominent prosodic features; which are Pitch, Intensity and Voice quality. The following sections elaborate on the statistical methods used to highlight roles of individual features, in particular, Pitch and Voice quality. The following section elaborates first on the role of pitch for identification of different attitudes.

4.4.1 Role of Pitch

Prosodic features are selected to train the classifier to understand different attitude states from speaker's vocal activity. From the previous section, Pitch is among the prosodic features that brings most important information in identifying different attitude states. Pitch activity, through the rising and falling of different tones of the voice, seems to give relevant information when speakers express different types of attitude states. As seen from the previous section, significance of pitch is tested through a significance test using the one-tailed t-test. Another relevant statistical procedure, the precision and recall measurement is conducted to measure the predictive performance of the features. As mentioned in Section 4.3.1, precision and recall is typically computed to measure the quality of an unordered set of data. The main purpose of this measure is not to select features, rather it is used to optimise the performance of the classifier. For this purpose, analysis using this method is examined to better understand the role of features. For the classification task, Support Vector Machine (SVM) was performed to better understand the contribution of the individual feature sets. A one vs. rest classification model was applied on the data collected and evaluated against the ground truth annotation. A standard measure for precision and recall is calculated using f-score, which measures the average percentage of precision and recall. Figure 4.3 summarises the f-score of the prosodic features with strongest correlates to the attitude states:

Features	**Precision**	**Recall**	**Fscore**
Pitch	0.63	0.57	**0.57**
Intensity	0.46	0.40	0.41
Voice quality	0.25	0.27	0.22

Table 4.3: Prediction performance of Prosodic features

Figure 4.3 reports on the predictive performance of individual prosodic features. It is observed that Pitch obtained highest performance rate during the classification task with an f-score of 0.57 compared to other prosodic features. This finding highlights the notion that information on speakers' pitch activity when stating attitudes brings most importance to the classification task. This result is in agreement with past research that supports the idea of Pitch as a determining factor in recognising speakers' attitude, affect and emotion [47].

Pitch influence is further supported with a subsequent experiment. This analysis is conducted using a larger sample size of vlog data attributed to 3 attitude classes; Positive, Neutral and Negative attitudes. A measure of variance (ANOVA) is tested on the sample data because this experiment contains multiple groups of samples (prosodic features). Hence, this measure is most appropriate to use in the given condition. Findings suggest that Pitch, again, contributed most to the prediction rate. Figure 4.6 illustrates the findings:

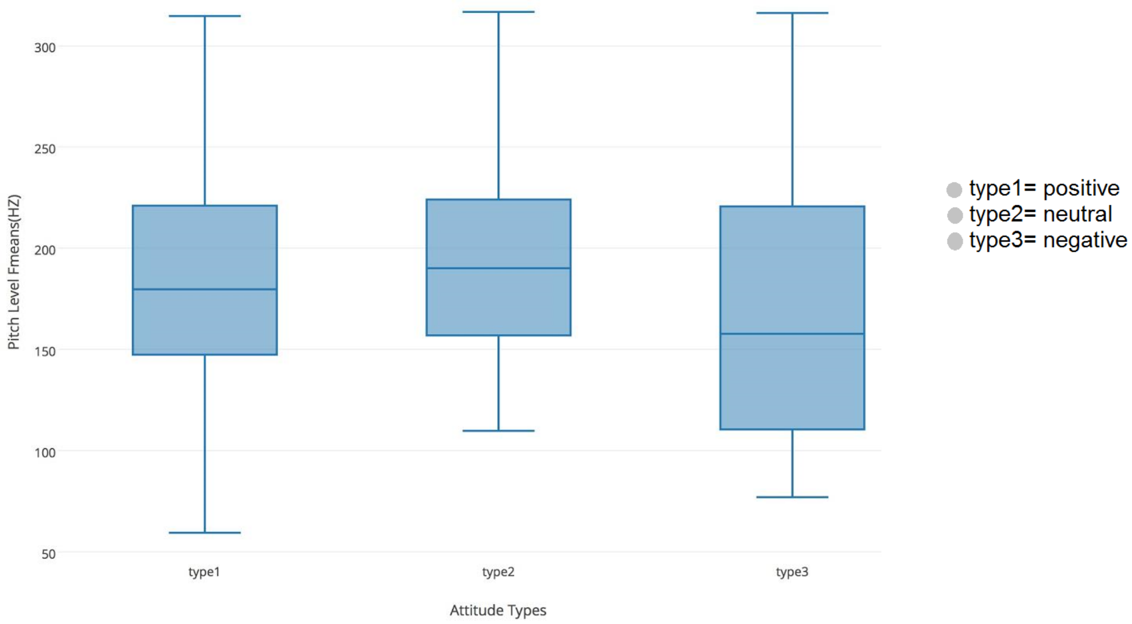


Figure 4.6: Boxplot indicating distribution of Pitch

Figure 4.6 indicates Pitch distribution across 3 attitude categories. The y-axis represents the pitch level measured in hertz (Hz), while the x-axis shows the attitude classes. Results from ANOVA test was conducted over several speech features and results showed that there was a significant difference in Pitch as compared to other features with a p-value of 0.001

(p<0.005) [171]. It is found that Neutral Attitude (Type 2) showed a normal distribution of pitch range while Negative Attitude (Type 3) showed the largest varying degree of pitch level. Neutral attitude showed a symmetric distribution as “Friendliness” (associated with neutral attitude) which is expressed with a relatively level pitch of the voice. In contrast, Negative Attitudes (Type 3), associated with “Impatience and “Frustration” showed greatest variance in pitch range. This observation is not surprising as Impatience (shown in Figure 4.5) shows a contrastive pitch range compared to Frustration. As indicated in Figure 4.5, Impatience is expressed with a high pitch range whilst Frustration is expressed in low pitch. Due to this observation, this might lead to the varying pitch levels of the Negative Attitudes [171].

Findings from the analyses above described the relevance of Pitch as a prosodic feature that brings contribution to the attitude classification task. In other words, the classifier is able to distinguish between attitude categories by the information given on speaker’s pitch range. High pitch indicates “Impatience” while low pitch indicates expression of “Frustration”. The influence of pitch in identifying attitude, affect and emotion is not a novel finding, but this result agrees with the pertinence of this feature for developing a reliable attitude classification model.

4.4.2 Role of Voice Quality

Voice quality refers to properties like breathiness, creakiness, harshness and so on. It seems clear that this feature plays an important role in communicating attitude: for example, saying ’What are you talking about?’ might be perceived as friendly, impatient or excited depending on whether it was uttered with a modal, breathy or harsh voice. So we now turn to the performance of the attitude identifier using the variable of voice quality. Previous research studies the role of voice quality across gender [156], [172], [65] and for detection of speaking styles and speech acts [66]. These works highlight the relevance of voice quality in identifying different speaker characteristics. This study further investigates whether voice quality plays an essential role in determining speaker attitudes. Preliminary analysis was conducted over a sample of vlog data. To accomodate an unbalanced sample size of attitude classes, the Welch T-Test measure of significance is selected over a conventional Student t-test. Treating measure of the amplitude of the first harmonic relative to the third formant (h1-a3) as a characteristic of Voice Quality, results from the test is pictured in Figure 4.7:

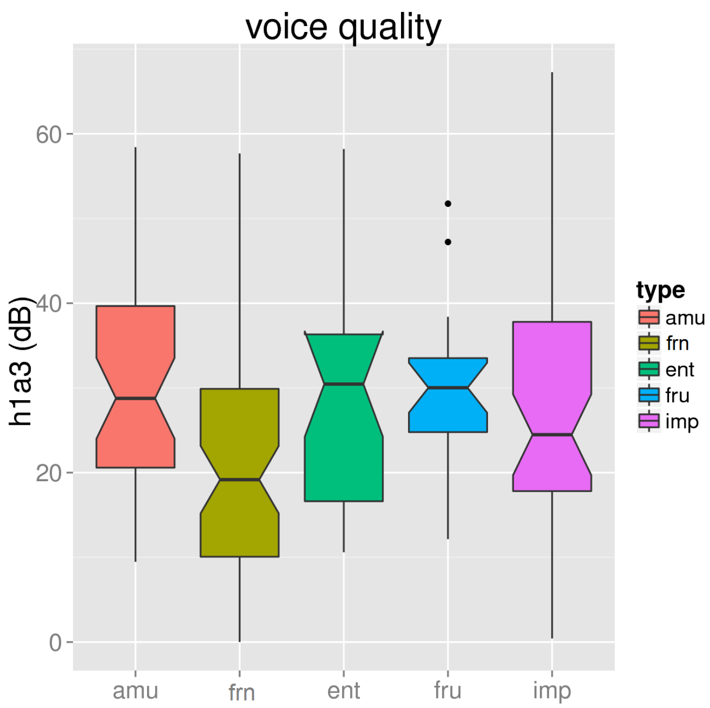


Figure 4.7: Boxplot indicating Voice Quality distribution

Figure 4.7 shows a boxplot of voice distribution across 5 attitude classes, with the y-axis representing value of h1a3 measured in decibels and the x-axis stating three attitude categories. From observation, the attitude class “Friendliness” significantly differs from the others with respect to voice quality, with a p-value of less than 0.005 ($p<0.005$). This is interesting as speakers possibly produce different glottal characteristics when showing their friendly state. They may indicate their friendly state using breathy voice and at other times, in a harsher voice. This finding is in agreement with findings depicted by Gobl and Ni Chasaide [19] whereby speaker affect may not necessarily be attributed to one glottal characteristic. A combination of lax-creaky voice has strong attribution to positive valence, which among them is friendliness [19].

Further analysis is conducted to understand the role of voice quality in a larger sample of vlog data. Another statistical method of understanding relevance of features is the measure of Principal Component Analysis (PCA). As previously mentioned in Section 4.3.1, PCA functions to reduce dimensionality of features. After feature reduction, new data is transformed into orthogonal (independent) variables called Principal Components (PC). By conducting this analysis, only features that have maximum variability are selected as they represent the most important features. In this section, analysis using PCA only pays attention to the identification of dominant features, and not on the data reduction analysis. By evaluating different principal components, PCA identifies features that are most useful to the classification task of different attitude categories. Table 4.4 indicates results of PCA on prosodic features, categorised from the first Principal Component (PC1) to the sixth Principal Component (PC6).

Features	PC1	PC2	PC3	PC4	PC5	PC6
fmean	**-0.4619**	0.0933	-0.2163	0.1073	-0.1163	-0.1322
fmax	-0.4261	0.0718	-0.2800	0.1396	0.0574	-0.1989
fmin	-0.3228	0.1207	-0.1077	-0.0719	-0.4520	0.0902
fpct	-0.0380	0.1291	0.3197	0.2918	0.3457	-0.1685
fvcd	-0.2217	-0.0604	-0.2196	-0.0700	0.4573	0.5451
pmean	-0.4360	-0.1889	0.2314	-0.0126	0.0551	-0.0149
pmax	-0.4247	-0.0800	0.1439	0.0464	0.0100	-0.2031
pmin	-0.1882	-0.3971	0.4112	-0.0844	0.0782	-0.0753
ppct	0.0176	0.1172	0.2240	0.6043	0.3005	0.0735
h1h2	-0.0730	-0.2851	-0.3935	-0.0809	0.3471	0.2545
h1a3	0.1025	-0.5642	-0.1057	0.2377	-0.1628	-0.0568
h1	0.0699	-0.5696	0.0425	-0.0470	-0.0628	-0.0859
a3	-0.0908	0.1146	0.3318	-0.6446	0.2439	-0.0477
dn	0.1220	-0.0141	-0.3755	-0.1350	0.3741	-0.6920

Table 4.4: PC values of prosodic features

As mentioned earlier, the first principal component explains most of the feature variations. Results from Table 4.4 shows fmean (mean value of pitch) with the highest number of variation. This information is nothing new as the previous section elaborated on the contributing role of pitch towards the classifier. What is useful in conducting PCA is by evaluating the remaining principal components (PC2 onwards), information on other feature variations can be retrieved. When examining the feature variations from the second principal component (PC2), voice quality, through the representation of (h1-A3) shows most variance. Figure 4.8 illustrates prosodic features pictured from Principal Component 2:

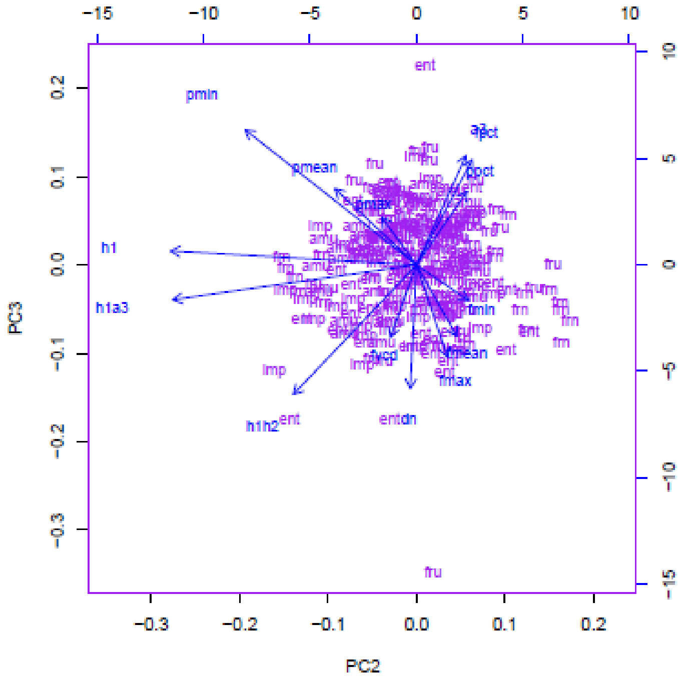


Figure 4.8: Scatterplot showing PCs for prosodic features

Figure 4.8 depicts prosodic parameters from PC2 of PCA. PC1 is largely influenced by pitch variations, thus PC2 is used to identify influences from other prosodic parameters. By examining the y-axis of the scatterplot, the feature pictured at the left quadrant marked as (h1a3) is clearly distinguished from other prosodic parameters. This can be interpreted that the feature (h1a3) is one of a voice quality feature that shows variation from other features. The results of this finding seems to agree with past work on the role of voice quality as a relevant prosodic feature for attitude recognition [19], [170].

This section describes the role of prominent prosodic features, namely Pitch and Voice Quality. The relevance of these features for the classification task is measured through several statistical models. These measures will also be applied in understanding visual feature characteristics, which is described in the next section.

4.5 Prominent Visual Features

Different attitudes are expressed dynamically through facial information. One could decipher a person getting excited or impatient merely by looking at the movement of the lips or eyebrows, and through eye gaze. Measurement of visual characteristics of a person can be obtained using several facial tracking models. The use of AAM in this study allows for the extraction of 67 visual features from landmarks of the face (illustrated in Figure 4.2). To evaluate the relevance of facial features towards the performance of the classifier, data processing of these data points is analysed through statistical models. When processing the visual data points, the absolute mean value of the movement of each data point is computed. This is automatically obtained when running the AAM software, giving absolute values of pixels for the movement of each data point.

To derive information on prominent visual features, the first analysis involves the measure of precision and recall. The procedure is similar to the analysis conducted for prosodic features in Section 4.4.1, where the f-score measures the average value of precision and recall. Specific visual features show most prominent information for the prediction task. With reference to the visual labels in Appendix G, the Nose is interpreted from values of V4 (Nose tip), the Mouth from values of V63 (top inner region of the mouth), the Eyes from values of V29 (lower part of the eye) and the Jaw represented by V9 (face contour). A summary of the result is shown in Table 4.5:

Features	**Precision**	**Recall**	**Fscore**
Nose	0.31	0.27	0.27
Mouth	0.25	0.26	0.25
Eyes	0.24	0.21	0.21
Eyebrow	0.21	0.21	0.19
Jaw	0.32	0.29	**0.30**
Video all	0.21	0.23	0.20

Table 4.5: Predictive performance of Visual features

Table 4.5 reports individual visual features that correlate to the attitude states. Although results do not show significant difference between all visual features, it is found that the jaw shows highest predictability rate with an f-score of 0.30. The Nose also shows a high predictability rate of 0.27 followed by the mouth region with an f-score of 0.25. Table 4.5 presents pertinent regions of the facial contour, specifically the nose, mouth, eyes, eyebrow and jaw. The following sections highlight some of the most prominent visual features that give influence to determine different expressions of attitudes.

4.5.1 Role of Jaw

As described in Table 4.5, movement of the jaw (represented through pixel values of the face contour) is one of the facial landmark that is a contributing factor for attitude recognition. The influence of the jaw brings forth the indication of speech activity [173]. This means that speech activity through the jaw movement is a major deciding factor for the classifier to discriminate between attitude states. For instance, "Amusement" and "Impatience" is discriminated through high and low jaw movements. So when a speaker is impatient or angry, the jaw shows highest movement as the speaker rants in a fast-paced and energetic manner. Contrastively, motion activity of the jaw presents low values when speakers are in the state of frustration, when a person may not show such energetic speech activity. Past research also supports this finding where the jaw is found to be among the contributing features for audio-visual feature extraction for automatic speech recognition (ASR) [86] [174].

4.5.2 Role of Eyebrows

Apart from the movement of the Jaw, one interesting facial feature that shows contribution to the attitude classification task is the movement of Eyebrows. Again, PCA is used to conduct feature selection and analysis of visual features. Using the same statistical approach as in Section 4.4.2 where PCA is used to find the influence of voice quality, PCA is used to derive pertinent visual features, as illustrated in the principal components. Table 4.6 illustrates the analysis obtained based on principal components:

Features	PC1	PC2	PC3	PC4	PC5	PC6
Head	**0.1439**	-0.0205	-0.1740	0.1099	-0.1252	0.0120
Left Eye	0.1304	-0.0657	-0.0061	0.1810	0.0830	-0.0513
Right Eye	0.0607	-0.0667	0.0414	0.2058	0.4156	-0.3911
Nose Tip	-0.0213	-0.0536	-0.1907	0.2367	0.2883	-0.3359
Corner Left Mouth	0.0777	0.1163	0.1070	0.1750	0.1214	0.4312
Corner Right Mouth	0.0960	0.0974	0.0730	0.2340	0.2947	0.3611
Left Face Contour	0.0611	0.2354	0.0169	-0.1142	-0.1276	-0.2134
Right Face Contour	0.0754	0.2284	-0.0550	0.0064	0.0196	-0.0487
Left Chin	0.0826	0.1917	-0.1111	-0.0491	-0.0228	0.0022
Right Chin	0.0665	0.2163	-0.0900	-0.1696	-0.0874	-0.0785
Bottom Chin	0.0611	0.2509	0.0234	-0.0876	-0.0577	-0.1840
Eye Brow Corner Outer Left	0.0733	0.2487	-0.0123	0.0081	0.0331	-0.0345
Eye Brow Corner Inner Left	0.0719	0.2359	-0.0790	-0.0339	0.0366	0.0634
Eye Brow Corner Inner Right	0.0635	0.2430	-0.0410	-0.1052	0.0044	-0.0081
Eye Brow Corner Outer Right	0.0578	0.2603	0.0495	-0.0509	0.0164	-0.1069
Eye Brow Middle Left	0.0679	0.2550	-0.0080	0.0400	0.0490	-0.0366
Eye Brow Middle Right	0.0626	0.2597	-0.0322	0.0042	0.0473	0.0310
Eye Brow Middle Left Left	0.0582	**0.2645**	-0.0178	-0.0018	0.0779	0.0063
Eye Brow Middle Right Left	0.0574	0.2632	0.0207	0.0185	0.0793	-0.0172
Eye Brow Middle Left Right	0.0802	0.2237	0.0826	0.0310	0.1143	-0.0815
Eye Brow Middle Right Right	0.1359	-0.0262	0.1699	-0.0973	-0.0311	-0.0270

Table 4.6: PC values of visual features

Table 4.6 indicates Principal Component 1 being heavily influenced by the head movement, other visual features are portrayed as less important. In order to find influences and contributions from other facial features, Figure 4.9 illustrates analysis using PC2 and PC3:

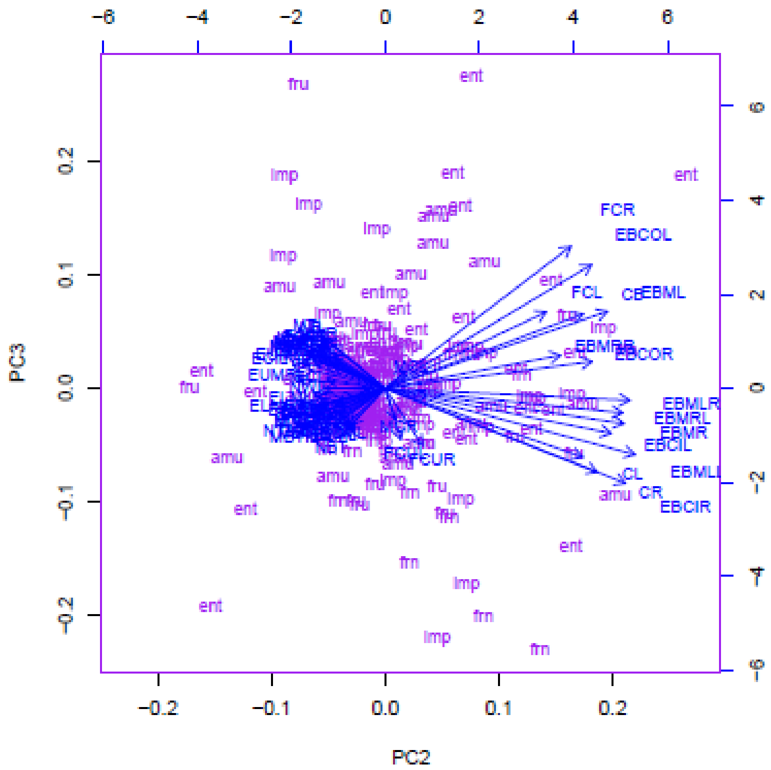


Figure 4.9: Scatterplot showing PCs for visual features

Figure 4.9 shows the analysis of visual features using PCA. It is apparent that the second and third components show a distinct separation of eyebrow features from the rest of the facial features. This is seen from the lower right quadrant (see Figure 4.9) with labels marked in initial EB letters indicating EyeBrows. This finding is in agreement with past literature on the role of eyebrows in emotion recognition where different emotions have strong correlates with different eyebrow movements [15]. Although the present research focuses directly on attitude, because of the similar attribute to emotion, this association to Ekman's findings on the role of eyebrows may present similar patterns for distinguishing attitude expressions of speakers through eyebrow movements.

Apart from PCA, the role of eyebrow movements can be presented by conducting analysis of variance for visual features using ANOVA. The measure of variance was conducted on a subset of attitude classes, which are Positive, Neutral and Negative attitudes. Figure 4.10 shows a boxplot indicating distribution of Eyebrow movement:

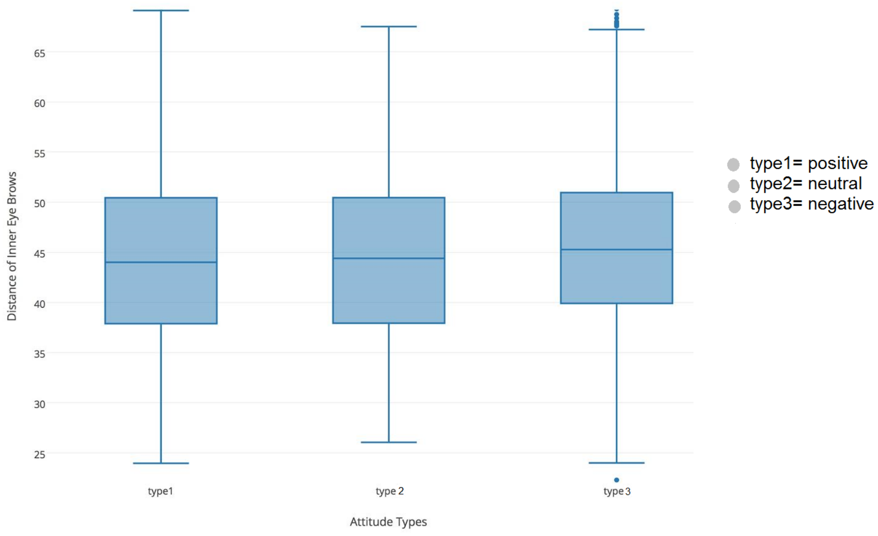


Figure 4.10: Distribution of Eyebrows

Figure 4.10 illustrates distribution of Eyebrows across three attitude states of vlog speakers. Findings suggest movement of the Eyebrows as the most salient signal with a p-value of 0.001 ($p<0.005$). It is found that there is little variance in the distribution of Eyebrows for all attitude classes. Negative Attitude (Type 3) however showed greater variance compared to the other attitude classes. This occurs because subsets within the Type 3 attitudes (Frustration and Impatience) themselves have varying prosodic and visual characteristics [171]. For example, when a speaker is impatient, it might appear that the eyebrows show higher activity of motion, where the eyebrows are likely to converge closer together to indicate anger or impatient. Whereas Frustration may be indicated with slower movement of the eyebrows, where there is less likelihood for the eyebrows to move in a fast-paced motion.

Findings from analyses on eyebrow movement as a discriminating factor for attitudes of vlog speakers is in agreement with past research. Ekman, for instance describes categories of eyebrow movements as indicators of happiness, sadness and other emotional expressions [15]. Our findings, although studies attitudes and not emotions, support the works by Ekman [15] and Haq et.al [175]. Haq et.al for example, conducted analysis of emotion recognition using the upper region of the face (inclusive of eyebrows) which proved to contribute highly to their recogniser. Although the role of eyebrows is not a novel contribution to the pertinent visual features, it supports the theory that eyebrows do provide relevant information and contribution to the prediction of people's psychological and affective states [15],[76],[77].

4.6 Fusion of Multimodal Features

This research's core interest is the idea of combining different modalities for the development of an automatic attitude classification model. The general observation is that a fusion of different modalities provides more information about the attitude state of speakers. For instance, a combination of harsh voice and furrowed eyebrows may indicate the state of Impatience. This combination of audio and visual information is useful for multimodal communication, not only in human-human communication [56], [58] but also in human-computer interaction [87],[88],[89]. This section elaborates on the integration of multimodalities in attitude perception as well as in automatic attitude recognition.

4.6.1 Multimodal Perception

This section describes a fusion of multimodalities towards perception of attitude states of speakers in the vlog dataset. As reported in 3.4.2, a perception test was conducted to measure the validity of attitude categories as expressed by vloggers in the vlog dataset [148]. However, this study also aims to determine participants' attitude perception based on several modalities.

Treating different modalities as the stimuli in the online survey, participants are required to select relevant attitude categories presented in three sections. Each section represents one modality, whereby Section A consists of Audio only stimuli, Section B consists of Video only stimuli and Section C contains both Audio and Video stimuli. Participants are asked to identify an attitude for each stimulus in each modality. A visual representation of the sections is illustrated in Figure 3.8 (see Section 3.4.2). Observation was done to notice the total values of agreed answers between participants and expert annotators for each section. Figure 4.11 illustrates the observation for each modality:

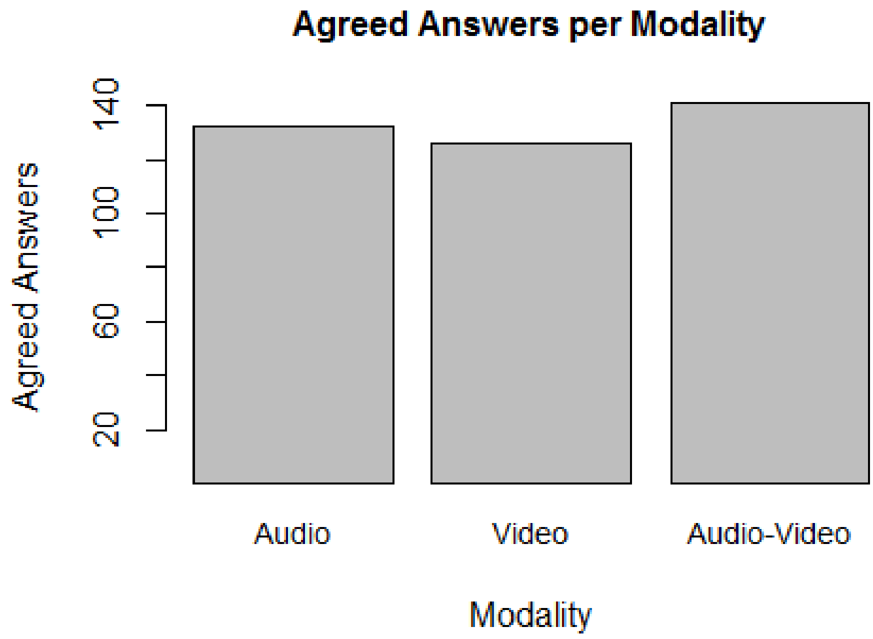


Figure 4.11: Agreed answers per modality

Figure 4.11 shows the observation of consistent answers between expert / non-expert agreement raters per modality. It is observed that there is a higher agreement between expert and non-expert raters in Section C, which is a fusion of audio and video stimuli. This means that participants perceive attitude categories better through audio-visual information. This is followed closely by Audio modality. The percentage of agreement for audio modality is 33.08%, while video modality is 31.58% and audio-video modality yields 35.34%.

This finding supports other studies concerning audio-visual perception of attitudes and affect. Allwood et.al [44] conducted perception of affective-epistemic states (AES) through audio-visual cues and found audio-visual cues to be helpful for some AES categories while other categories are perceived unimodally. Shochi et.al [91] conducted a similar perceptual study on Japanese participants. They also found that some attitudes are perceived better through audio modality while other attitudes are better perceived through visual modality. Findings from their perceptual results further indicate that Audio-Visual modality is closely related to Audio only modality.

Findings from the present work, as illustrated in Figure 4.11 is also in agreement with Shochi et al. [91] whereby Audio-Visual and Audio modalities show stronger influence in attitude perception. Although this work did not analyse trends on whether specific attitudes differ for each modality, this is an aspect of great interest for this research's future direction. To summarise, findings from this perception study agrees with the assumption that a fusion of modalities is helpful for identification of speaker's attitude expressions.

4.6.2 Multimodalities in Attitude Classification

The previous section discusses relevance of multimodalities through a perceptual analysis of attitude expressions. This section further reports on the influence of multimodal features, namely from the combinations of prosodic and visual features for attitude discrimination through automatic classification techniques. There are several techniques used to classify attitudes through multimodal features. One method of automatic classification of different attitudes is through conducting analysis using decision trees.

Decision tree is one method of classification through the representation of most prominent features in the form of a tree. This is an interesting approach as it provides a detailed visualisation of features that give most influence towards deciding and discriminating different attitudes. There are several stages of classification using decision trees. The algorithm functions to detect all features that provide most information. The second stage of classification involves the building of multiple trees of features that gives most relevant information and ignores the irrelevant ones. This gives an overall view of which features that are most important. Seeing that the features are discretely categorised, they only appear once in the decision path. The next stage of classification involves pruning of nodes whereby only information of features that is of most importance will be kept, hence giving a reliable classification task [176].

Since this section explains feature selection, and not on classification, the present work focuses on the second stage of classification using the decision tree. This stage involves construction of the tree using an algorithm to identify the most relevant information without conducting further classification tasks, such as pruning. Figure 4.12 outlines feature selection based on a fusion of prosodic and visual features for the attitude classification task:

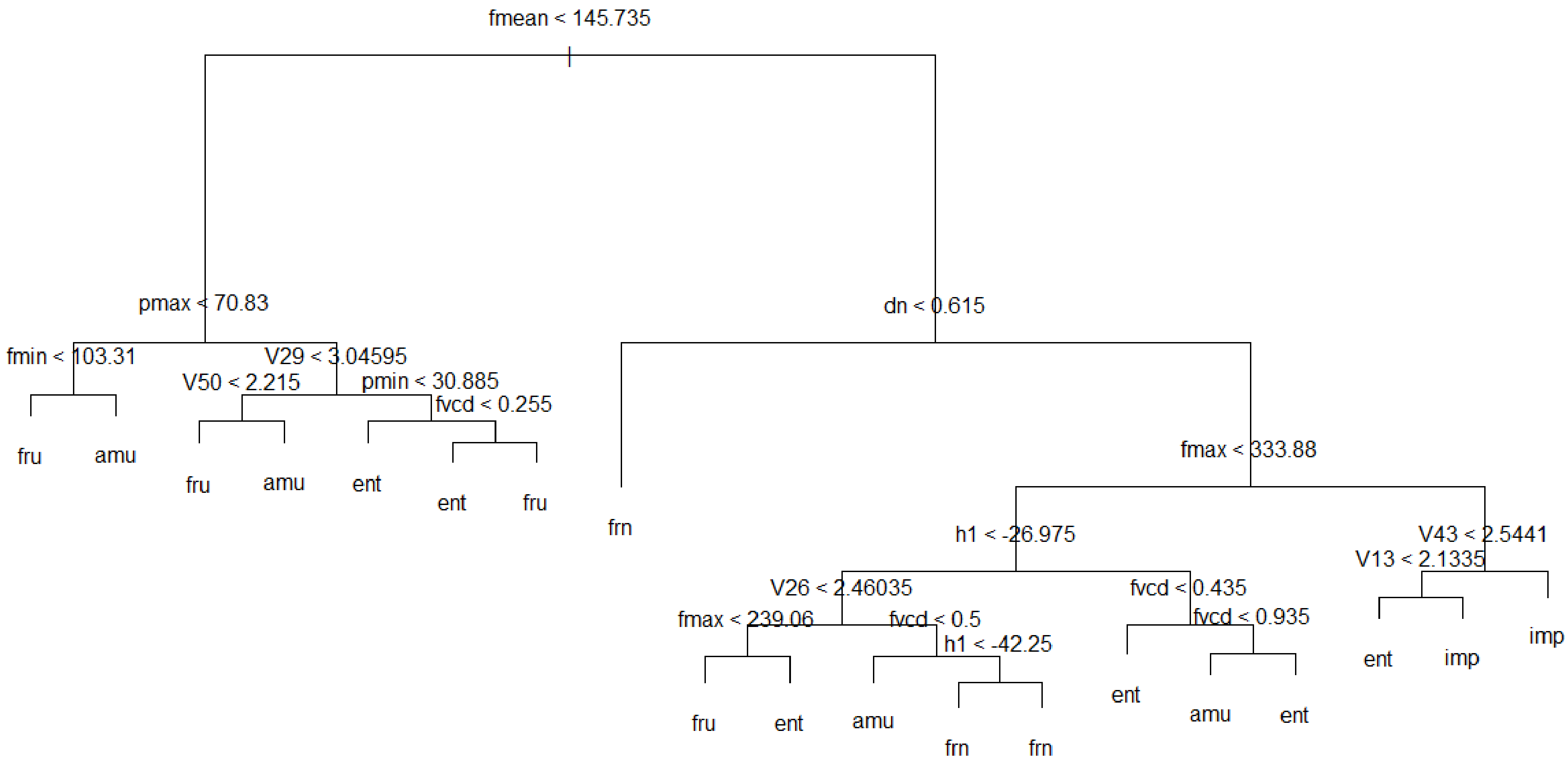


Figure 4.12: Feature selection using Decision Tree

Figure 4.12 indicates the feature selection process through a fusion of prosodic and visual parameters using the decision tree. It is observed that the mean value of Pitch of the voice acts as the main contributor to the classification task of identifying different attitude states. It is also found that a combination of the minimum value of Pitch and movement of the Lower Left Eye (indicated as V29) play a significant role in determining attitude states of "Frustration" and "Amusement". This means that the classifier is able to predict these two attitudes in particular based on low pitch range and lower left eye movement of the speakers. The left eye movement may possibly indicate speakers' aversion of the eyes from the camera. This is a possible assumption that indicates the state of being frustrated or amused, as speakers may avert their gaze from the camera when they are frustrated with something. The analysis using decision tree highlights two distinct features that achieved highest individual contribution for each modality, which are Pitch and Eye Movement. This observation is in line with the aforementioned analyses, particularly on prosodic modality, where Pitch stands out as a prominent feature, whereas Eye movement shows contribution in this particular analysis, compared to the other analyses conducted. From this finding, the fusion of prosodic and visual information in the classification task marks distinctive attitude traits as expressed by the vloggers.

This initial task of classification sheds light to the role of multimodal features as indicators of attitudes. This seems to support the notion from previous literature that a fusion of multimodalities is useful in developing interfaces for affective recognition. Busso et.al [57], for instance elaborates on a classification model for achieving accurate emotion recognition through a fusion of speech and facial features. Considering attitudes as similar to emotional traits, results from this work on attitude classification provides sufficient claim that a fusion of multimodalities is helpful for achieving a robust classification model for affective and behavioural conditions of interlocutors.

Although there are several techniques for classification, utilising the algorithm of the decision tree merely involves observation of features that are derived based on relevant information that each feature provides. The main research work of this thesis applies other techniques for classification using different classifiers and algorithms. Elaboration of the processes, methods and analyses will be best explained in the next chapter.

4.7 Discussion

The analysis conducted in this chapter describes the second stage of building an affective modeling technique, which involves feature extraction and selection. Methods and tools used to extract prosodic and visual features are elaborated in the initial part of the chapter. Notable prosodic characteristics, which include pitch, intensity, voice quality and duration of the speakers' voice are extracted. These features are considered sufficient to better understand speakers' vocal activity.

Besides that, this work applies the use of Luxand FaceSDK facial tracking application to track motions of the facial region, which is especially relevant for the vlog data as speakers are naturally talking to the camera without any obstructions. Visual features are easily derived by using this fast-paced facial tracker. So, using this tool is suitable for this vlog data. For the pre-processing stage, the values of visual features were calculated by using the absolute mean values for each data point of the facial region in each video segment containing the attitudes. However, pre-processing of facial regions from other research is conducted differently. For example, Tome et.al [158] conducted analysis on facial detection in forensics scenarios. They applied Face SDK application for automatic face recognition to determine a person's identity. They selected 15 out of 65 facial regions focusing more on the eyes, eyebrows, nose, mouth and chin. Using equal size of facial proportions, the extractor was able to derive values for each of the 15 facial regions at frame level. This present study however, analysed the videos at segment level, and not at frame level. This choice of extracting values of the facial region at a segmental level instead of looking at the frame level is a poor choice as more valuable information could have been discarded. This limitation could be overcome when we conduct visual analysis at a frame level. This analysis however is reserved for future work.

The next stage of data processing prior to the classification task is feature selection. Features are selected so as to avoid any redundancy between similar features, which may affect the overall performance of the classifier. Features that are highly correlated are identified and further processed to minimise redundancy. Only features that provide most relevant information for discriminating different attitudes are kept while features that are least relevant are discarded. This is achieved by testing the correlates between features through measures of significance, PCA, f-scores and decision tree. These methods are typically utilised to measure the performance of the classifier. However, they can also be conducted to provide information on relevant features. Through these analyses, several observations were made about prominent features. This work contributes to the knowledge that pitch and voice quality have prominent roles in terms of providing sufficient prosodic information for identifying different attitude states. This is also observed for visual features whereby eyebrows and the jaw movement show prominent roles in identifying attitudes. This present work also highlights the relevance of multimodal feature contributions where eye movement and pitch of the voice may suggest the speaker's state of amusement or frustration. The relevance of multimodalities is also supported by the findings from a perception test which suggests that different attitudes are best perceived through the combination of audio and visual modalities. This notion of fusing together multimodal information contributes to the body of knowledge on multimodal communication.

4.8 Conclusion

The chapter elaborates on the data processing techniques used prior to arriving to the last stage of modeling a reliable attitude recognition interface. This data processing task involves feature extraction and selection. This pre-processing stage is essential to ensure the reliability of the features used to train the classifier during the classification task. The purpose of presenting several statistical analyses of the features in this chapter is mainly to explore the data, whereas machine-learning experiments constitute a different analysis perspective, which will be described further in Chapter 5. Combining these methods (i.e data exploration and machine-learning experiments) is left for future research. In the next chapter, an elaborate description of the classification task is explained in detail. Several methods are used to train the classifier using machine-learning techniques and an elaboration for each stage of experimentation is discussed at length.

Chapter 5

Automatic Attitude Classification

5.1 Introduction

The previous chapters have described the methods and processes involved in obtaining data labels as well as feature identification, extraction and processing. These processes are necessary for the final stage of developing a recognition interface. This chapter elaborates on the discussion of a computational framework through several stages of experiments. As previously highlighted in the first chapter, supervised machine-learning techniques are utilised for the purpose of developing an attitude recognition interface. The first part of this chapter details conceptual applications of machine-learning techniques as well as the methods and algorithms used in this study. The final part of the chapter discusses the various levels of classification tasks and analyses in developing the system.

5.2 Conceptual Applications

Prior to discussion on the computational analyses on attitude recognition, this section briefly describes concepts used in the machine-learning process. Application of a supervised machine learning procedure seeks to examine the predictive performance of the classifier based on pre-determined set of data labels. This work applies this procedure after associating attitudes to five distinct labels (detailed in Chapter 3). Machine-learning requires us to firstly determine what "Object" is used, as well as the data "Labels" and "Feature Vectors" that are identified. In the present work, these conceptual underpinnings are summarised in Table 5.1:

Concept	Implication
Object	Attitude
Labels	Amusement Enthusiasm Friendliness Frustration Impatience
Feature Vectors	14 prosodic features 67 visual features

Table 5.1: Concepts for Supervised Learning

Table 5.1 describes the concepts used in supervised machine learning. The object relevant to this study is the “Attitude”, while the set of labels from the dataset is defined based on the N5 attitude annotation scheme as explained in Chapter 3. These labels are “Amusement, “Enthusiasm, “Friendliness”, “Frustration” and “Impatience”. The feature vectors involved in this process is the combination of multimodal features, which are 14 prosodic features and 67 features visual features, as elaborated in Chapter 4. Now that the conceptual applications are visualised, the following section elaborates on the important concepts and tools used for the present computational framework.

5.3 Classification technique using SVM

There are several known machine-learning algorithms used for data classification and learning tasks. One of these algorithms that is utilised in the study is Support Vector Machine (SVM). SVM is a powerful tool used for data classification as it gives easier and faster results. Although results from this approach do not guarantee highest accuracy from its predictions, SVM typically gives acceptable results, when used correctly.

Data classification task using SVM involves separating data into training and testing sets. Training sets contain instances of the data labels, and these are called target values. And these target values are associated with feature vectors, called attributes. SVM functions to produce a model that predicts the target values of the test data based on the test attributes [177]. For example, one instance of the label “Amusement” is associated with several attributes obtained from the feature vectors (prosodic and visual features). The SVM algorithm trains on the given training set (comprising of attributes) to determine whether it is able to identify the target value as “Amusement”.

There are a number of procedures used in order for SVM to function appropriately. Data should be pre-processed beforehand to ensure that they are presented in a form that is readable by SVM algorithm. Attributes or features should be presented numerically. This is not a major concern of this study as features are extracted in numerical values. However, when addressing the values of these attributes, SVM functions best when these values are not too large. In order to derive normalised values, where the average range of values is scaled down to an acceptable value, the scaling or normalisation procedure is to be conducted. The main purpose of normalisation is to avoid the condition where attributes (features) with larger values dominate other attributes with smaller values, and also to avoid calculation difficulties [177]. A typical scaling method is through linear scaling for each attribute to the range [-1,+1] or [0,1]. To avoid specific feature dominance over others, the present experiments conducted in this study conduct normalisation of feature values prior to training the feature sets during the classification task.

After normalisation, SVM is trained based on the algorithm of a suitable parameter. For model selection, a typical kernel parameter used is Radial Basis Function (RBF) kernel. A kernel is a similarity function that determines how similar the samples are to the labels. This RBF kernel has its advantage compared to other kernels as it accommodates non-linear class samples. Seeing that the present work has five class labels, this kernel is the most suitable to be used for classification of non-linear samples. This kernel functions to map samples into higher dimensional space in non-linear fashion. The use of kernels is also advantageous as kernels can easily compute over large feature spaces.

A cross-validation technique is a standard procedure used to evaluate the performance of classifiers, including SVM. SVM classification typically involves a binary classification. Since this study attempts to develop multi-class classification of five attitudes, SVM could also be used for this task by applying a one vs. rest approach. The concurrent experiments for attitude classification are conducted using this approach. Training using the cross-validation technique aims to determine whether the prediction accuracy from the unknown dataset precisely reflects the classification performance on an independent dataset [177]. The present work applies a 10-fold cross-validation technique. What this entails is that the training set is equally divided into 10 subsets. Then one subset is trained on the remaining subsets and each instance of the training set is predicted once. The outcome of this training will give an accuracy percentage of which data is correctly classified. After finding the subsets that give higher accuracy, then these subsets are used to train on the complete dataset to obtain higher prediction accuracy.

This section describes an overview of methods used for data classification. This also includes optimisation procedure for the classifier to give higher predictability rate during the classification task using a cross-validation technique. The following section discusses stages of experimentations using this machine-learning approach for the modelling of an automatic attitude recognition system.

5.4 Experimentation

In order to derive a reliable recognition model for attitude detection, a series of experiments were conducted using different number of samples and feature sets. Figure 5.1 outlines the stages of experimentation:

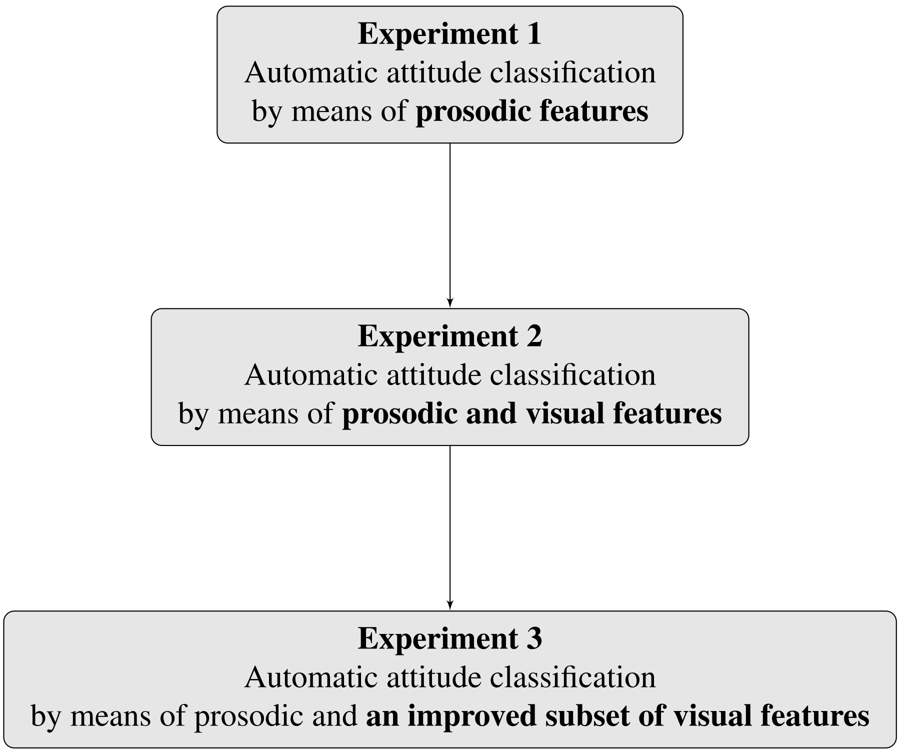


Figure 5.1: Stages of Experimentation

Figure 5.1 outlines phases of experimentation for the development of an automatic attitude recognition system using a fusion of multimodal features, namely prosodic and visual features. The following sections provide detailed elaborations for each phase. Note that data samples differ across all stages of experimentation. The total of attitudes described during data collection stage in Chapter 3 summarises the complete set of data samples, while these subsequent experiments use samples collected during the initial stages of analysis. The complete set of data samples is used in the the final stage of experimentation, as will be described in Experimentation 3.

5.4.1 Experiment 1

The goal of this experiment is to develop a classification model for recognising attitudes through prosodic feature extraction. Five attitude classes as expressed by vloggers are identified and annotated by expert annotators [170]. At this stage of experimentation, a total of 100 videos of vlogger speech was collected. After annotation, a total of 194 attitude instances was identified. The number of labels categorised according to the respective attitudes is structured in Table 5.2:

Attitude Label	No. of Instances
Amusement	42
Enthusiasm	25
Friendliness	62
Frustration	22
Impatience	43
TOTAL	194

Table 5.2: Total instances for each attitude label

Table 5.2 describes the total number of instances for the five attitude labels for each attitude category for the initial stage of experimentation. These labels are derived through expert annotation and segmentation. The next phase required in the classification task is feature extraction to form feature vectors for the classifier to train on. Prosodic feature extraction is conducted and 16 prosodic features are derived. This experiment makes use of 14 prosodic features (seen in Table 4.1) and included 2 additional features, which are median value of the fundamental frequency and the median value of Intensity. The inclusion of the median values at this phase of experimentation was crucial to determine its relevance compared to merely depending on the mean values. Median values are relevant as they are not affected by outlier values, hence providing more representative values compared to what the mean values could report on. Although median values are highly regarded over mean values, due to some inevitable feature extraction difficulties, whereby calculation of the median values within the script returned errors, the median values are not included for the subsequent experiments. Results of the prosodic characteristics for each attitude category is summarised in Table 5.3. This result generally provides an overview of relevant prosodic features and their association to the attitude states. Table 5.3 lists the prosodic characteristics for each attitude category. From observation, the attitudinal category Impatience shows the highest Pitch and Frustration the lowest.

After obtaining information on the amount of sample size of class labels and the relevant prosodic features, the classification task is performed with SVM using R programming language. SVM, with the implementation of radial basis function (RBF) kernel algorithm, was used for the classification task of predicting attitude categories. To evaluate the accuracy of the trained model, a 10 fold cross-validation approach was performed. In the evaluation phase, two feature sets are derived, which are, a feature set that includes all 16 prosodic features (Feature set ALL), and a feature set that contains only values that are not highly correlated (Feature set SEL_1). These correlated values were measured using Pearson's Correlation Coefficient. The reason for selecting values that are not highly correlated is that this may be an indication that these features are distinctive and are largely varying from other features. The two feature sets are summarised below:

Feature set ALL: All 16 features described in Table 5.3.

Feature set SEL_1: Only the features that are not highly correlated (with Pearson's correlation coefficient $r < 0.7$): fmean, fmin, fpct, fvcd, pmean, ppct, h1h2, h1a3, h1, a3, dn (refer to Table 4.1). A correlation study of the feature set revealed that some of the features are highly correlated (correlation coefficient $r > 0.7$ with $p < 0.01$ in T-test). Only one of the features in the highly correlated feature pairs was selected and a new 11 dimensional feature set: SEL_1 was generated.

Type	fmean	fmed	fmax	fmin	fpct	fvcd	pmean	pmed	pmax	pmin	ppct	h1h2	h1a3	h1	a3	dn
Amused	162.97 *(36.09)*	172.24 *(50.94)*	228.04 *(59.84)*	103.01 *(30.43)*	0.43 *(0.25)*	0.55 *(0.19)*	59.86 *(5.75)*	60.51 *(5.88)*	75.37 *(5.38)*	36.71 *(10.43)*	0.49 *(0.27)*	6.49 *(4.74)*	30.00 *(12.17)*	-25.28 *(9.53)*	-55.28 *(5.81)*	1.07 *(0.31)*
Impatient	234.39 *(39.72)*	242.39 *(49.92)*	311.76 *(47.04)*	139.05 *(44.12)*	0.43 *(0.27)*	0.61 *(0.16)*	64.09 *(4.13)*	66.05 *(3.99)*	77.90 *(2.67)*	36.48 *(13.09)*	0.38 *(0.25)*	6.00 *(6.34)*	27.96 *(14.88)*	-26.06 *(11.18)*	-54.03 *(7.41)*	1.26 *(0.50)*
Casual	196.14 *(32.2)*	193.97 *(30.02)*	248.40 *(43.85)*	147.33 *(40.80)*	0.26 *(0.16)*	0.73 *(0.13)*	63.23 *(4.69)*	66.75 *(3.28)*	76.60 *(2.97)*	32.08 *(15.61)*	0.32 *(0.20)*	4.34 *(4.67)*	20.65 *(15.28)*	-34.72 *(17.54)*	-55.37 *(5.53)*	0.55 *(0.16)*
Enthusiastic	220.1 *(38.65)*	242.62 *(40.23)*	310.56 *(47.55)*	115.04 *(50.65)*	0.39 *(0.28)*	0.72 *(0.18)*	64.70 *(4.7)*	66.50 *(4.51)*	79.28 *(3.08)*	35.66 *(14.28)*	0.35 *(0.24)*	6.08 *(6.74)*	29.54 *(13.54)*	-26.72 *(11.87)*	-56.27 *(6.28)*	1.20 *(0.52)*
Frustrated	124.78 *(35.15)*	136.84 *(40.16)*	175.89 *(55.96)*	86.67 *(33.09)*	0.40 *(0.30)*	0.49 *(0.23)*	53.65 *(7.32)*	54.49 *(7.24)*	70.35 *(5.64)*	30.62 *(14.80)*	0.57 *(0.26)*	3.59 *(5.02)*	29.22 *(9.81)*	-27.43 *(9.72)*	-56.64 *(4.51)*	1.27 *(0.57)*

Table 5.3: Mean values for each attitude category with standard deviation in brackets

For clearer observation, Table 5.4 shows the results of 10 fold cross-validation SVMs with the different feature sets. Results show that the feature set selected after removing the highly correlated features attained the best prediction accuracy. Thus, feature set SEL_1 (derived after removing highly correlated features) is shown to be the feature set that provides better predictive performance with a 65.46% accuracy rate.

Feature Set	Accuracy
ALL	61.85
SEL_1	**65.46**

Table 5.4: Results for the different feature sets.

Table 5.4 shows the percentage of correct classification of the class labels using two feature subsets. It is found that using a prosodic feature subset of only 11 dimensional feature space indicated higher prediction rate with a 65% accuracy rate while training using all 16 feature vectors indicated a lower percentage of 61.8%. This is an interesting result seeing that the threshold value is 20%, hence this result exceeds the above-chance value. Given five classes of data labels (100% / 5), the threshold value is set to 20%. So when the probability for each of the five labels of the classifier was performing less that the chance value, then the accuracy rate would show a percentage of less than 20%, which means that the classifier did not learn successfully from the feature inputs in recognising the class labels. In this case, with a result that is of above-chance value, the classifier was able to predict the attitude classes based on the training of the prosodic feature vectors.

When comparing the accuracy rate of prosodic features using SVM and human agreement (see Figure 4.11, page 114), it is observed that the accuracy rate of the classification system is higher with a percentage of 65% (all prosodic features), while human agreement for the perception of attitudes using audio modality is merely 33%. This implies that, from the information gathered on prosodic features, an automated system would be able to distinguish attitudes better than human perception.

In retrospect to the above experiment, there are several points of observation worth mentioning. Firstly, the number of attitude segments is unequal and insufficient to form general conclusions. The classifier gives better predictive performance when there is a balanced set of sample data. A possible solution for this evaluation is to include an increased number of data instances for the classifier to learn from. Increased instances of attitude labels are included in the following experiments. Then we could understand the significance of additional labels in contributing to an increased performance rate of the classifier.

Another observation is the use of a subset of the prosodic features. The use of 11 features out of all 16 features shows a relevant contribution to the performance of the classifier. This is because features that are highly correlated are removed and training was conducted with features that are co-varied and distinctive. With a result of 65% prediction rate, this machine learning technique is a good method to be used for automatic classification and recognition of attitudes, as expressed by speakers in the vlog data.

5.4.2 Experiment 2

The objective of the second experimentation phase is to examine methods for achieving an improved recognition system of attitudes through a combination of multimodal, namely prosodic and visual features [173]. From the first phase of experimentation, the classifier was trained using prosodic features to detect five attitude classes from vlogger' speech. The classifier's initial performance showed a reasonable accuracy rate. In this stage of experimentation, the aim is to investigate other features that could contribute to, and better improve the classifier's performance. This can best be achieved by providing an increased amount of instances for the five attitude labels and feature vectors, which includes the addition of visual feature vectors.

For this experiment, an increased number of vlogs was collected and annotated. The total number of vlogs collected for this purpose is 134 vlogs and annotation is conducted using the N5 attitude annotation scheme. The total number of annotated attitudes is 256 instances. A breakdown of instances in accordance to the five attitude labels is detailed in Table 5.5:

Attitude Labels	No. of Instances
Amusement	58
Enthusiasm	51
Friendliness	49
Frustration	50
Impatience	48
TOTAL	**256**

Table 5.5: Total instances for each attitude label

Table 5.5 summarises amount of instances for the five attitude classes collected after the annotation process. During the first stage of experimentation, the total number of instances was 194 with imbalanced numbers for each category (refer to Table 5.2). For example, "Frustration" label only amounted to 22 instances while "Friendliness" label totalled to slightly over 60 instances. This imbalance in sample size is addressed in this current experiment where an increased amount of attitude instances are collected. A total of 256 instances for the five attitude labels are collected to provide relatively balanced instances of attitude categories for the classification task. The number of instances in each attitude class ranges from 48 to 58.

Machine learning also requires information from various features to train the classifier. For this experiment, machine learning was performed for attitude classification using a combination of non-verbal features that includes 14 prosodic parameters and 65 facial movements (refer to Table 4.2 and Appendix G for full list, values of the Head and Mouth Bottom Inner Right are excluded). The prosodic values are extracted and similarly used as detailed in Experiment 1 (refer to Table 4.1).

Initial analysis of prosodic features are outlined in Table 5.6, where there is a description of the mean and standard deviation values of the extracted prosodic features with relation to each attitude category. This description provides a general overview of the extracted prosodic features. Relating to Experiment 1, the attitudinal category showed similar trends where Impatience, again, shows the highest Pitch while Frustration showed the lowest mean value of pitch.

Type	**fmean**	**fmax**	**fmin**	**fpct**	**fvcd**	**pmean**	**pmax**	**pmin**	**ppct**	**h1-h2**	**h1-a3**	**h1**	**a3**	**dn**
Unit	*Hz*	*Hz*	*Hz*	*%*	*%*	*dB*	*dB*	*dB*	*%*	*dB*	*dB*	*dB*	*dB*	*sec*
Amu	168.4 *(35.42)*	231.5 *(57.23)*	108.14 *(32.58)*	45 *(25)*	58 *(20)*	59.28 *(6.18)*	74.46 *(5.74)*	35.92 *(11.10)*	45 *(28)*	6.70 *(6.08)*	30.43 *(12.61)*	-24.19 *(9.47)*	-54.61 *(6.79)*	1.04 *(0.30)*
Ent	220.21 *(49.10)*	298.00 *(59.23)*	131.48 *(57.20)*	35 *(28)*	70 *(27)*	63.62 *(6.68)*	78.03 *(4.21)*	36.56 *(13.18)*	34 *(23)*	6.25 *(5.65)*	27.91 *(11.74)*	-27.32 *(0.75)*	-55.24 *(5.21)*	1.25 *(0.73)*
Frn	197.6 *(32.22)*	248.1 *(45.50)*	148.63 *(40.72)*	26 *(14)*	74 *(12)*	64.71 *(4.10)*	76.72 *(2.38)*	37.78 *(16.43)*	24 *(20)*	4.08 *(4.80)*	20.69 *(16.11)*	-32.32 *(18.21)*	-54.52 *(5.66)*	0.52 *(0.12)*
Fru	116.53 *(28.81)*	155.4 *(42.21)*	84.86 *(27.91)*	32 *(25)*	48 *(23)*	51.87 *(8.26)*	67.71 *(7.72)*	29.22 *(13.09)*	43 *(28)*	3.57 *(5.07)*	29.27 *(10.74)*	-26.53 *(8.89)*	-55.80 *(5.24)*	1.22 *(2.35)*
Imp	234.4 *(40.61)*	312.1 *(47.05)*	142.4 *(49.39)*	39 *(27)*	55 *(23)*	63.77 *(4.24)*	78.96 *(2.69)*	35.84 *(12.43)*	35 *(25)*	5.35 *(5.97)*	28.55 *(14.81)*	-26.12 *(11.61)*	-54.66 *(6.27)*	1.37 *(0.65)*

Table 5.6: Prosodic Mean values for each attitude category with standard deviation in brackets

In this study, visual information is obtained by performing visual feature extraction using FaceSDK AAM (mentioned in 4.2.2). Seeing that the algorithm used was from an older version during this experimentation phase, it only allows extraction of 65 facial landmarks. Appendix G lists 67 visual features from the FaceSDK AAM algorithm, but points V1 (Head) and V67 (Mouth Bottom Inner Right) are excluded for this experiment. Due to this limitation, only 65 values of the facial points were included for this experiment. However, a recent version of the FaceSDK AAM feature extraction software, using 67 facial points is conducted in Experiment 3 and will be further explained in the next section. An example of the facial tracking process is displayed in Figure 5.2:

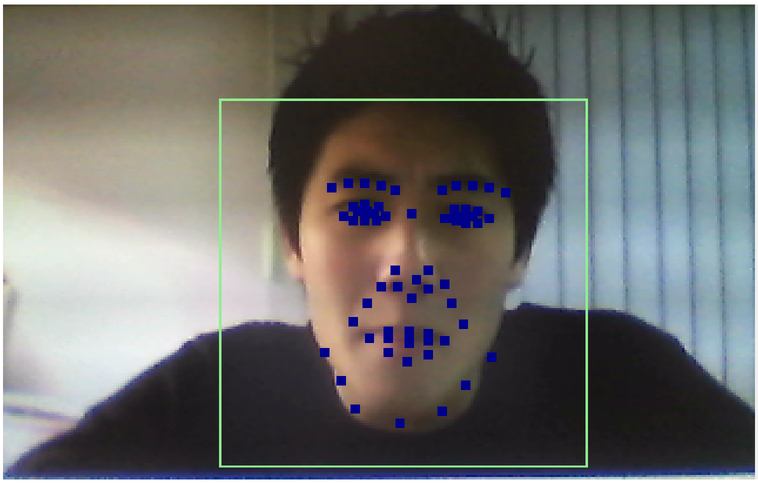


Figure 5.2: Facial Tracking with FaceSDK AAM

Figure 5.2 pictures visual extraction of the 65 facial data landmarks from a vlogger's facial region. To obtain all visual data information, videos with clear visuals of the vloggers' face are pre-selected. Videos by which there exist blurry or rapid movements that cause errors for facial tracking are discarded and replaced with other workable videos. The average movement of each data point is calculated and the values are used to be representatives of the visual feature vectors in the classification task.

Similar to past experiments, data classification was performed using SVM to train the classifier to automatically recognise speaker attitudes through prosodic and visual feature information. The experiment is conducted using a 10 fold cross-validation technique using a library called LibSVM [128]. LibSVM is a package utilised for SVMs and is a standardly used software that provides better and faster classification [128].

The use of 10-fold cross-validation is similarly conducted in Experiment 1 whereby this technique is used to prove the consistency of the predictive model. The cross validation algorithm uses one fold as the test set and the remaining folds as train sets. The process is repeated 10 cycles to find variations in the performance of the classifier. Results from this experiment highlight interesting points of discussion. Based on results of the classifier, it is apparent that attitudes are recognised most through prosodic information compared to the visual parameters. The classifier's predictive performance is summarised in Table 5.7:

Features	**Prediction Rate**
PROSODIC	61.4%
VISUAL	29.2%
PROSODIC + VISUAL	60.9%

Table 5.7: Performance of the Trained Classifier

The trained classifier shows greatest prediction rate using prosodic features, as indicated in Table 5.7. This is a reliable result as this agrees with past studies on the relevance of prosodic information as indicators for distinguishing speaker attitude, affect, speech act and emotion [47],[66], [22], [48], [153].

It is observed however that visual feature information contributed least to the classification task. One reason for this outcome is because of the constrained visual data processing technique. This method merely takes into account the average movement of the visual landmarks of the face. This information alone is insufficient to indicate the visual representations for recognising different attitude categories. Directionality, which refers to the horizontal and vertical direction of motion of the visual data points [171], may provide a more comprehensive representation of the visual feature vectors for each attitude category. Further exploration of the data manipulation of the visual features, particularly in the directionality of facial points is elaborated in the following experiment.

5.4.3 Experiment 3

The main objective of the third experiment is to discover different techniques for data processing to achieve an improved prediction rate of the attitude classification model. Past experiments have established a promising predictive power of the attitude recognition system. The attitude classifier trained on prosodic features achieved a promising accuracy result, with 65% accurate classification of 5 attitude labels for the first experiment while 61% accurate prediction of 5 attitude labels for the second experiment. Both results show higher rates than the baseline threshold (20%), which means that the classifier did not simply make a random guess in the classification task. Automatic assignment of attitudes is performed through a learned task of the input feature vectors. It is therefore worthy to state that automatic classification of attitudes based on non-verbal information is relatively successful. Although visual information too performed better than above-chance level, exceeding the baseline threshold of 20%, it could give better predictive power to the classifier. The aim of this experiment is to explore other methods of visual data processing to improve the predictive power of the classifier.

In this stage of analysis, the collection of videos is increased from 134 videos to 250 videos from 10 vlog speakers. In this experiment, similar techniques are again used from Experiment 1 and 2, to annotate and segment these videos according to the five attitude categories. Based on annotation and segmentation, 513 instances of attitude labels are collected. Table 5.8 summarises total number of instances for each attitude category:

Attitude Label	**No. of Instances**
Amusement	100
Enthusiasm	107
Friendliness	101
Frustration	103
Impatience	102
TOTAL	**513**

Table 5.8: Total instances for each attitude label

Table 5.8 indicates a total number of attitude instances for each attitude category. With a balanced and increased number of attitude instances, this can be helpful during the classification task by which the classifier is able to distinctively predict different attitude classes through trends derived from balanced data label instances. Following the collection of attitude labels, the classification task requires input from feature vectors.

The current experiment explores classification of prosodic and visual features using LibSVM, using a one vs. rest approach, with an extended dataset (see Table 5.8). Having observed that now the number of instances for each class is quite balanced, there is a reservation in the outcome of the result with regards to unseen data during the training process. As indicated in Table 3.6 (see Chapter 3), the number of speakers representing each attitude class is relatively imbalanced. Thus, when using 10-fold cross validation for evaluating the consistency of the classifier, the same speakers might appear in both training and testing sets, which could lead to errors in accuracy rate of the classifier. In order to clarify whether this imbalance might affect the classification result, a speaker-independent attitude classification analysis was conducted with LibSVM using the process of leave-one group-out 10-fold cross validation. Results from this analysis are indicated in Table5.9:

Features	**Prediction Rate**
PROSODIC	52.64%
VISUAL	31.42%

Table 5.9: Performance of Speaker-Independent Classification Task

Results showed that classification of attitudes based on prosodic features indicate 52.6%, while classification based on visual features totals to 31.4%. This result is quite low for both modalities, in comparison to the classification rate for experiment 2. Despite the low accuracy rate, the performance of the classifier exceeds the discriminate threshold, hence this implies that the classifier was able to a certain degree, discriminate between the five attitude classes, based on the audio and visual features. A factor for such low prediction of the classifier is the imbalance of speaker representation for each attitude class. The prediction accuracy of the classifier can be improved in future research, by re-sampling the data to obtain a more balanced and equal number of speakers per class in the dataset.

After analysing classification results that are speaker independent, this experiment extends further to find improvements in the feature vectors for the training of the classifier. As mentioned from past experiments, prosodic features performed well in the attitude classification tasks. Similar to Experiment 1 and 2, the use of prosodic features are maintained for this experiment. Classification was conducted using Random Forest and LibSVM. As observed in the previous experiments, LibSVM was selected as the classifier, but the present work explores another classifier called Random Forest (RF) to examine whether improved classification results could be obtained.

As stated in Chapter 4, the use of a decision tree is one method to be used for data classification. RF is directly related to decision tree whereby RF is an extension of the decision tree classification, that combines a multitude of decision trees. Decision tree algorithms are typically used for classification as they are extremely straight-forward to build and provide fast results [178]. However, there are certain limitations in terms of optimising the predictive power of the classifier. Unseen data can easily be generalised through the pruning technique, which affects the accuracy rate of the classifier [178]. In order to overcome this limitation, RF is developed to construct a systematic build of multiple decision trees [178]. RF is developed by adding the discriminant function in its algorithm. This discriminant preserves the accuracy of the training set by combining multiple classifiers from the individual decision tree classifier.

The present experiment conducts a classification task using RF algorithms and the classifier obtained an accuracy of 54.78% for correct classification of five attitude classes. A detailed analysis of the classifier's accuracy is measured using the precision and recall percentage for each attitude label. Precision and recall analysis is represented through the average measure of the f-score. This is stated in Table 5.10:

Attitude	**Precision**	**Recall**	**F-score**
Amusement	0.40	0.42	0.41
Enthusiasm	0.43	0.43	0.43
Friendliness	0.76	0.70	**0.73**
Frustration	0.67	0.76	0.71
Impatience	0.48	0.43	0.45

Table 5.10: Precision and Recall per Attitude Class

Table 5.10 indicates the f-score of the classification task by means of prosodic features. It is observed that the classifier was able to correctly predict “Friendliness” and “Frustration” better than the other attitude categories. A confusion matrix analysis was conducted. Table 5.11 shows the confusion matrix to obtain a clearer picture of the performance of the classifier. This is shown in Table 5.11:

Classified \ Annotated	**Amusement**	**Enthusiasm**	**Friendliness**	**Frustration**	**Impatience**
Amusement	42	19	2	27	10
Enthusiasm	15	46	9	5	32
Friendliness	14	7	71	5	4
Frustration	18	1	4	78	2
Impatience	16	34	7	1	44

Table 5.11: Confusion Matrix of Attitudes

Table 5.11 shows the number of classifications and misclassifications made by the classifier. It is found that “Frustration” is the attitude category that is most correctly classified by the trained predictor. This result further supports the findings, for which are evident in Experiments 1 and 2, that prosodic information gives sufficient information to the classifier when predicting different attitude states of speakers.

After conducting the classification task using prosodic features, this study further explores techniques for processing visual data extracted from FaceSDK AAM facial tracking software. Experiment 2 of this research work involves visual feature processing by calculating the average movement of the facial regions from attitude segments in the vlog dataset. When conducting the classification task, the predictive power of the visual features is 29.2%. This predictive power, although slightly exceeds the above-chance level of 20%, is considered quite low. The following experiment in conducted by using information on the 67 visual data landmarks (see Appendix G for full list) and creating a subset of measurements to give an improved feature vector for the classifier to train on. The subset of features is obtained by calculating the values that have highest levels of variance, which provide the most useful information to the classifier. The following information in Table 5.12 describes a subset of the visual features:

Visual Features	Percentage
All facial landmarks	36.16%
Head and Eyebrow Movement	**38.07%**
Head Movement	35.9%
Head Direction	34.7%
ALL	33.59%

Table 5.12: Result of Visual Prediction

The subset features, shown in Table 5.12 are values that contain all 67 facial landmarks, values of the Head movement, Head and Eyebrow movement, Head direction and combination of all subset features. It is observed that training on all the subset of visual features improved the performance of the classifier, with a predictive rate of about 34% compared to the previous analysis with a relative accuracy rate of 29%. The subset feature that shows highest classification accuracy is the combination of Head and Eyebrow movement, with a total percentage of over 38%. It is also found that the addition of directionality of the visual data point as a subset feature improves the classification rate. Directionality is measured by calculating the horizontal and vertical motions of the data points. The previous work focused entirely on the absolute values of the difference in the movements of the visual data points, [V_1 - V_0] [179]. Thus, adding directionality (by calculating the difference of the absolute values) as a subset feature is one way of improving the classification system [171].

For ease of reference, Table 5.13 provides a summary of results obtained for each of the experiments conducted in this research:

Experiment	Features	Result
1	All prosodic features	61.85%
	Selected prosodic features	65.46%
2	All prosodic features	61.33%
	All facial features	29.33%
	Prosodic and facial features	60.93%
3	All prosodic features	52.64%
	All facial features (LibSVM)	31.42%
	All facial features (RF)	36.16%
	Head and eyebrow (RF)	38.07%
	Selected facial features (RF)	33.59%

Table 5.13: Summary of Experiments

5.5 Discussion

The development of an automatic attitude classification model requires a systematic classification process. Supervised machine learning is conducted using statistical classification algorithms. In order to develop a robust classification system, several components need particular attention. Accurate attitude labels for instance need to be identified and annotated according to the N5 attitude annotation scheme. Treating these annotated labels as the ground truth, a considerable amount of instances for the attitude labels is necessary to ensure they provide sufficient instances for machine learning and classification task. In retrospect to the previous experiments, we notice at each stage, attitude instances are increased to provide sufficient instances for machine learning purposes. Thus it is integral to note that the number of attitude instances has relevance in developing a functioning classification system for attitude recognition. This work steadily increases the amount of balanced data instances from 194 in Experiment 1, to 513 instances of five attitude labels in Experiment 3. Observation from this task found that increased accuracy levels of the classifiers may be influenced by the amount of data instances from the class labels. Another similar point worth noting is the imbalanced number of speakers represented in each attitude class. This imbalance, as displayed in Experiment 3, greatly influenced the performance of the classifier. Speaker-independence during the validation process is essential to address the issue of unseen data, which affects the predictability rate of the classifier. Hence, the number of speaker representation for each attitude class needs to be balanced so that speakers would not affect the cross-validation process of the classification task.

Another point of observation is that Prosody is a good source of information to recognise attitudes. Prosodic features in particular demonstrate sufficient information for the classifier to correctly identify and recognise different attitude states of speakers. Hence, findings from analyses agree with other literature [47],[66], [22], [48], [153] that suggests the relevance of prosody as a quintessential factor for recognising attitudinal states of speakers during speech. The prosodic features used for machine learning in Experiments 2 and 3 were from the whole list, not from Feature Set SEL 1 from Experiment 1 (indicated in Page 117). The reason for not using the same feature set for the subsequent experiments was that more data was added into each experiment, and this might change the prosodic feature contributions, hence it was a conscious decision to retain all the prosodic features. However, applying Feature Set SEL 1 as a representative subset of the prosodic features in Experiments 2 and 3 might yield more favourable outcome to the classification task. This work is however reserved for future research.

Non-verbal signals are a potential source of good information to improve the recognition of attitudes. However data processing using 67 visual data landmarks of the speaker's face proved to be a challenging task for this study. Measuring the average movement of all facial data points proved to have contributed to low predictive power of the classifier, which was not in line with what was expected. This challenging task of using AAM algorithm is addressed by Neti et al [86] where they also achieved limited success in doing classification when applying AAM. They further suggested using a 3D visual tracking model to obtain greater visual information. This suggestion is interesting to apply in potential future work. Tome et.al [158] also agree that the use of facial tracking tools for automatic face recognition may not always bring accurate results. Fast-paced movements, particularly inherent in real-world data, are one of the many challenges faced when attempting to obtain accurate automatic extractions of facial regions. Misalignments of the region to the face images could lead to low precision. Therefore, Tome et.al [158] proposed an alternative method to reach higher precision, which is by conducting manual tagging of the facial landmarks, and aligning this to automatic tagging. This method, however, needs careful consideration as high accuracy of localisation at a manual level is required.

Furthermore, Neti et al. [86] also suggest the method of making visually meaningful groupings to obtain better additional visual information. The present study agrees with this notion by exploring other suitable methods for visual data processing. A subset of visual features, which includes the average movement of all facial data points, only head movements, combination of head and eyebrow movement as well as directionality of the head are derived and measured. This approach achieves an improved predictive power of the visual features towards the classification task of attitudes. Although the predictive power of the visual features is not as powerful as the prosodic features, the prediction rate still exceeds the 20% confidence threshold, and improvements of the visual features' performance are also achieved through deeper visual processing and analysis. Thus, the overall performance of the automatic attitude classification system by means of combining a fusion of prosodic and visual features is quite reliable.

Another limitation of the visual data processing stage of experimentation is the incorrect choice, in hindsight, of measuring absolute mean values for the visual features. More valuable information is lost when only absolute mean values are measured. Tome et.al [158] applied the facial proportions rules to their automatic face recognition task. By applying this measure of proportionality and using the eye centers as reference point, more accurate and valuable information may be attained. This could be a good alternative to obtain more precise feature values of the facial regions.

The last stage of analysis conducted for building a classification system for automatic attitude recognition presents several approaches of data processing for training on visual feature vectors. Results show that the combination of head and eyebrow movements yield better results for influencing the predictive power of the classifier. This analysis suggests the importance of selecting the best features to be used as a feature vector. Feature reduction and selection is a crucial process in machine learning. This process is integral to prevent overfitting, which refers to an enormous amount of features dimensions incongruent with the number of data samples, that could disrupt the prediction task and ultimately gives inaccurate predictability rates. Conducting feature reduction prior to classification could generally improve the performance of the classifier in recognising attitudes.

Attitudes, as a part of affective states of humans are relatively challenging to computationally evaluate and recognise. D'Mello and Graesser [180] state that it is highly unlikely for affective recognisers to be perfectly functional under real-world conditions where several challenges persist. Hence, they believe that a moderate degree of recognition accuracy is deemed sufficient if the model framework is conducted correctly and appropriately. The present study, despite several challenges, considers this attitude recognition system as a computational framework that provides moderate and sufficient accuracy rate.

5.6 Conclusion

Chapter 5 describes at length the processes of developing an automatic attitude classification system. Several concepts involved in this development are introduced, such as supervised machine learning, statistical algorithms used for classification, labels as well as feature vectors used to train the classifier.

This chapter also elaborates on several analyses conducted to develop an automatic attitude classification system based on a fusion of non-verbal modalities, namely prosodic and visual features. Results from analyses show a reliable performance of the attitude classification system. Prosodic features in particular contributed well to the predictive power of the classifier. Visual features contributed less reliably to the classification task compared to the prosodic features. This is most probably a result of incorrect, in hindsight, feature extraction and selection of the visual features. A method of improving the selection of features was by creating a subset of visual features, which presented improvements to the predictive power of the classifier. Although results can be further improved, the development of this automatic classification system is still quite reliable given the subjectivity of attitudes themselves.

Shifting to the following chapter, the section highlights the overall conclusion of this dissertation, highlighting research contributions, its applications as well as suggestions for improvement in future work.

Chapter 6

Conclusion

6.1 Research Contributions

The main objective of this study is to create a collection of attitude categories in vlog speech for the purpose of information retrieval through development of an automatic attitude recognition system. Previous chapters elaborate on the processes involved in building this model.

Contribution 1 This research introduces a novel corpus of vlogs. The vlog corpus consists of a collection of 250 vlogs from 10 male American English speakers, extracted from YouTube. These vlogs are downloaded using a freeware tool and are annotated and segmented according to attitude categories. The vlog corpus consists of a total of 517 attitude segments.

Contribution 2 This research introduces an attitude annotation scheme for attitude transcription. An adaptation of a standard A10 attitude annotation scheme is developed: the N5 attitude annotation scheme, which is created using a subset of the A10 attitude annotation scheme. These attitudes are Amusement, Enthusiasm, Friendliness, Frustration and Impatience. Attitudes from the N5 annotation scheme are considered to be most representative of the vlog corpus.

Contribution 3 This research analyses attitudes prevalent in multimodal settings. Behavioural expressions such as attitude and affect are not only detected through verbal speech and semantic content, but they are recognised through several non-verbal modalities.This study proposes a combination of multimodal features in the machine-learning process of attitude classification. The use of several modalities, namely prosodic and visual features contribute to a reliable attitude classification system. Training of the classifier is conducted using 14 prosodic features and 67 visual features. In multimodal feature selection, this study agrees with previous literature on the significance of pitch of the voice in recognising different attitudes expressions of speakers [47]. Findings from this study suggests that Pitch is the main contributor to the attitude prediction model, while voice quality also plays a significant role in predicting the speaker's attitude expressions. Visual features, extracted from the AAM algorithm, such as eyebrows also play a significant role in recognising attitudes. Movements of eyebrows and the jaw contribute to the classification task in identifying different attitude classes.

Contribution 4 A combination of multimodal information facilitates the classification task in recognising attitudes. Several stages of experimentation are conducted to develop a reliable attitude classification model.

Stage 1 involves building an attitude classification model through means of prosody. By using Support Vector Machine (SVM) as the classifier, five attitudes are correctly recognised over 60% of the time by training the classifier with 16 prosodic features.

Stage 2 involves the inclusion of visual features to recognise the attitudes of vlog speakers. Training was conducted using 14 prosodic features and 65 visual parameters. Results of the classifier's performance maintained over 60% accuracy rate.

Stage 3 involves creation of a subset of the visual features which involves movement and direction of the visual landmarks. The use of this subset of visual features for the classification task resulted to an improved accuracy rate for the visual features, with a prediction rate of over 30% compared to just over 20% during the second experiment.

6.2 Applications

With relation to this study, there are several applications of attitude recognition in several areas. One contribution of this research involves a collection of multimodal attitudinal states of vloggers. This corpus collection is beneficial to create a metadata involving attitudinal states of speakers in spontaneous speech. To develop a standard for multimodal signal processing and retrieval, this corpus can be applied as a training set in the recognition task. Automatic Speech Recognisers (ASR) for example can benefit from this dataset. Recognising different attitudes of speakers through prosodic information is interesting to further enhance systems in speech technology. In fact, this technology is not only beneficial for speech, but also for research in visual recognition systems. This dataset is helpful for detecting attitudes through facial movements. Hence, this vlog corpus can be applied and utilised for retrieval of speech and facial information.

Technology of attitude recognition is applicable to be integrated in training and security. Integrating a system for recognition of attitudes is useful for training purposes as it enables information processing for security reasons. For instance, organisations such as the National Security Act (NSA) can benefit from this system to investigate people's attitude expressions and derive conclusions on their behaviour.

Automatic attitude recognition system brings benefit for research in Human Computer Interaction (HCI). Producing socially intelligent robots as artificial communicative agents is largely dependent on robust recognition systems. With further improvements and optimisation, this attitude recognition system is useful to be integrated into embodied communicative agents. Agents such as robots and avatars can perform the task of identifying how humans communicate their attitudinal states. To communicate with humans, robots need to firstly understand the dynamics of human-human communication. Integration of this system into robots allows them to interpret signals that humans use. Socially intelligent robots like Pepper, the robot that could read emotions [181] and Paro, the therapeutic baby seal robot that encourages social interaction between the elderly [182], are some of the state-of-the-art technologies in Artificial Intelligence that utilise affective behaviour (including emotions and attitudes). Hence, this source of communicative content is useful for social robots to better understand people and produce appropriate responses.

Affective recognition systems can also be applied to educational fields of teaching and learning. Knowledge of students' attitudinal states is beneficial for teachers to devise learning strategies in classrooms. Teachers should be conscious of students' attitudes during lessons. When teachers realise students portrayal of the example expressions of enthusiasm or frustration in class, teachers can respond appropriately and change strategies to capture their students' attention. Analysis of student learning behaviour and motivation is typically measured through questionnaires and survey. This method is useful but it may cause interruptions during the lessons [183]. This drawback can potentially be solved by conducting automatic analysis of student's attitude states. This approach can assist teachers in formulating suitable teaching and learning strategies in classroom settings.

6.3 Research Constraints

This dissertation presents a detailed description of the processes involved in developing an automatic attitude recognition system by means of prosodic and visual feature selections. Although findings from analysis suggest a reliable performance of the classification system in detecting attitudes, the researcher identifies several limitations of this study.

In building an attitude recognition system, a solid scheme of attitude annotation should be developed. In this study, the researcher developed a subset of attitude categories based on a standard A10 attitude annotation scheme [47]. Although the level of agreement between expert annotators was high (k=0.75), agreement between a larger public opinion was low. This is probably due to the fact that different people perceive attitudes differently. Hence, differences in public perception may contribute to such low agreement of attitude perception. Apart from that, the research design for the perception test can be further improved, particularly with regards to the attitudes from the "Other" category. The design was created so the attitudes are somewhat hidden in the drop-down menu. Hence, this menu should have been discarded and instead, all the attitudes should have been presented to the participants.

Another research constraint in building a robust attitude recognition system is the limited number of features to train from. For instance, using 14 prosodic features may not provide sufficient information for the classifier to learn from. Aspects of vocal energy and syllabic duration were not considered. Visual features, on the other hand were quite elaborate with 67 features. However, only facial aspects of the speakers were considered. Other visual gestures such as hand movements and shoulder movements were not further explored nor included as part of the visual characteristics worth considering as visual representation of speakers in expressing different attitudes.

One main process of developing a robust classification model is the data preparation stage. From the experiments conducted in this thesis, the preparation of data could have been conducted in a more detailed manner. For instance, there is an imbalance of speaker representations for each attitude segment. Some speakers over-represent a certain attitude while others were under-represented. This affected the validation process and predictability outcome of the classification system.

Another integral process in machine learning and development of a reliable classification system is feature selection. Data processing techniques are used to select the best features for training of the classifiers. This stage is crucial to achieve a high classification performance. In this study, data processing, especially for the visual parameters, is still in the preliminary stage. The decision, incorrect in hindsight, of processing and extracting absolute mean values of visual data affected the classification task. This is noticeable in the first experiment, where low results were obtained with just over 20% accuracy rate of attitude prediction. However, improving on the data processing technique by developing a subset of visual parameters in the third experiment resulted in an increased accuracy rate of 30%. This result, although promising, is still unsatisfactory to create a robust attitude recognition model based on visual features.

6.4 Future Work

This research highlights development of an automatic classification system of attitude expressions. Results from analyses found that the performance of the classification system is reliable but robustness can also be improved by the following suggestions:

1. **Attitude labels in the N5 attitude annotation scheme**. There is a need for deeper understanding of the relationships between each attitude category. Looking at the five attitude classes in vlogger expression, there may be similarities and differences between each attitude. For example, "Amusement" contains broader meaning in terms of the semantics behind appearing amused. Speakers may express low and high levels of amusement. A high level of amusement may also indicate expression of "Enthusiasm", which stands as a category of its own in the N5 attitude annotation scheme. Hence, a cluster analysis or the creation of confusion matrices of the attitude categories, and also a semantics analysis of these attitudes can be conducted to make further conclusions on the N5 attitude classes in this annotation scheme.

2. **Modes of multimodal feature extraction and selection**. One of this study's constraints is the limited amount of features used to train on the attitude classification system. To improve on this limitation, a larger number of prosodic features that provides valuable information for machine-learning can be used by utilising a software called OpenSmile. This software is valuable for prosodic feature extraction and selection as it contains multiple numbers of prosodic features, ranging from pitch contour, energy, syllabic duration to other speech information for speech analysis. Providing a comprehensive feature for explaining speech activity of speakers is beneficial for a more robust attitude classification system. For visual features in this attitude classification model, the constraint in visual feature extraction is the limited knowledge in processing raw visual data. One possible solution to overcome this limitation is to use measurements in Active Shape Model (ASM) to measure the shape of the visual data points extracted from the FaceSDK AAM algorithm. Clustering facial dots to create shapes of the lips, for instance, could be included as a subset of visual features

to further understand the visual movements of the speakers when expressing different attitudes. Another alternative to processing raw visual data is by conducting proportionality rules of the facial region. By focusing on a selected number of visual regions, the proportions of these regions can be calculated to represent the movement activity of the facial contour.

3. **Machine-learning techniques**. To develop a robust attitude classification system using machine-learning techniques, the choice of the classifier is integral. Results from analyses suggest that different classifiers performed differently. Training with Support Vector Machine (SVM) performed better when using the prosodic features. Classification using Random Forest (RF) provided better results when identifying attitudes based on visual features compared to using SVM. It is helpful to see comparisons in terms of classifier performance as different classifiers may present different classification results, depending on the dataset presented for the learning task. That said, the classification task would be able to yield greater results given the right choice of features presented in the dataset. Future work involves the exploration of other feature processing and selection techniques to deduce an improved set of multimodal features. In extension to that, the use of different classifiers in future analyses might be helpful to obtain an improved accuracy rate of speakers' attitude expression.

Appendix A

Full List of Videos

Name of vlogger	**Channel URL**	**Name of video**	**Length of video**	**Video No.**
Niga Higa	https://www.youtube.com/user/nigahiga	Off the Pill - 2009	03:00	V001
Niga Higa	https://www.youtube.com/user/nigahiga	Off the Pill - Arrogant People	02:06	V002
Niga Higa	https://www.youtube.com/user/nigahiga	Off the Pill - Bieber Fever	02:16	V003
Niga Higa	https://www.youtube.com/user/nigahiga	Censorship Makes No Sense	03:21	V004
Niga Higa	https://www.youtube.com/user/nigahiga	Off the Pill - Christmas Spirit	03:27	V005
Niga Higa	https://www.youtube.com/user/nigahiga	Off the Pill - Dancing	02:41	V006
Niga Higa	https://www.youtube.com/user/nigahiga	Do You Love Animals	02:57	V007
Niga Higa	https://www.youtube.com/user/nigahiga	Expectations vs Reality: Romance	04:57	V008
Niga Higa	https://www.youtube.com/user/nigahiga	Famous Lazy People	03:43	V009
Niga Higa	https://www.youtube.com/user/nigahiga	Off the Pill - Farts	03:12	V010
Niga Higa	https://www.youtube.com/user/nigahiga	Off the Pill - Feminist	02:25	V011
Niga Higa	https://www.youtube.com/user/nigahiga	F*** The Police!	04:11	V012

Niga Higa	https://www.youtube.com/user/nigahiga	Immature Guys	07:21	V013
Niga Higa	https://www.youtube.com/user/nigahiga	Off the Pill - Judgement	02:59	V014
Niga Higa	https://www.youtube.com/user/nigahiga	Legally Blind	03:27	V015
Niga Higa	https://www.youtube.com/user/nigahiga	Milking!?	01:42	V016
Niga Higa	https://www.youtube.com/user/nigahiga	Most Annoying People On The Internet	04:50	V017
Niga Higa	https://www.youtube.com/user/nigahiga	Rant On Music (Remade)	02:47	V018
Niga Higa	https://www.youtube.com/user/nigahiga	New Puppy	03:02	V019
Niga Higa	https://www.youtube.com/user/nigahiga	Off the Pill - Nosy People	03:02	V020
Niga Higa	https://www.youtube.com/user/nigahiga	Off the Pill - Rebecca Black (Friday)	02:29	V021
Niga Higa	https://www.youtube.com/user/nigahiga	Off the Pill - Stink People	03:01	V022
Niga Higa	https://www.youtube.com/user/nigahiga	The Worst Boyfriend Ever	04:13	V023
Niga Higa	https://www.youtube.com/user/nigahiga	VIOLENCE!	03:55	V024
Niga Higa	https://www.youtube.com/user/nigahiga	Off the Pill - The Olympics	03:48	V025
Niga Higa	https://www.youtube.com/user/nigahiga	Off the Pill - Weird People	02:28	V026
KevJumba	https://www.youtube.com/user/kevjumba	Asians Aren’t Short!	03:16	V027
KevJumba	https://www.youtube.com/user/kevjumba	Asians Just Aren’t Cool Enough?	02:52	V028
KevJumba	https://www.youtube.com/user/kevjumba	AWKWARD	02:43	V029
KevJumba	https://www.youtube.com/user/kevjumba	BDAY	03:44	V030
KevJumba	https://www.youtube.com/user/kevjumba	BFF?	03:45	V031
KevJumba	https://www.youtube.com/user/kevjumba	Biggest Cockblock	03:06	V032
KevJumba	https://www.youtube.com/user/kevjumba	BULLY	02:32	V033
KevJumba	https://www.youtube.com/user/kevjumba	CAUGHT	02:48	V034

KevJumba	https://www.youtube.com/user/kevjumba	COLLEGE	02:27	V035
KevJumba	https://www.youtube.com/user/kevjumba	College Here I Come!	04:02	V036
KevJumba	https://www.youtube.com/user/kevjumba	Confidence	03:15	V037
KevJumba	https://www.youtube.com/user/kevjumba	DRUGS	03:18	V038
KevJumba	https://www.youtube.com/user/kevjumba	Elbow Zit	02:09	V039
KevJumba	https://www.youtube.com/user/kevjumba	Gifts from my Parents	02:46	V040
KevJumba	https://www.youtube.com/user/kevjumba	Girls are Like M&M's	03:26	V041
KevJumba	https://www.youtube.com/user/kevjumba	I'm not cool	03:11	V042
KevJumba	https://www.youtube.com/user/kevjumba	Living Alone	03:53	V043
KevJumba	https://www.youtube.com/user/kevjumba	Mama's Boy	03:12	V044
KevJumba	https://www.youtube.com/user/kevjumba	Most Exciting Job EVER!	04:35	V045
KevJumba	https://www.youtube.com/user/kevjumba	Mr. Lonely	02:46	V046
KevJumba	https://www.youtube.com/user/kevjumba	Mullet boy	01:52	V047
KevJumba	https://www.youtube.com/user/kevjumba	No More Mr. Nice Guy	02:11	V048
KevJumba	https://www.youtube.com/user/kevjumba	Real Men Trick-or-Treat	02:32	V049
KevJumba	https://www.youtube.com/user/kevjumba	Ridiculous Lyrics	03:57	V050
KevJumba	https://www.youtube.com/user/kevjumba	That's not gay!	02:04	V051
KevJumba	https://www.youtube.com/user/kevjumba	The Good and the Bad	05:01	V052
KevJumba	https://www.youtube.com/user/kevjumba	The Next Big Music Artist	04:06	V053
KevJumba	https://www.youtube.com/user/kevjumba	Totem Poles are Stupid	02:28	V054
KevJumba	https://www.youtube.com/user/kevjumba	What I Hate About College	03:44	V055
KevJumba	https://www.youtube.com/user/kevjumba	Wingman	02:41	V056

Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Twitter (*D) [1]	02:52	V057
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	10 Things I Love About Thanksgiving(*D)	01:56	V058
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	10 Things I Look For In A Woman (*D)	01:42	V059
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	69 Things I Hate(d) About High School	03:28	V060
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Boobs, Booze, Boo's (*D)	04:28	V061
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Quitting (*D)	01:55	V062
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Gifts for Father's Day (*D)	02:50	V063
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Second Channel (*D)	02:58	V064
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Things to do for Mother's Day (*D)	02:00	V065
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	How to Talk Trash (*D)	03:27	V066
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Recap of 2012 (*D)	01:56	V067
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	If I Were a Celebrity (*D)	02:25	V068
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Spring Fashion (*D)	04:24	V069
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	My Childhood (*D)	04:08	V070
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Relationships (*D)	04:03	V071
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Mother's Day (*D)	01:42	V072

[1] Deleted

Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	5 Rules for Better Resolutions (*D)	03:57	V073
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Stories by the Sticks: Blind Date (*D)	04:55	V074
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Stories by the Sticks: Embarrassing Moment (*D)	05:06	V075
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Stories by the Sticks: Worst Halloween (*D)	02:55	V076
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Internet Icon (*D)	06:53	V077
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Stories by the Sticks: St. Patrick's Day(*D)	02:47	V078
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Video Haul (*D)	03:40	V079
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Birthday (*D)	03:26	V080
Joey Engelman	https://www.youtube.com/user/HashbrownLIVE	Anything but the Laughter!	01:46	V081
Joey Engelman	https://www.youtube.com/user/uncuthashbrown	BOOM, Things Happen	02:28	V082
Joey Engelman	https://www.youtube.com/user/uncuthashbrown	Cloud of Plastic	03:39	V083
Joey Engelman	https://www.youtube.com/user/uncuthashbrown	Hey, cool shirt Joey.	01:14	V084
Joey Engelman	https://www.youtube.com/user/uncuthashbrown	How to Hate Yourself	03:31	V085
Joey Engelman	https://www.youtube.com/user/uncuthashbrown	I Eat Bird Poop	03:32	V086
Joey Engelman	https://www.youtube.com/user/uncuthashbrown	I have superpowers.	03:42	V087
Joey Engelman	https://www.youtube.com/user/HashbrownLIVE	Mittens = Kittens(to me at least)	01:54	V088

Joey Engelman	https://www.youtube.com/user/uncuthashbrown	My Camera Gave Me a Haircut	02:43	V089
Joey Engelman	https://www.youtube.com/user/uncuthashbrown	My T-Shirt Adventure	03:32	V090
Joey Engelman	https://www.youtube.com/user/HashbrownLIVE	One Year Ballooniversary	01:11	V091
Joey Engelman	https://www.youtube.com/user/uncuthashbrown	Girst Fuess	02:44	V092
Joey Engelman	https://www.youtube.com/user/uncuthashbrown	Staring is Not Caring	02:00	V093
Joey Engelman	https://www.youtube.com/user/HashbrownLIVE	Thank You, Luke Skywalker!	01:22	V094
Joey Engelman	https://www.youtube.com/user/uncuthashbrown	The Flaw in the Grocery Gameplan	02:59	V095
Joey Engelman	https://www.youtube.com/user/uncuthashbrown	The Quiet Buffer	03:14	V096
Joey Engelman	https://www.youtube.com/user/uncuthashbrown	The World is Scary	03:08	V097
Joey Engelman	https://www.youtube.com/user/uncuthashbrown	When it all fades out.	04:26	V098
Joey Engelman	https://www.youtube.com/user/uncuthashbrown	YOLO!	02:29	V099
Joey Engelman	https://www.youtube.com/user/HashbrownLIVE	You See What I'm Getting At?!	01:00	V100
Niga Higa	https://www.youtube.com/user/nigahiga	My Addiction	06:38	V101
Niga Higa	https://www.youtube.com/user/nigahiga	Lessons to Learn From 2013	05:30	V102
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	CONFESSION (Unedited)	09:49	V103
Tyler Oakley	https://www.youtube.com/user/tyleroakley	Why 2011 Was Fricking Amazing	03:16	V104
Tyler Oakley	https://www.youtube.com/user/tyleroakley	I'm Gonna Kill Him	04:11	V105
Tyler Oakley	https://www.youtube.com/user/tyleroakley	Serving Sizes & Calorie Counts	04:23	V106

Tyler Oakley	https://www.youtube.com/user/tyleroakley	One Night Stand (*D)	04:06	V107
Niga Higa	https://www.youtube.com/user/nigahiga	How To Survive A Horror Movie!	07:55	V108
Tyler Oakley	https://www.youtube.com/user/tyleroakley	How I Met Chris Colfer	06:58	V109
Tyler Oakley	https://www.youtube.com/user/tyleroakley	I'M JOINING GLEE!?	02:54	V110
Tyler Oakley	https://www.youtube.com/user/tyleroakley	Why I Hate The Entire World (*D)	01:28	V111
Tyler Oakley	https://www.youtube.com/user/tyleroakley	HARRY STYLES RETWEETED ME—Tyler Oakley	03:24	V112
Joey Engelman	https://www.youtube.com/user/uncuthashbrown	A Voice	01:10	V113
Tyler Oakley	https://www.youtube.com/user/tyleroakley	12 Ways To Tell Someone to Sit On Your Face—Tyler Oakley	04:37	V114
David So	https://www.youtube.com/user/DavidSoComedy	VLOG 92: Fuck FLAPPY BIRDS!	6:11	V115
David So	https://www.youtube.com/user/DavidSoComedy	VLOG 91: Common Manners!	5:00	V116
David So	https://www.youtube.com/user/DavidSoComedy	VLOG 37: What women do that men hate...	4:06	V117
David So	https://www.youtube.com/user/DavidSoComedy	VLOG 26: MY TRAM!	4:42	V118
David So	https://www.youtube.com/user/DavidSoComedy	VLOG 63: Duke Racist Party (racist rager)	3:13	V119
David So	https://www.youtube.com/user/DavidSoComedy	VLOG 34: Memoirs of an Angry Fat Man...	3:42	V120
David So	https://www.youtube.com/user/DavidSoComedy	vlog 45: Friend Zone	3:35	V121

David So	https://www.youtube.com/user/DavidSoComedy	VLOG 49: Freshman Survival	4:00	V122
David So	https://www.youtube.com/user/DavidSoComedy	Funny Fat Fit: FUCK YOUR SCALE!	5:23	V123
David So	https://www.youtube.com/user/DavidSoComedy	Funny Fat Fit: I Hate Being Sore	5:58	V124
David So	https://www.youtube.com/user/DavidSoComedy	VLOG 70: I Hate Diets	3:41	V125
David So	https://www.youtube.com/user/DavidSoComedy	vlog 20: White people are NOT THAT racist!	3:50	V126
David So	https://www.youtube.com/user/DavidSoComedy	P.S.A : Stinky People	2:49	V127
Tyler Oakley	https://www.youtube.com/user/tyleroakley	Scratch & Sniff & Evil Cat Boxes	4:57	V128
Connor Franta	https://www.youtube.com/user/ConnorFranta	My First Time	5:09	V129
Connor Franta	https://www.youtube.com/user/ConnorFranta	Things Girls Should Know About Guys	3:56	V130
Connor Franta	https://www.youtube.com/user/ConnorFranta	10 Things Girls Hate That Guys Do	4:31	V131
Connor Franta	https://www.youtube.com/user/ConnorFranta	Things I Don’t Understand About Parents	4:44	V132
Connor Franta	https://www.youtube.com/user/ConnorFranta	Things Teachers Don’t Understand	2:10	V133
Connor Franta	https://www.youtube.com/user/ConnorFranta	HORRIBLE HABITS!	4:37	V134
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	10 RANDOM FACTS about ME	04:19	V135
Mikey Bolts	https://www.youtube.com/user/michaelbalalis	17 Random Accents — Mikey Bolts	2:35	V136
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	69 THINGS I Hate about People (*D)	03:33	V137

Shane Dawson	https://www.youtube.com/user/shane	A Message to Haters	9:16	V138
Shane Dawson	https://www.youtube.com/user/shane	ALL ABOUT THAT BASS!	5:51	V139
Shane Dawson	https://www.youtube.com/user/shane	AM I A VIRGIN?	6:44	V140
Shane Dawson	https://www.youtube.com/user/shane	EMBARRASSING MASTURBATION STORY!	6:37	V141
Mikey Bolts	https://www.youtube.com/user/michaelbalalis	Amanda Bynes: A Symbol of Hope	3:25	V142
Mikey Bolts	https://www.youtube.com/user/michaelbalalis	ANNOYING COUPLES	3:10	V143
Timothy DeLaGhetto	https://www.youtube.com/user/TimothyDeLaGhetto2	Annoying People I Hate	5:45	V144
Timothy DeLaGhetto	https://www.youtube.com/user/TimothyDeLaGhetto2	Annoying People I Hate #2	5:59	V145
Timothy DeLaGhetto	https://www.youtube.com/user/TimothyDeLaGhetto2	Asians In the LIBRARY?! perspective on UCLA Girl Alexandra Wallace	6:50	V146
Timothy DeLaGhetto	https://www.youtube.com/user/TimothyDeLaGhetto2	Be a Gentleman, Get the Booty	4:17	V147
Timothy DeLaGhetto	https://www.youtube.com/user/TimothyDeLaGhetto2	Bitches Be In DENIAL!	4:08	V148
KevJumba	https://www.youtube.com/user/kevjumba	Commitment Issues	03:20	V149
Shane Dawson	https://www.youtube.com/user/shane	DANCE RECESS!	2:45	V150
Timothy DeLaGhetto	https://www.youtube.com/user/TimothyDeLaGhetto2	Dear DeLaGhetto #48- The BEST Pickup Line!	9:14	V151
Niga Higa	https://www.youtube.com/user/nigahiga	Dear Ice Bucket Challenge Haters...	05:42	V152

Shane Dawson	https://www.youtube.com/user/shane	DEAR SUICIDAL TEENS...	6:04	V153
Shane Dawson	https://www.youtube.com/user/shane	DEPRESSION & CUTTING	7:12	V154
Mikey Bolts	https://www.youtube.com/user/michaelbalalis	Different Types of Girlfriends	3:38	V155
Niga Higa	https://www.youtube.com/user/nigahiga	EBOLO!	05:07	V156
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Valentine's Day Gift Ideas (*D)	03:21	V157
Tyler Oakley	https://www.youtube.com/user/tyleroakley	Fake Cops & Nude Beaches—Tyler Oakley	09:59	V158
David So	https://www.youtube.com/user/DavidSoComedy	FFF: Consistency Is Key!	5:08	V159
Mikey Bolts	https://www.youtube.com/user/michaelbalalis	FRIEND ZONE FELLAS	4:26	V160
Shane Dawson	https://www.youtube.com/user/shane	GET IN MY BED!	8:00	V161
Timothy DeLaGhetto	https://www.youtube.com/user/TimothyDeLaGhetto2	Girls Don't Make Sense! Part 2	6:02	V162
KevJumba	https://www.youtube.com/user/kevjumba	Growing Up [VLOG]	02:42	V163
Mikey Bolts	https://www.youtube.com/user/michaelbalalis	GUYS AT THE GYM	4:00	V164
Mikey Bolts	https://www.youtube.com/user/michaelbalalis	Herbert the Pervert Prank Calls CHUCK E. CHEESE'S!	2:50	V165
Mikey Bolts	https://www.youtube.com/user/michaelbalalis	Holy Cow.	3:32	V166
Mikey Bolts	https://www.youtube.com/user/michaelbalalis	How Girls Act Around Their Friends	3:17	V167
Mikey Bolts	https://www.youtube.com/user/michaelbalalis	How To Avoid Taking a Final	2:00	V168

Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	How To Be A Successful YouTuber (*D)	04:46	V169
Tyler Oakley	https://www.youtube.com/user/tyleroakley	How To Fix Your Bad Memory—Tyler Oakley	05:55	V170
Tyler Oakley	https://www.youtube.com/user/tyleroakley	How To Pick Up Girls (*D)	03:48	V171
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	How To Get Girls To Like You (*D)	03:28	V172
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	How To Have More Self-Confidence (*D)	03:18	V173
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	How To Kick Summer's Ass (*D)	04:04	V174
Niga Higa	https://www.youtube.com/user/nigahiga	How To Know If You Have ADHD	05:11	V175
Tyler Oakley	https://www.youtube.com/user/tyleroakley	HOW TO: Live The Dream	06:55	V176
Shane Dawson	https://www.youtube.com/user/shane	HOW TO *THROW SHADE*!	6:34	V177
Connor Franta	https://www.youtube.com/user/ConnorFranta	I've Been Lying To You	5:23	V178
Tyler Oakley	https://www.youtube.com/user/tyleroakley	I've Officially Gone Insane	05:14	V179
Connor Franta	https://www.youtube.com/user/ConnorFranta	I AM A HORRIBLE HUMAN BEING	7:16	V180
Connor Franta	https://www.youtube.com/user/ConnorFranta	I AM DEAD	6:30	V181
Mikey Bolts	https://www.youtube.com/user/michaelbalalis	I am Stewie Griffin — @MikeyBolts	1:49	V182
Connor Franta	https://www.youtube.com/user/ConnorFranta	I Can't Believe I Did That	5:14	V183
Shane Dawson	https://www.youtube.com/user/shane	I F**KING LOVE U!	4:12	V184

Shane Dawson	https://www.youtube.com/user/shane	I Have a Stalker!	6:46	V185
Shane Dawson	https://www.youtube.com/user/shane	I SLAY BIG BROTHER 16	14:27	V186
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Is Justin Bieber a Douche? (*D)	04:29	V187
Mikey Bolts	https://www.youtube.com/user/michaelbalalis	Impressions of All 44 U.S. Presidents — Mikey Bolts	5:47	V188
Timothy DeLaGhetto	https://www.youtube.com/user/TimothyDeLaGhetto2	Instagram Deleted My Account (*D)	5:12	V189
Niga Higa	https://www.youtube.com/user/nigahiga	Is Justin Bieber Racist?	05:24	V190
Shane Dawson	https://www.youtube.com/user/shane	Collabs with Chris and Jessica (*D)	5:50	V191
Niga Higa	https://www.youtube.com/user/nigahiga	Lazy Halloween Costume Ideas	04:52	V192
Mikey Bolts	https://www.youtube.com/user/michaelbalalis	LETS GET AWKWARD	5:22	V193
Connor Franta	https://www.youtube.com/user/ConnorFranta	Awkward Things My Body Does	5:12	V194
Tyler Oakley	https://www.youtube.com/user/tyleroakley	Meeting One Direction & Selling My Body—Tyler Oakley	06:08	V195
Shane Dawson	https://www.youtube.com/user/shane	My Apology (Blackface & Offensive Videos)	11:46	V196
Shane Dawson	https://www.youtube.com/user/shane	MY AWKWARD CHILDHOOD	6:32	V197
Shane Dawson	https://www.youtube.com/user/shane	My Body Dysmorphia Disorder	8:20	V198
Shane Dawson	https://www.youtube.com/user/shane	MY FIRST TIME!	7:31	V199

Mikey Bolts	https://www.youtube.com/user/michaelbalalis	MY FRIENDS ARE SLUT BUCKETS?!	3:06	V200
Shane Dawson	https://www.youtube.com/user/shane	Christmas Tag (*D)	5:44	V201
Mikey Bolts	https://www.youtube.com/user/michaelbalalis	OH BOY	5:09	V202
Shane Dawson	https://www.youtube.com/user/shane	QUESTIONS ABOUT MY SEX LIFE!	4:55	V203
Shane Dawson	https://www.youtube.com/user/shane	REACTING TO OLD INSTAGRAM PICS!	4:54	V204
Connor Franta	https://www.youtube.com/user/ConnorFranta	Reacting To Old Profile Photos	8:58	V205
Shane Dawson	https://www.youtube.com/user/shane	READING *STRIP CLUB* YELP REVIEWS	5:12	V206
Timothy DeLaGhetto	https://www.youtube.com/user/TimothyDeLaGhetto2	Selena Dumped Bieber! BEST DAY EVER!!!	5:11	V207
Shane Dawson	https://www.youtube.com/user/shane	SELENA GOMEZ - THE HEART WANTS WHAT IT WANTS	4:23	V208
Shane Dawson	https://www.youtube.com/user/shane	SO MANY TAMPONS! (FAN MAIL VLOG)	7:58	V209
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Stories By The Sticks: Bad Tinder Date (*D)	04:59	V210
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Stories By The Sticks: Elevator Fart (*D)	04:50	V211
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Sunday Social: Who's asking? (*D)	03:33	V212

Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Sunday Social: Hate Comments (*D)	07:51	V213
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Sunday Social: Nickelback (*D)	06:45	V214
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Sunday Social: Headboards (*D)	04:38	V215
Tyler Oakley	https://www.youtube.com/user/tyleroakley	SUPERNOTE 2012: Team Fangirl—Tyler Oakley	05:29	V216
Shane Dawson	https://www.youtube.com/user/shane	TEACHING YOU HOW TO FLIRT	5:59	V217
Mikey Bolts	https://www.youtube.com/user/michaelbalalis	The 2014 VMA's were ASS-TASTIC!	4:23	V218
Connor Franta	https://www.youtube.com/user/ConnorFranta	The Most Angry I've Ever Been	6:17	V219
Connor Franta	https://www.youtube.com/user/ConnorFranta	The Most Annoying Thing Ever	5:04	V220
Connor Franta	https://www.youtube.com/user/ConnorFranta	THE WORST LUCK	8:16	V221
Tyler Oakley	https://www.youtube.com/user/tyleroakley	Bedroom Makeover (*D)	05:09	V222
Connor Franta	https://www.youtube.com/user/ConnorFranta	Thirsty	7:30	V223
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Tips for VidCon (*D)	03:46	V224
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Top Ten Movies(*D)	05:56	V225
Niga Higa	https://www.youtube.com/user/nigahiga	Treat People Like They're Dyeing	04:00	V226
Niga Higa	https://www.youtube.com/user/nigahiga	Unpopular Opinion: Cyber Bullying	07:48	V227
Shane Dawson	https://www.youtube.com/user/shane	VIDCON VLOG!	6:03	V228
Shane Dawson	https://www.youtube.com/user/shane	VINE TRIED TO KILL ME!	6:40	V229

David So	https://www.youtube.com/user/DavidSoComedy	VLOG 105: Teenage Relationships	4:19	V230
David So	https://www.youtube.com/user/DavidSoComedy	VLOG 106: LET GO OF YOUR BAGGAGE!	4:09	V231
David So	https://www.youtube.com/user/DavidSoComedy	VLOG 111: Robin Williams - Comedy Is Pain	4:39	V232
David So	https://www.youtube.com/user/DavidSoComedy	VLOG 112: THINGS I HATE ABOUT COLLEGE!	6:20	V233
David So	https://www.youtube.com/user/DavidSoComedy	VLOG 113: ASIAN WOMEN ARE SUBMISSIVE!	4:36	V234
David So	https://www.youtube.com/user/DavidSoComedy	VLOG 119: I'm So Proud of You?	3:04	V235
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Europe Trip (*D)	03:07	V236
Shane Dawson	https://www.youtube.com/user/shane	We Broke Up...	3:52	V237
Mikey Bolts	https://www.youtube.com/user/michaelbalalis	WHAT GRINDS MY GEARS	4:03	V238
Tyler Oakley	https://www.youtube.com/user/tyleroakley	When Did You Lose Your Viginity?—Tyler Oakley	06:14	V239
Shane Dawson	https://www.youtube.com/user/shane	WHY AM I SO GAY?	5:30	V240
Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	Being an Adult Sucks (*D)	04:13	V241
Timothy DeLaGhetto	https://www.youtube.com/user/TimothyDeLaGhetto2	Why Girls Love A**Holes	5:07	V242
Shane Dawson	https://www.youtube.com/user/shane	WHY I LOVE MY GIRLFRIEND	4:00	V243

Connor Franta	https://www.youtube.com/user/ConnorFranta	How To Approach The Guy You Like	5:27	V244
Shane Dawson	https://www.youtube.com/user/shane	ANACONDA!	5:18	V245
Mikey Bolts	https://www.youtube.com/user/michaelbalalis	LETS GET AWKWARD 2	4:09	V246
David So	https://www.youtube.com/user/DavidSoComedy	VLOG 14: Kanye's Fashion Future	3:56	V247
Shane Dawson	https://www.youtube.com/user/shane	My Thoughts On Zayn Malik Leaving One Direction	5:08	V248
David So	https://www.youtube.com/user/DavidSoComedy	VLOG 129: New Years Bullshit!	3:33	V249
Tyler Oakley	https://www.youtube.com/user/tyleroakley	#SELFIE — Tyler Oakley	05:43	V250

Table A.1: List of videos

Appendix B

Full list of attitude segments

No.	Speaker	Video URL	Video	Start Time	End Time	Attitude Name
1	Niga Higa	https://www.youtube.com/watch?v=7sz5cI51enE	V002	0:58	0:59	Amusement
2	Niga Higa	https://www.youtube.com/watch?v=r56jqb-fWVM	V003	1:26	1:27	Amusement
3	Niga Higa	https://www.youtube.com/watch?v=qeFZUFX95kE	V004	2:59	3:00	Amusement
4	Niga Higa	https://www.youtube.com/watch?v=EvAJt3VM0uk	V008	0:20	0:21	Amusement
5	Niga Higa	https://www.youtube.com/watch?v=r56jqb-fWVM	V009	1:23	1:24	Amusement
6	Niga Higa	https://www.youtube.com/watch?v=5pZbpUy7i_s	V010	0:26	0:27	Amusement
7	Niga Higa	https://www.youtube.com/watch?v=OOXZb0P1tsA	V015	1:10	1:11	Amusement
8	Niga Higa	https://www.youtube.com/watch?v=Tkm5K2fu0AI	V017	1:02	1:03	Amusement
9	Niga Higa	https://www.youtube.com/watch?v=Tkm5K2fu0AI	V017	1:09	1:10	Amusement
10	Niga Higa	https://www.youtube.com/watch?v=r56jqb-fWVM	V021	1:04	1:05	Amusement
11	Niga Higa	https://www.youtube.com/watch?v=r56jqb-fWVM	V021	1:05	1:06	Amusement
12	Niga Higa	https://www.youtube.com/watch?v=r56jqb-fWVM	V021	1:30	1:32	Amusement
13	Niga Higa	https://www.youtube.com/watch?v=NTnG4OFaFYc	V023	0:21	0:22	Amusement
14	Niga Higa	https://www.youtube.com/watch?v=K1aLtgEjzPk	V026	0:08	0:09	Amusement
15	KevJumba	https://www.youtube.com/watch?v=aIrbxxsLgTk	V027	1:14	1:15	Amusement
16	KevJumba	https://www.youtube.com/watch?v=SAbJgXUM4o4	V028	1:32	1:33	Amusement
17	KevJumba	https://www.youtube.com/watch?v=evdljyJYIPQ	V050	3:29	3:30	Amusement
18	KevJumba	https://www.youtube.com/watch?v=9pG2HmQiB0U	V052	2:25	2:26	Amusement
19	KevJumba	https://www.youtube.com/watch?v=9pG2HmQiB0U	V052	2:48	2:49	Amusement
20	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V061	2:21	2:22	Amusement
21	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V064	0:21	0:22	Amusement
22	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V064	0:23	0:24	Amusement
23	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V068	0:25	0:27	Amusement
24	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V068	1:07	1:08	Amusement
25	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V069	0:34	0:35	Amusement
26	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V069	0:35	0:36	Amusement
27	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V069	2:48	2:50	Amusement
28	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V069	3:16	3:17	Amusement

29	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V069	3:18	3:20	Amusement
30	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V072	0:31	0:32	Amusement
31	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V072	0:37	0:38	Amusement
32	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V072	1:36	1:37	Amusement
33	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V074	2:43	2:44	Amusement
34	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V075	1:49	1:50	Amusement
35	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V075	1:50	1:51	Amusement
36	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V076	2:14	2:15	Amusement
37	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V076	2:22	2:23	Amusement
38	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V080	0:17	0:18	Amusement
39	Joey Engelman	https://www.youtube.com/watch?v=sU2e4Xeuqpw	V081	1:15	1:16	Amusement
40	Joey Engelman	https://www.youtube.com/watch?v=kM99wsh-4aM	V085	2:13	2:14	Amusement
41	Joey Engelman	https://www.youtube.com/watch?v=S44ZnSPS8g0	V086	1:15	1:16	Amusement
42	Niga Higa	https://www.youtube.com/watch?v=SuEeK-Z94qE	V101	4:38	4:39	Amusement
43	Niga Higa	https://www.youtube.com/watch?v=rITp0sCFNik	V102	0:08	0:09	Amusement
44	Niga Higa	https://www.youtube.com/watch?v=rITp0sCFNik	V102	0:48	0:49	Amusement
45	Niga Higa	https://www.youtube.com/watch?v=rITp0sCFNik	V102	2:30	2:31	Amusement
46	Niga Higa	https://www.youtube.com/watch?v=rITp0sCFNik	V102	2:50	2:51	Amusement
47	Justin James Hughes	https://www.youtube.com/watch?v=URhYdy0iUqg	V103	4:55	4:56	Amusement
48	Tyler Oakley	https://www.youtube.com/watch?v=v1nQuJUpkrU	V104	2:40	2:42	Amusement
49	Tyler Oakley	https://www.youtube.com/user/tyleroakley	V107	0:01	0:03	Amusement
50	Tyler Oakley	https://www.youtube.com/user/tyleroakley	V107	1:19	1:20	Amusement
51	Tyler Oakley	https://www.youtube.com/watch?v=SxeS3q32yow	V109	3:06	3:07	Amusement
52	Tyler Oakley	https://www.youtube.com/watch?v=SxeS3q32yow	V109	3:50	3:51	Amusement
53	Tyler Oakley	https://www.youtube.com/watch?v=SxeS3q32yow	V109	4:03	4:04	Amusement
54	Tyler Oakley	https://www.youtube.com/watch?v=SxeS3q32yow	V109	4:29	4:30	Amusement
55	Tyler Oakley	https://www.youtube.com/watch?v=SxeS3q32yow	V109	4:30	4:31	Amusement
56	Tyler Oakley	https://www.youtube.com/watch?v=4trF2NqAKaA	V110	1:45	1:46	Amusement
57	Tyler Oakley	https://www.youtube.com/watch?v=HwnVk9B22xI	V114	3:14	3:16	Amusement
58	Connor Franta	https://www.youtube.com/watch?v=DwdDJx9mhto	V132	1:35	1:36	Amusement
59	Mikey Bolts	https://www.youtube.com/watch?v=3L3qZLmYyc0	V136	2:02	2:04	Amusement
60	Mikey Bolts	https://www.youtube.com/watch?v=3L3qZLmYyc0	V136	2:06	2:08	Amusement
61	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V137	1:30	1:31	Amusement
62	Shane Dawson	https://www.youtube.com/watch?v=3hOCo2Rg8JE	V138	2:10	2:13	Amusement
63	Shane Dawson	https://www.youtube.com/watch?v=XyrxQ1ne5Jc	V140	0:59	1:02	Amusement
64	Shane Dawson	https://www.youtube.com/watch?v=XyrxQ1ne5Jc	V140	5:07	5:09	Amusement
65	Niga Higa	https://www.youtube.com/watch?v=yB8K7sVGYh4	V152	2:09	2:10	Amusement
66	Shane Dawson	https://www.youtube.com/watch?v=cAZRXDF-Gfk	V153	4:13	4:15	Amusement
67	Shane Dawson	https://www.youtube.com/watch?v=F0lD9OHaWLk	V154	0:29	0:30	Amusement
68	Shane Dawson	https://www.youtube.com/watch?v=F0lD9OHaWLk	V154	6:49	6:50	Amusement
69	Niga Higa	https://www.youtube.com/watch?v=AzrDFykzvZM	V156	0:07	0:08	Amusement
70	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V157	0:47	0:49	Amusement
71	KevJumba	https://www.youtube.com/watch?v=jjISCt0vwVM	V163	0:13	0:17	Amusement
72	Mikey Bolts	https://www.youtube.com/watch?v=xhj2C38WBKI	V167	3:05	3:07	Amusement

73	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V169	0:36	0:38	Amusement
74	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V172	1:09	1:12	Amusement
75	Connor Franta	https://www.youtube.com/watch?v=m-0Ywa_0scY	V181	0:03	0:05	Amusement
76	Shane Dawson	https://www.youtube.com/watch?v=tasZxBMuKWM	V185	4:18	4:21	Amusement
77	Shane Dawson	https://www.youtube.com/watch?v=5oxRxwJZaj4	V186	12:12	12:14	Amusement
78	Shane Dawson	https://www.youtube.com/watch?v=5oxRxwJZaj4	V186	13:42	13:45	Amusement
79	Niga Higa	https://www.youtube.com/watch?v=yu7nmRITpaM	V192	0:14	0:15	Amusement
80	Niga Higa	https://www.youtube.com/watch?v=yu7nmRITpaM	V192	3:45	3:46	Amusement
81	Shane Dawson	https://www.youtube.com/watch?v=4YMHyJrnoGo	V197	2:10	2:12	Amusement
82	Shane Dawson	https://www.youtube.com/watch?v=EPSxheNL2yk	V198	1:32	1:34	Amusement
83	Shane Dawson	https://www.youtube.com/watch?v=rEVk551L-9I	V199	1:00	1:02	Amusement
84	Shane Dawson	https://www.youtube.com/watch?v=rEVk551L-9I	V199	2:40	2:42	Amusement
85	Shane Dawson	https://www.youtube.com/user/shane	V201	2:11	2:14	Amusement
86	Connor Franta	https://www.youtube.com/watch?v=hx8nifx8m9Y	V205	0:08	0:10	Amusement
87	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V210	4:42	4:44	Amusement
88	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V213	0:04	0:06	Amusement
89	Connor Franta	https://www.youtube.com/watch?v=Ywf6XyQAUpw	V219	2:34	2:37	Amusement
90	Connor Franta	https://www.youtube.com/watch?v=Ywf6XyQAUpw	V219	3:57	3:59	Amusement
91	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V224	2:54	2:55	Amusement
92	Niga Higa	https://www.youtube.com/watch?v=ktWHlSRRbMY	V226	0:54	0:56	Amusement
93	Niga Higa	https://www.youtube.com/watch?v=SAWoGBuJMT0	V227	6:00	6:02	Amusement
94	Shane Dawson	https://www.youtube.com/watch?v=UWdMYKtVjfs	V237	0:20	0:22	Amusement
95	Shane Dawson	https://www.youtube.com/watch?v=UWdMYKtVjfs	V237	0:43	0:45	Amusement
96	Shane Dawson	https://www.youtube.com/watch?v=UWdMYKtVjfs	V237	2:37	2:39	Amusement
97	Shane Dawson	https://www.youtube.com/watch?v=UWdMYKtVjfs	V237	3:24	3:27	Amusement
98	Shane Dawson	https://www.youtube.com/watch?v=-XPXQpdwdvM	V240	4:51	4:53	Amusement
99	Shane Dawson	https://www.youtube.com/watch?v=-XPXQpdwdvM	V240	4:55	4:57	Amusement
100	Shane Dawson	https://www.youtube.com/watch?v=eoKr5CrvFxQ	V243	1:43	1:44	Amusement
101	Connor Franta	https://www.youtube.com/watch?v=nJPNGSR7LDo	V244	1:15	1:18	Amusement
102	Connor Franta	https://www.youtube.com/watch?v=nJPNGSR7LDo	V244	1:58	2:00	Amusement
103	Connor Franta	https://www.youtube.com/watch?v=nJPNGSR7LDo	V244	4:27	4:29	Amusement
1	Niga Higa	https://www.youtube.com/watch?v=gErOFu61v-A	V001	0:10	0:11	Enthusiasm
2	Niga Higa	https://www.youtube.com/watch?v=gErOFu61v-A	V001	1:31	1:32	Enthusiasm
3	Niga Higa	https://www.youtube.com/watch?v=gErOFu61v-A	V001	1:34	1:35	Enthusiasm
4	Niga Higa	https://www.youtube.com/watch?v=cKfiurUtglA	V005	0:02	0:04	Enthusiasm
5	Niga Higa	https://www.youtube.com/watch?v=cKfiurUtglA	V005	0:58	1:00	Enthusiasm
6	Niga Higa	https://www.youtube.com/watch?v=cKfiurUtglA	V005	1:34	1:35	Enthusiasm
7	Niga Higa	https://www.youtube.com/watch?v=-pszRSF4qwc	V006	0:59	1:00	Enthusiasm
8	Niga Higa	https://www.youtube.com/watch?v=OOXZb0P1tsA	V015	1:09	1:10	Enthusiasm
9	Niga Higa	https://www.youtube.com/watch?v=KwdYZ3upT4c	V021	0:49	0:50	Enthusiasm
10	Niga Higa	https://www.youtube.com/watch?v=K1aLtgEjzPk	V026	0:31	0:32	Enthusiasm
11	KevJumba	https://www.youtube.com/watch?v=Hk-VrU0FoKw	V030	2:10	2:13	Enthusiasm
12	KevJumba	https://www.youtube.com/watch?v=clXERkuQmXM	V037	0:52	0:53	Enthusiasm
13	KevJumba	https://www.youtube.com/watch?v=lxLjpl_Tp54	V043	0:30	0:31	Enthusiasm
14	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V064	1:08	1:09	Enthusiasm
15	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V074	3:18	3:20	Enthusiasm
16	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V076	1:27	1:28	Enthusiasm
17	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V077	0:27	0:28	Enthusiasm

18	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V079	2:04	2:06	Enthusiasm
19	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V080	1:09	1:10	Enthusiasm
20	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V080	1:47	1:48	Enthusiasm
21	Joey Engelman	https://www.youtube.com/watch?v=_Bm9XzFrs0I	V099	0:54	0:55	Enthusiasm
22	Niga Higa	https://www.youtube.com/watch?v=rITp0sCFNik	V102	1:38	1:40	Enthusiasm
23	Justin James Hughes	https://www.youtube.com/watch?v=URhYdy0iUqg	V103	0:29	0:31	Enthusiasm
24	Justin James Hughes	https://www.youtube.com/watch?v=URhYdy0iUqg	V103	9:34	9:35	Enthusiasm
25	Tyler Oakley	https://www.youtube.com/watch?v=v1nQuJUpkrU	V104	0:01	0:03	Enthusiasm
26	Tyler Oakley	https://www.youtube.com/watch?v=v1nQuJUpkrU	V104	0:03	0:07	Enthusiasm
27	Tyler Oakley	https://www.youtube.com/user/tyleroakley	V111	0:40	0:42	Enthusiasm
28	Tyler Oakley	https://www.youtube.com/watch?v=BEECmzn7kJg	V112	0:35	0:38	Enthusiasm
29	Tyler Oakley	https://www.youtube.com/watch?v=BEECmzn7kJg	V112	2:08	2:13	Enthusiasm
30	Tyler Oakley	https://www.youtube.com/watch?v=BEECmzn7kJg	V112	3:01	3:04	Enthusiasm
31	Joey Engelman	https://www.youtube.com/watch?v=we7jePRmLGA	V113	0:09	0:11	Enthusiasm
32	David So	https://www.youtube.com/watch?v=-D18CZYbrGc	V115	0:09	0:11	Enthusiasm
33	David So	https://www.youtube.com/watch?v=-D18CZYbrGc	V115	0:23	0:24	Enthusiasm
34	David So	https://www.youtube.com/watch?v=-D18CZYbrGc	V115	1:02	1:03	Enthusiasm
35	David So	https://www.youtube.com/watch?v=36_KN5tJWbI	V116	1:10	1:11	Enthusiasm
36	David So	https://www.youtube.com/watch?v=r3ufPb6AP6g	V117	0:32	0:33	Enthusiasm
37	David So	https://www.youtube.com/watch?v=T7e1jW9r2ps	V121	0:34	0:35	Enthusiasm
38	David So	https://www.youtube.com/watch?v=_f5XgiK1mNc	V124	0:50	0:52	Enthusiasm
39	David So	https://www.youtube.com/watch?v=cqJE1xgpbLM	V126	0:29	0:30	Enthusiasm
40	Tyler Oakley	https://www.youtube.com/watch?v=MwRI8i1QA6I	V128	2:15	2:17	Enthusiasm
41	Joey Engelman	https://www.youtube.com/watch?v=we7jePRmLGA	V129	0:09	0:11	Enthusiasm
42	Connor Franta	https://www.youtube.com/watch?v=V02wBggJiCs	V130	2:00	2:01	Enthusiasm
43	Connor Franta	https://www.youtube.com/watch?v=V02wBggJiCs	V130	3:31	3:32	Enthusiasm
44	Connor Franta	https://www.youtube.com/watch?v=DwdDJx9mhto	V132	3:40	3:41	Enthusiasm
45	Connor Franta	https://www.youtube.com/watch?v=-rlwrZ_Zg4E	V133	0:15	0:16	Enthusiasm
46	Connor Franta	https://www.youtube.com/watch?v=-rlwrZ_Zg4E	V133	0:50	0:51	Enthusiasm
47	Connor Franta	https://www.youtube.com/watch?v=-rlwrZ_Zg4E	V133	1:38	1:39	Enthusiasm
48	Connor Franta	https://www.youtube.com/watch?v=vJE5EMbKCTw	V134	0:11	0:12	Enthusiasm
49	Connor Franta	https://www.youtube.com/watch?v=vJE5EMbKCTw	V134	0:19	0:20	Enthusiasm
50	Connor Franta	https://www.youtube.com/watch?v=vJE5EMbKCTw	V134	3:34	3:35	Enthusiasm
51	Connor Franta	https://www.youtube.com/watch?v=vJE5EMbKCTw	V134	4:15	4:16	Enthusiasm
52	Shane Dawson	https://www.youtube.com/watch?v=u1o5-H0bYbw	V139	5:10	5:12	Enthusiasm
53	Mikey Bolts	https://www.youtube.com/watch?v=xR2L4loWPa8	V142	0:07	0:08	Enthusiasm
54	Timothy DeLaGhetto	https://www.youtube.com/watch?v=pY7n5mpG5mU	V145	2:15	2:17	Enthusiasm
55	Timothy DeLaGhetto	https://www.youtube.com/watch?v=pY7n5mpG5mU	V145	2:56	2:58	Enthusiasm
56	Timothy DeLaGhetto	https://www.youtube.com/watch?v=pY7n5mpG5mU	V145	3:25	3:28	Enthusiasm
57	Shane Dawson	https://www.youtube.com/watch?v=A2G1TKQ0o04	V161	1:33	1:35	Enthusiasm
58	Shane Dawson	https://www.youtube.com/watch?v=A2G1TKQ0o04	V161	1:35	1:37	Enthusiasm
59	Shane Dawson	https://www.youtube.com/watch?v=A2G1TKQ0o04	V161	2:43	2:45	Enthusiasm
60	KevJumba	https://www.youtube.com/watch?v=jjISCt0vwVM	V163	1:24	1:27	Enthusiasm
61	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V169	1:21	1:22	Enthusiasm
62	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V172	0:49	0:51	Enthusiasm
63	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V173	0:26	0:30	Enthusiasm

64	Shane Dawson	https://www.youtube.com/watch?v=jW75n5qoRz0	V177	0:59	1:01	Enthusiasm
65	Connor Franta	https://www.youtube.com/watch?v=ALspMqxZ9dM	V178	0:03	0:07	Enthusiasm
66	Tyler Oakley	https://www.youtube.com/watch?v=Sfr64ys3o0M	V179	0:05	0:08	Enthusiasm
67	Tyler Oakley	https://www.youtube.com/watch?v=Sfr64ys3o0M	V179	2:51	2:52	Enthusiasm
68	Connor Franta	https://www.youtube.com/watch?v=m-0Ywa_0scY	V181	0:32	0:34	Enthusiasm
69	Connor Franta	https://www.youtube.com/watch?v=uTOLEEVZYCM	V183	3:53	3:55	Enthusiasm
70	Shane Dawson	https://www.youtube.com/watch?v=Q5n7IscTXBY	V184	4:01	4:03	Enthusiasm
71	Shane Dawson	https://www.youtube.com/watch?v=tasZxBMuKWM	V185	1:05	1:07	Enthusiasm
72	Shane Dawson	https://www.youtube.com/watch?v=tasZxBMuKWM	V185	1:08	1:10	Enthusiasm
73	Shane Dawson	https://www.youtube.com/watch?v=tasZxBMuKWM	V185	1:11	1:13	Enthusiasm
74	Shane Dawson	https://www.youtube.com/watch?v=tasZxBMuKWM	V185	1:21	1:23	Enthusiasm
75	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V187	3:52	3:55	Enthusiasm
76	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V187	4:13	4:16	Enthusiasm
77	Shane Dawson	https://www.youtube.com/user/shane	V191	5:19	5:21	Enthusiasm
78	Shane Dawson	https://www.youtube.com/user/shane	V201	0:02	0:05	Enthusiasm
79	Mikey Bolts	https://www.youtube.com/watch?v=IitpMkLqJMs	V202	3:27	3:30	Enthusiasm
80	Shane Dawson	https://www.youtube.com/watch?v=42-6i8T5IVI	V203	0:02	0:03	Enthusiasm
81	Shane Dawson	https://www.youtube.com/watch?v=42-6i8T5IVI	V203	0:18	0:20	Enthusiasm
82	Shane Dawson	https://www.youtube.com/watch?v=42-6i8T5IVI	V203	0:65	0:66	Enthusiasm
83	Shane Dawson	https://www.youtube.com/watch?v=F9X9fbz_fzk	V206	3:34	3:36	Enthusiasm
84	Timothy DeLaGhetto	https://www.youtube.com/watch?v=ffbZy7PAreg	V207	0:00	0:08	Enthusiasm
85	Timothy DeLaGhetto	https://www.youtube.com/watch?v=ffbZy7PAreg	V207	0:20	0:25	Enthusiasm
86	Timothy DeLaGhetto	https://www.youtube.com/watch?v=ffbZy7PAreg	V207	4:35	4:37	Enthusiasm
87	Shane Dawson	https://www.youtube.com/watch?v=_B999RpnJ-g	V208	0:19	0:20	Enthusiasm
88	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V210	4:44	4:46	Enthusiasm
89	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V211	4:20	4:21	Enthusiasm
90	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V215	3:11	3:12	Enthusiasm
91	Shane Dawson	https://www.youtube.com/watch?v=Dgem-1aW85Y	V217	3:57	3:59	Enthusiasm
92	Connor Franta	https://www.youtube.com/watch?v=Ywf6XyQAUpw	V219	4:51	4:52	Enthusiasm
93	Connor Franta	https://www.youtube.com/watch?v=TIg0B_fcF40	V223	0:18	0:19	Enthusiasm
94	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V225	0:08	0:10	Enthusiasm
95	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V225	0:12	0:14	Enthusiasm
96	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V225	0:42	0:44	Enthusiasm
97	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V225	4:55	4:56	Enthusiasm
98	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V225	5:19	5:20	Enthusiasm
99	Shane Dawson	https://www.youtube.com/watch?v=q5ntOvbi9UU	V228	0:02	0:05	Enthusiasm
100	Shane Dawson	https://www.youtube.com/watch?v=N2i2f5C7g74	V229	0:02	0:03	Enthusiasm
101	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V236	2:34	2:36	Enthusiasm
102	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V236	2:39	2:41	Enthusiasm
103	Tyler Oakley	https://www.youtube.com/watch?v=ZKkCUnWzEpk	V239	0:004	0:07	Enthusiasm
104	Tyler Oakley	https://www.youtube.com/watch?v=ZKkCUnWzEpk	V239	0:17	0:19	Enthusiasm
105	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V241	3:43	3:45	Enthusiasm

106	Shane Dawson	https://www.youtube.com/watch?v=zf5M1nLmtc0	V245	0:41	0:44	Enthusiasm
107	Shane Dawson	https://www.youtube.com/watch?v=zf5M1nLmtc0	V245	4:49	4:52	Enthusiasm
1	Niga Higa	https://www.youtube.com/watch?v=7sz5cI51enE	V002	0:00	0:01	Friendliness
2	Niga Higa	https://www.youtube.com/watch?v=r56jqb-fWVM	V003	0:00	0:01	Friendliness
3	Niga Higa	https://www.youtube.com/watch?v=qeFZUFX95kE	V004	0:27	0:28	Friendliness
4	Niga Higa	https://www.youtube.com/watch?v=cKfiurUtglA	V005	0:00	0:01	Friendliness
5	Niga Higa	https://www.youtube.com/watch?v=-pszRSF4qwc	V006	0:08	0:09	Friendliness
6	Niga Higa	https://www.youtube.com/watch?v=vPLgar2zcI4	V007	0:00	0:01	Friendliness
7	Niga Higa	https://www.youtube.com/watch?v=EvAJt3VM0uk	V008	0:00	0:01	Friendliness
8	Niga Higa	https://www.youtube.com/watch?v=r56jqb-fWVM	V009	0:00	0:01	Friendliness
9	Niga Higa	https://www.youtube.com/watch?v=5pZbpUy7i_s	V010	0:00	0:01	Friendliness
10	Niga Higa	https://www.youtube.com/watch?v=V_jKNxM65Nw	V011	0:00	0:01	Friendliness
11	Niga Higa	https://www.youtube.com/watch?v=wg8FAoFa0gU	V012	0:00	0:01	Friendliness
12	Niga Higa	https://www.youtube.com/watch?v=H2PtjWEUdKw	V013	0:00	0:01	Friendliness
13	Niga Higa	https://www.youtube.com/watch?v=_oNyp4RCU-Q	V014	0:00	0:01	Friendliness
14	Niga Higa	https://www.youtube.com/watch?v=OOXZb0P1tsA	V015	0:00	0:01	Friendliness
15	Niga Higa	https://www.youtube.com/watch?v=BY-ot74cpZQ	V016	0:00	0:01	Friendliness
16	Niga Higa	https://www.youtube.com/watch?v=Tkm5K2fu0AI	V017	0:00	0:01	Friendliness
17	Niga Higa	https://www.youtube.com/watch?v=oRJvbEej_Is	V018	0:00	0:01	Friendliness
18	Niga Higa	https://www.youtube.com/watch?v=PimSVfMgplQ	V019	0:00	0:01	Friendliness
19	Niga Higa	https://www.youtube.com/watch?v=r56jqb-fWVM	V021	0:00	0:01	Friendliness
20	Niga Higa	https://www.youtube.com/watch?v=96c–mQ_rP4	V022	0:00	0:01	Friendliness
21	Niga Higa	https://www.youtube.com/watch?v=NTnG4OFaFYc	V023	0:00	0:01	Friendliness
22	Niga Higa	https://www.youtube.com/watch?v=dIa-M5qoOHE	V024	0:00	0:01	Friendliness
23	Niga Higa	https://www.youtube.com/watch?v=0GmfIsOLJ6k	V025	0:00	0:01	Friendliness
24	Niga Higa	https://www.youtube.com/watch?v=K1aLtgEjzPk	V026	0:00	0:01	Friendliness
25	KevJumba	https://www.youtube.com/watch?v=aIrbxxsLgTk	V027	0:12	0:13	Friendliness
26	KevJumba	https://www.youtube.com/watch?v=WgV89H4jevI	V029	0:16	0:17	Friendliness
27	KevJumba	https://www.youtube.com/watch?v=Hk-VrU0FoKw	V030	0:00	0:01	Friendliness
28	KevJumba	https://www.youtube.com/watch?v=hzuX95bPgv8	V031	0:30	0:31	Friendliness
29	KevJumba	https://www.youtube.com/watch?v=1qk23jdUT1g	V032	0:17	0:18	Friendliness
30	KevJumba	https://www.youtube.com/watch?v=q7ASGB-uWTk	V034	0:00	0:01	Friendliness
31	KevJumba	https://www.youtube.com/watch?v=MrXvK2BrT9A	V035	0:21	0:22	Friendliness
32	KevJumba	https://www.youtube.com/watch?v=clXERkuQmXM	V037	0:00	0:01	Friendliness
33	KevJumba	https://www.youtube.com/watch?v=kOvMLF650k0	V038	0:00	0:01	Friendliness
34	KevJumba	https://www.youtube.com/watch?v=EB7LWOwSPOs	V042	0:00	0:01	Friendliness
35	KevJumba	https://www.youtube.com/watch?v=mZBX9nU3DRk	V044	0:04	0:05	Friendliness
36	KevJumba	https://www.youtube.com/watch?v=UENgfA6N8nA	V045	0:02	0:03	Friendliness
37	KevJumba	https://www.youtube.com/watch?v=_gdCjQw6Lb8	V051	0:01	0:02	Friendliness
38	KevJumba	https://www.youtube.com/watch?v=RN4Zk7sUR_A	V054	0:14	0:15	Friendliness
39	KevJumba	https://www.youtube.com/watch?v=lon_IKOEXzA	V056	0:00	0:01	Friendliness
40	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V070	0:00	0:01	Friendliness
41	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V080	2:47	2:48	Friendliness
42	Joey Engelman	https://www.youtube.com/watch?v=kM99wsh-4aM	V085	0:00	0:01	Friendliness
43	Joey Engelman	https://www.youtube.com/watch?v=S44ZnSPS8g0	V086	0:00	0:01	Friendliness
44	Joey Engelman	https://www.youtube.com/watch?v=dwX9jwdLYk4	V087	0:00	0:01	Friendliness
45	Joey Engelman	https://www.youtube.com/watch?v=8oUHxT1nAds	V090	0:00	0:01	Friendliness
46	Joey Engelman	https://www.youtube.com/watch?v=YtGVLHCffMU	V091	0:00	0:01	Friendliness
47	Joey Engelman	https://www.youtube.com/watch?v=4ZPijNHGxtU	V092	0:00	0:01	Friendliness
48	Joey Engelman	https://www.youtube.com/watch?v=GBa-8fsdoZo	V097	0:00	0:01	Friendliness
49	Joey Engelman	https://www.youtube.com/watch?v=_Bm9XzFrs0I	V099	0:00	0:01	Friendliness
50	Joey Engelman	https://www.youtube.com/watch?v=HNnPnA-whEg	V100	0:00	0:01	Friendliness
51	Justin James Hughes	https://www.youtube.com/watch?v=b0dBDjprLFw	V135	0:05	0:06	Friendliness

52	Shane Dawson	https://www.youtube.com/watch?v=3hOCo2Rg8JE	V138	0:20	0:21	Friendliness
53	KevJumba	https://www.youtube.com/watch?v=JpWIjbcYemk	V149	3:19	3:20	Friendliness
54	Shane Dawson	https://www.youtube.com/watch?v=9nZ3O7sRu_M	V150	0:11	0:12	Friendliness
55	Niga Higa	https://www.youtube.com/watch?v=yB8K7sVGYh4	V152	0:00	0:01	Friendliness
56	Shane Dawson	https://www.youtube.com/watch?v=F0lD9OHaWLk	V154	0:03	0:05	Friendliness
57	Tyler Oakley	https://www.youtube.com/watch?v=KbnZLFz6c4I	V158	0:00	0:01	Friendliness
58	Mikey Bolts	https://www.youtube.com/watch?v=AQ4Mk_8kNeQ	V160	3:16	3:18	Friendliness
59	KevJumba	https://www.youtube.com/watch?v=jjISCt0vwVM	V163	0:03	0:04	Friendliness
60	Mikey Bolts	https://www.youtube.com/watch?v=AqZhIoLwjMY	V164	0:05	0:06	Friendliness
61	Mikey Bolts	https://www.youtube.com/watch?v=hHgPB2ROX7w	V165	0:00	0:01	Friendliness
62	Mikey Bolts	https://www.youtube.com/watch?v=6oBZUfR_P8Y	V168	0:00	0:01	Friendliness
63	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V169	0:03	0:05	Friendliness
64	Tyler Oakley	https://www.youtube.com/watch?v=QxBx7q3jQsg	V170	0:00	0:02	Friendliness
65	Tyler Oakley	https://www.youtube.com/user/tyleroakley	V171	0:00	0:02	Friendliness
66	Niga Higa	https://www.youtube.com/watch?v=5GBMS7WPFSs	V175	0:00	0:01	Friendliness
67	Tyler Oakley	https://www.youtube.com/watch?v=CZlgl2b6Lg8	V176	0:00	0:01	Friendliness
68	Tyler Oakley	https://www.youtube.com/watch?v=Sfr64ys3o0M	V179	0:00	0:02	Friendliness
69	Mikey Bolts	https://www.youtube.com/watch?v=3WNLOmb9eVo	V182	0:00	0:01	Friendliness
70	Shane Dawson	https://www.youtube.com/watch?v=Q5n7IscTXBY	V184	0:04	0:05	Friendliness
71	Shane Dawson	https://www.youtube.com/watch?v=tasZxBMuKWM	V185	0:19	0:20	Friendliness
72	Shane Dawson	https://www.youtube.com/watch?v=5oxRxwJZaj4	V186	0:04	0:06	Friendliness
73	Mikey Bolts	https://www.youtube.com/watch?v=AyirRfMKIE0	V188	0:00	0:01	Friendliness
74	Mikey Bolts	https://www.youtube.com/watch?v=AyirRfMKIE0	V188	5:09	5:11	Friendliness
75	Niga Higa	https://www.youtube.com/watch?v=yu7nmRITpaM	V192	0:00	0:01	Friendliness
76	Mikey Bolts	https://www.youtube.com/watch?v=Lqs5_8bBRZ0	V193	0:00	0:01	Friendliness
77	Mikey Bolts	https://www.youtube.com/watch?v=KNhuqN60_x8	V200	0:05	0:07	Friendliness
78	Shane Dawson	https://www.youtube.com/watch?v=42-6i8T5IVI	V203	0:02	0:03	Friendliness
79	Shane Dawson	https://www.youtube.com/watch?v=42-6i8T5IVI	V203	1:03	1:04	Friendliness
80	Shane Dawson	https://www.youtube.com/watch?v=lkvRrhA_yoA	V204	0:12	0:13	Friendliness
81	Shane Dawson	https://www.youtube.com/watch?v=OQspFNYUPIA	V209	0:28	0:29	Friendliness
82	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V211	0:07	0:08	Friendliness
83	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V215	0:00	0:01	Friendliness
84	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V215	0:13	0:14	Friendliness
85	Tyler Oakley	https://www.youtube.com/watch?v=Bk-CLMZJItI	V216	0:00	0:01	Friendliness
86	Mikey Bolts	https://www.youtube.com/watch?v=OplfGXwX5BY	V218	0:00	0:02	Friendliness
87	Connor Franta	https://www.youtube.com/watch?v=Ywf6XyQAUpw	V219	0:03	0:05	Friendliness
88	Connor Franta	https://www.youtube.com/watch?v=SIA-l3wtge0	V220	0:01	0:02	Friendliness
89	Connor Franta	https://www.youtube.com/watch?v=gwsX9e8uYU4	V221	0:16	0:18	Friendliness
90	Tyler Oakley	https://www.youtube.com/user/tyleroakley	V222	0:00	0:02	Friendliness
91	Niga Higa	https://www.youtube.com/watch?v=SAWoGBuJMT0	V227	0:00	0:01	Friendliness
92	David So	https://www.youtube.com/watch?v=WyV2U4IX1nA	V232	0:06	0:07	Friendliness
93	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V236	0:04	0:05	Friendliness
94	Mikey Bolts	https://www.youtube.com/watch?v=91yfKTmkxD8	V238	0:05	0:06	Friendliness
95	Shane Dawson	https://www.youtube.com/watch?v=-XPXQpdwdvM	V240	0:02	0:03	Friendliness
96	Timothy DeLaGhetto	https://www.youtube.com/watch?v=W8tvdiRPAA8	V242	0:00	0:03	Friendliness
97	David So	https://www.youtube.com/watch?v=XgVvvGCM37k	V247	0:12	0:14	Friendliness
98	Shane Dawson	https://www.youtube.com/watch?v=-WivPGPXA0s	V248	0:20	0:21	Friendliness
99	David So	https://www.youtube.com/watch?v=gh3CuDPDL1o	V249	0:13	0:15	Friendliness
100	Tyler Oakley	https://www.youtube.com/watch?v=QdKzaWmxCLI	V250	0:00	0:03	Friendliness
1	Niga Higa	https://www.youtube.com/watch?v=gErOFu61v-A	V001	0:58	0:59	Frustration
2	Niga Higa	https://www.youtube.com/watch?v=qeFZUFX95kE	V004	2:57	2:58	Frustration

3	Niga Higa	https://www.youtube.com/watch?v=r56jqb-fWVM	V009	2:37	2:38	Frustration
4	Niga Higa	https://www.youtube.com/watch?v=Tkm5K2fu0AI	V017	4:46	4:47	Frustration
5	Niga Higa	https://www.youtube.com/watch?v=NTnG4OFaFYc	V023	0:06	0:07	Frustration
6	KevJumba	https://www.youtube.com/watch?v=1qk23jdUT1g	V032	1:11	1:12	Frustration
7	KevJumba	https://www.youtube.com/watch?v=EB7LWOwSPOs	V042	0:08	0:09	Frustration
8	KevJumba	https://www.youtube.com/watch?v=s51eNYOqAaA	V048	0:25	0:26	Frustration
9	Justin James Hughes	https://www.youtube.com/watch?v=tsPLEvBbSwk	V060	2:24	2:26	Frustration
10	Justin James Hughes	https://www.youtube.com/watch?v=tsPLEvBbSwk	V060	2:41	2:42	Frustration
11	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V067	1:32	1:33	Frustration
12	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V070	0:21	0:23	Frustration
13	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V071	3:23	3:24	Frustration
14	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V074	3:58	4:00	Frustration
15	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V077	0:31	0:33	Frustration
16	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V077	0:40	0:42	Frustration
17	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V080	0:34	0:37	Frustration
18	Joey Engelman	https://www.youtube.com/watch?v=HNnPnA-whEg	V100	0:41	0:42	Frustration
19	Justin James Hughes	https://www.youtube.com/watch?v=URhYdy0iUqg	V103	0:18	0:20	Frustration
20	Justin James Hughes	https://www.youtube.com/watch?v=URhYdy0iUqg	V103	0:20	0:21	Frustration
21	Justin James Hughes	https://www.youtube.com/watch?v=URhYdy0iUqg	V103	0:31	0:48	Frustration
22	Justin James Hughes	https://www.youtube.com/watch?v=URhYdy0iUqg	V103	0:49	0:50	Frustration
23	Justin James Hughes	https://www.youtube.com/watch?v=URhYdy0iUqg	V103	1:15	1:16	Frustration
24	Justin James Hughes	https://www.youtube.com/watch?v=URhYdy0iUqg	V103	2:12	2:24	Frustration
25	Justin James Hughes	https://www.youtube.com/watch?v=URhYdy0iUqg	V103	2:48	2:50	Frustration
26	Justin James Hughes	https://www.youtube.com/watch?v=URhYdy0iUqg	V103	6:18	6:19	Frustration
27	Justin James Hughes	https://www.youtube.com/watch?v=URhYdy0iUqg	V103	7:55	7:56	Frustration
28	Tyler Oakley	https://www.youtube.com/user/tyleroakley	V107	0:34	0:36	Frustration
29	Tyler Oakley	https://www.youtube.com/user/tyleroakley	V107	2:33	2:34	Frustration
30	Niga Higa	https://www.youtube.com/watch?v=zAXIxlVjuA8	V108	7:35	7:36	Frustration
31	Niga Higa	https://www.youtube.com/watch?v=zAXIxlVjuA8	V108	7:43	7:45	Frustration
32	Tyler Oakley	https://www.youtube.com/watch?v=4trF2NqAKaA	V110	0:54	0:55	Frustration
33	Tyler Oakley	https://www.youtube.com/user/tyleroakley	V111	1:18	1:19	Frustration
34	David So	https://www.youtube.com/watch?v=r3ufPb6AP6g	V117	0:05	0:06	Frustration
35	David So	https://www.youtube.com/watch?v=r3ufPb6AP6g	V117	3:39	3:41	Frustration
36	David So	https://www.youtube.com/watch?v=_0pkkOkLQ5k	V118	3:11	3:12	Frustration
37	David So	https://www.youtube.com/watch?v=_0pkkOkLQ5k	V118	3:13	3:15	Frustration
38	David So	https://www.youtube.com/watch?v=_0pkkOkLQ5k	V118	3:43	3:44	Frustration
39	David So	https://www.youtube.com/watch?v=NNZwjifEiKg	V120	3:06	3:06	Frustration
40	David So	https://www.youtube.com/watch?v=T7e1jW9r2ps	V121	0:26	0:28	Frustration
41	David So	https://www.youtube.com/watch?v=M60RSBbYrHY	V122	3:13	3:14	Frustration
42	David So	https://www.youtube.com/watch?v=i0YoUSo1adM	V125	3:17	3:18	Frustration
43	Tyler Oakley	https://www.youtube.com/watch?v=MwRI8i1QA6I	V128	3:01	3:04	Frustration

44	Connor Franta	https://www.youtube.com/watch?v=ZoQTwC2W8Yc	V129	1:32	1:33	Frustration
45	Connor Franta	https://www.youtube.com/watch?v=DZoMbe1M2LA	V131	1:23	1:25	Frustration
46	Connor Franta	https://www.youtube.com/watch?v=DZoMbe1M2LA	V131	1:27	1:29	Frustration
47	Connor Franta	https://www.youtube.com/watch?v=DwdDJx9mhto	V132	1:13	1:14	Frustration
48	Connor Franta	https://www.youtube.com/watch?v=DwdDJx9mhto	V132	1:44	1:45	Frustration
59	Connor Franta	https://www.youtube.com/watch?v=DwdDJx9mhto	V132	2:34	2:35	Frustration
50	Connor Franta	https://www.youtube.com/watch?v=-rlwrZ_Zg4E	V133	1:36	1:38	Frustration
51	Justin James Hughes	https://www.youtube.com/watch?v=b0dBDjprLFw	V135	2:32	2:34	Frustration
52	Justin James Hughes	https://www.youtube.com/watch?v=b0dBDjprLFw	V135	3:36	3:37	Frustration
53	Justin James Hughes	https://www.youtube.com/watch?v=b0dBDjprLFw	V135	3:53	3:55	Frustration
54	Shane Dawson	https://www.youtube.com/watch?v=3hOCo2Rg8JE	V138	8:17	8:18	Frustration
55	Timothy DeLaGhetto	https://www.youtube.com/watch?v=pY7n5mpG5mU	V145	4:47	4:49	Frustration
56	Timothy DeLaGhetto	https://www.youtube.com/watch?v=GHyChAzTCTA	V146	3:28	3:31	Frustration
57	Timothy DeLaGhetto	https://www.youtube.com/watch?v=XMYq_YrFfv0	V148	2:40	2:42	Frustration
58	Shane Dawson	https://www.youtube.com/watch?v=cAZRXDF-Gfk	V153	5:34	5:36	Frustration
59	Shane Dawson	https://www.youtube.com/watch?v=cAZRXDF-Gfk	V153	5:43	5:47	Frustration
60	Shane Dawson	https://www.youtube.com/watch?v=F0lD9OHaWLk	V154	3:31	3:34	Frustration
61	David So	https://www.youtube.com/watch?v=f94BYrmNeGQ	V159	0:42	0:44	Frustration
62	David So	https://www.youtube.com/watch?v=f94BYrmNeGQ	V159	1:31	1:33	Frustration
63	Mikey Bolts	https://www.youtube.com/watch?v=K2FdwLwl37o	V166	0:04	0:05	Frustration
64	Shane Dawson	https://www.youtube.com/watch?v=3hOCo2Rg8JE	V166	0:05	0:09	Frustration
65	Mikey Bolts	https://www.youtube.com/watch?v=xhj2C38WBKI	V167	2:30	2:33	Frustration
66	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V173	2:17	2:19	Frustration
67	Connor Franta	https://www.youtube.com/watch?v=EEvAWhwloj8	V180	2:08	2:10	Frustration
68	Connor Franta	https://www.youtube.com/watch?v=m-0Ywa_0scY	V181	0:42	0:43	Frustration
69	Shane Dawson	https://www.youtube.com/watch?v=Q5n7IscTXBY	V184	3:04	3:08	Frustration
70	Shane Dawson	https://www.youtube.com/watch?v=5oxRxwJZaj4	V186	12:20	12:24	Frustration
71	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V187	2:33	2:35	Frustration
72	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V187	2:47	2:49	Frustration
73	Timothy DeLaGhetto	https://www.youtube.com/user/TimothyDeLaGhetto2	V189	5:00	5:03	Frustration
74	Niga Higa	https://www.youtube.com/watch?v=A-MFTuWcFNs	V190	4:53	4:54	Frustration
75	Mikey Bolts	https://www.youtube.com/watch?v=Lqs5_8bBRZ0	V193	3:55	3:56	Frustration
76	Connor Franta	https://www.youtube.com/watch?v=1QQHkuQcYfg	V194	0:00	0:05	Frustration
77	Shane Dawson	https://www.youtube.com/watch?v=3jZDt5zQWsE	V196	0:07	0:09	Frustration
78	Shane Dawson	https://www.youtube.com/watch?v=3jZDt5zQWsE	V196	0:29	0:31	Frustration
79	Shane Dawson	https://www.youtube.com/watch?v=3jZDt5zQWsE	V196	1:27	1:29	Frustration
80	Shane Dawson	https://www.youtube.com/watch?v=3jZDt5zQWsE	V196	2:05	2:07	Frustration
81	Shane Dawson	https://www.youtube.com/watch?v=3jZDt5zQWsE	V196	2:58	3:00	Frustration
82	Shane Dawson	https://www.youtube.com/watch?v=3jZDt5zQWsE	V196	3:26	3:27	Frustration
83	Shane Dawson	https://www.youtube.com/watch?v=3jZDt5zQWsE	V196	5:35	5:38	Frustration
84	Shane Dawson	https://www.youtube.com/watch?v=3hOCo2Rg8JE	V196	7:52	7:55	Frustration
85	Shane Dawson	https://www.youtube.com/watch?v=3jZDt5zQWsE	V196	9:40	9:43	Frustration
86	Shane Dawson	https://www.youtube.com/watch?v=EPSxheNL2yk	V198	0:50	0:57	Frustration
87	Shane Dawson	https://www.youtube.com/watch?v=EPSxheNL2yk	V198	1:16	1:19	Frustration
88	Shane Dawson	https://www.youtube.com/watch?v=EPSxheNL2yk	V198	1:24	1:28	Frustration
89	Shane Dawson	https://www.youtube.com/watch?v=EPSxheNL2yk	V198	1:52	1:54	Frustration
90	Shane Dawson	https://www.youtube.com/watch?v=EPSxheNL2yk	V198	3:16	3:19	Frustration

91	Shane Dawson	https://www.youtube.com/watch?v=EPSxheNL2yk	V198	4:49	4:53	Frustration
92	Shane Dawson	https://www.youtube.com/watch?v=EPSxheNL2yk	V198	5:05	5:10	Frustration
93	Shane Dawson	https://www.youtube.com/watch?v=EPSxheNL2yk	V198	5:42	5:43	Frustration
94	Shane Dawson	https://www.youtube.com/watch?v=EPSxheNL2yk	V198	6:55	6:58	Frustration
95	Connor Franta	https://www.youtube.com/watch?v=hx8nifx8m9Y	V205	0:37	0:38	Frustration
96	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V213	1:08	1:09	Frustration
97	Mikey Bolts	https://www.youtube.com/watch?v=OplfGXwX5BY	V218	2:22	2:24	Frustration
98	Connor Franta	https://www.youtube.com/watch?v=gwsX9e8uYU4	V221	0:01	0:05	Frustration
99	David So	https://www.youtube.com/watch?v=5Vz1yoCg9Cs	V230	1:16	1:17	Frustration
100	David So	https://www.youtube.com/watch?v=G_YKurWpDkk	V231	0:12	0:14	Frustration
101	Shane Dawson	https://www.youtube.com/watch?v=UWdMYKtVjfs	V237	1:41	1:43	Frustration
102	Shane Dawson	https://www.youtube.com/watch?v=UWdMYKtVjfs	V237	3:03	3:06	Frustration
103	Shane Dawson	https://www.youtube.com/watch?v=UWdMYKtVjfs	V237	3:10	3:11	Frustration
104	Mikey Bolts	https://www.youtube.com/watch?v=tdIT94HzHCw	V246	3:55	3:56	Frustration
1	Niga Higa	https://www.youtube.com/watch?v=gErOFu61v-A	V001	0:12	0:13	Impatience
2	Niga Higa	https://www.youtube.com/watch?v=qeFZUFX95kE	V004	3:14	3:16	Impatience
3	Niga Higa	https://www.youtube.com/watch?v=cKfiurUtglA	V005	1:59	2:00	Impatience
4	Niga Higa	https://www.youtube.com/watch?v=cKfiurUtglA	V005	2:00	2:01	Impatience
5	Niga Higa	https://www.youtube.com/watch?v=cKfiurUtglA	V005	2:01	2:02	Impatience
6	Niga Higa	https://www.youtube.com/watch?v=vPLgar2zcI4	V007	1:27	1:29	Impatience
7	Niga Higa	https://www.youtube.com/watch?v=r56jqb-fWVM	V009	0:19	0:20	Impatience
8	Niga Higa	https://www.youtube.com/watch?v=oRJvbEej_Is	V018	1:06	1:07	Impatience
9	Niga Higa	https://www.youtube.com/watch?v=PimSVfMgplQ	V019	0:41	0:42	Impatience
10	Niga Higa	https://www.youtube.com/watch?v=PimSVfMgplQ	V019	2:18	2:19	Impatience
11	Niga Higa	https://www.youtube.com/watch?v=PimSVfMgplQ	V019	2:33	2:35	Impatience
12	Niga Higa	https://www.youtube.com/watch?v=CbFr_fUdRLI	V020	1:18	1:20	Impatience
13	Niga Higa	https://www.youtube.com/watch?v=r56jqb-fWVM	V021	0:23	0:24	Impatience
14	Niga Higa	https://www.youtube.com/watch?v=r56jqb-fWVM	V021	1:10	1:12	Impatience
15	Niga Higa	https://www.youtube.com/watch?v=r56jqb-fWVM	V021	1:32	1:33	Impatience
16	Niga Higa	https://www.youtube.com/watch?v=r56jqb-fWVM	V021	1:35	1:36	Impatience
17	Niga Higa	https://www.youtube.com/watch?v=NTnG4OFaFYc	V023	2:14	2:16	Impatience
18	Niga Higa	https://www.youtube.com/watch?v=NTnG4OFaFYc	V023	2:17	2:18	Impatience
19	Niga Higa	https://www.youtube.com/watch?v=NTnG4OFaFYc	V023	2:19	2:21	Impatience
20	Niga Higa	https://www.youtube.com/watch?v=dIa-M5qoOHE	V024	0:30	0:31	Impatience
21	Niga Higa	https://www.youtube.com/watch?v=0GmfIsOLJ6k	V025	1:35	1:37	Impatience
22	Niga Higa	https://www.youtube.com/watch?v=0GmfIsOLJ6k	V025	3:04	3:06	Impatience
23	Niga Higa	https://www.youtube.com/watch?v=K1aLtgEjzPk	V026	1:57	1:58	Impatience
24	KevJumba	https://www.youtube.com/watch?v=Hk-VrU0FoKw	V030	0:44	0:46	Impatience
25	KevJumba	https://www.youtube.com/watch?v=q7ASGB-uWTk	V034	1:37	1:38	Impatience
26	KevJumba	https://www.youtube.com/watch?v=MrXvK2BrT9A	V035	0:39	0:40	Impatience
27	KevJumba	https://www.youtube.com/watch?v=MrXvK2BrT9A	V035	0:52	0:53	Impatience
28	KevJumba	https://www.youtube.com/watch?v=lxLjpl_Tp54	V043	2:05	2:06	Impatience
29	KevJumba	https://www.youtube.com/watch?v=RN4Zk7sUR_A	V054	1:30	1:31	Impatience
30	KevJumba	https://www.youtube.com/watch?v=RN4Zk7sUR_A	V054	2:00	2:03	Impatience
31	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V064	0:59	1:00	Impatience
32	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V075	4:24	4:26	Impatience
33	Joey Engelman	https://www.youtube.com/watch?v=sU2e4Xeuqpw	V081	0:38	0:39	Impatience
34	Joey Engelman	https://www.youtube.com/watch?v=GBa-8fsdoZo	V097	1:23	1:24	Impatience
35	Niga Higa	https://www.youtube.com/watch?v=SuEeK-Z94qE	V101	1:19	1:23	Impatience
36	Niga Higa	https://www.youtube.com/watch?v=zAXIxlVjuA8	V108	1:48	1:49	Impatience
37	David So	https://www.youtube.com/watch?v=-D18CZYbrGc	V115	0:54	0:55	Impatience
38	David So	https://www.youtube.com/watch?v=-D18CZYbrGc	V115	1:04	1:06	Impatience
39	David So	https://www.youtube.com/watch?v=-D18CZYbrGc	V115	1:07	1:08	Impatience

40	David So	https://www.youtube.com/watch?v=-D18CZYbrGc	V115	2:01	2:02	Impatience
41	David So	https://www.youtube.com/watch?v=36_KN5tJWbI	V116	3:44	3:45	Impatience
42	David So	https://www.youtube.com/watch?v=r3ufPb6AP6g	V117	0:48	0:50	Impatience
43	David So	https://www.youtube.com/watch?v=r3ufPb6AP6g	V117	2:41	2:43	Impatience
44	David So	https://www.youtube.com/watch?v=lzUQOtyQ7Ts	V119	0:30	0:31	Impatience
45	David So	https://www.youtube.com/watch?v=M60RSBbYrHY	V122	1:06	1:08	Impatience
46	David So	https://www.youtube.com/watch?v=M60RSBbYrHY	V122	1:27	1:29	Impatience
47	David So	https://www.youtube.com/watch?v=aeA-Fw7XYdk	V123	2:07	2:09	Impatience
48	David So	https://www.youtube.com/watch?v=aeA-Fw7XYdk	V123	3:11	3:12	Impatience
49	David So	https://www.youtube.com/watch?v=_f5XgiK1mNc	V124	0:36	0:37	Impatience
50	Justin James Hughes	https://www.youtube.com/watch?v=b0dBDjprLFw	V135	0:10	0:12	Impatience
51	Justin James Hughes	https://www.youtube.com/watch?v=b0dBDjprLFw	V135	2:50	2:52	Impatience
52	Shane Dawson	https://www.youtube.com/watch?v=XyrxQ1ne5Jc	V140	3:25	3:28	Impatience
53	Mikey Bolts	https://www.youtube.com/watch?v=ltRl19Uw_fE	V143	0:24	0:26	Impatience
54	Timothy DeLaGhetto	https://www.youtube.com/watch?v=8FNb1abOhF0	V144	1:01	1:04	Impatience
55	Timothy DeLaGhetto	https://www.youtube.com/watch?v=8FNb1abOhF0	V144	2:10	2:14	Impatience
56	Timothy DeLaGhetto	https://www.youtube.com/watch?v=8FNb1abOhF0	V144	2:40	2:43	Impatience
57	Timothy DeLaGhetto	https://www.youtube.com/watch?v=pY7n5mpG5mU	V145	0:17	0:19	Impatience
58	Timothy DeLaGhetto	https://www.youtube.com/watch?v=pY7n5mpG5mU	V145	0:42	0:46	Impatience
59	Timothy DeLaGhetto	https://www.youtube.com/watch?v=GHyChAzTCTA	V146	1:23	1:24	Impatience
60	Timothy DeLaGhetto	https://www.youtube.com/watch?v=GHyChAzTCTA	V146	6:02	6:05	Impatience
61	Timothy DeLaGhetto	https://www.youtube.com/watch?v=jbqW7SUjuT0	V147	0:10	0:14	Impatience
62	Timothy DeLaGhetto	https://www.youtube.com/watch?v=jbqW7SUjuT0	V147	2:27	2:30	Impatience
63	KevJumba	https://www.youtube.com/watch?v=JpWIjbcYemk	V149	2:50	2:51	Impatience
64	Timothy DeLaGhetto	https://www.youtube.com/watch?v=60r-5P7FI1M	V151	0:18	0:21	Impatience
65	Timothy DeLaGhetto	https://www.youtube.com/watch?v=60r-5P7FI1M	V151	5:45	5:47	Impatience
66	Mikey Bolts	https://www.youtube.com/watch?v=42XK4sxePEI	V155	1:00	1:02	Impatience
67	Mikey Bolts	https://www.youtube.com/watch?v=42XK4sxePEI	V155	2:44	2:45	Impatience
68	Timothy DeLaGhetto	https://www.youtube.com/watch?v=hbH0StiODcw	V162	2:18	2:20	Impatience
69	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V169	1:44	1:46	Impatience
70	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V174	0:17	0:19	Impatience
71	Connor Franta	https://www.youtube.com/watch?v=ALspMqxZ9dM	V178	0:12	0:13	Impatience
72	Shane Dawson	https://www.youtube.com/watch?v=tasZxBMuKWM	V185	3:14	3:16	Impatience
73	Shane Dawson	https://www.youtube.com/watch?v=5oxRxwJZaj4	V186	1:59	2:00	Impatience
74	Shane Dawson	https://www.youtube.com/watch?v=5oxRxwJZaj4	V186	2:00	2:01	Impatience
75	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V187	2:01	2:02	Impatience
76	Timothy DeLaGhetto	https://www.youtube.com/user/TimothyDeLaGhetto2	V189	1:27	1:29	Impatience
77	Tyler Oakley	https://www.youtube.com/watch?v=v_4vcKn_ASU	V195	0:19	0:20	Impatience
78	Tyler Oakley	https://www.youtube.com/watch?v=v_4vcKn_ASU	V195	1:06	1:07	Impatience
79	Shane Dawson	https://www.youtube.com/watch?v=rEVk551L-9I	V199	0:41	0:42	Impatience
80	Shane Dawson	https://www.youtube.com/watch?v=rEVk551L-9I	V199	2:18	2:19	Impatience

81	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V210	0:12	0:13	Impatience
82	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V211	3:14	3:16	Impatience
83	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V212	1:59	2:00	Impatience
84	Connor Franta	https://www.youtube.com/watch?v=Ywf6XyQAUpw	V219	2:00	2:01	Impatience
85	Connor Franta	https://www.youtube.com/watch?v=SIA-l3wtge0	V220	2:01	2:02	Impatience
86	Connor Franta	https://www.youtube.com/watch?v=SIA-l3wtge0	V220	1:27	1:29	Impatience
87	Connor Franta	https://www.youtube.com/watch?v=SIA-l3wtge0	V220	0:19	0:20	Impatience
88	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V225	1:06	1:07	Impatience
89	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V225	0:41	0:42	Impatience
90	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V225	2:18	2:19	Impatience
91	Niga Higa	https://www.youtube.com/watch?v=SAWoGBuJMT0	V227	0:12	0:13	Impatience
92	David So	https://www.youtube.com/watch?v=vp2yYgvDjuY	V233	3:14	3:16	Impatience
93	David So	https://www.youtube.com/watch?v=vp2yYgvDjuY	V233	1:59	2:00	Impatience
94	David So	https://www.youtube.com/watch?v=vp2yYgvDjuY	V233	2:00	2:01	Impatience
95	David So	https://www.youtube.com/watch?v=053XPAqVmiU	V234	2:01	2:02	Impatience
96	David So	https://www.youtube.com/watch?v=053XPAqVmiU	V234	1:27	1:29	Impatience
97	Shane Dawson	https://www.youtube.com/watch?v=-XPXQpdwdvM	V240	0:19	0:20	Impatience
98	Justin James Hughes	https://www.youtube.com/user/JustinJamesHughes	V241	1:06	1:07	Impatience
99	David So	https://www.youtube.com/watch?v=XgVvvGCM37k	V247	0:41	0:42	Impatience
100	Shane Dawson	https://www.youtube.com/watch?v=-WivPGPXA0s	V248	2:18	2:19	Impatience
101	Shane Dawson	https://www.youtube.com/watch?v=-WivPGPXA0s	V248	0:12	0:13	Impatience
102	Niga Higa	https://www.youtube.com/watch?v=qeFZUFX95kE	V249	3:14	3:16	Impatience
103	David So	https://www.youtube.com/watch?v=gh3CuDPDL1o	V249	1:59	2:00	Impatience

Table B.1: List of Attitude Segments

Appendix C

Guide to Creating a Vlog

STEP 1: PLANNING A SCRIPT

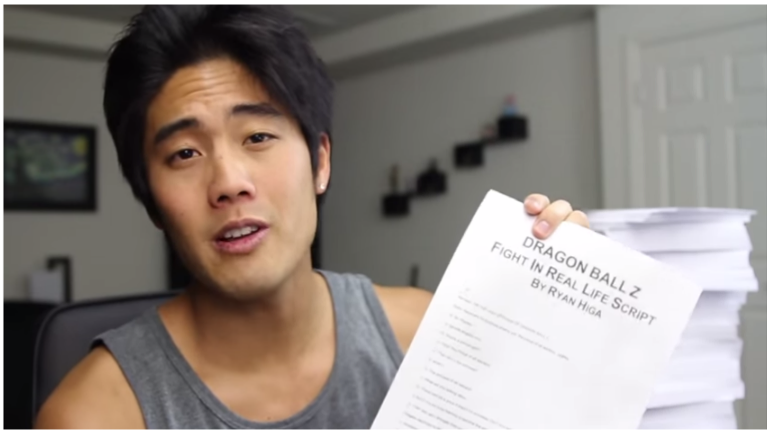


Figure C.1: Showing a vlog script

The first stage of creating a vlog is deciding on a topic and creating a script to help the speaker deliver the topic effectively. The content needs to be interesting and appealing to the viewers so a considerable amount of time needs to be dedicated to planning the best topic. Topics range from daily moods and events, recent happenings, thoughts on politics, religion, culture, lifestyle, fashion, celebrity news and others. Speakers need to select relevant topics and present themselves with dynamic emotions to capture viewer attention.

STEP 2: PREPARING NECESSARY EQUIPMENT

The second stage involves setting up proper equipment for filming. People use a variety of cameras for filming purposes, ranging from camera phones to high performance cameras such as Canon C300 or SONY NEX A5100. Other optional equipment for filming vlogs ranges from lightings to tripods

Figure C.2: Related equipment needed

as seen in the Figure C.2. The equipment is necessary to present the best filming quality where the speaker's face is clearly captured with proper lighting and camera angles.

STEP 3: FILMING

Figure C.3: Filming a vlog

The third stage is the actual recording of the vlog. This is the stage where the content of the video is recorded. The speaker sits or stands in front of the camera and talks about the topic being addressed. Filming time varies depending on the length of the script, speaking mistakes like saying the wrong words or unintentional laughing, and equipment failures. All raw footages are saved in the actual duration. Parts of the video that are irrelevant will be processed at the next stage.

STEP 4: EDITING

Figure C.4: Editing raw film footage

The fourth stage is the most crucial part of vlog production. The speaker begins to edit the raw videos by deleting irrelevant footage, adding text or music to enhance the effectiveness of the presentation, and other editing works. Most videos are shortened to 2-5 minutes. The length of the vlog and creative delivery methods are crucial so as to capture and maintain the interest of the viewers.

STEP 5: UPLOADING

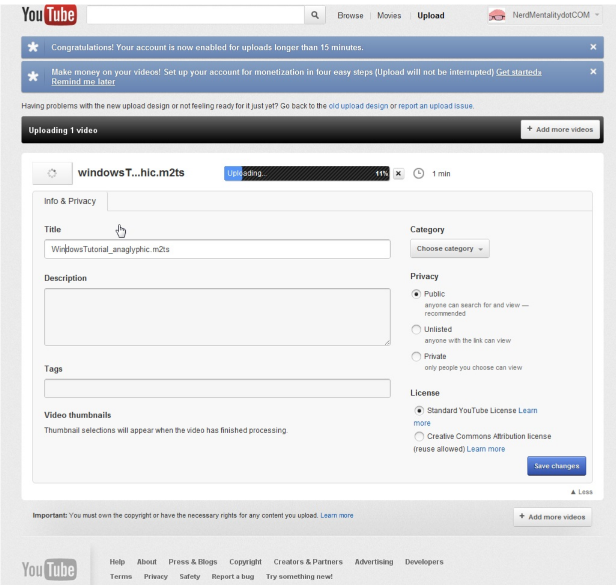


Figure C.5: Uploading video on YouTube

The fifth stage involves uploading the video content on a media website. As seen in the figure above, YouTube is the platform typically used by vloggers to publish their videos [184].

STEP 6: PUBLISHING

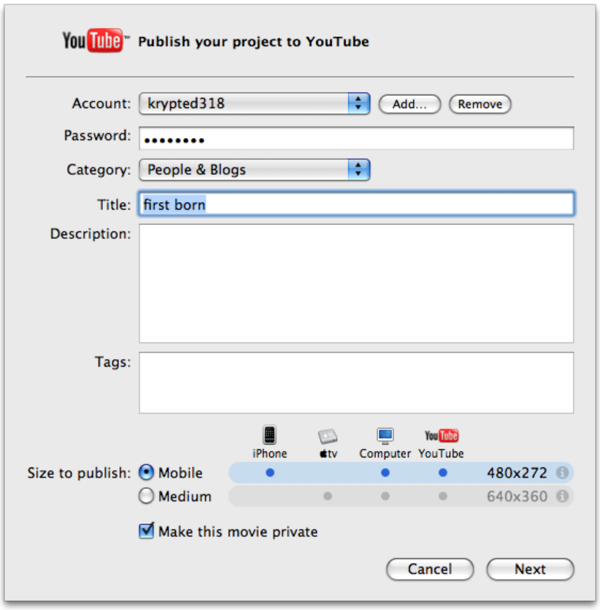


Figure C.6: Publishing video on YouTube

The final stage of vlog production is publishing the video on YouTube. After uploading video content on YouTube, users need to provide more information about the vlog. Relevant information such as subtitles and tags about the video is given for ease of video searching and viewing. Now the vlog is published on YouTube and is available for public viewing.

Appendix D

Annotation with Wavesurfer

STEP 1: OPEN WAV FILE

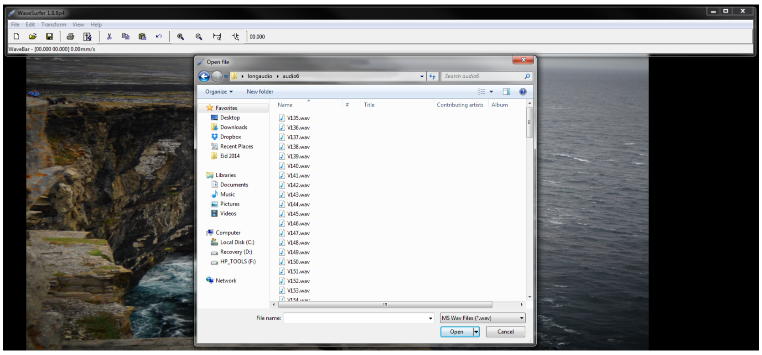


Figure D.1: Open related wav file

After downloading, installing and running this application, open the required WAV file from the directory containing all audio files. This software only identifies WAV files so it is essential to have all the audio files in the required format.

STEP 2: CHOOSE CONFIGURATION

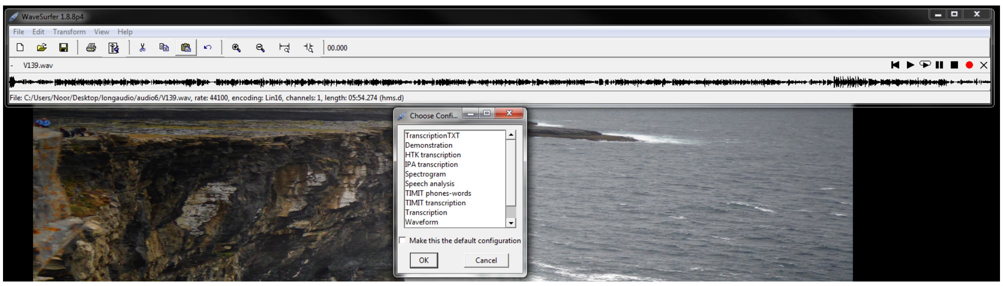


Figure D.2: Configure settings

Select “TranscriptionTXT” from the configuration setting. This configuration setting can be customised according to the user’s needs. For this study, the transcription setting is customised into pre-determined attitude labels and saved in a TXT format.

STEP 3: INSERT LABEL

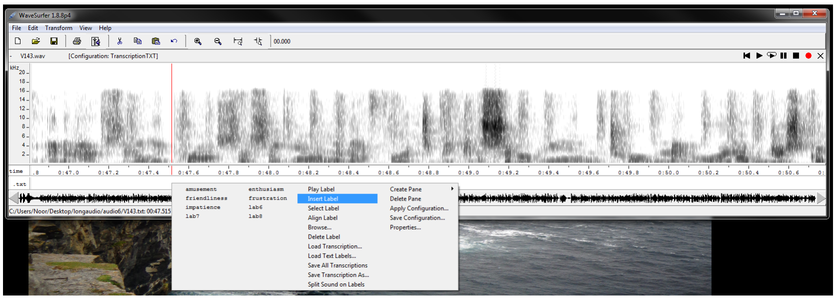


Figure D.3: Insert attitude label

The next stage is the labelling process. Annotators right-click on the transcription pane and select “Insert Label” to indicate the start time of the attitude label.

STEP 4: SELECT ATTITUDE

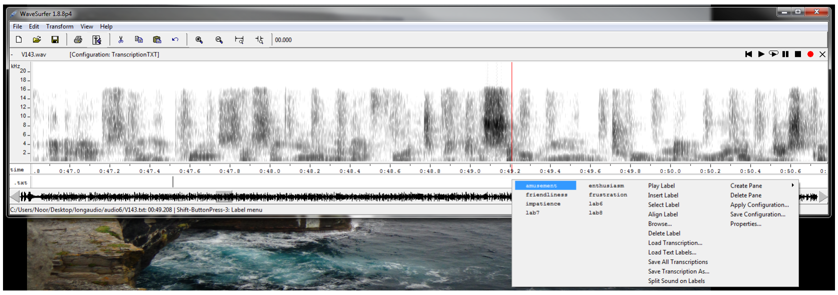


Figure D.4: Select related attitude label

After marking the starting time of the attitude, annotators navigate through the transcription pane and mark the end time. Then right-click on the pane and assign the suitable attitude for that particular segment. Text of the attitude label will appear on the pane.

STEP 5: SAVE TRANSCRIPTION

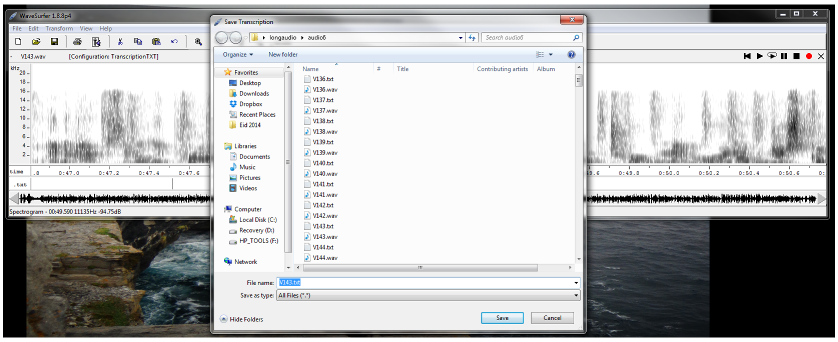


Figure D.5: Save transcription in directory

After transcribing attitude segments for that particular video, annotators proceed to the menu bar or right-click on the transcription pane and select "Save transcription as". Annotation is saved to the necessary directory in a TXT format. This format is required for the attitude segmentation process.

Appendix E

Segmentation with Windows Live Movie Maker

STEP 1: OPEN MP4 FILE

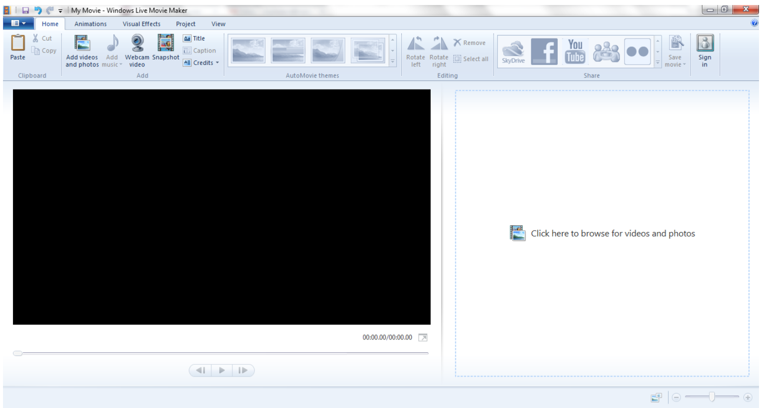


Figure E.1: Open MP4 video file from directory

Open related video by clicking "Click here to browse for videos and photos" on the right-hand pane. Videos that are of use in this stage are the full-length videos downloaded from YouTube.

STEP 2: SELECT MP4 FILE

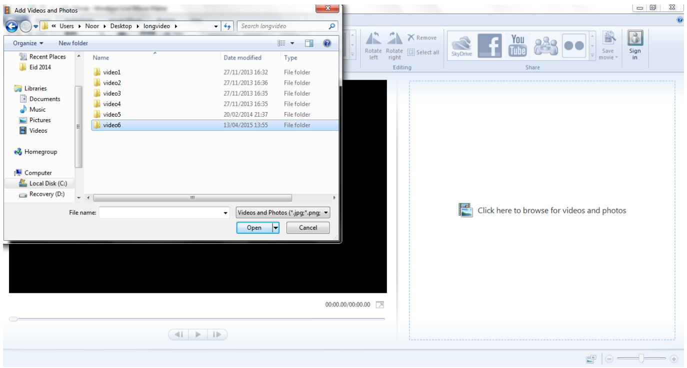


Figure E.2: Select MP4 video file in relevant directory

Select the required video from the user directory.

STEP 3: SELECT “EDIT” ON MENU BAR

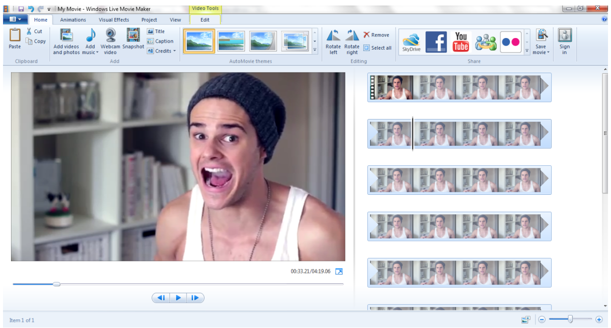


Figure E.3: Edit on menu bar

Click “Edit” on the Menu bar to navigate to the Trim Tool option.

STEP 4: SELECT TRIM TOOL

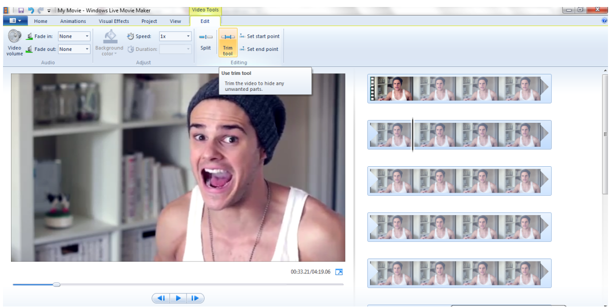


Figure E.4: Trim tool

Select the “Trim Tool” option.

STEP 5: MARK START TIME

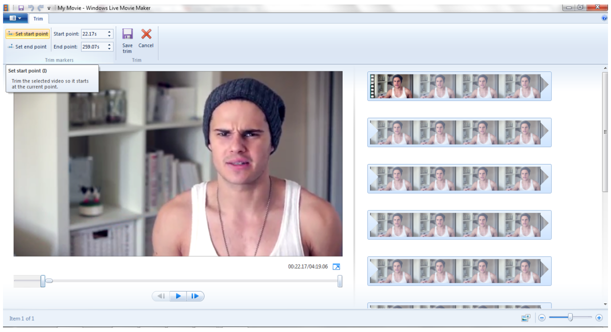


Figure E.5: Mark start time

Mark the start time of the annotated attitude label by clicking the “Set Start Point”. The start time is required to precisely be the same as the annotated time indicated in the TXT file from Wavesurfer during the annotation stage.

STEP 6: MARK END TIME

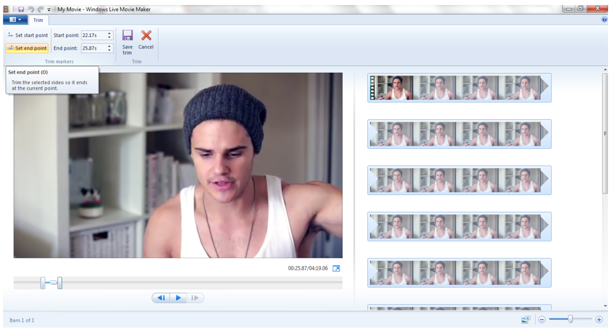


Figure E.6: Mark end time

Mark the end time of the attitude label. This is done by clicking the “Set End Point” button. Again, the end time must be exactly the same time as indicated in the TXT file produced in Wavesurfer during the annotation stage.

STEP 7: SAVE TRIM

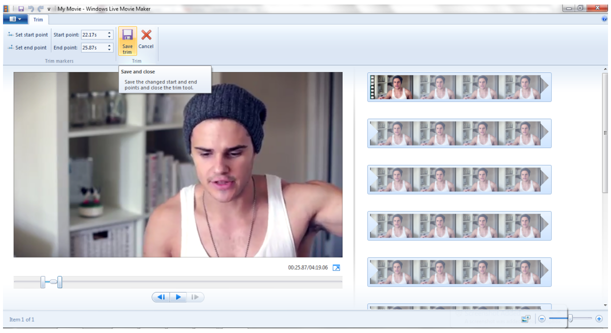


Figure E.7: Save trimmed video

Save the selected segment by clicking the "Save Trim" button.

STEP 8: SAVE MOVIE

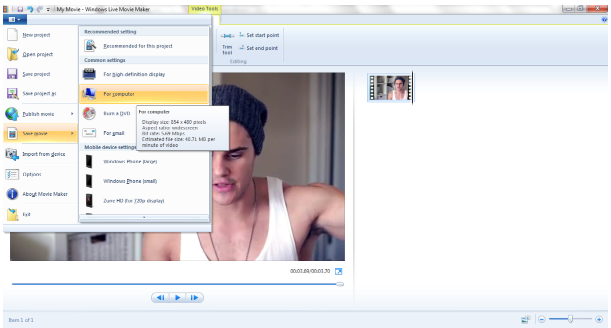


Figure E.8: Save movie

In order to save the trimmed video segment, it is insufficient to merely click the save trim button as pictured in Figure E.7. Click on "File" at the taskbar menu, select "Save movie" and then select "For Computer".

STEP 9: SAVE IN RELEVANT FOLDER

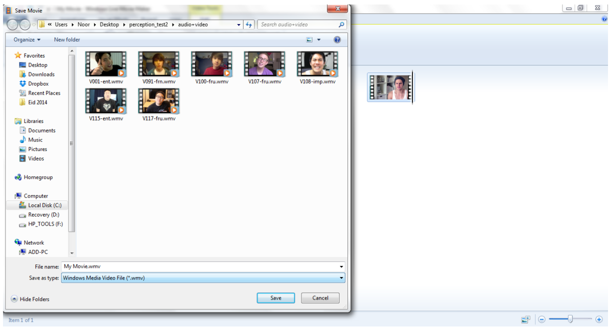


Figure E.9: Save to directory

After clicking on “For Computer” as stated in Figure E.8, the user will see a pop-up directory option. The user will need to rename the file and select the relevant folder directory. All segmented video clips will be stored in the same user directory for ease of further analysis and future reference.

Appendix F

Research Ethics Committee Approval Letter

COLÁISTE NA TRÍONÓIDE, BAILE ÁTHA CLIATH | TRINITY COLLEGE DUBLIN
Ollscoil Átha Cliath | The University of Dublin

Academic Year 2014/15

Applicant: TT67 Noor Madzlan

Supervisor: Dr Breffni O'Rourke

Project title: Multimodal Perception Study of Attitudes in Video Blogs

Dear Noor,

Your submission for ethics approval for the research project above was considered by the Research Ethics Committee, School of Linguistic, Speech and Communication Sciences, Trinity College Dublin, on Wednesday 24 June 2015, and has been approved in full. We wish you the very best in your research activities.

CENTRE FOR LANGUAGE AND COMMUNICATION STUDIES
TRINITY COLLEGE DUBLIN

Dr Lorna Carson
Chair, Research Ethics Committee
School of Linguistic, Speech and Communication Sciences
Trinity College Dublin

An Lárionad Léinn Teanga agus Cumarsáide,
Scoil na nEolaíochtaí Teangeolaíochta, Urlabhra agus Cumarsáide,
Coláiste na Tríonóide,
Baile Átha Cliath 2. Éire

Centre for Language & Communication Studies,
School of Linguistic, Speech & Communication Sciences,
Trinity College,
Dublin 2, Ireland

T 353 (0)1 896 1560
F 353 (0)1 896 2941
clcsinfo@tcd.ie
www.tcd.ie/slscs/clcs

Appendix G

Full list of 67 facial landmarks in AAM

Visual Feature	**Code**	**Label**
V1	H	Head
V2	EL	Eye Left
V3	ER	Eye Right
V4	NT	Nose Tip
V5	MCL	Mouth Corner Left
V6	MCR	Mouth Corner Right
V7	FCUL	Face Contour Upper Left
V8	FCUR	Face Contour Upper Right
V9	FCL	Face Contour Left
V10	FCR	Face Contour Right
V11	CL	Chin Left
V12	CR	Chin Right
V13	CB	Chin Bottom
V14	EBCOL	Eye Brow Corner Outer Left
V15	EBCIL	Eye Brow Corner Inner Left
V16	EBCIR	Eye Brow Corner Inner Right
V17	EBCOR	Eye Brow Corner Outer Right
V18	EBML	Eye Brow Middle Left
V19	EBMR	Eye Brow Middle Right

V20	EBMLL	Eye Brow Middle Left Left
V21	EBMRL	Eye Brow Middle Right Left
V22	EBMLR	Eye Brow Middle Left Right
V23	EBMRR	Eye Brow Middle Right Right
V24	NB	Nose Bridge
V25	ECOL	Eye Corner Outer Left
V26	ECIL	Eye Corner Inner Left
V27	ECIR	Eye Corner Inner Right
V28	ECOR	Eye Corner Outer Right
V29	ELML	Eye Lower Middle Left
V30	EUML	Eye Upper Middle Left
V31	IECLL	Iris Eye Corner Left Left
V32	IECRL	Iris Eye Corner Right Left
V33	ELMR	Eye Lower Middle Right
V34	EUMR	Eye Upper Middle Right
V35	IECLR	Iris Eye Corner Left Right
V36	IECRR	Iris Eye Corner Right Right
V37	EUML	Eye Upper Middle Left Left
V38	EUMRL	Eye Upper Middle Right Left
V39	ELMLL	Eye Lower Middle Left Left
V40	ELMRL	Eye Lower Middle Right Left
V41	EUMLR	Eye Upper Middle Left Right
V42	EUMRR	Eye Upper Middle Right Right
V43	ELMLR	Eye Lower Middle Left Right
V44	ELMRR	Eye Lower Middle Right Right
V45	NWL	Nose Wing Left
V46	NWR	Nose Wing Right
V47	NWOL	Nose Wing Outer Left
V48	NWOR	Nose Wing Outer Right
V49	NWIL	Nose Wing Inner Left
V50	NWIR	Nose Wing Inner Right

V51	NBT	Nose Bottom
V52	NLFUL	Nasolabial Fold Upper Left
V53	NLFUR	Nasolabial Fold Upper Right
V54	NLFLL	Nasolabial Fold Lower Left
V55	NLFLR	Nasolabial Fold Lower Right
V56	MT	Mouth Top
V57	MBT	Mouth Bottom
V58	MTL	Mouth Top Left
V59	MTR	Mouth Top Right
V60	MBTL	Mouth Bottom Left
V61	MBTR	Mouth Bottom Right
V62	MTIL	Mouth Top Inner Left
V63	MTI	Mouth Top Inner
V64	MTIR	Mouth Top Inner Right
V65	MBTIL	Mouth Bottom Inner Left
V66	MBTI	Mouth Bottom Inner
V67	MBTIR	Mouth Bottom Inner Right

Table G.1: List of Visual Features

Appendix H

Process of Visual Feature Extraction with AAM

STEP 1: OPEN FaceSDK SOFTWARE

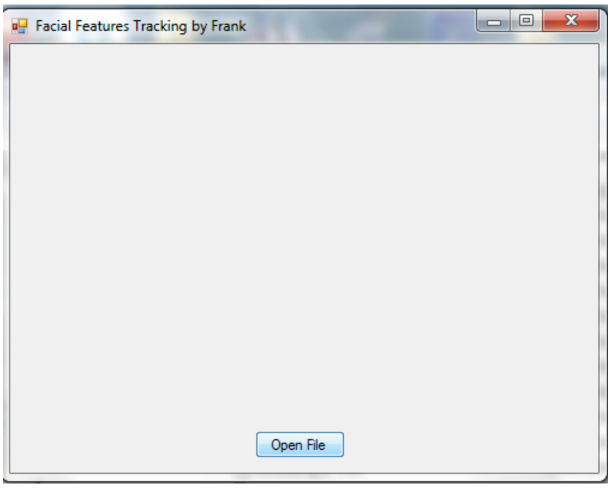


Figure H.1: Open AAM software

STEP 2: SELECT RELATED VIDEO

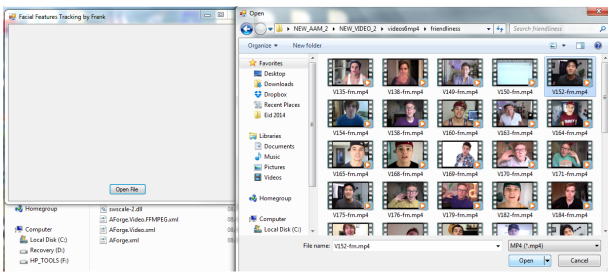


Figure H.2: Select Video

STEP 3: RUN FACIAL TRACKER

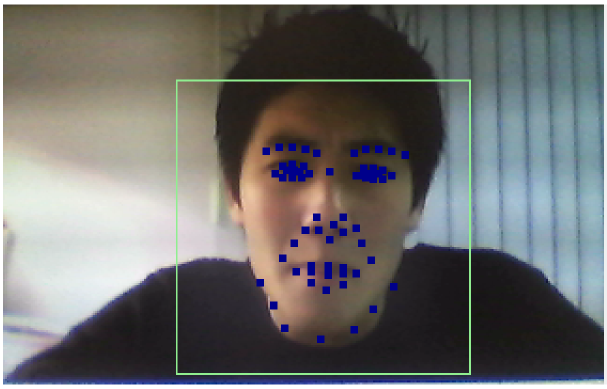


Figure H.3: Facial Tracker

STEP 4: AUTOMATED CREATION OF OUTPUT FOLDER

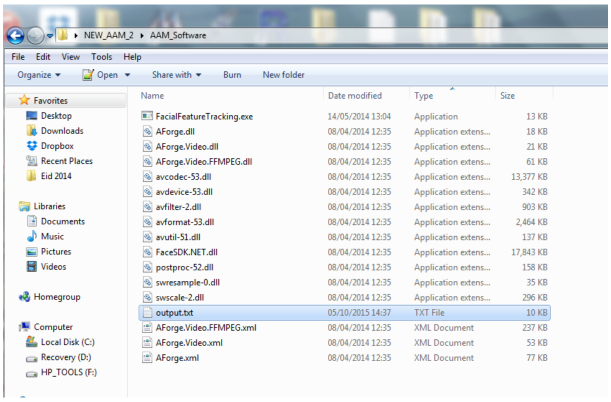


Figure H.4: Output folder

STEP 5: RENAME OUTPUT FOLDER

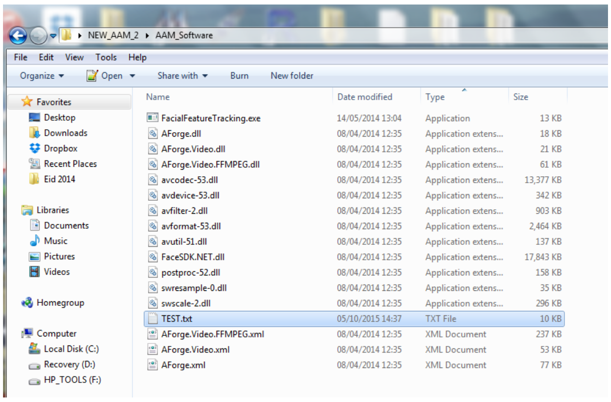
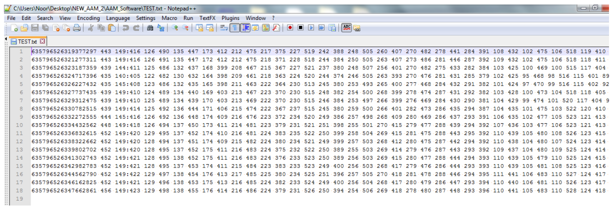


Figure H.5: Rename Output folder

STEP 6: STRUCTURE OF THE RAW VISUAL DATA




```
635796526319377297 443 149:416 126 490 135 447 173 412 212 475 217 375 227 519 242 388 248 505 260 407 270 482 278 441 284 391 108 432 102 475 106 518 119 410
635796526321277311 443 149:416 126 491 135 447 172 412 212 475 218 371 228 518 244 384 250 505 263 407 273 486 281 446 287 392 109 432 102 475 106 518 118 411
635796526323187359 439 144:411 125 486 132 437 168 399 208 467 215 367 227 521 237 380 248 507 256 401 270 482 275 433 282 384 103 425 100 469 100 515 117 404
635796526324717396 435 140:405 122 482 130 432 164 398 209 461 218 363 224 520 244 374 246 505 263 393 270 476 281 431 285 379 102 425 95 468 98 516 115 401 89
635796526326227432 435 145:408 123 486 132 435 165 398 211 463 222 364 230 513 245 380 253 493 265 400 277 468 284 432 291 382 101 424 97 470 99 516 115 402 92
635796526327737435 439 149:410 124 489 134 440 169 403 213 467 223 370 230 515 248 382 254 500 268 399 278 474 287 431 292 382 103 428 100 473 104 518 118 405
635796526329312475 439 149:410 125 489 134 439 170 403 213 469 222 370 230 515 246 384 253 497 266 399 276 469 284 430 290 381 104 429 99 474 101 520 117 404 9
635796526330782515 439 149:414 125 492 136 444 171 406 215 474 222 367 237 515 245 380 259 500 266 401 282 473 286 435 294 387 104 435 101 475 103 522 120 410
635796526332272555 444 145:416 126 492 136 448 174 409 216 476 223 372 234 520 249 386 257 498 268 409 280 469 286 437 293 391 106 435 102 477 105 523 121 413
635796526334432562 448 149:418 126 494 137 450 173 411 214 481 223 379 231 521 251 398 255 501 270 415 279 477 288 439 294 392 107 436 103 477 106 523 121 413
635796526336832615 452 149:420 129 495 137 452 174 410 216 481 224 383 235 522 250 399 258 504 269 415 281 475 288 443 295 392 110 439 105 480 108 526 123 415
635796526338322662 452 149:420 128 494 137 451 174 409 215 482 224 380 234 521 249 399 257 503 268 412 280 475 287 442 294 392 110 438 104 480 107 524 123 414
635796526339802702 452 149:420 128 495 137 452 175 411 216 483 224 375 232 522 250 389 255 503 269 414 279 476 287 443 293 392 109 437 104 480 109 525 124 414
635796526341302743 452 149:421 128 495 138 452 175 411 216 483 224 376 233 523 250 389 256 503 269 415 280 477 288 444 294 393 110 439 105 479 110 525 124 415
635796526342982783 452 149:421 128 495 137 453 174 411 215 484 223 383 233 523 249 400 256 503 268 417 279 476 286 444 293 393 110 439 105 481 108 525 123 416
635796526344562790 452 149:422 129 497 138 454 176 413 217 485 225 380 234 525 251 396 257 505 270 418 281 478 288 446 294 395 111 441 106 483 110 527 124 417
635796526346162825 452 149:421 129 496 138 453 175 413 216 485 224 382 233 524 249 400 256 504 268 417 280 479 286 447 293 394 110 440 106 481 110 526 123 417
635796526347662861 456 149:423 129 498 138 455 176 414 216 486 224 379 231 526 250 394 254 506 269 418 278 480 287 448 293 396 110 441 105 483 110 528 124 418
```

Figure H.6: Structure of data

Bibliography

[1] Jens Allwood. A framework for studying human multimodal communication. *Coverbal Synchrony in Human-Machine Interaction*, page 17, 2013.

[2] Antonio Damasio and Raymond J Dolan. The feeling of what happens. *Nature*, 401(6756):847–847, 1999.

[3] Antonio Damasio. *Descartes' error: Emotion, reason and the human brain*. Random House, 2008.

[4] James A Russell and Doreen Ridgeway. Dimensions underlying children's emotion concepts. *Developmental Psychology*, 19(6):795, 1983.

[5] M. Wetherell. *Affect and Emotion: A New Social Science Understanding*. SAGE Publications, 2012.

[6] Mark P Zanna and John K Rempel. Attitudes: A new look at an old concept. 1988.

[7] Eric Shouse. Feeling, emotion, affect. *M/c journal*, 8(6):26, 2005.

[8] Stuart Oskamp and P Wesley Schultz. *Attitudes and opinions*. Psychology Press, 1977.

[9] Martin Fishbein and Icek Ajzen. Attitudes and opinions. *Annual review of psychology*, 23(1):487–544, 1972.

[10] Véronique Aubergé. A gestalt morphology of prosody directed by functions: the example of a step by step model developed at icp. In *Speech Prosody 2002, International Conference*, 2002.

[11] Jens Allwood. Multimodal corpora. *Corpus linguistics: An international handbook*, 1:207–224, 2008.

[12] Alessandro Vinciarelli and Gelareh Mohammadi. Towards a technology of nonverbal communication: Vocal behavior in social and affective phenomena. Technical report, igi-global, 2010.

[13] Mark Knapp, Judith Hall, and Terrence Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013.

[14] Virginia P Richmond, James C McCroskey, and Steven K Payne. *Nonverbal behavior in interpersonal relations*. Prentice Hall Englewood Cliffs, NJ, 1991.

[15] Paul Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993.

[16] Alessandro Vinciarelli and Fabio Valente. Social signal processing: Understanding nonverbal communication in social interactions. In *Proceedings of Measuring Behavior 2010, Eindhoven (The Netherlands)*, number EPFL-CONF-163182, 2010.

[17] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176. ACM, 2011.

[18] Veronica Rosas, Rada Mihalcea, and L Morency. Multimodal sentiment analysis of spanish online videos. 2013.

[19] Christer Gobl, Ailbhe Nı, et al. The role of voice quality in communicating emotion, mood and attitude. *Speech communication*, 40(1):189–212, 2003.

[20] Ginevra Castellano, Santiago D Villalba, and Antonio Camurri. Recognising human emotions from body movement and gesture dynamics. In *Affective computing and intelligent interaction*, pages 71–82. Springer, 2007.

[21] Nadia Bianchi-Berthouze, Paul Cairns, Anna Cox, Charlene Jennett, and Whan Woong Kim. On posture as a modality for expressing and recognizing emotions. In *Emotion and HCI workshop at BCS HCI London*, 2006.

[22] Dang-Khoa Mac, Véronique Aubergé, Albert Rilliard, and Eric Castelli. Cross-cultural perception of vietnamese audio-visual prosodic attitudes. In *Speech Prosody*, 2010.

[23] João Antônio de Moraes, Albert Rilliard, Bruno Alberto de Oliveira Mota, and Takaaki Shochi. Multimodal perception and production of attitudinal meaning in brazilian portuguese. In *Proc. of Speech Prosody*, 2010.

[24] Yann Morlec, Grard Bailly, and Vronique Auberg. Generating prosodic attitudes in french: Data, model and evaluation. *Speech Communication*, 33(4):357–371, 2001.

[25] Bonnie A Nardi, Diane J Schiano, and Michelle Gumbrecht. Blogging as social activity, or, would you let 900 million people read your diary? In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 222–231. ACM, 2004.

[26] Diane J Schiano, Bonnie A Nardi, Michelle Gumbrecht, and Luke Swartz. Blogging by the rest of us. In *CHI'04 extended abstracts on Human factors in computing systems*, pages 1143–1146. ACM, 2004.

[27] Lilia Efimova and Aldo De Moor. Beyond personal webpublishing: An exploratory study of conversational blogging practices. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, pages 107a–107a. IEEE, 2005.

[28] Nicole B Ellison, Charles Steinfield, and Cliff Lampe. The benefits of facebook friends: social capital and college students use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168, 2007.

[29] Alessandro Acquisti and Ralph Gross. Imagined communities: Awareness, information sharing, and privacy on the facebook. In *Privacy enhancing technologies*, pages 36–58. Springer, 2006.

[30] Yusuke Yanbe, Adam Jatowt, Satoshi Nakamura, and Katsumi Tanaka. Can social bookmarking enhance search in the web? In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 107–116. ACM, 2007.

[31] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

[32] Xintian Yang, Amol Ghoting, Yiye Ruan, and Srinivasan Parthasarathy. A framework for summarizing and analyzing twitter feeds. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 370–378. ACM, 2012.

[33] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[34] Joan-Isaac Biel, Oya Aran, and Daniel Gatica-Perez. You are known by how you vlog: Personality impressions and nonverbal behavior in youtube. In *ICWSM*, 2011.

[35] Oya Aran, Joan-Isaac Biel, Daniel Gatica-Perez, et al. You are known by how you vlog: Personality impressions and nonverbal behavior in youtube. In *Proceedings of AAAI International Conference on Weblogs and Social Media*, number EPFL-CONF-165900, 2011.

[36] Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and L Morency. Youtube movie reviews: In, cross, and open-domain sentiment analysis in an audiovisual context. 2013.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[38] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques, 2007.

[39] Naresh K Malhotra. Attitude and affect: new frontiers of research in the 21st century. *Journal of Business Research*, 58(4):477–482, 2005.

[40] I. Ajzen. *Attitudes, Personality, and Behavior*. Mapping social psychology. McGraw-Hill Education, 2005.

[41] John A Bargh and Tanya L Chartrand. The unbearable automaticity of being. *American psychologist*, 54(7):462, 1999.

[42] James Agarwal and Naresh K Malhotra. An integrated model of attitude and affect: Theoretical foundation and an empirical investigation. *Journal of Business research*, 58(4):483–493, 2005.

[43] Yan Lu, Véronique Aubergé, Albert Rilliard, et al. Do you hear my attitude? prosodic perception of social affects in mandarin. *Proceedings of Speech Prosody 2012*, pages 685–688, 2012.

[44] Jens Allwood, Stefano Lanzini, and Elisabeth Ahlsén. Contributions of different modalities to the attribution of affective-epistemic states. In *Proceedings from the 1st European Symposium on Multimodal Communication University of Malta*, pages 1–6.

[45] Massimo Chindamo, Julian Allwood, and Elisabeth Ahlsen. Some suggestions for the study of stance in communication. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International*

Conference on and 2012 International Confernece on Social Computing (SocialCom), pages 617–622. IEEE, 2012.

[46] Patrizia Paggio, Jens Allwood, Elisabeth Ahlsén, and Kristiina Jokinen. The nomco multi-modal nordic resource–goals and characteristics. 2010.

[47] Peter Juel Henrichsen and Jens Allwood. Predicting the attitude flow in dialogue based on multi-modal speech cues. *NEALT PROCEEDINGS SERIES*, 2012.

[48] Yann Morlec, Gérard Bailly, and Veronique Aubergé. Generating the prosody of attitudes. In *Intonation: Theory, Models and Applications*, 1997.

[49] Jean-Marc Blanc and Peter Ford Dominey. Identification of prosodic attitudes by a temporal recurrent network. *Cognitive Brain Research*, 17(3):693–699, 2003.

[50] Albert Rilliard, Jean-Claude Martin, Véronique Aubergé, Takaaki Shochi, et al. Perception of french audio-visual prosodic attitudes. *Speech Prosody, Campinas, Brasil*, 2008.

[51] Plínio Barbosa and Gérard Bailly. Characterisation of rhythmic patterns for text-to-speech synthesis. *Speech Communication*, 15(1-2):127–137, 1994.

[52] Vasiliki Orgeta and Louise H Phillips. Effects of age and emotional intensity on the recognition of facial emotion. *Experimental aging research*, 34(1):63–79, 2007.

[53] Alberto Di Domenico, Rocco Palumbo, Nicola Mammarella, and Beth Fairfield. Aging and emotional expressions: is there a positivity bias during dynamic emotion recognition? *Frontiers in psychology*, 6, 2015.

[54] Judith A Hall and David Matsumoto. Gender differences in judgments of multiple emotions from facial expressions. *Emotion*, 4(2):201, 2004.

[55] D Matsumoto and P Ekman. Japanese and caucasian brief affect recognition test (jacbart), i, ii, iii [videotapes]. *Available from Culture and Emotion Research Laboratory, Department of Psychology, San Francisco State University*, 1600, 1992.

[56] Jens Allwood. Cooperation and flexibility in multimodal communication. In *Cooperative Multimodal Communication*, pages 113–124. Springer, 2001.

[57] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. Analysis of emo-

tion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211. ACM, 2004.

[58] Christina Regenbogen, Daniel A Schneider, Raquel E Gur, Frank Schneider, Ute Habel, and Thilo Kellermann. Multimodal human communicationtargeting facial expressions, speech content and prosody. *Neuroimage*, 60(4):2346–2356, 2012.

[59] David Crystal. *The English tone of voice: Essays in intonation, prosody and paralanguage*. Hodder Arnold, 1975.

[60] Peter Roach. *English Phonetics and Phonology Fourth Edition: A Practical Course*. Ernst Klett Sprachen, 2010.

[61] Sieb Nooteboom. The prosody of speech: melody and rhythm. *The handbook of phonetic sciences*, 5:640–673, 1997.

[62] John Laver. *Principles of phonetics*. Cambridge University Press, 1994.

[63] Peter Roach. Techniques for the phonetic description of emotional speech. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[64] Petri Laukka, Patrik Juslin, and Roberto Bresin. A dimensional approach to vocal expression of emotion. *Cognition & Emotion*, 19(5):633–653, 2005.

[65] Helen M Hanson and Erika S Chuang. Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *The Journal of the Acoustical Society of America*, 106(2):1064–1077, 1999.

[66] Nick Campbell and Parham Mokhtari. Voice quality: the 4th prosodic dimension. In *15th ICPhS*, pages 2417–2420, 2003.

[67] Frank Dellaert, Thomas Polzin, and Alex Waibel. Recognizing emotion in speech. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1970–1973. IEEE, 1996.

[68] Tin Lay Nwe, Foo Say Wei, and Liyanage C De Silva. Speech based emotion classification. In *TENCON 2001. Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology*, volume 1, pages 297–301. IEEE, 2001.

[69] Irena Yanushevskaya, Christer Gobl, and AILBHE NI CHASAIDE. Voice quality and loudness in affect perception. 2008.

[70] Sylvie Mozziconacci. Emotion and attitude conveyed in speech by means of prosody. In *2nd Workshop on Attitude, Personality and Emotions in User-Adapted Interaction*. Citeseer, 2001.

[71] Julia Hirschberg, Diane Litman, and Marc Swerts. Prosodic and other cues to speech recognition failures. *Speech Communication*, 43(1):155–175, 2004.

[72] Paul Ekman and Harriet Oster. Facial expressions of emotion. *Annual review of psychology*, 30(1):527–554, 1979.

[73] Klaus R Scherer. *The Functions of Nonverbal Signs in Conversation*, chapter 8, pages 225–244. Hillsdale: L. Erlbaum, 1980.

[74] Juergen Luettin and Neil A Thacker. Speechreading using probabilistic models. *Computer Vision and Image Understanding*, 65(2):163–178, 1997.

[75] Paul Ekman. About brows: Emotional and conversational signals. *Human ethology*, pages 169–202, 1979.

[76] Paul Ekman. *Telling lies: Clues to deceit in the marketplace, politics, and marriage*. WW Norton & Company, 2009.

[77] Javid Sadrô, Izzat Jarudi, and Pawan Sinhaô. The role of eyebrows in face recognition. *Perception*, 32(3):285–293, 2003.

[78] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[79] Javier R Movellan. Visual speech recognition with stochastic networks. *Advances in neural information processing systems*, pages 851–858, 1995.

[80] Beat Fasel and Juergen Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003.

[81] Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75, 1976.

[82] Marian Stewart Bartlett, Gwen C Littlewort, Mark G Frank, Claudia Lainscsek, Ian R Fasel, and Javier R Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of multimedia*, 1(6):22–35, 2006.

[83] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 46–53. IEEE, 2000.

[84] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. In *Computer VisionECCV98*, pages 484–498. Springer, 1998.

[85] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Comparing active shape models with active appearance models. In *BMVC*, volume 99, pages 173–182, 1999.

[86] Chalapathy Neti, Gerasimos Potamianos, Juergen Luettin, Iain Matthews, Herve Glotin, Dimitra Vergyri, June Sison, and Azad Mashari. Audio visual speech recognition. Technical report, IDIAP, 2000.

[87] Minh Tue Vo and Alex Waibel. Multimodal human-computer interaction. *Proceedings of ISSD*, 93, 1993.

[88] Maja Pantic and Leon JM Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

[89] Matthew Turk. Multimodal human-computer interaction. In *Real-time vision for human-computer interaction*, pages 269–283. Springer, 2005.

[90] Matthew Turk and Mathias Kölsch. Perceptual interfaces. *Emerging Topics in Computer Vision, Prentice Hall*, 2004.

[91] Takaaki Shochi, Donna Erickson, Albert Rilliard, Véronique Aubergé, Jean-Claude Martin, et al. Recognition of japanese attitudes in audio-visual speech. In *Speech prosody*, volume 2008, pages 689–692, 2008.

[92] Loic Kessous, Ginevra Castellano, and George Caridakis. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces*, 3(1-2):33–48, 2010.

[93] Heather Molyneaux, Kerri Gibson, Susan O'Donnell, and Janice Singer. New visual media and gender: A content, visual and audience analysis of youtube vlogs. 2008.

[94] James Trier. cool engagements with youtube: Part 1. *Journal of Adolescent & Adult Literacy*, 50(5):408–412, 2007.

[95] Prince Deh. Promoting information-sharing in ghana using video blogging. *Participatory Learning and Action*, 59(1):40–43, 2009.

[96] Wen Gao, Yonghong Tian, Tiejun Huang, and Qiang Yang. Vlogging: A survey of videoblogging technology on the web. *ACM Computing Surveys (CSUR)*, 42(4):15, 2010.

[97] Elisabetta Adami. we/youtube: exploring sign-making in video-interaction. *Visual Communication*, 8(4):379–399, 2009.

[98] Joan-Isaac Biel and Daniel Gatica-Perez. Vlogcast yourself: Nonverbal behavior and attention in social media. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, page 50. ACM, 2010.

[99] Joan-Isaac Biel and Daniel Gatica-Perez. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *Multimedia, IEEE Transactions on*, 15(1):41–55, 2013.

[100] Dairazalia Sanchez-Cortes, Joan-Isaac Biel, Shiro Kumano, Junji Yamato, Kazuhiro Otsuka, and Daniel Gatica-Perez. Inferring mood in ubiquitous conversational video. In *Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia*, page 22. ACM, 2013.

[101] Joan-Isaac Biel and Daniel Gatica-Perez. Wearing a youtube hat: directors, comedians, gurus, and user aggregated behavior. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 833–836. ACM, 2009.

[102] Joan-Isaac Biel and Daniel Gatica-Perez. Vlogsense: Conversational behavior and social attention in youtube. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 7(1):33, 2011.

[103] Joan-Isaac Biel and Daniel Gatica-Perez. The good, the bad, and the angry: Analyzing crowdsourced impressions of vloggers. *Ethnicity*, 16(4.8):0–7, 2012.

[104] Joan-Isaac Biel, Vagia Tsiminaki, John Dines, and Daniel Gatica-Perez. Hi youtube!: personality impressions and verbal content in social video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 119–126. ACM, 2013.

[105] Robert R McCrae and Oliver P John. An introduction to the five-factor model and its applications. *Personality: critical concepts in psychology*, 60:295, 1998.

[106] Paul Boersma. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345, 2001.

[107] Ligia Maria Batrinca, Nadia Mana, Bruno Lepri, Fabio Pianesi, and Nicu Sebe. Please, tell me about yourself: automatic personality assessment using short self-presentations. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 255–262. ACM, 2011.

[108] Rosalind W Picard and Roalind Picard. *Affective computing*, volume 252. MIT press Cambridge, 1997.

[109] Cindy L Bethel and Robin R Murphy. Survey of non-facial/non-verbal affective expressions for appearance-constrained robots. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 38(1):83–92, 2008.

[110] Ashish Kapoor, Winslow Burleson, and Rosalind W Picard. Automatic prediction of frustration. *International journal of human-computer studies*, 65(8):724–736, 2007.

[111] Ashish Kapoor and Rosalind W Picard. Multimodal affect recognition in learning environments. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 677–682. ACM, 2005.

[112] Laurel D Riek, Maria F Oconnor, and Peter Robinson. Guess what? a game for affective annotation of video using crowd sourcing. In *Affective computing and intelligent interaction*, pages 277–285. Springer, 2011.

[113] Eva Schultze-Berndt. Linguistic annotation. *Essentials of language documentation*, 178:213, 2006.

[114] Stefanie Nowak and Stefan Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566. ACM, 2010.

[115] Jens Allwood, Loredana Cerrato, Laila Dybkjaer, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. The mumin multimodal coding scheme. In *Proc. Workshop on Multimodal Corpora and Annotation*, 2005.

[116] Michael Kipp. Graduate college for cognitive sciences university of the saarland, germany. 2001.

[117] Magnus Gunnarsson. User manual for multitool. Technical report, Technical Report availale from http://www. ling. gu. se/mgunnar/multitool/MT-manual. pdf, 2002.

[118] Niels Ole Bernsen, Laila Dybkjaer, and Mykola Kolodnytsky. The nite workbench-a tool for annotation of natural interactivity and multimodal data. In *Las Palmas*. Citeseer, 2002.

[119] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

[120] Jacob Cohen et al. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[121] Joseph L Fleiss, Jacob Cohen, and BS Everitt. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323, 1969.

[122] Ron Kohavi and Foster Provost. Glossary of terms. *Machine Learning*, 30(2-3):271–274, 1998.

[123] Eva Székely, John Kane, Stefan Scherer, Christer Gobl, and Julie Carson-Berndsen. Detecting a targeted voice style in an audiobook using voice quality features. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4593–4596. IEEE, 2012.

[124] Chu-Hsing Lin, Jung-Chun Liu, and Chia-Han Ho. Anomaly detection using libsvm training tools. In *Information Security and Assurance, 2008. ISA 2008. International Conference on*, pages 166–171. IEEE, 2008.

[125] Guido W Imbens and Richard Spady. Confidence intervals in generalized method of moments models. *Journal of econometrics*, 107(1):87–98, 2002.

[126] Marian Stewart Bartlett, Gwen Littlewort, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Machine learning methods for fully automatic recognition of facial expressions and facial actions. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 1, pages 592–597. IEEE, 2004.

[127] Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. Distinguishing the communicative functions of gestures. In *Machine Learning for Multimodal Interaction*, pages 38–49. Springer, 2008.

[128] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[129] Hatice Gunes, Caifeng Shan, Shizhi Chen, and YingLi Tian. Bodily expression for automatic affect recognition. *Emotion Recognition: A Pattern Analysis Approach*, pages 343–377, 2015.

[130] I Mccowan, G Lathoud, M Lincoln, A Lisowska, W Post, D Reidsma, and P Wellner. The ami meeting corpus. In *In: Proceedings Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research. LPJJ Noldus, F. Grieco, LWS Loijens and PH Zimmerman (Eds.), Wageningen: Noldus Information Technology*, 2005.

[131] Shannon Hennig, Ryad Chellali, and Nick Campbell. The d-ans corpus: the dublin-autonomous nervous system corpus of biosignal and multimodal recordings of conversational speech. In *Proceedings of the ELRA, the 9th Edition of the Language Resources and Evaluation Conference. Reykjavik, Iceland*, pages 26–31, 2014.

[132] Joan-Isaac Biel, Daniel Gatica-Perez, et al. Voices of vlogging. In *ICWSM*, 2010.

[133] Casey Neistat. How To Vlog. `https://www.youtube.com/watch?v=dGLEEZZ15N4`, 2015. [Online; accessed 07-July-2015].

[134] Miss Fenderr. HOW I MAKE MY VIDEOS!(Step-By-Step Tutorial). `https://www.youtube.com/watch?v=dGLEEZZ15N4`, 2014. [Online; accessed 07-July-2015].

[135] Ruchard Dufour, Vincent Jousse, Yannick Estève, Fréderic Béchet, and Georges Linarès. Spontaneous speech characterization and detection in large audio database. *SPECOM, St. Petersburg*, 2009.

[136] Timothy Delaghetto. Annoying People I Hate #2. `https://www.youtube.com/watch?v=pY7n5mpG5mU`, 2014. [Online; accessed 07-March-2014].

[137] Timothy Delaghetto. Be a Gentleman, Get the Booty. `https://www.youtube.com/watch?v=jbqW7SUjuT0`, 2012. [Online; accessed 07-March-2014].

[138] British Psychological Society. Report of the working party on conducting research on the internet: guidelines for ethical practice in psychological research online. British Psychological Society Leicester, 2007.

[139] Paul Reilly. The battle of stokes crofton youtube: The development of an ethical stance for the study of online comments. 2013.

[140] Fabian Neuhaus and Timothy Webmoor. Agile ethics for massified research and visualization. *Information, Communication & Society*, 15(1):43–65, 2012.

[141] Stephen Pihlaja. *Antagonism on YouTube: Metaphor in Online Discourse*. Bloomsbury Publishing, 2014.

[142] Youtube copyright centre. `https://www.youtube.com/yt/copyright/fair-use.html#yt-copyright-resources`. Accessed: 2012-12-14.

[143] Niga Higa. Off The Pill - Arrogant People). `https://www.youtube.com/watch?v=7sz5cI51enE`, 2011. [Online; accessed 07-March-2014].

[144] Sound systems: Mono versus stereo. `http://www.mcsquared.com/mono-stereo.htm`. Accessed: 2012-12-14.

[145] James A Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145, 2003.

[146] Kre Sjlander and Jonas Beskow. Wavesurfer - an open source speech tool, 2000.

[147] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. Elan: a professional framework for multimodality research. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, 2006.

[148] Noor A Madzlan, J Reverdy, Francesca Bonin, Loredana Sundberg Cerrato, and Nick Campbell. Multimodal perception of attitudes: A study on video blogs. In *3rd European Symposium on Multimodal Communication*, 2015.

[149] Ekaterina P Volkova, Betty J Mohler, Detmar Meurers, Dale Gerdemann, and Heinrich H Bülthoff. Emotional perception of fairy tales: achieving agreement in emotion annotation of text. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 98–106. Association for Computational Linguistics, 2010.

[150] Björn Schuller. Multimodal affect databases: Collection, challenges, and chances. *The Oxford Handbook of Affective Computing*, pages 323–333, 2014.

[151] Judith A Hall. Gender effects in decoding nonverbal cues. *Psychological bulletin*, 85(4):845, 1978.

[152] Hillary Anger Elfenbein and Nalini Ambady. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128(2):203, 2002.

[153] Aoju Chen. *Universal and language-specific perception of paralinguistic intonational meaning*. Utrecht: LOT, 2005.

[154] Loredana Sundberg Cerrato. Investigating communicative feedback phenomena across languages and modalities. 2007.

[155] Marc Schröder, Roddy Cowie, Ellen Douglas-Cowie, Machiel Westerdijk, and Stan CAM Gielen. Acoustic correlates of emotion dimensions in view of speech synthesis. In *INTERSPEECH*, pages 87–90, 2001.

[156] James Hillenbrand, Ronald A Cleveland, and Robert L Erickson. Acoustic correlates of breathy vocal quality. *Journal of Speech, Language, and Hearing Research*, 37(4):769–778, 1994.

[157] Yen-Liang Shue, Gang Chen, and Abeer Alwan. On the interdependencies between voice quality, glottal gaps, and voice-source related acoustic measures. In *INTERSPEECH*, pages 34–37, 2010.

[158] Pedro Tome, Julian Fierrez, Ruben Vera-Rodriguez, and Daniel Ramos. Identification using face regions: Application and assessment in forensic scenarios. *Forensic science international*, 233(1):75–83, 2013.

[159] Hua Wang, Heng Huang, and Fillia Makedon. Emotion detection via discriminant laplacian embedding. *Universal access in the information society*, 13(1):23–31, 2014.

[160] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[161] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[162] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, pages 240–242, 1895.

[163] Bernard L Welch. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, pages 28–35, 1947.

[164] R. Mankiewicz. *The story of mathematics*. The story of mathematics. Princeton University Press, 2000.

[165] Graeme D Ruxton. The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test. *Behavioral Ecology*, 17(4):688–690, 2006.

[166] Jarkko Isotalo. Basics of statistics. *University of Tampere, Tampere*, 2001.

[167] CORRELATION DEDUCED FROM A SMALL SAMPLE. correlation coefficients covering the cases (i)the frequency dis-tribution of the values of the correlation coefficient in samples from an indefinitely large population, biometrika, vol. 10, pp. 507v521, 1915. here the method of defining the sample by the coordinates of. 1921.

[168] David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011.

[169] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.

[170] Noor Alhusna Madzlan, Jingguang Han, Francesca Bonin, and Nick Campbell. Towards automatic recognition of attitudes: Prosodic analysis of video blogs. *Speech Prosody, Dublin, Ireland*, pages 91–94, 2014.

[171] Noor Alhusna Madzlan, Yuyun Huang, and Nick Campbell. *Speech and Computer: 17th International Conference, SPECOM 2015, Athens, Greece, September 20-24, 2015, Proceedings*, volume 9319, chapter Automatic Classification and Prediction of Attitudes: Audio-Visual Analysis of Video Blogs. Springer, 2015.

[172] Helen M Hanson. Glottal characteristics of female speakers: Acoustic correlates. *The Journal of the Acoustical Society of America*, 101:466, 1997.

[173] Noor Alhusna Madzlan, JingGuang Han, Francesca Bonin, and Nick Campbell. Automatic recognition of attitudes in video blogs-prosodic and visual feature analysis. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[174] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326, 2003.

[175] Sanaul Haq, Philip JB Jackson, and James Edge. Audio-visual feature selection and reduction for emotion classification. In *Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP08), Tangalooma, Australia*, 2008.

[176] Wei Fan. Systematic data selection to mine concept-drifting data streams. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 128–137. ACM, 2004.

[177] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification, 2003.

[178] Tin Kam Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE, 1995.

[179] J Hohnsbein and S Mateeff. The time it takes to detect changes in speed and direction of visual motion. *Vision Research*, 38(17):2569–2573, 1998.

[180] SK DMello and AC Graesser. Feeling, thinking, and computing with affect-aware learning technologies. *The Oxford Handbook of Affective Computing*, pages 419–434, 2014.

[181] Stephen Brown. Meet pepper, the emotion reading robot. *TECHNOLOGY*, 2014.

[182] Cory D Kidd, Will Taggart, and Sherry Turkle. A sociable robot to encourage social interaction among the elderly. In *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*, pages 3972–3976. IEEE, 2006.

[183] Rosalind W Picard, Seymour Papert, Walter Bender, Bruce Blumberg, Cynthia Breazeal, David Cavallo, Tod Machover, Mitchel Resnick, Deb Roy, and Carol Strohecker. Affective learninga manifesto. *BT Technology Journal*, 22(4):253–269, 2004.

[184] J. Burgess, J. Green, H. Jenkins, and J. Hartley. *YouTube: Online Video and Participatory Culture*. DMS - Digital Media and Society. Wiley, 2013.