Improving Social Intelligence of Machines in the Context of Public Speaking Situations

Fasih Haider

Thesis Submitted for the Degree of Doctor of Philosophy

School of Computer Science and Statistics

University of Dublin

Trinity College Dublin

Declaration

I hereby declare that this thesis, submitted in candidature for the degree of Doctor of Philosophy at the University of Dublin, Trinity College Dublin, is entirely my own work and has not been previously submitted for a degree at this or any other university. Wherever there is published or unpublished work included, it is duly acknowledged. 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 the university library conditions of use and acknowledgement.

Fasih Haider

Dated: September 12, 2018

Acknowledgement

Firstly, I would like to thank my primary supervisor Prof. Saturnino Luz for his guidance, support and encouragement over the years. Prof. Luz deserves special acknowledgement because even though he went to a different university and a different city, he still found time to guide me and made sure my research kept on track, despite the pressures of his new role he always made a special effort to conduct meetings remotely to guide me whenever I needed it. Speaking of my primary supervisor's departure it fell on Prof. Carl Vogel to take me on as my official supervisor which was a bit challenging considering that he was not aware of my research beforehand. But he took the challenge and caught up quickly and provided valuable feedback particularly very last time corrections that helped me greatly in shaping the thesis. I would also like to thank my secondary PhD supervisor and principle investigator of *EU-FP7 METALOGUE* project Prof. Nick Campbell for his philosophical insights about the subject matter and for introducing me to his vast network of academics at multiple conferences. Much of the work presented in this thesis is done in collaboration with others and it was a privilege to work with all of you. Thanks to Akira Hayakawa, Loredana Cerrato for all those data collection sessions and the madness that came with it. Thanks to Kevin El-Haddad for the time spent during many eNTERFACE workshops and conferences. Thanks to

Fahim A. Salim for the spirited discussions about what can and cannot be done with TED talks. The journey of my PhD was not confined to my desk, it took me around the world. It was a fun journey filled with good food, jet lags and fun company of my friends and colleagues e.g. raiding Loredana's new office in Stockholm, all the site seeing with Kevin and using Fahim as my food delivery boy on rainy Seattle evenings. Finally I would like to thank all my colleagues, friends and family members. Last but not the least I would like to thank my parents for all the love and prayer they have given me throughout my life.

Abstract

The thesis is a collection of various studies which contribute to human-machine multimodal systems in different ways by measuring cognition. Some of the studies address basic system capabilities such as the identification of who is talking and to whom in multiparty and multimodal dialogue systems, while other studies focus on aspects, which can be relevant in systems for training presentation styles (attitudes recognition, user engagement etc.). The focus of the thesis is on the use of technologies (speech analysis, biometry and computer vision) to make machines able to understand the human social signals and behaviour cues which can help in automating the process of public speaking training. This thesis presents novel systems and empirical studies that contribute towards the development of a multimodal multi-party spoken dialogue system which can be used for training humans for public speaking. The thesis is structured in two main parts: The first part of thesis deals with automatic recognition of speaking abilities in four different kinds of public speaking situations: students' presentations, formal talks (TED talks), informal talks (video blogs) and simulated political debates. The second part of the thesis tackles some of the challenges encountered when using a multimodal multi-party dialogue system for training of these different types of public speaking situations. The first public speaking situation is students' presentations where the

proposed novel system can predict the level of 'self confidence and body language' of a student during a presentation. The second public speaking situation is formal talks (TED talks), where the proposed system can predict the level of engagement of on-line viewers and extract engaging and non-engaging parts of a talk which can be used either as a feedback to the presenter or as a help in building talk summarization or talk search tools. The third public speaking situation is informal talks (video blogs), where the proposed system can predict the presenter's attitude. This can help a user to train his/her attitude for video blogging. The fourth public speaking situation is simulated debates which were specifically collected. In these debates, subjects are arguing on a proposed smoking ban. The method proposed in this thesis can be helpful to train public speaking abilities. However using automatic methods as a part of a multimodal multiparty dialogue system, which is designed to train users for public speaking, poses many challenges. That is why, the second part of thesis focuses on some of those challenges. The first challenge is about the active speaker detection (who is speaking among a set of person), so that a multiparty multimodal dialogue system can manage multiple people. The second challenge is about the cognitive states detection (while interacting with a machine using Automatic Speech Recognition (ASR)) which can help a machine to switch between different ASR models to improve performance of ASR and sense the users' experience. The third challenge is about the detection of system directed speech which can help a machine in detecting if a user speaks to machine or not. This thesis summarizes the proposed methods and their evaluation. These methods can help in developing a spoken dialogue system for public speaking training with some human like characteristics.

Contents

List of Figures

List of Tables

Associated Publications

Hayakawa, A., **Haider, F.**, Cerrato, L., Campbell, N., & Luz, S. (2015). **Detection of Cognitive States and Their Correlation to Speech Recognition Performance in Speech-to-Speech Machine Translation Systems**. *In Proceedings of INTERSPEECH'15: the 16th Annual Conference of the International Speech Communication Association* (pp. 2539 – 2543). Dresden, Germany: ISCA. Retrieved from `http://www.isca-speech.org/archive/interspeech_2015/i15_2539.html`

Salim, F A., **Haider, F.**, Conlan, O.,Luz, S., & Campbell, N. (2015). **Analyzing Multimodality of Video for User Engagement Assessment**. *In Proceedings of International Conference on Multimodal Interaction ICMI '15.* Seattle, Washington, USA: Retrieved from `http://dl.acm.org/citation.cfm?id=2820775`

Haider, F., Salim, F A., ,Luz, S., Conlan, O. & Campbell, N. (2015). **High Level Visual and Paralinguistic Features Extraction and Their Correlation with User Engagement**. *In Proceedings of IEEE International Symposium on Signal Processing and Infor-*

mation Technology (ISSPIT 2015) . (pp. 326 – 331) Abu Dhabi, UAE

Haider, F., Cerrato, L.,Luz, S., & Campbell, N. (2016).
Presentation quality assessment using acoustic information and hands movements.
In Proceeding of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (pp. 2812 – 2816). IEEE.

Haider, F., Luz, S., & Campbell, N. (2016)
METALOGUE: Data Collection Using a Real Time Feedback Tool for Non Verbal Presentation Skills Training.
Proceedings of the LREC 2016 Workshop Just talking – casual talk among humans and machines,Portoroz, Slovenia,(pp. 13 – 15)

Hayakawa, A., **Haider, F.**, Luz, S., Cerrato, L., & Campbell, N. (2016).
Talking to a system and oneself: A study from a Speech-to-Speech, Machine Translation mediated Map Task.
In Proceedings of Speech Prosody 2016 (SP8). Boston, Massachusetts, USA: ISCA.

Haider, F., Luz, S., & Campbell, N. (2017).
Data Collection and Synchronisation: Towards a Multiperspective Multimodal Dialogue System with Metacognitive Abilities.
In Dialogues with Social Robots (pp. 245 – 256). Springer Singapore.

Haider, F., Cerrato, L.,Luz, S., & Campbell, N. (2016).

Attitude recognition of video bloggers using audio-visual descriptors

Proceedings of the Workshop on Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction. (pp. 38 – 42) Tokyo, Japan ACM, 2016.

Haider, F., Cerrato, L.,Luz, S., & Campbell, N. (2016).

Active Speaker Detection in Human Machine Multiparty Dialogue using Visual Prosody Information

Proceedings of the IEEE Global Conference on Signal and Information Processing (GlobalSIP) Washington, D.C., USA

Haider, F., Salim, F A., ,Luz, S., Vogel, C., Conlan, O. & Campbell, N. (2017).

Visual, Laughter, Applause and Spoken Expression Features for Predicting Engagement within TED Talks.

In Proceedings of the INTERSPEECH 2017, Stockholm Sweden

Haider, F., Akira, H., Luz, S., Carl, V., & Campbell, N. (2018).

On-talk and off-talk detection: A discrete wavelet transform analysis of electroencephalogram.

In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 960 – 964). Calgary, Canada.

Petukhova, V., Malchanau, A., Oualil, Y., Klakow, D., Luz, S., **Haider, F.**, Campbell, N., Koryzis D., Spiliotopoulos D., Albert P., Linz N, & Alexandersson,

J. (2018).

The Metalogue Debate Trainee Corpus: Data Collection and Annotations.

Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA).

Haider, F., Luz, S., Vogel, C., & Campbell, N. (2018).

Improving Response Time of Active Speaker Detection using Visual Prosody Information Prior to Articulation.

In Proceedings of the INTERSPEECH 2018, (pp. 1736 – 1740), Hyderabad India

Haider, F., Hayakawa, A., Vogel, C., & Campbell, N ,Luz, S.(2018).

Analysing patterns of right brain-hemisphere activity prior to speech articulation for identification of system-directed speech.

(under review)

Other Publications

Cabral, J.P., Campbell, N., Ganesh, S., Gilmartin, E., **Haider, F.**, Kenny, E., Kheirkhah, M., Murphy, A., Chiaráin, N.N., Pellegrini, T. & Orozko, O.R (2014). **MILLA – Multimodal Interactive Language Learning Agent**. *In Proceedings of DialWatt — Semdial 2014. The 18th Workshop on the Semantics and Pragmatics of Dialogue.* (pp 164-166). Edinburgh, UK

Cabral, J.P., Campbell, N., Ganesh, S., Gilmartin, E., **Haider, F.**, Kenny, E., Kheirkhah, M., Murphy, A., Chiaráin, N.N., Pellegrini, T. & Orozko, O.R (2014). **MILLA – Multimodal Interactive Language Learning Agent**. *In Proceedings of eNTERFACE '14 Workshop, June 9th – July 4th*, Bilbao, Spain

Cabral, J. P., Saam, C., Vanmassenhove, E., Bradley, S., & **Haider, F.** (2016) **The ADAPT entry to the Blizzard Challenge 2016**. Proceedings of the Blizzard Challenge, Cupertino, CA, USA

Vanmassenhove, E., Cabral, J. P., & **Haider, F.** (2016). **Prediction of emotions from text using sentiment analysis for expressive**

speech synthesis.

In 9th ISCA Speech Synthesis Workshop, Sunnyvale, USA, (pp=22-27) September 13 – 15.

López A., Ratni, A., Trong, T. N., Olaso, J. M., Montenegro, S., Lee, M. **Haider, F.**, Schlögl, S., Chollet, G., Jokinen, K., Petrovska-Delacrétaz D., Sansen, H., & Torres, M. I. (2016).

Lifeline Dialogues with Roberta.

In Proceedings of the FETLT International Workshop on Future and Emerging Trends in Language Technologies, Machine Learning and Big Data. (pp. 73-85), Springer, Cham

Lee M., Schlögl, S., Montenegro S., López A., Trong T. N., Olaso J. M., Haider F., Chollet G., Jokinen K., Petrovska-Delacretaz D., Sansen H., & Torres, M. I. (2017).

First time encounters with Roberta: a humanoid assistant for conversational autobiography creation.

In eNTERFACE'16, July 18th-August 12th 2016, Enschede, Netherlands 2017.

Salim, F. A., **Haider, F.**, Conlan, O., & Luz, S. (2017, September).

An Alternative Approach to Exploring a Video.

In International Conference on Speech and Computer (pp. 109 – 118). Springer, Cham.

Debattista J., Salim F. A., **Haider F.**, Conran C., Conlan O., Curtis K., Wei

W., Junior A. C., & O'Sullivan D. (2018)

Expressing Multimedia Content Using Semantics — A Vision.

In Semantic Computing (ICSC), 2018 IEEE 12th International Conference on 2018 Jan 31 (pp. 302 – 303). IEEE.

Cakmak H., Haddad K. E., Riche N., Leroy J. , Marighetto P., Türker B. B., Khaki H., Pulisci R., Gilmartin E., **Haider F.**, Cengiz K., Sulir M., Torre I., Marzban S., Yazici R., Bâgci F. B., Kili V. G., Sezer H., & Yengec S. B. (2018)

EASA: Environment Aware Social Agent

In Proceedings of eNTERFACE 2015 Workshop on Intelligent Interfaces, Mons, Belgium.

Salim, F. A., **Haider, F.**, Conlan, O., & Luz, S. (2018).

An approach for exploring a video via multimodal feature extraction and user interactions.

Journal on Multimodal User Interfaces, (pp. 1 – 12).

Haider, F., Salim, F. A., Conlan, O., & Luz, S. (2018).

An Active Feature Transformation Method For Attitude Recognition of Video Bloggers

In Proceedings of the INTERSPEECH 2018, (pp. 431 – 435), September 2 – 6, Hyderabad, India.

Chapter 1

Introduction

This thesis is a collection of various studies which contribute to human-machine multimodal systems in different ways. Some of the studies address basic system capabilities such as the identification of who is talking and to whom in multiparty and multimodal dialogue systems, while other studies focus on aspects, which can be relevant in systems for training presentation styles (attitudes recognition, user engagement etc.) The objective of the research presented in this thesis is to propose models and methods which could help machines in giving the social intelligence abilities that can be used in a multimodal multiparty spoken dialogue systems which are designed for training humans for public speaking. The proposed models and methods are designed using technologies like speech analysis, computer vision and biometry.

Previous studies suggest that the use of multimodal information is helpful for machines to understand the social signals and behaviours of humans (Vinciarelli, Pantic, & Bourlard, 2009; Pentland, 2007). However, fewer efforts have been given to the use of multimodal information for proposing models and methods

that can be helpful for machines in training humans for teaching and instructional advice. Applications of teaching and instructional advice systems are training students/users to deliver oral presentations and training call centre agents to talk according to their audience (teachers, 'on-line viewers' and customers). The social skills of a presenter/instructor have some special characteristics, such as good body language (standing straight, gestures etc) and voice tone that can result in a positive feedback from the audience (Lucas, 2008). To predict the audience's reaction to the presentations using machines, it is required to model the information (e.g. social signals and behaviour cues) that lead to a positive/negative judgement by an audience. However, providing instructional advice using a machine (multimodal spoken dialogue system) with verbal and non-verbal interactive elements presents many challenges. One such challenge is system component failure (e.g. Automatic Speech Recognition (ASR)), which may result in a kind of behaviour that is not as common as in human-human interaction (e.g. self-speaking and frustration). Here the social signals and behaviours cues of human can also help in identifying situations that require certain strategies to be deployed (e.g. switching between different ASR models that are designed for behaviour cues, and handling the self-speaking talks). Compared to human-human interaction, less research has been conducted in analysing and modelling the human social signals and behaviour cues in a human-machine interaction that can give the machines an ability or intelligence to take possible actions for improving the spoken interaction. The task (public speaking skills training through machines) stated above requires some form of intelligence for machines and to highlight that some literature is consulted, and it is found that the definitions of intelligence are diverse and some of them are as follows:

1. According to Oxford dictionary, intelligence is "*the ability to acquire and apply knowledge and skills*" [1].

2. " *the aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment*" (Wechsler, 1958, p.7)

In the literature, it is also argued that there are different measures and forms of intelligence . Some of them are as follow:

1. Intelligent Quotient (IQ): it is the ratio between mental age and chronological age multiplied by 100. The ratio (mental age/ chronological age) is suggested by William Stern (Stern, 1921) and Later Lewis Terman multiplied it by 100 to introduce the IQ. Later David Wechsler introduces Deviation Intelligent Quotient (IQ) which is the deviation of IQ from one's age peers.

2. Experiential intelligence is the ability of someone to use past experience in solving a novel problem (Sternberg, 1990).

3. "*Emotional Intelligence is the ability to identify and manage your own emotions and the emotions of others. It is generally said to include three skills: emotional awareness; the ability to harness emotions and apply them to tasks like thinking and problem solving; and the ability to manage emotions, which includes regulating your own emotions and cheering up or calming down other people.*" [2]

4. "*Social Intelligence encompasses our abilities to interpret others' behaviour in terms of mental states (thoughts, intentions, desires and beliefs), to interact both in complex social groups and in close relationships, to empathize*

[1] https://en.oxforddictionaries.com/definition/intelligence – last verified July 2017
[2] https://www.psychologytoday.com/basics/emotional-intelligence – last verified July 2017

with others' state of mind, and to predict how other feel, think and behave" (Baron-Cohen et al., 1999, p.1).

According to these distinctions social intelligence has its own definition and measures that are different from the other measures of intelligence. To highlight the measure and importance of social intelligence, some literature is explored as described below.

1.1 Social Intelligence

Social signal can be used to express, interpret and recognise the human's intentions. Vinciarelli et al. state that someone's ability to recognise, express and manage social signals and behaviour cues in a communication (dialogue, conversation or presentation, etc.) is core of the Social Intelligence (SI) (Vinciarelli et al., 2009). Social signals and behaviours are also manifested in the non-verbal behaviour cues (e.g. body posture, gestures, facial expressions and prosody). SI is independent of general intelligence and has a vital importance for a person to communicate effectively in his daily life (Vinciarelli et al., 2009). Vinciarelli et al. highlighted two main component of SI, one is behaviour/social cue (physical appearance, gesture and posture, face and eye behaviour, vocal behaviour, Space and Environment) and the other is social behaviour/signal (emotions, personality, status, dominance, rapport etc). However, a social signal/behaviour may produced using the multiple behaviour cues (Vinciarelli et al., 2009).

1.1.1 Why Do Humans Need Social Intelligence?

Humans need Social Intelligence (SI) to effectively deal with each other. SI is required in many job positions, like leaders, teachers, instructor, crisis managers, call centre agents etc.

Goleman and Boyatzis highlighted the importance of social intelligence and define seven metrics to measure the SI of managers and leaders that are:

1. Empathy
2. Attunement
3. Organizational Awareness
4. Influence
5. Developing others
6. Inspiration
7. Teamwork (Goleman & Boyatzis, 2008).

However, the metrics of social intelligence depend on the situations. For example in public speaking situation, a presenters will be evaluated on a subset of the metrics for managers and leaders. It is due to the reason that a manager or leader should be a good public speaker but not all public speakers are managers or leaders. For example, teachers may evaluate a student's public speaking abilities based on a metrics defined in Section 2.1.1 and 3.2.1.

1.1.2 Why Do Machines Need Social Intelligence?

For the past few decades, researchers in the field of signal processing and artificial intelligence have been investigating how to build machines with social intelligence. These researchers are trying to achieve this objective by understanding social aspects of human-human interactions in multiple settings so that they can use these interactions as a model for human-machine interaction (Hjalmarsson, 2010; Edlund, Gustafson, Heldner, & Hjalmarsson, 2008). Currently, many machines are using human-human interaction as a model to make human-machine interaction more natural to some extent (Al Moubayed, Beskow, Skantze, & Granström, 2012). However, it is not yet possible to build a machine that can understand and interact with humans as other humans do. Many dialogue systems can help users in filling out a form (book tickets, buy goods, etc.), but these systems do not require social intelligence in the context of public speaking (e.g. recognising self-confidence using prosody and gestures) to function. Moreover, instruction based spoken dialogue systems can function without the ability of understanding human social cues and behaviour, particularly non-verbal aspects of social cues and behaviours (gestures, voice quality, facial expressions, etc.), where they are just following the verbal instructions given by a user such as booking a ticket (Zue et al., 1994). However an instruction based dialogue system can also get benefit from non-verbal cues like gestures. Where the gestures (e.g. yes or no head gestures) can improve the interaction with an instruction based spoken dialogue system for users. The inference drawn by instruction based spoken dialogue system and dialogue system for public speaking training can be different. In summary, the SI is present in both systems but a public speaking training system should have an ability to draw

more complex inference (e.g. self-confidence of a presenter through gestures) than an instructional spoken dialogue system (e.g. simple yes or no head gestures).

Recently, efforts have been given to training machines for complex tasks like acting as an instructor (Hashimoto, Kato, & Kobayashi, 2011) or receptionist (Hashimoto, Hiramatsu, Tsuji, & Kobayashi, 2007). For these kinds of tasks, machines need more social intelligence than instruction based spoken dialogue systems, as they, need to teach the humans using multimodal cues (eye gaze, gestures, human-like face and body, etc.) and to evaluate the social behaviour and signals of users in the context of public speaking. For example, a machine which is designed to train students for delivering presentations is not able to provide feedback to the users about their appropriate use of 'body gestures and posture', and 'speech and facial expressions' without some form of social intelligence ability (Helvert, Rosmalen, Börner, Petukhova, & Alexandersson, 2015). However, using machines for instructional advice and training purposes is still an area under research, and there is no general agreement as to how systems can use the users and instructor social behaviour as a model. For this purpose, the system should be able to understand a user's behaviour in the context of the instructional advice (goals and objectives).

1.1.3 Research Issues

To be able to study the use of multimodal information for social intelligence in the context of public speaking, a wide range of scenarios of interactions (detailed and motivated in Section 1.1.4) are considered. The social cues and technologies associated with each research issue are depicted in Table 1.1. The specific research

Table 1.1: Social cues and technologies associated with each research issue.

Research Issue	**Technology**	**Social Cue Type**
Active Speaker Detection	Computer Vision	Face Gestures
On-Talk Detection	Biometry, Speech Analysis	Prosody
Attitude Recognition	Computer Vision, speech Analysis	Prosody, Body Gestures
User Engagement Detection	Speech Analysis, Computer Vision	Prosody, Body Gestures
Cognitive State Detection	Biometry, Speech Analysis	Prosody
Presentation Quality	Speech Analysis, Computer Vision	Prosody, Hand Gestures

issues are as follows:

1. Presentation Quality Detection: What are the low and high-level multimodal features that contribute towards the successful delivery of a student's presentation, and how can machines automatically detect the level of a student's presentation skills (e.g. use of adequate voice level, body gestures and self-confidence) using multimodal information? Success in this research improves the social intelligence of machines by processing prosody and gestural information of humans and making explicit parameters that may be incorporated into algorithmic models.

2. User Engagement Detection: Which low and high-level multimodal features contribute to engagement of on-line users for formal talks (TED talks). How can machines automatically detect whether a public talk is engaging or not for on-line audience using multimodal information? Success in this research improves the social intelligence of machines by processing prosody and face information of humans and enhancing computational models accordingly. It can also provide a feedback to users by detecting which parts of the talk are engaging and which are not. It can also help building a recommender, talk search and talk summarization tool for potential viewers.

3. Attitude Recognition: How can machines automatically detect attitudes (states that may permeate strong emotions (Zanna & Rempel, 1988)) of video bloggers (informal talks) using audio-visual information? Success in this research can improve the social intelligence of machines by processing prosody and gestural information of humans. It could help someone to train his/her attitudes for video blogging.

4. System Directed Speech Detection: How can machines automatically detect the system directed speech (On-Talk: user is speaking to the machine) using acoustic and physiological signals? Success in this research can improve the social intelligence of machine by processing prosody information of humans and presumably enabling the machine to identify the addressee of the speech as being itself or not.

5. Cognitive State Detection: How can machines automatically detect the cognitive states using acoustic and physiological signals? Success in this research can improve the social intelligence of machine by processing prosody information of humans for identifying situations that require certain interaction strategies to be deployed (e.g. switching between different automatic speech recognition models which are designed for different cognitive states) and sensing of users' experience.

6. Active Speaker Detection for Attunement: How can the visual prosody (head and lip movements) help a machine to detect an active speaker (who is speaking to machine out of a set of persons) in a human-machine multiparty dialogue? This research issue improves the social intelligence of machine by processing face (lip and head movements) information of humans.

The above stated research issues help in developing a multimodal multiparty spoken dialogue system for public speaking training which can evaluate one's presentation from different perspective (teachers, on-line viewers etc.), helps users in training their non-verbal behaviours (attitudes, spoken expressions, etc.), able to manage multiple users at a time and improve the spoken interaction.

1.1.4 Scenarios

Developing a multimodal spoken dialogue system for public speaking training requires many components like automatic speech recognition, speech synthesis, dialogue manager, a human like body or avatar, prosodic analysis, face detection, voice activity detection, speaker recognition, recognising someone presentation skills etc. However research issues of the thesis focus on two main parts, first one is recognising someone's public speaking abilities using social signals, and the second part focuses on proposing methods and computational models of cognitive processing components (e.g. active speaker detection) which could help a public speaking training system (multimodal spoken dialogue system) to interact with humans. The data used to investigate these research issues comes from experiments conducted on both human-human and human-machine communication as described below.

Public Speaking

There are different scenarios of human-human communication including a casual talk with a friend or stranger, getting information from a receptionist, speaking on the phone with a customer care representative and public speaking (presentations,

talks, debates). However, this study mainly focuses on the three different kinds of public speaking situations where there is no live verbal interactive element, but in two situations a live non-verbal interactive element of an audience is present. It helps in training machines to recognise public speaking abilities in three different kinds of public speaking situations as described below.

1. Students presentations (Classroom settings): In this scenario, a student is presenting in front of an audience, and at the end of the presentation each student is graded by a tutor. This scenario is related to the research issue of "Presentation Quality Detection".

2. TED Talks: These talks are more professional than students' presentations. The subjects prepare and practise before presenting in front of an audience. Later, the talks are edited and made available for the on-line community for feedback. This scenario is related to the research issue of "Engagement Detection".

3. Informal talks (Video Blogs): These kinds of talks are informal and delivered in front of a camera/computer without having a live audience feedback. The subjects can have a possibility to improve and gauge their presentations before making them available to the on-line community for feedback. This scenario is related to the research issue of "Attitude Detection".

The motivation (detailed in chapter 3) behind using these scenarios is to propose models and methods which can help machines in evaluating public speaking abilities of humans in class room, formal and informal settings.

Cognitive Processing Components

This study focuses on two human-machine scenarios that have a live interactive element (both verbal and non-verbal) between the participants of interaction. These scenarios help in proposing models and methods which can be used as cognitive processing components in public speaking training systems for improving machine interaction abilities. The scenarios are as follows:

1. Human-machine multi-party dialogue, where subjects are conversing with a machine (simulated using video conference software). This scenario is related to the research issue of "active speaker detection". It could help a public speaking training system to manage multiple users at a time.

2. Interlingual map-task mediated by an automatic speech-to-speech translation system, where participants are solving an interlingual map task using speech-to-speech machine translation system. It is a simulation of call centre settings (a public speaking situation) where the call centre agent (information giver in map task who has the complete map) has the full information and the customer has incomplete information (information receiver in map task who has the incomplete map). This scenario is related to the research issue of "Cognitive State and System Directed Speech Detection". Cognitive states detection could help a public speaking training training system to switch between different ASR models which are trained for different cognitive states to improve the performance of ASR and to sense the user experience. System directed speech detection could help a public public speaking training system, whether the user is speaking to the system or not.

The motivation (detailed in chapter 4) behind using these scenarios is to propose models and methods which can improve machines interaction abilities.

1.2 Contribution of The Thesis

The work conducted in answering the above-mentioned research questions has resulted in many novel methods and findings. A brief description of author's contribution is described below.

1. **Presentation Quality Assessment:** This study validated hypotheses that relate self-confidence and use of body language during presentation to prosodic and gestural features, and propose a novel system for automatic inference of presentation quality using audio and video descriptors. The proposed system is able to predict the students' self confidence and use of body language during a presentation. The system and evaluation results are described in Section 3.2. The data collection and annotation is not the author's contribution but the proposed methods and their evaluation as follow:

   (a) Validation of multiple hypothesis detailed in Section 3.2.2.

   (b) Automatic presentation quality detection system using audio and visual information detailed in Section 3.2.3.

2. **User Engagement Detection within TED Talks:** This study explores how multimodal characteristics of a video, such as prosodic, visual and paralinguistic features, can help create novel models for user engagement detection of TED talks. It proposes novel models to predict the user's (on-line viewers) engagement using high-level visual features (camera angles), the

audiences laughter and applause, and the presenter's speech expressions features which are extracted directly from video recordings, demonstrating the potential of this method in identifying engaging TED talks and automatic identification of engaging video segments within TED talks. The system and evaluation results are described in Section 3.3. The author contribution here is system designing for user engagement detection, experimentation and evaluation. The collection of dataset and extraction of high level features is not the author's contribution but the proposed methods and their evaluation as follow:

(a) Validation of multiple hypothesis detailed in Section 3.3.

(b) Automatic user engagement detection system using audio and visual information detailed in Section 3.3.

3. **Attitude Recognition of Video Bloggers:** In the scope of this study the attitude represents affects. This study uses the acoustic and visual features (body movements that are captured by low-level visual descriptors) to propose novel models which can predict the six different attitudes (amusement, enthusiasm, friendliness, frustration, impatience and neutral) annotated in the speech of 10 video bloggers. The automatic detection of attitude can be helpful in a scenario where a machine has to automatically provide feedback to bloggers about their performance in terms of the extent to which they manage to engage the audience by displaying certain attitudes. The system and evaluation results are described in Section 3.4. The data collection and annotation is not the author's contribution but the proposed methods and their evaluation as follow:

(a) Validation of multiple hypothesis detailed in Section 3.4.

(b) Automatic attitude detection system using audio and visual information detailed in Section 3.4.

4. **Political Debates: Data Collection and Synchronisation** This study describes the data collection system and methods. The data collection scenario consists of debates where two students are exchanging views and arguments on a social issue, such as a proposed ban on smoking in public areas, and delivering their presentations in front of an audience. Approximately 3 hours of data have been recorded to date, and all recorded streams have been precisely synchronised and pre-processed for statistical learning. The data consists of audio, video and 3-dimensional skeletal movement information of the participants. The data collection system and methods are described in Section 3.5 The data collection is the author contribution.

5. **System Directed Speech Detection:** This study proposes novel models to automatically detect On-Talk (system directed speech) and Off-Talk (speaking to oneself and speaking to another person) using acoustic and physiological signals. This study also focuses on proposing novel models which have a low response time for detection using Electroencephalography (EEG). The EEG based models can help in a system as there are many less intrusive wireless EEG headsets available in the market e.g. EPOC[3] . The system and evaluation results are described in Section 4.2. The data collection methods and system are not the author contribution but the proposed methods and their evaluation as follow.

[3]https://www.emotiv.com/epoc/ – Last verified July 2018

(a) Validation of multiple hypothesis detailed in Section 4.2.

(b) Automatic attitude detection system using audio and physiological information detailed in Section 4.2.

6. **Cognitive States Detection:** This study proposes novel models for automatic detection of cognitive states (i.e. temporary psychological states which are annotated using facial expressions) by using features of the two physiological signals, features of the speech signal, and combinations of speech and physiological features (heart rate and skin conductance). In the scope of this study, the cognitive states represents affects. The system and evaluation results are described in Section 4.3. The data collection methods, system and annotation are not the author contribution but the proposed methods and their evaluation as follow.

   (a) Automatic cognitive state detection system using audio and physiological information detailed in Section 4.3.

7. **Active Speaker Detection:** This study proposes a novel active speaker detection system using visual prosody information (i.e. head and lip movements) for human-machine multiparty dialogue that could help a robot in generating multimodal output (e.g. moving his head or gaze) towards the speaker. This study also focuses on proposing novel models which have a low response time for detection. The system and evaluation results are described in section 4.4. The data collection and its annotation is not the thesis contribution but the proposed methods and their evaluation.

   (a) Validation of multiple hypothesis detailed in Section 4.4.

(b) Automatic active speaker detection system using audio and visual information detailed in Section 4.4.

The contribution of the thesis is towards developing multimodal multiparty spoken dialogue systems for public speaking training. The presentation quality, user engagement and attitude detection systems help a machine (multimodal multiparty spoken dialogue system for public speaking training) in detecting humans non-verbal presentation skills like (attitude, self-confidence, body language and engagement). Without them it is not possible for machines to recognise someone's presentation skills which are manifested in the social cues (e.g. prosody). The system directed speech detection system helps a machine in improving the spoken interaction with humans. Without the system directed speech detection, a machine is not able to detect either the user is speaking to machine or not which may results in processing of spoken utterances that are not directed towards the machine. The cognitive state detection system helps a machine in improving automatic speaker recognition performance. Without the cognitive state detection system, a machine (multimodal multiparty spoken dialogue system for public speaking training) is not able to switch between different ASR models, which are designed for different cognitive states which may results in poor performance of ASR as cognitive states effects the ASR performance, and not able to sense the user experience. The active speaker detection system helps a machine in detecting who is speaking to the machine. Without the active speaker detection system, a machine (multimodal multiparty spoken dialogue system for public speaking training) can not manage multiple users.

1.3 Organization of The Thesis

Chapter 1 provides an overall introduction, thesis focus, research aim, research issues and contribution of the thesis.

Chapter 2 focuses on the theoretical background and literature review of research issue described in Chapter 1.

Chapter 3 describes the systems which are able to recognise someone's presentation skills. This chapter focuses on recognising students' 'self-confidence and body language' during a presentation, on-line viewer engagement detection within TED talks and recognising attitudes of video bloggers. The dataset used and collected is also the part of this chapter and it also describes the empirical studies, the evaluation of proposed systems and conclusion along with signal processing and machine learning methods.

Chapter 4 describes the systems which could help a machine (multimodal multiparty spoken dialogue system) designed for public speaking training. This chapter focuses on active speaker detection, system directed speech detection and cognitive states detection. The description of datasets used are also the part of this chapter and it also describes the empirical studies, the evaluation of propose systems and conclusion along with signal processing and machine learning methods.

Chapter 5 describes the conclusion and future work of the thesis.

1.4 Conclusion

This chapter describes the different measures of intelligence particularly social intelligence. It also describes the research issues, scenarios, contribution and organisation of thesis. The thesis mainly focuses on two parts. The first one is to automatically recognise public speaking abilities in different situations and second part deals with the challenges which a multimodal multiparty spoken dialogue system may face for training someone for public speaking. The political debates data collection and synchronisation is the contribution of the thesis. The rest of the datasets are not the contribution of the thesis but the proposed models and methods. The most of the machine learning models are evaluated using the F-Score (harmonic mean) of each class or the A-weighted F-Score (average harmonic mean of all classes) and the rest of the machine learning models are evaluated using the accuracy. In the scope of this thesis, the attitude and cognitive states represents affects.

Chapter 2

Theoretical Background and Literature Review

This chapter covers theoretical background and literature of research issues of the thesis. It is divided into two main sections for clarity of presentation. Section 2.1 focuses on automatic recognition of public speaking abilities and provides a review of literature on public speaking abilities, methods used for automatically predicting the oral presentations skills and attitudes. Section 2.2 focuses on the some of the challenges which a multimodal multiparty spoken dialogue system encounters while interacting with users during a public speaking training session. It also provides a literature review on On-Talk & Off-Talk, active speaker and cognitive states detection.

2.1 Public Speaking Metrics and Systems

This section describes the different pubic speaking situations and the metrics that contribute towards the successful delivery of public speeches. This section also provides a background and literature review on three different public speaking situations as follow:

1. Student presentations which are graded by teachers.

2. TED Talks (formal talks) which are viewed by on-line viewers.

3. Video Blogs (informal talks) which are viewed by on-line viewers.

2.1.1 Student Presentations

Prosody is believed to be of fundamental importance in contributing to the success of a public speech. Several manuals on public speaking advise the presenter to speak with a lively voice, where by lively voice is meant a voice that varies in intonation, rhythm and loudness (Lamerton, 2001; Grandstaff, 2004). Liveliness has also been associated with enthusiasm (Sinclair, 1995). Previous studies have formulated and tested the hypothesis that the higher the variability (or standard deviation) of fundamental frequency (F0), the more a spoken utterance is perceived as lively (Traunmüller & Eriksson, 1995; Hincks, 2005). However F0 deviation alone might not always be an optimal feature discriminating lively speech from monotonic speech (typical of depressive states) (Stassen et al., 1993). Another aspect that seems to contribute to the success of public speech is speaking rate. This has shown to be more strongly correlated than pitch variation with perceptions of liveliness (Traunmüller & Eriksson, 1995) and has been considered,

together with voice level and intensity, as an indicator of self confidence. Fast rate of speech, lower voice level and high speech intensity are listed among the characteristics of self confident voices in several studies (Grandstaff, 2004; Lamerton, 2001). Other characteristics believed to contribute to the success of a presentation include the speaker's ability to establish contact with their listeners (e.g. eye contact) and be aware of their body language. Specific postures that supposedly denote self confidence, such as standing straight with feet aligned under the shoulders, are recommended by public speaking guides. Other postures that denote the lack of self confidence, such as fidgeting, crossing the legs, gesturing widely without purpose are considered inappropriate (DeCoske & White, 2010). Automatic detection of someone's presentation abilities is a challenging task. Few studies have been conducted in this field. In one study on automatic detection of self confidence (Krajewski, Batliner, & Kessel, 2010), the authors compared multiple classifiers using a set of prosodic and spectral features, on a very limited dataset consisting of fourteen females speakers giving regular lectures ranked by 5 experts judging self confidence. The classifiers are able to detect two classes (low self confidence and high self confidence) with a maximum accuracy of 87.7% and 75.2% for speaker-dependent and speaker-independent settings respectively. There are other studies conducted on the MLA dataset (Ochoa, Worsley, Chiluiza, & Luz, 2014a).

Luzardo et al. (Luzardo, Guamán, Chiluiza, Castells, & Ochoa, 2014) employ features extracted from presentation slides to predict overall presentation quality (2-class problem), obtaining up to 65% accuracy. When audio features are used, pitch and filled-pause related features improve accuracy to 69%. Chen et al. (L. Chen, Leong, Feng, & Lee, 2014) propose a different approach, performing

a clustering of presentation ratings to derive two principal components (roughly corresponding to delivery skills and slide quality) which they use as the target functions of a regression task. Finally, Echeverria et al. (Echeverría, Avendaño, Chiluiza, Vásquez, & Ochoa, 2014) employed machine learning models to classify presentations according to performance (good vs. poor), achieving accuracy scores of 68% and 63%.

2.1.2 TED Talks

There is an enormous amount of audio-visual content (videos) also available on-line in the form of talks and presentations. The prospective users of the content face difficulties in finding the right content for them. Take *YouTube* as an example: over a billion hours of video content are watched daily[1]. Because of the amount of video content available, it is becoming increasingly difficult for users to find desired content. A recent study reports that an average American would spend more than a year over a lifetime looking for something to watch on TV[2]. One criterion for filtering content is how engaging a video is. Hence, a model to detect user perceptions of engaging verses non-engaging talks would be beneficial for any number of applications, including video recommendation and video searching.

The set of videos available to viewers is very diverse, and each kind of video engages users differently or to put it differently, users watch different types of videos for various reasons, i.e. engagement with content is context dependent (Attfield, Piwowarski, & Kazai, 2011). This section focuses on one video genre which is

[1]https://www.engadget.com/2017/02/27/youtube-one-billion-hours-watched-daily/ – – last verified: March 2017

[2]http://gizmodo.com/is-it-a-bad-thing-that-we-spend-1-3-years-of-our-lives-1788632578 – – last verified: May 2017

video presentations such as TED talks. To distinguish between engaging versus non-engaging TED talks, it is necessary to define the meaning of user engagement within the context of TED talks and how to quantify it. In the literature, the quality of user (human) experience with a system is called user engagement (Albers & Mazur, 2014; O'Brien & Toms, 2008) and a six-factor based matrix is proposed by OBrien et al. (O'Brien & Toms, 2013) for user engagement. In terms of video content, researchers have described user engagement with video content in a variety of ways, e.g. duration for which a user watches a video (Dobrian et al., 2013; Guo, Kim, & Rubin, 2014) and subjective evaluation of user engagement through questionnaires (Benini, Migliorati, & Leonardi, 2010; Haesen et al., 2011). It can be seen that there is not much agreement in measuring engagement due to its highly context dependent nature. For this study, an elaborate feedback system (described in detail in Section 3.3.1) is used to define the engagement.

Wernicke conducted statistical analysis on TED talks and proposed a metric for creating an optimal TED talk based on user ratings (TED, 2010). A major difference between his study and the current work is that the TED user ratings on which the former study is based were considerably simpler i.e. viewers could simply 'like' or 'dislike' a particular TED talk. Recommender system development for viewers based on their viewing/listening preferences and commenting patterns has attracted considerable interest. For example, Tan, Bu, Qin, Chen, and Cai use heterogeneous data from different sources to create a recommender system based on user video preferences (Tan et al., 2014). Brezeale and Cook use movie subtitles and low-level visual descriptors to cluster the data (videos), and then use Hidden Markov Model (HMM) to learn the sequence of clusters to predict the users' preferences (Brezeale & Cook, 2009). Anwar, Salama, and Abdelhalim

proposed to file videos into different categories using the caption text and visual features (Anwar et al., 2013).

A significant amount of research is currently being conducted within the field of video summarization based on user engagement. In video summarization, importance is mostly attributed to visual features (Benini et al., 2010; F. Chen, De Vleeschouwer, & Cavallaro, 2014). However, multimodal features are also receiving considerable attention due to the added value they bring in terms of identifying engaging chunks. An example of comprehensive multimodal feature extraction is the work of Evangelopoulos et al. who take advantage of all three visual, audio and linguistic modalities to create video summaries (Evangelopoulos et al., 2013). Other interesting examples of multimodal feature extraction are work of Dong and Li (Dong & Li, 2008) and Haesen et al. (Haesen et al., 2011). Extracting all these multimodal features and indexing them would make videos more searchable and would also help correlate videos with user feedback , and thereby user engagement. In terms of assessing the quality of a presentation, there have been many studies which focus on the analysis of speech utterances and body gestures (Haider, Cerrato, Campbell, & Luz, 2016; Curtis, Jones, & Campbell, 2016). The Multimodal Learning Analytics (MLA) dataset contains student presentations graded by teachers in terms of body language, self-confidence, loudness level, eye contact and slides content (Ochoa, Worsley, Chiluiza, & Luz, 2014b). While presentation rating in the educational context is a well-researched area, the factors affecting presentation ratings by ordinary viewers and listener have hardly been investigated.

2.1.3 Video Blogs

A vlog is a form of unidirectional communication where the vlogger (video blogger) does not receive feedback from the viewers in real time, but the viewer can provide their feedback later in the form of textual comments. Vloggers can also improve their presentation quality and message before uploading it for the audience. In previous studies conducted on video blogs, it was found that the non-verbal behaviour influences the level of attention gained by a video (Biel, Aran, & Gatica-Perez, 2011). So, an automatic detection of non-verbal behaviour for video bloggers could be useful for video bloggers in gauging their behaviour by providing them with feedback. Usually, the non-verbal behaviour is modelled using the prosody and visual features in terms of facial movements. However, gestures are also a form of non-verbal behaviour, in particular, hand gestures (McNeill, 2008) which correlate with the semantic concept and rhythm of speech (McNeill, 1992). In the field of affective computing, different methodologies are proposed to detect the affective and emotional states in different contexts ranging from human-human to human-machine communication (Liu et al., 2014; Akira, Haider, Cerrato, Campbell, & Luz, 2015; Vogel & Mamani Sanchez, 2012). This study investigates attitude detection in video blogs (vlogs).

The analysis of vlogs has not been explored extensively in the literature. In one study, the facial expression, acoustic and the multimodal information is used to predict the personality traits in vlog using regression analysis (Biel, Teijeiro-Mosquera, & Gatica-Perez, 2012). In another study, a perceptual and acoustic analysis is performed for 12 different attitudes expressed by Portuguese speakers. The results show that the audio-visual information provides a better perception of

attitude than any of the single modality (De Moraes, Rilliard, de Oliveira Mota, & Shochi, 2010). An analysis of speaking time, pitch energy, voice rate, speech turn along with head motions, looking time and proximity to camera (Biel & Gatica-Perez, 2011) in terms of Pearson's correlation (between non-verbal cues and the median number of log views) shows that the audio-visual cues are significantly correlated with the median number of log views. In another study, Allwood and Henrichsen propose an automatic attitude detection system for multimodal dialogue using acoustic features (Allwood & Henrichsen, 2013).

In a study conducted on a subset of the data used in this thesis, the N. A. Madzlan, Huang, and Campbell (N. A. Madzlan et al., 2015) analysed the acoustic and high-level visual features to train a classifier to detect the attitude automatically. The authors propose a three-class problem grouping the attitudes in the following three classes: positive, negative and neutral attitudes. They defined friendliness attitude as neutral. 'Amusement and Enthusiasm' as positive attitudes, and 'Frustration and Impatience' as negative attitudes. The results show that the acoustic features (63.63%) provide better results than the visual features (50.6%), but the authors do not perform the fusion of features (N. A. Madzlan et al., 2015). In one study, N. A. Madzlan, Han, Bonin, and Campbell analyse the prosodic features of vlogger and found that the prosodic features (pitch, voice quality and intensity) are correlated with a vlogger attitude (N. Madzlan, Han, Bonin, & Campbell, 2014). The vloggers' audio-visual features are analysed for the attitude recognition (N. A. Madzlan et al., 2014). There is no study conducted on the multimodal (audio-visual) attitude recognition system proposed for vloggers.

2.2 Cognitive Processing Components for Interactive systems

The social cues (prosody, head and lip movements) can be used for the purpose of improving machine interaction ability with humans. Social cues can also be used to detect the emotions, engagement and improve the understanding of verbal aspect of communication. This section focuses on three kind of improvements as follow:

1. Detection of active speaker in a human-machine multiparty dialogue. This ability can help a machines in managing a dialogue with multiple humans.

2. On-Talk & Off-Talk Detection: Either a subject speaking to a system or not. This ability allows a machine to not process the speech utterance which are not directed to the machine, hence allowing the machine to keep silent if the speech utterance is not directed to it.

3. Detection of cognitive states in a human-machine interaction. This ability can help a machine to switch between different machine learning models of ASR which are trained for different cognitive states, and can also help a machine in sensing user experience with the system.

2.2.1 Active Speaker Detection

Dialogue has two main components, one is verbal, and other is non-verbal. In order for a machine to be able to manage these two components when engaged in a dialogue with several humans it needs to be able to detect which speaker

holds the floor. If the machine is to be seen as a believable participant in the communication, it should seem to turn its visual attention towards the current active speaker, and achieve realistic production and attunement to gaze, lip and head movements.

An active speaker detection system can be used in a robot to aid the generation of the multimodal output (moving its head or gaze towards the speaker) particularly in situated interactions (Han, Gilmartin, & Campbell, 2013; Christian & Avery, 1998; Breazeal, 2003; Cech et al., 2013; Sansen et al., 2016). In Human-Human interaction, it is observed that the listener turns their gaze towards the speakers around 30–80% of the time (Kendon, 1967). From the social robotics perspective, it is useful to detect the active speaker as soon as possible to enable the robot to gaze/head towards the speaker to show that it is attending to the speaker. In particular, it is useful if one may anticipate who the next active speaker will be, in order to speed this process.

Multiple studies explore the use of visual information and its fusion with acoustic information to increase the performance of voice/speaker detection. Takeuchi, Hashiba, Tamura, and Hayamizu extract the low-level visual descriptors (optical flow vectors) from the mouth region for speech activity detection (Takeuchi et al., 2009). Viola, Jones, and Snow use the appearance and motion cues of humans (Viola et al., 2005) for speaker detection, with a dataset collected in a distributed meeting setting using smart rooms (Zhang et al., 2008). Some studies use the audio and visual features to detect the speaker in videos and human-machine interaction (Chakravarty, Mirzaei, Tuytelaars, et al., 2015; Pavlović, Garg, Rehg, & Huang, 2000; Cutler & Davis, 2000; Cech et al., 2013).

The facial dynamics of a person can be used to detect the voice, and can also

be helpful in predicting an active speaker from a set of subjects facing a robot. In previous studies, lip movements are considered an important signal which can increase the speech intelligibility (Sumby & Pollack, 1954; McGurk & MacDonald, 1976; Bernstein, Tucker, & Demorest, 2000). While head movements have been less explored in this kind of task, perception studies have found that head movements are correlated with prosody and improve the speech intelligibility (Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004), as well as reveal prosodic structure (Graf, Cosatto, Strom, & Huang, 2002). Ishii, Kumano, and Otsuka use manually annotated mouth opening transition patterns after the subject stops speaking to predict the next speaker in a meeting (Ishii et al., 2016). Cech et al. report on an active speaker detection system for a humanoid robot that uses audio and visual information of four microphones and two cameras (Cech et al., 2013). Other studies propose models to detect speakers in videos (Chakravarty et al., 2015; Pavlović et al., 2000; Cutler & Davis, 2000), where the objective is not to generate a real-time multimodal output (gaze and head) of a robot.

2.2.2 System Directed Speech Detection

It has been observed that when people interact with computer systems, not only do they talk to the computer system but they also tend to talk to themselves and naturally to other people if present (Oppermann, Schiel, Steininger, & Beringer, 2001; Batliner, Hacker, & Nöth, 2006; Hayakawa, Haider, Luz, Cerrato, & Campbell, 2016a). Oppermann et al. coined this interaction, that is not addressed to the computer system as "*Off-Talk*" and utterances that are directed to the computer system, and therefore need to be understood by the system as "*On-Talk*". Batliner

et al. open their paper with an example from Shakespeare's *Hamlet*, where Hamlet seems change his speaking style when addressing his interlocutor to utterances that are spoken, but not directed towards his interlocutor, show that this is not a new phenomena, but part of human nature that Shakespeare expressed with his characters.

The definition of *Off-Talk*, as provided by Oppermann et al. is every utterance that is not directed to the system: e.g., (i) soliloquy/thinking aloud, (ii) swearing, (iii) reading from displayed text aloud, (iv) conversation with other person(s) present, (v) telephone conversation (e. g. with cellular phone) and (vi) extrinsic speech (e. g. video player, TV set, etc.) (Oppermann et al., 2001, p. 1).

The objective of this research is not to redefine these concepts of *On-Talk* and *Off-Talk*, but to propose a system which can automatically detect these two types of talk.

Previous studies by Oppermann et al. report that the loudness difference between *On-Talk* and *Off-Talk* can be used as a significant indicator of *Off-Talk* (Oppermann et al., 2001). However, a system that is trained using audio which is recorded in a controlled acoustic environment (minimising the reverberation, background noise and competing speech, etc.) may perform poorly in an uncontrolled acoustic environment . Another disadvantage of a system which is only based on acoustic modality is that it needs a speaker recognition (Gerl & Herbig, 2008; Reynolds, 2002) module so that the system only process the user's speech utterances and not the subjects speaking behind or with the user. The visual information may also be helpful in an uncontrolled acoustic environment, but then a subject needs to face a camera, and visual environment (e.g. lightning conditionings) also affect the results. However, physiological information can help in

uncontrolled acoustic and visual conditions. The user does not need to face a camera and the system doesn't need a speaker recognition system because the person who is wearing the electronic device (wireless EEG cap or smart watch for heart rate and skin conductance) is the user.

The electroencephalogram (EEG) signal and its different frequency bands have been employed in some applications, such as seizure detection, emotion recognition, and even speech recognition. Ocak analyses the frequency bands between 0–86.8 Hz using wavelet transform, and reports that the higher bands between 43.4–86.8 Hz provides the optimum accuracy for detection of epileptic seizures (Ocak, 2009). Adeli, Ghosh-Dastidar, and Dadmehr use a wavelet chaos methodology to detect seizure using EEGs and EEG sub-bands and analyse EEG signals between 0–60 Hz (Adeli et al., 2007). However, for both studies, the data is collected in very controlled settings. Petrantonakis and Hadjileontiadis use the lower frequency bands between 8–12 Hz and 13–30 Hz for emotion recognition (Petrantonakis & Hadjileontiadis, 2010). The EEG signal has also been used for the speech recognition of unspoken words (Porbadnigk, Wester, & Jan-p Calliess, 2009) where Porbadnigk et al. recorded the 16 EEG signal channels with a 128 cap montage and recognised five words with an average accuracy of 45.50%. The most prominent band of the EEG signal lies in the lower frequencies (Alpha band for attentional demands and Beta band for emotional and cognitive processes (Ray & Cole, 1985)).

The EEG signal is quite susceptible to artefacts caused by talk-related muscle activity, including head movement and eye blinks. This problem is commonly approached by recording signals on several different positions on the scalp, instructing the subjects to avoid moving and to keep calm during recordings, and subsequently employing independent component analysis on the EEG data in or-

der to remove the artefacts (Delorme & Makeig, 2004). However, in an interactive setting, it is not possible to restrict subjects' movements; they have to move their heads, speak, display emotions, gesture and laugh. Therefore, removing all these artefacts becomes even harder if any amount of naturalness in human-computer interaction is to be preserved.

Muscle activity can introduce noise in EEG signals (e.g. peak frequency of masseter muscles movements are in 50–60 Hz range, and frontalis muscles movements are between 30–40 Hz), and the noise band limit is between 15–100 Hz (D O'Donnell, Berkhout, & Adey, 1974). Kumar, Narayan, and Amell (Kumar et al., 2003) also report a noise range for frontalis muscles between 20–30 Hz and temporal muscles between 40–80 Hz. Posterior head muscle movements have a higher peak frequency close to 100 Hz, but it depends on many factors (e.g. sex, force and direction of contraction, etc.) (Kumar et al., 2003). Muscle activity may introduce artefacts in EEG signal in a frequency range (≈20–300 Hz) where the most artefacts are at the lower end (Criswell, 2010). However, the use of physiological signals (including the EEG signal) for speech related task in noisy and competing speech environment is well recognised. For example, an EEG-based voice activity detection helps in recording/processing the speech utterances of the system's user only (Von Borstel, Esquivel, & Meyer, 2015). It is also found that the right hemisphere of the brain is responsible for the speech prosodic characteristics (Shapiro & Danly, 1985; Weintraub, Mesulam, & Kramer, 1981; Ross & Mesulam, 1979) and Heart Rate (HR) and Skin Conductance (SC) also help in predicting the cognitive states (Hayakawa, Haider, Cerrato, Campbell, & Luz, 2015), emotions (Matejka et al., 2013), and On-Talk and Off-Talk (Hayakawa, Haider, et al., 2016a).

Muscle artefact noise is the main reason why the EEG signal has not been used for speech related applications. However, some studies addressed this problem by analysing the EEG signal using overt and covert speech settings to minimise muscle artefacts noise. The EEG signal for a speech related task is evaluated using the covert speech production settings (van Turennout, Hagoort, & Brown, 1997; Schmitt, Münte, & Kutas, 2000; Schmitt, Schiltz, Zaake, Kutas, & Münte, 2001; Abdel Rahman, van Turennout, & Levelt, 2003) to minimise the muscle activities, where the subject is thinking about a word instead of articulating the word. A limitation of this methodology is that it can not be verified whether the subject followed the task instruction or not. However, for overt speech (Duncan-Johnson & Kopell, 1981; Liotti, Woldorff, Perez, & Mayberg, 2000), the EEG signal can be analysed after a stimulus is presented to the subject until the start of articulation to avoid the noise due to muscle activity related to speech articulation.

The Broca part of the human brain plays a role in speech production (Stone, 1991; Whitaker, 1970). The seminal research by Broca, Wernicke and others on the relationship between neural activity and speech production, which highlighted parts of the brain responsible for speech production has been supported by a number of studies (Blank, Scott, Murphy, Warburton, & Wise, 2002). It has been observed that the speech signal is preceded by low variation in EEG Signal up to one second before the articulation (McAdam & Whitaker, 1971). The cognitive processes that lead to speech articulation (activate the speech production areas in a brain) are of three main types (Bock, 1982; Dell, 1986; Garrett, 1975, 1988; Kempen, 1977; Kempen & Hoenkamp, 1987; Levelt, 1989). 1. Conceptualization: the content and pre linguistics representation of intended speech 2. Formulation: retrieval of the best match between linguistic representation and conceptual struc-

ture 3. Grammatical and Phonological encoding: Selection of lexical items and intonation pattern (van Turennout et al., 1997)

Electro-physiological evidence of phonological encoding that leads to articulation is found. M. Van Turennout et al analysed the EEG signal from the midline frontal (Fz), central (Cz), and parietal (Pz) sites of the 10-20 system (for EEG electrode placement (Jasper, 1958)) in a picture naming task (van Turennout et al., 1997). Some studies highlighted the right hemisphere of brain for the control of speech prosody (Shapiro & Danly, 1985; Weintraub et al., 1981; Ross & Mesulam, 1979). The above reviewed literature suggests that EEG information is useful for modelling the characteristics of speech prior to articulation, and may help distinguish on- and off-talk and anticipate prosodic differences in intonation level, speech rate and lexical words.

2.2.3 Cognitive States Detection

Over the past 15 years, computer and speech scientists have explored various methodologies to automate the process of emotion, affective and cognitive state recognition. Past research has mostly focused on emotion recognition from one single sensorial source, or modality: mainly the face (Pantic & Rothkrantz, 2003). Given the fact that emotion, affective and cognitive states of a user can influence the unfolding of the interaction with the system, increasing effort has been spent to test methods for recognition and detection of different affective and cognitive states in human-machine communication. Detecting the cognitive reactions of a user could be a step forward in the process of designing proactive systems capable of adapting to the user's needs (Picard, 2000). While it is true that the face is

the main display of a person's affective and cognitive state, other sources such as body movements and gestures have been shown to increase the recognition accuracy (Hudlicka, 2003), (Balomenos, Raouzaiou, Karpouzis, Kollias, & Cowie, 2003), (Burgoon, Jensen, Meservy, Kruse, & Nunamaker, 2005), (Gunes & Piccardi, 2005), (Kapoor & Picard, 2005), (Martin, Niewiadomski, Devillers, Buisine, & Pelachaud, 2006) and achieve better results in the prediction of user's affective and cognitive reactions. Similarly, features of the speech signal itself have been employed in inferring what has been loosely termed "emotion" or "affect" in the literature (El Ayadi, Kamel, & Karray, 2011). While great progress has been made in recent years on detecting such cognitive states from speech and other modalities on a number of speech datasets, the data used have mostly come from acted speech collected in non-interactive settings (El Ayadi et al., 2011). Studies involving the dynamics of cognitive states in interactive systems, specially systems where the interaction is mediated by automatic speech recognition (ASR) have been far less common. In speech-to-speech machine translation (S2S MT), the limitations of the technologies involved both ensure the elicitation of certain linguistic and cognitive reactions (in response to ASR and MT errors) and require careful design to address communication issues that might arise from those reactions (Schneider & Luz, 2011; Lee & Narayanan, 2005). A cognitive state (i.e. temporary psychological states) detection system for machine translation system can help in sensing the user experience with the system.

2.3 Conclusion

There are many spoken dialogue systems available but using them as an instructor of delivering presentation require a considerable amount of research. An important component of the system is to automatically recognize someone's non-verbal presentation skills (like body language, attitude, engagement level and tone of voice etc.) which are manifested in humans' social signals and behavioural cues. So, to overcome this problem for a spoken dialogue system, this thesis proposes models and methods which can be used in combination with a spoken dialogue system to monitor a user presentation's performance (good voice tone or not) during a presentation. The above-stated problem solution (recognizing someone's public speaking abilities) can help a spoken dialogue system to monitor humans but providing verbal feedback or simulating a practice session using computer (spoken dialogue system) brings many challenges. That is why, the second part of the thesis focuses some of those challenges which are modelling human behaviour during a session with a machine (simulated by a human with the help of technologies like Automatic Speech Recognition (ASR), speech synthesis, etc.). It could help a machine in deploying certain repair strategies in case some of the machine components (e.g., ASR) fail, and manage multiple users (trainees) at a time.

Chapter 3

Recognising Public Speaking Abilities

3.1 Introduction

This chapter describes the datasets, methodology, results and discussion of the research issues which are related to public speaking. It also presents a system which can automatically detect the level of self-confidence and body language of a student using audio-visual information. A system is also proposed which can automatically detect the user engagement of formal presentations and also able to detect which segments of talk are engaging or not and provide them (talk segments) as a feedback to users. An automatic attitudes detection system is also the part of this chapter which can help a user to train his/her attitudes for informal presentations (video blogs). At the end, methods and system are described which are used to collect a dataset of political presentations (debates). The validation of multiple hypothesis and systems' evaluation are also the part of this chapter.

3.2 Student Presentations

The use of multimodal information (prosody, visual, etc.) can help in automatic detection of good and poor presentations. This automatic process, for instance, can help teachers in distinguishing good and bad presenters without having to review many hours of video-recorded presentations, thus freeing teachers to focus more on the weak students. Moreover, it can also help a multimodal dialogue system that is meant to serve as a tutor to train students for presentations. In the context of this study, two presentation quality factors are considered, the first one is the self-confidence, and the other is body language. This study focuses on prosodic and gestural features that contribute to the positive judgement of students' oral presentations.The general hypothesis is that certain prosodic characteristics, such as high pitch variation and perceived loudness, together with the production of natural hand gestures, influence the audience's perception of the speaker as a good presenter. Being able to identify features that can give an indication of a good presenter is useful for applications in the field of skills training, where automatic feedback could be provided to trainees at the end of their presentation about the extent to which they have been able to use their voices and gestures to keep the audience engaged. For this reason, novel models are proposed based on prosodic and visual features.

3.2.1 Dataset

A subset of the presentations contained in a corpus of the Multimodal Learning Analytics (MLA) dataset (Ochoa et al., 2014a) is used as a dataset to run the experiment: 416 oral presentations given by Spanish-speaking students presenting

projects about entrepreneurship ideas, literature reviews, research designs, software design, etc. The dataset contains speech, facial expressions and physical movements in the video, skeletal data gathered from Kinect[1] for each individual and slides of presentations, making up a total of 19 hours of multimodal data. In addition, individual ratings for each presentation, and group ratings related to the quality of the slides used when doing each presentation are available. Each presentation has a rating based on the following performance factors:

1. structure and connection of ideas,
2. presentation of relevant information with good pronunciation,
3. maintenance of adequate voice volume for the audience,
4. usage of language according to the audience,
5. grammar of the slides,
6. readability of the slides,
7. the impact of the visual design of slides,
8. posture and body language,
9. eye contact,
10. self-confidence and enthusiasm.

[1]https://developer.microsoft.com/en-us/windows/kinect – last verified Aug 2017

3.2.2 Hypotheses

In the MLA dataset, each student is judged by the audience on a scale ranging from four to one. In this analysis, it is assumed that a presentation factor (such as self-confidence) is considered good if the rating assigned to it is >= 2.5; otherwise, the presentation factor is considered poor. The number of students present in each class (poor vs good) is depicted in Figure 3.1.

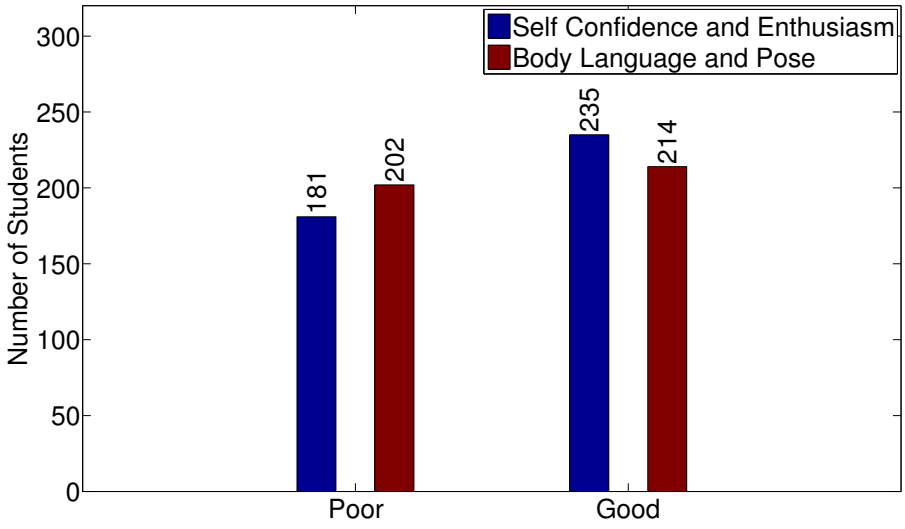


Figure 3.1: Number of students present in each class (good vs poor)

Based on the assumptions and results found in the literature, the following null hypotheses for the investigation of prosodic features are formulated (Lamerton, 2001; Grandstaff, 2004; Sinclair, 1995; Traunmüller & Eriksson, 1995; Hincks, 2005; Lamerton, 2001; Stassen et al., 1993):

1. Standard deviation (std) of F0 is an indication of liveliness and enthusiasm. So a higher value is an indication of a lively and enthusiastic voice. So the null hypothesis is the poor and good presenter have the same std of F0.

2. The harmonic-to-noise ratio (HNR) may indicate abnormality in the voice, so that speakers with the lack of self-confidence tend to exhibit high values of HNR (Yumoto, Gould, & Baer, 1982). So the null hypothesis is the poor and good presenter have the same value of HNR.

3. High values of perceived loudness reflect a loud voice, which is considered an indication of a good presenter. So the null hypothesis is the poor and good presenter have the same value of loudness.

4. A fast speech rate is an indication of a fluent speaker. So the null hypothesis is the poor and good presenter have the same value of speech rate.

Regarding the last hypothesis, speech rate is usually measured in number of words spoken per minute. The dataset do not contain the transcription that's why vocalisation to pause (i.e. voice to silence ratio) ratio is used as an alternative measure to indicate the fluency of speech. Pauses and vocalisation lengths are known to play a significant role in structuring both discourse and interactive speech (Oliveira, 2002; Luz, 2012), so it is expected that this feature provides a reasonable index of fluency in presentations.

For the analysis of visual features, the following hypothesis is formulated: production of hand gestures in the upper part of a body is assumed to be an indication of fluent gestures produced by good presenters. This hypothesis is based on the observations of behaviour of the top ten speakers who obtained good ratings for their body language and the top ten speakers who received poor ratings for their body language. The two groups follow a clear trend: the good speakers produce fluent arm and hand gestures concentrated in the upper part of the body. Their gestures have the following main functions:

1. provide discourse with continuity and coherence (McNeill, 1992),

2. mark stress and rhythm of utterances'

3. point at the slides'

4. describe something.

The speakers who received the lowest ratings for body gestures seem, in general, to produce fewer gestures in the upper part of the body. They tend to keep their arms down, parallel to their body, or keep their hands at the level of their belly. When they produce gestures, they produce particular types of hand movements which are not connected to the co-occurring discourse (semantically nor structurally), as described by Ekman (Ekman & Friesen, 1981). These gestures seem to be produced by the speaker to manage particular emotional states, such as tension or anxiety. An attempt is made to find a measure that could detect the position of the arm gestures in relation to the shoulder centre used as a reference by analysing the skeletal data. The mean value of the Euclidean distance between the hands and shoulder is selected as the basic measure. Despite its relative simplicity, this measure provides a good indication of hand and arm movements concentrated in the upper part of the body and thus can be used in testing the null hypothesis which is the poor and good presenter have the same mean value of the Euclidean distance between the hands and shoulder during a presentation.

3.2.3 Experimentation

The first step of experimentation is to analyse correlations among the different categories of ratings in order to estimate how visual and prosodic features might

contribute to the prediction of overall presentation quality.

Figure 3.2 depicts the correlations of all ratings (correlation matrix) as a corrgram (Friendly, 2002), where blue indicates positive correlation and red indicates negative correlation, with darker hues indicating stronger correlations. It is observed that ratings that appear to be motivated by voice feature (e.g. self-confidence and enthusiasm) are sometimes highly correlated to visually-motivated ratings (e.g. body language and pose). This might imply that either visual or voice features alone might suffice to distinguish some aspects of presentation quality. Alternatively, it is possible that combined feature sets might be more useful overall.

In order to investigate these issues in more detail, in the following sections, it is demonstrated that how the various prosodic and visual features vary according to two broad performance categories (poor vs. good presentation), in line with the hypotheses formulated in Section 3.2.2, and then describe a method for automatic categorisation of presentation quality level based on a rich set of such features, presenting its results on the MLA dataset.

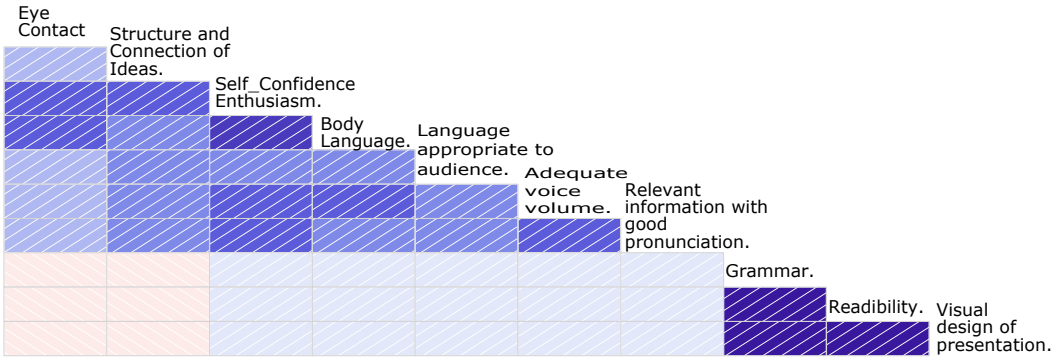


Figure 3.2: Correlation matrix for rating categories

Feature Extraction

In total, 6376 audio features are used for the classification tasks: the complete audio set of the ComParE challenge (Schuller et al., 2013a) (6,373 features) with the addition of perceived loudness (sd and mean) and V/P (vocalisation to pause ratio).

To extract the features related to the speakers' hand movements the Euclidean distance (ED) between wrist joint and shoulder centre joint (tracked by Kinect) in each frame of the video (presentation of a student) is calculated. Finally, the mean, standard deviation, maximum, minimum, median, maximum ratio and minimum ratio of the ED, its first (velocity) and second (acceleration) order derivative for each video/speaker are calculated. In total, 42 features are extracted for both hands. The maximum ratio for a speaker is measured by counting the number of frames which have higher ED compared to their preceding and following frames and then averaged over the total number of frames in that video. Similarly, the minimum ratio of a speaker is measured by counting the number of frames which have lower ED compared to their preceding and following frames and then averaged over the total number of frames in that video. The feature set is z-score normalise and then scaled in the range of [0 1].

Hypothesis Testing

In order to validate the hypotheses formulated for prosodic and gestural features, analysis of variance (ANOVA) test is performed on the values of several features with respect to presentations rated as poor, as compared to presentations rated as good. The results show that a statistically significant difference exists between

the poor and good groups of speakers for the different measures considered: a significant difference is shown between fundamental frequency standard deviation values (p = 0:00) for good and poor presenters. The box plots in Figure 3.3 depict the quartiles for the respective distributions of values. The higher the values of the standard deviation, the higher the pitch variation of the student during the presentation. This is in line with the results of previous studies (Traunmüller & Eriksson, 1995; Hincks, 2005) that show that the higher the standard deviation of fundamental frequency, the more a speech sample is perceived as lively. Since this study assume that liveliness is associated with enthusiasm, this result is an indication that speakers rated as good presenters are very likely to have lively and enthusiastic voices. The log HNR (in Figure 3.3) has higher values for good speakers (p = 0:0002) which might be due to the fact that the presenters perceived as good speakers do not present obvious abnormalities in their voice quality (e.g. roughness of sound (Sousa & Ferreira, 2008)). The results summarised in Figure 3.3 show higher values of perceived loudness for speakers judged as good ones ($p = 0.009$). This is because loudness plays an important role in expressing self-confidence and enthusiasm and speaking loud is generally considered a characteristic of good presenters, consistently with our hypothesis and the literature on presentation quality.

As for the results of the vocalisation to pause ratio, a statistically significant difference ($p = 0.1148$) between speakers judged as good versus poor is not observed. This might depend on the fact that pauses can also be used for rhetorical purposes and in our calculation of vocalisation to pause ratio (as explained in Section 3.2.3) the filled pauses and hesitations are not taken into account since they were not annotated in the audio files. As for visual gestures, our hypothesis is

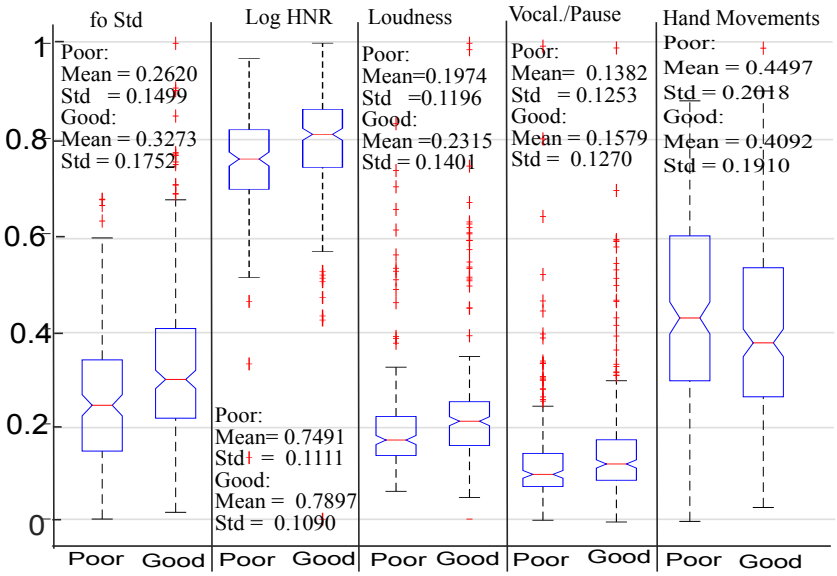


Figure 3.3: ANOVA Test Results.

that hand gesture produced in the upper part of the body are an important factor that characterises a good presenter. To perform hand gestures in the upper part of the body, speakers need to move their hands in that region for a relatively long period. A good presenter can also move his/her hands in the lower body region (to relax) or overhead (to point out the slides), but these gestures should not be maintained for long periods of time. So, it is decided to choose the mean value of ED as a measure of these gestures. The results show that the good presenters have statistically significantly lower ED values ($p = 0.036$) as shown in Figure 3.3. Although it is true that the visual features are, in some sense, 'optimised' (i.e. designed by us rather than discovered automatically from data), they come not only from simply watching the videos, but are also informed by the literature on gestures and presentations (McNeill, 1992; Ekman & Friesen, 1981).

Classification Method

To reduce the high dimensionality of features, the PCA (Principle Component Analysis) is used on the feature set to reduce the number of dimensions (6376 audio features + 42 visual features to 416 principle components). From the statistical significance ($\boldsymbol{p}$) of the transformed feature set with the rating (poor or good), a subset of the transformed features with $p < 0.5$ is selected. The classification method was implemented in MATLAB [2] and employed discriminant analysis in 10-fold cross-validation experiments. The classification method works by assuming that the feature sets of the classes to be discerned are drawn from different Gaussian distributions and adopting a pseudo-linear discriminant analysis (i.e. using the pseudo-inverse of the covariance matrix (Raudys & Duin, 1998)).

3.2.4 Results and Discussion

Prosodic and visual features are analysed to predict presentation quality. The correlation test results (Figure 3.2) show that the presentation quality factors under consideration are highly correlated with each other. Therefore, in principle, it should be possible to detect the body language rating ('Body') with prosodic features, and the self-confidence rating ('Conf.') with skeletal features.

The motivation for this automatic inference task is to be able to distinguish those students who present really poorly (and therefore might need expert attention and extra tutoring), from those who present really good and might be selected as examples of how to present from the average presentations. The very good speakers do not necessarily need extra attention from the tutor, while the

[2] https://uk.mathworks.com/products/matlab.html – last verified Aug 2017

Table 3.1: 2-Class Experiment Results (F-Score of both class (Poor and Good))

Feature	Rank	Poor	Good
Audio	Conf.	92.64%	94.19 %
Audio	Body	94.32%	94.61%
Visual	Conf.	60.06%	73.35%
Visual	Body	64.18%	66.51%
Fusion	Conf.	94.02%	95.26%
Fusion	Body	95.80%	96.02%

poor presenters might benefit from advice. Therefore, two experiments are conducted. In the first (2-Class) experiment, there are 3 types of feature vectors, 2 types of ratings and 2 types of groups of speakers (poor and good). The results are shown in Table 3.1. In the second experiment (3-Class), the same settings are repeated, but the students are divided into three groups: poor (rating range is $1-2$), average (rating range is $2-3$) and good (rating $>= 3$). The results (harmonic mean) are shown in Table 3.2 and the number of students present in each class is depicted in Figure 3.4.

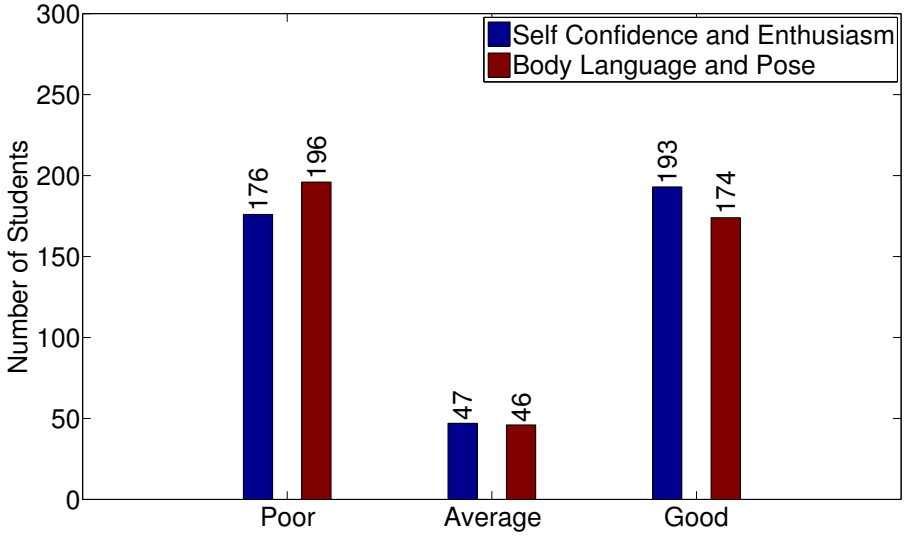


Figure 3.4: Number of students present in each class (poor, average and good)

Table 3.2: 3-Class Experiment Results (F-Score of all three class (Poor, Average and Good))

Feature	Rank	Poor	Average	Good
Audio	Conf.	83.76%	84.54%	85.42%
Audio	Body	84.32%	80.85%	84.24%
Visual	Conf.	60.00%	18.46%	66.19%
Visual	Body	62.35%	07.02%	56.98%
Fusion	Conf.	82.49%	82.00%	83.07%
Fusion	Body	84.62%	76.60%	81.03%

Our study uses an extended dataset including both male and female students, in contrast to the limited dataset used in a similar study (Krajewski et al., 2010). Our approach is tested in speaker-independent settings, and the student presentations are ranked by an audience. It yields maximum F scores of 95.26% (good) and 94.02% (poor) in detecting self-confidence. Moreover, in the two-class problem, the F-Score of prosodic features indicates that the prosodic features are not only able to predict the rating of self-confidence and enthusiasm but also the rating of body language and pose. This may be due to the impact of good posture on speaking style. The visual features show the same behaviour, but with less accuracy. However, the fusion of prosody and visual features does, in fact, improve overall categorisation performance.

The promising results for three-class rating detection is also obtained, with Fscores as high as 83.76% (poor) 84.54% (average) and 85.42% (good) in detection of self-confidence and enthusiasm. In the three-class problem, the features show the same behaviour as for the two-class problem, except for the fusion which causes a slight decrease in performance. At the same time, they cause a slight increase in poor (body language rating) class detection, while visual features alone are almost unable to detect average class (07.02% and 18.46%). In the three-class problem,

the features show the same behaviour as for the two-class problem, except that feature fusion does not seem to improve performance in this case.

3.3 TED Talks

These days, several hours of new video content (in the form of talks) are uploaded to the internet every second. It is simply impossible for anyone to see every piece of video which could be engaging or even useful to them. Therefore, it is desirable to automatically identify videos (TED talks) that might be regarded as engaging for users, for a variety of applications such as recommendation services. This section describes validation of hypotheses that relate user (on-line viewer) engagement of TED Talks to prosodic, spoken expressions, high level visual and paralinguistic features. A novel system is proposed for automatic inference of user engagement and automatic identification of engaging video segments within TED talks.

3.3.1 TED Talks and User Feedback

"TED is a non-profit organization devoted to spreading ideas, usually in the form of short, powerful talks (18 minutes or less)". The TED website, instead of asking users to simply give like or dislike feedback, asks viewers to describe the video in terms of particular words. Among the 14 rating words provided to users, 9 words are identified as being positive words, 4 as being negative, and 1 as neutral. A user can choose up to 3 words to rate a video (Table 3.3).

By giving these choices to users, the TED website provides elaborate feedback on a given video as depicted in Figure 3.5 and 3.6. This makes the problem more interesting than a like or dislike based content detection. Instead of crisp binary

feedback to learn from, this study uses a fuzzy description of what viewers thought about a particular video. The rating system for user feedback thus provides a more nuanced characterisation of user engagement with the video presentation. Since the ratings consist of voluntarily information given by the users, in terms of semantically positive and negative words, it provides good basis for analysis of the relevant factors of engagement (O'Brien & Toms, 2013) for TED talks.

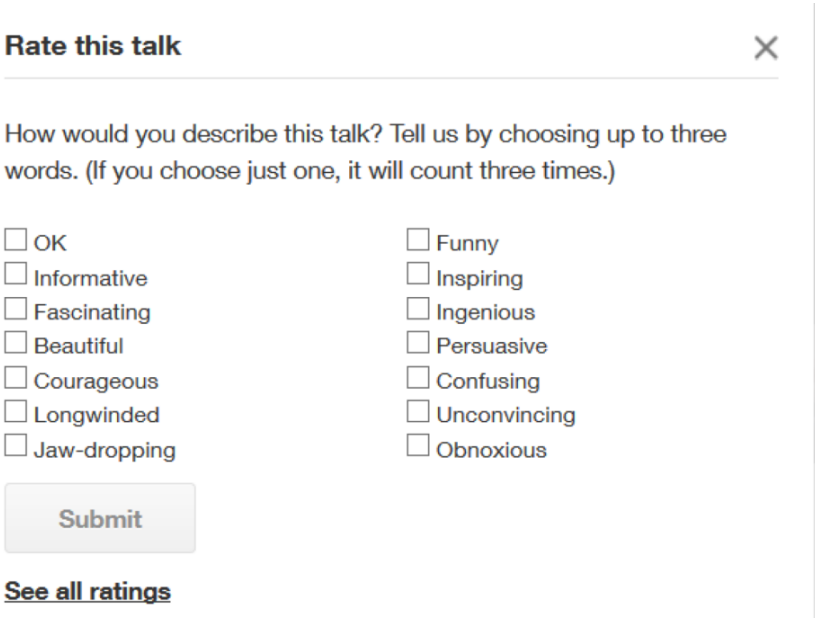


Figure 3.5: *Ted.com rating criterion*

3.3.2 Analysis of User Rating

The TED website reports what percentage of viewers rated the video as saying "Inspiring" or "Longwinded" etc. But simply relying on ratings given to an individual video is not a good idea because it may not give the whole story. As seen in Table 3.3, ratings tend to be overwhelmingly positive. Both count and percentage, positive criterion (e.g. beautiful) tend to score much higher than negative (e.g.

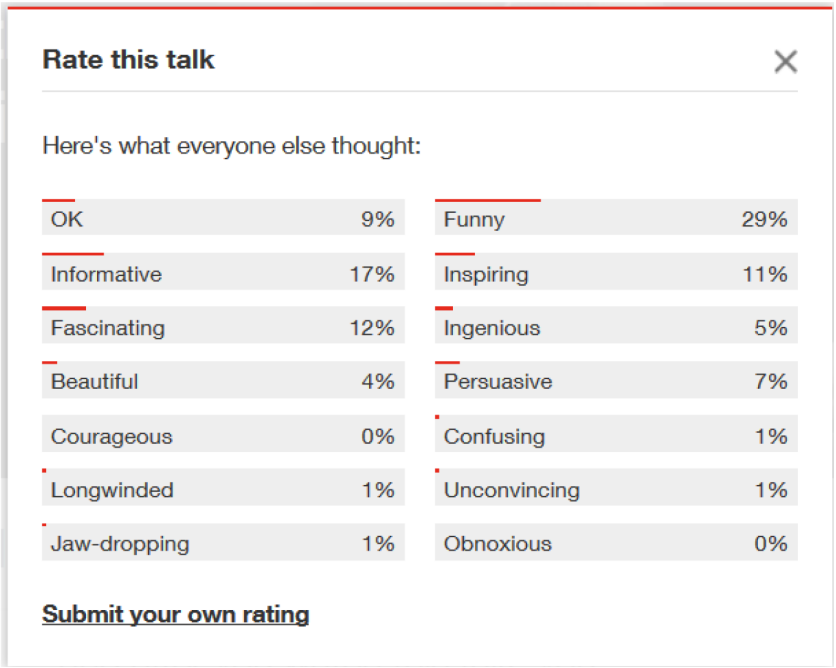


Figure 3.6: *Overall ratings of a TED video*

unconvincing) or neutral (i.e. ok) ones. Even the highest scoring negative ratings "Unconvincing" has average user count of 51 and percentage of users of 3.73% which are less than the lowest scoring positive rating "Funny" with average user count of 106 and average user percentage of 4.34%.

The 14 words rating (proposed by TED [3]) describe a video in a more detail than just like/dislike. However, if only the ratings (proposed by TED) of an individual video are considered then it would seem like all the videos only positively engage the viewers. There is no video to which a negative rating word got the highest count by the viewers. So to deduct which video is found to be "Obnoxious" or "Longwinded" by viewers, some kind of normalization is required.

In order to do that the following definitions for a video to be considered "Beau-

[3]www.ted.com – Last verified July 2017

Table 3.3: Average number of user ratings per each rating criteria for 1340 Ted videos across different topics.

Rating	**Avg. (Count)**	**Avg.(%)**
Beautiful	120	6.67
Confusing	15	1.17
Courageous	122	6.08
Fascinating	234	12.64
Funny	106	4.73
Informative	246	15.24
Ingenious	134	7.64
Inspiring	384	18.16
Jaw-dropping	118	5.45
Longwinded	28	2.23
Obnoxious	23	1.62
OK	65	4.88
Persuasive	188	9.70
Unconvincing	51	3.73

tiful" or "Persuasive" etc is used. It must have a rating count more than average rating count for that particular rating word. With this, TED talks were categorized as "Beautiful and not Beautiful", "Inspiring and not Inspiring", "Persuasive and not Persuasive" etc. giving two classes for classification for each of the 14 rating words. The details of user rating distribution (Yes, No) for videos is depicted in Figure 3.7.

3.3.3 Correlation Between User Ratings

The first step is to perform the correlation between different types of user engagement ratings to find the relationship between them. Figure 3.8 depicts the correlations of all ratings (correlation matrix) as a corrgram, where blue and red indicate positive and negative correlations respectively and the darker hues indicates stronger correlations. This section explains the dataset along with normalization

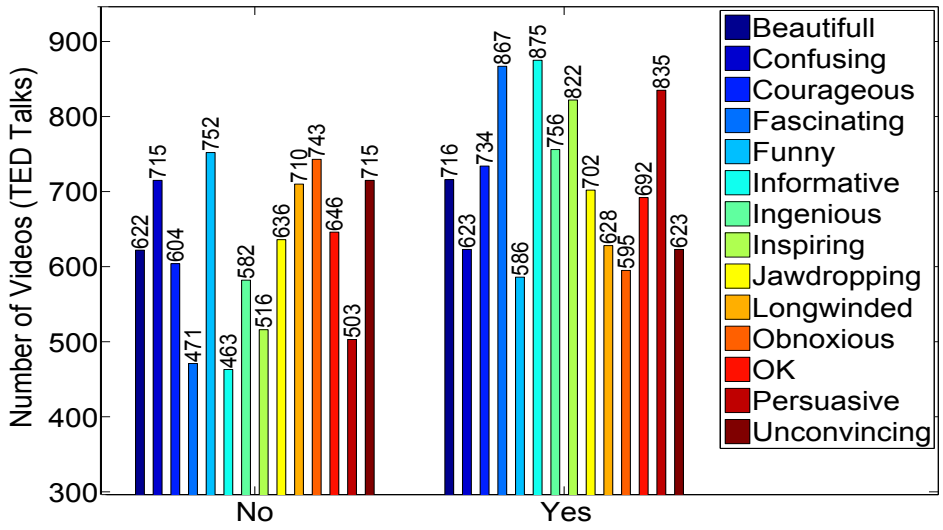


Figure 3.7: Number of videos present in each class (Yes/No).

and feature extraction approach.

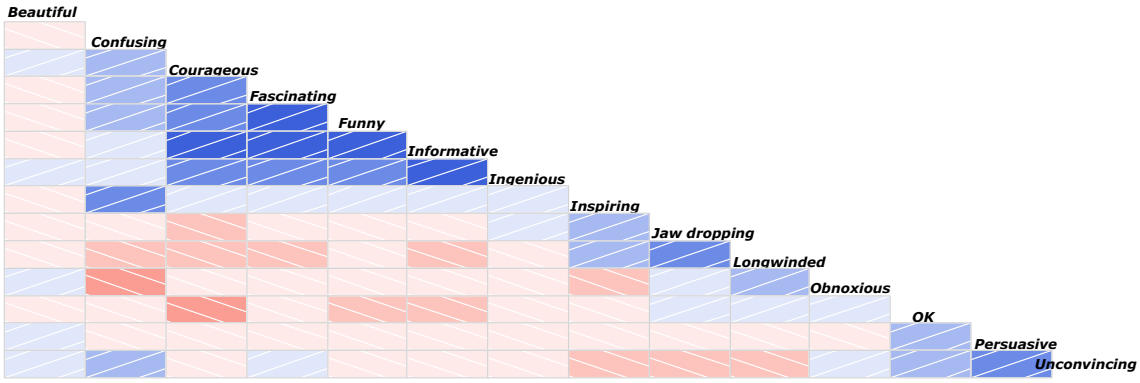


Figure 3.8: Correlation Matrix for User Engagement Ratings.

Features Extraction

For visual features, HAAR cascades (Lienhart, Kuranov, & Pisarevsky, 2003) is used, from the OpenCV library (Bradski, 2000), to detect whether the speaker is on the screen or not. For this study, it is calculated that for how many seconds there was a close up shot of the speaker and when there was a distant shot and when the

speaker was not on the screen. For non-visual features, number of laughters and applauses and laughter-to-applause ratio within TED talks are considered. Since TED talks come with subtitles, getting this information was a simple process and was obtained with a python script. For all extracted features, it is also measured whether for each video, the value of each feature was greater or less than the average value for that feature. E.g if the number of close up face seconds for a given video was greater than the average number of close up face seconds, the value 1 to the feature is assigned "Above average close up shots" and 0 otherwise. The same procedure is repeated for other features thereby doubling the number of visual and paralinguistic features to 12 for the experimentation.

For prosodic features, the openSMILE (Eyben, Wöllmer, & Schuller, 2010) tool kit is used. Prosodic features have been shown to correlate to structure in dialogue interaction and evidence of participant status and engagement levels (Campbell, 2008; Luz, 2012). It is hypothesized that these features might also occur in monologues such as TED talks.

3.3.4 High Level Visual and Paralinguistic Features Evaluation

The following sections analyse how various camera views and paralinguistic features contribute to the user engagement ratings by performing Analysis of Variance (ANOVA) test.

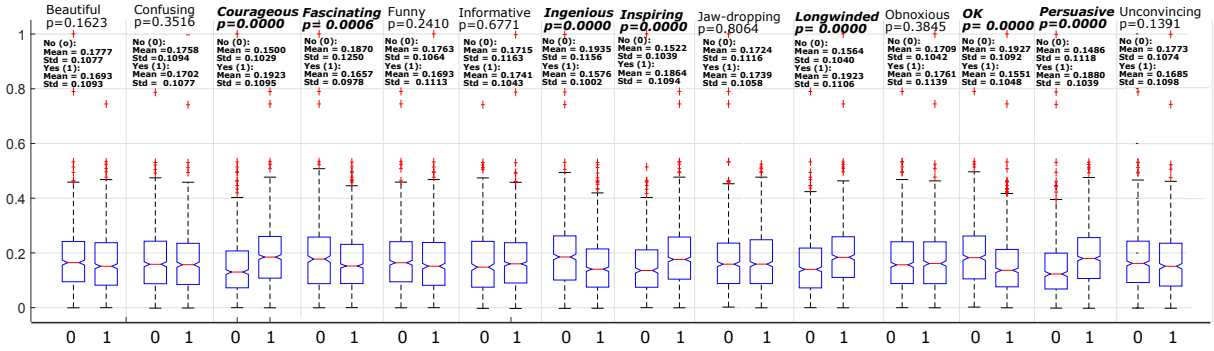


Figure 3.9: Close up Shots (ANOVA Results)

Close up Shots

The results depicted in Figure 3.9 show that some ratings are significantly influenced by 'Close up Shots'. Videos rated by users as courageous, inspiring and long winded have high mean value of 'Close up Shots' but at the same time videos which have less mean value of close up shot are perceived as fascinating, ingenious and OK. For rest of the ratings, no significant difference exists.

Distance Shots

The results depicted in Figure 3.10 show that some ratings are significantly influenced by the 'Distance Shots. The videos rated by users as courageous, fascinating, inspiring, jaw dropping, long windowed and persuasive have high mean value of 'Distance Shots' but at the same time videos which have less mean value of distance shot are perceived as confusing, ingenious, OK and unconvincing. For the rest of ratings, no significant difference exists.

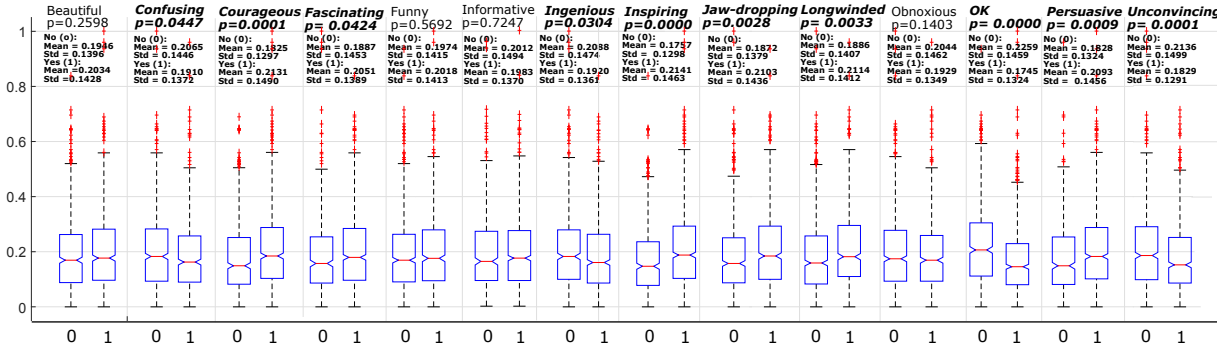


Figure 3.10: Distance Shots (ANOVA Results)

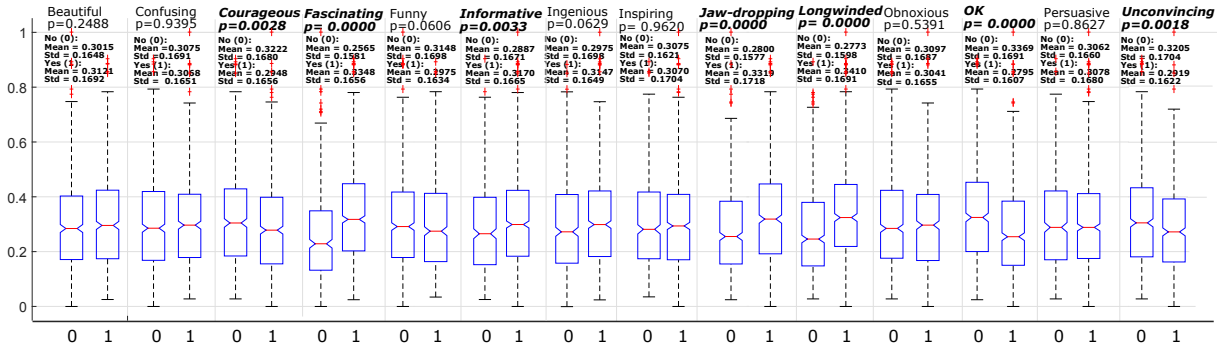


Figure 3.11: Person Not on Screen (ANOVA Results)

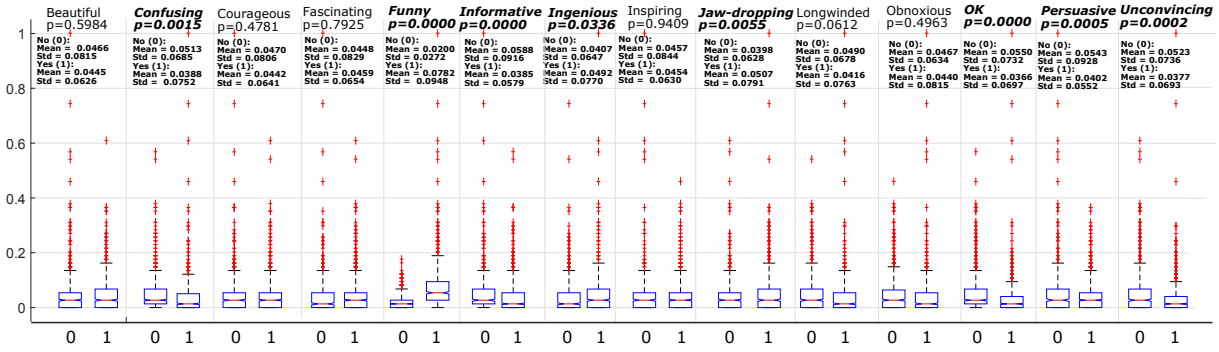


Figure 3.12: Laughter by Ted Audience (ANOVA Results)

Person Not on Screen

The results depicted in Figure 3.11 show that some ratings are significantly influenced by the person not being on screen. Videos rated by users as fascinating, informative, jaw dropping and long winded have high mean value of 'Person Not on Screen' but at the same time videos which have less mean value of 'Person Not on Screen' are perceived as courageous, OK and unconvincing. For the rest of ratings, no significant difference exists.

Laughter by Ted Audience

The results depicted in Figure 3.12 show that some ratings are significantly influenced by laughter by the TED video audience. Videos rated by users as funny, ingenious and jaw dropping have high number of laughter but at the same time videos which have fewer laughter are perceived as confusing, OK, persuasive and unconvincing. For the rest of ratings, no significant difference exists.

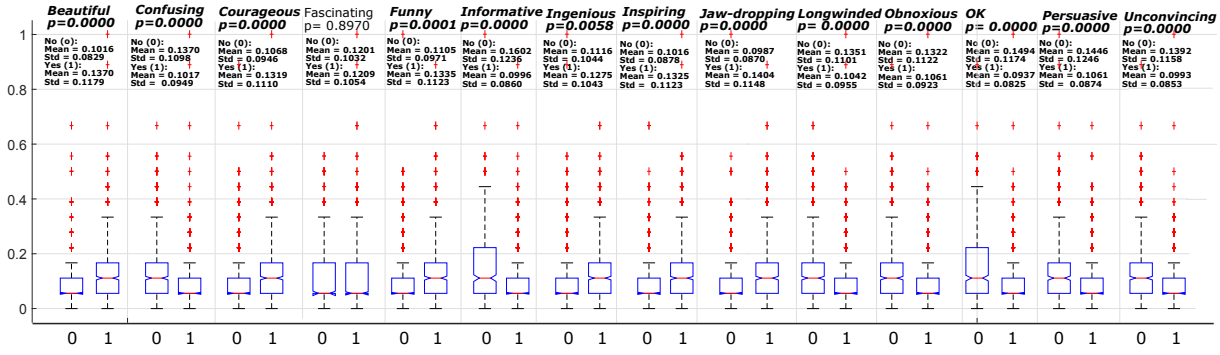


Figure 3.13: Applauses by Ted Audience (ANOVA Results)

Applauses by Ted Audience

The results depicted in Figure 3.13 show that some ratings are significantly influenced by the applauses by Ted Audience. Videos rated by users as beautiful, courageous, funny, ingenious, inspiring, and jaw dropping have high mean value of applauses but at the same time videos which have less mean value of applauses are perceived as confusing, informative, long winded, obnoxious, OK, persuasive and unconvincing. For the rest of ratings there, no significant difference exists.

Liveliness of Speech

The results depicted in Figure 3.14 show that some ratings are significantly influenced by the standard deviation of 'Fo' (fundamental frequecy) which is a measure of liveliness of speech (Traunmüller & Eriksson, 1995; Hincks, 2005). The videos rated by users as beautiful and funny have high mean value of Std Fo but at the same time videos which have less mean value of Std Fo are perceived as confusing, informative and persuasive. For the rest of ratings, no significant difference exists.

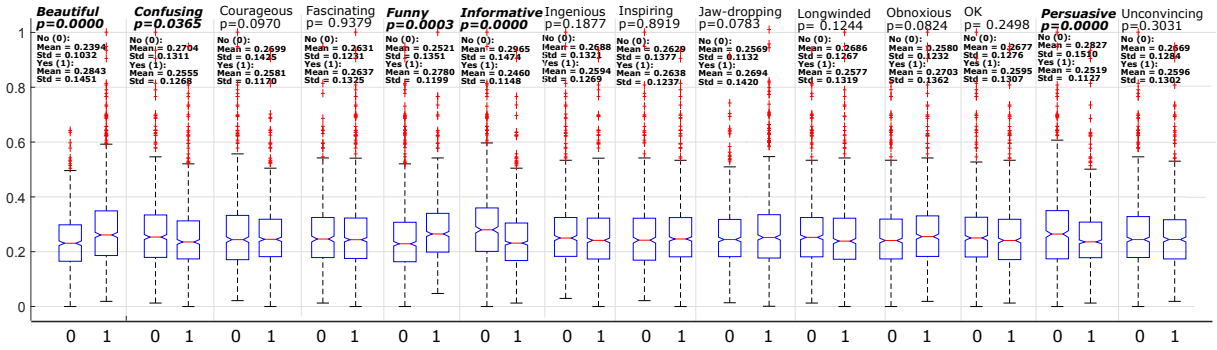


Figure 3.14: f0 Std (ANOVA Results)

Discussion

In film making, 'Close up Shots' are used to focus the subjects's emotional states and 'Distance Shots' are used to focus the full body of a person with some surrounding. In the TED talk settings, surrounding of a speakers are audience and content (slides). So it can be said that a higher number of 'Close up Shots' indicates more duration of emotional states. One of the purpose of distance shot is to focus on a person body to convey the body language message to the audience. So it can be said that a large number of 'Distance Shots' indicate the more use of body language/content by the speaker. Moreover 'Person not on the screen' is used to present content by the speaker. So it can be said that a higher value indicates more content presentation through slides. The laughter and applause are usually correlated with joy and appreciation respectively and this study has found that these (laughter and applause) has an influence on the user engagement rating.

In previous studies, it is stated that standard deviation of fundamental frequency is highly correlated with liveliness of speech. So the effect of liveliness of speech on these user engagement ratings are also analysed. The interesting

point to note about 'beautiful' rating is its strong correlation only with applause and liveliness of speech features. However the funny rating is strongly correlated with laughter, applause and liveliness of speech but the camera views have no effects on the rating funny. 'Person Not on Screen' is the only camera view which significantly differ for informative user engagement. It is assumed that a higher count means more content presentation (informative) and the results (as depicted in Figure 3.11 with $p = 0.0033$) are also in line with this assumption.

The results depicted in subsection 3.3.4 show that the feature under consideration have a relationship with user engagement. Camera views are not based on some random selection and strongly correlate with user engagement. But it is not possible to increase engagement level by just increasing or decreasing the number of camera views in a video. These views are depended on the speaker's way of speaking, if he /she shows emotions then a professional camera-operator/editor may take his/her close up shot and if he/she uses body language and slides then a camera-operator will focus through 'Distance Shots' and at the end if some content is really important then a camera-operator/editor just show the slides. Believing that the camera men/editors are highly professional, these views help in building some feedback to the speaker e.g. they need to show more content, use of body language and emotions etc. Moreover, these features can also be used to predict the engagement level.

3.3.5 Spoken Expression Evaluation

The following section analyses how various spoken expressions contribute to the user engagement ratings.

Data Pre-Processing and Feature Extraction

Speech segmentation is performed on all the audio files of TED videos using the Lium toolkit (Rouvier et al., 2013) with a minimum cluster size of 2 and the maximum possible duration of a segment is 20 seconds. The duration of chunks is between a few seconds to 20 seconds. As a result, dataset has 120,382 chunks of audio from 1338 videos for experimentation (clustering).

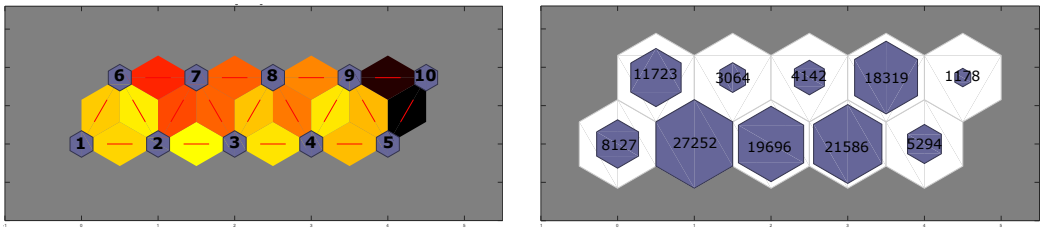
Acoustic feature extraction is performed using openSMILE toolkit (Eyben et al., 2010). The feature set is extracted using the openEAR configuration file. This set is also used for emotion and speech expression recognition (Eyben, Wöllmer, & Schuller, 2009) and consists of low-level descriptors as well as statistical functionals applied to these descriptors. A correlation test is also performed between duration of each audio chunk and its features, selecting those features which are less correlated with chunk duration ($R < 0.2$). As a result, 387 features have left for clustering. The feature set was further centred with mean value 0 and standard deviation 1.

Statistical Analysis

First, SOM (Self Organised Map) is employed to cluster the speech segments into 10 clusters. The motivation behind using clusters size of 10 for SOM is to separate the 6+1 universal spoken expressions (happiness, sadness, fear, surprise, disgust, anger and neutral) and any other non-speech segments (music, applause, laughter). Results of clustering are shown in 3.15a and 3.15b. Then the number of speech segments in each cluster for every video are calculated and then divided it by the total number of speech segments within a video. Later, to analyse the significance

of speech expressions, Kruskal–Wallis test and null hypothesis are used in the following manner:

H: The number of speech expressions in each group (e.g. beautiful and not beautiful) has the same mean value.



(a) Distance between neighbouring clusters (b) Sample hits

Figure 3.15: Left Figure (a) indicates the distance between clusters (darker colour indicates more distance between clusters than lighter colours) and the right Figure (b) indicates the number of speech segments present in each cluster

The Kruskal–Wallis test rejects the null hypothesis ($p < 0.05$) for many clusters (speech segments). For example, speech segments in clusters number 1,2,4,5,7 and 8 have a significant difference in their mean values for beautiful-YES and beautiful- NO. Speech segments from cluster number 1,4,5,7 and 8 have higher mean for beautiful-YES than beautiful-NO. Hence, the speech segments in these clusters (that also represents speech expression) are engaging. For example,the cluster number 2 has higher mean for beautiful-No than beautiful-YES, hence, the speech segments in this cluster are non-engaging. The details for all engagement ratings are depicted in Table 3.4.

3.3.6 Engagement Detection

Based on the statistical analysis results, most of the spoken expressions, high level visual and paralinguistic features are statistically different for engaging and

Table 3.4: Statistical significant clusters for each rating.

Rating	Cluster $p < 0.05$	YES	NO
Beautiful	1, 2, 4, 5, 7, 8	1, 4, 5, 7, 8	2
Confusing	7, 8	nil	7,8
Courageous	3, 4, 6, 7, 8, 9	3, 4, 6, 7, 8	9
Fascinating	6, 7, 9	7, 9	6
Funny	3 , 4, 5, 6, 7, 9	4, 5, 7, 9	3, 6
Informative	1 4 5 7 8	4	1 5 7 8
Ingenious	2 3 4 6 9	2,9	3, 4, 6
Inspiring	3 4 6 7 8	3, 4, 6,7, 8	nil
Jaw-dropping	2, 3, 7, 8, 9	2, 3, 7, 8, 9	nil
Longwinded	3 7	3	7
Obnoxious	4 6 7 9	4 6	7 9
OK	2 3 7 8 9	nil	2 3 7 8 9
Persuasive	1 3 7 8 10	3	1 7 8 10
Unconvincing	2 4 5 7 9	nil	2 4 5 7 9

non-engaging speech segments. So, a novel approach is proposed to detect user engagement with TED talks. The proposed approach is based on the hypothesis that multimodal features can be extracted automatically from TED videos and be correlated to user engagement criterion for a variety of applications. Segments of TED talks are extracted using speech segmentation using Lium Toolkit (Rouvier et al., 2013). Clustering was performed on the resulting dataset of segments. After clustering the dataset, classification test using LDA and statistical analysis was performed to identify the relationship between the clusters and user engagement.

In terms of contribution, this study proposes the following

- A classification model based on multimodal features to identify engagement in TED talks.
- A method to identify segments within a talk that are more engaging than others.

The proposed system architecture is depicted in the Figure 3.16, where the

user (a presenter, potential viewer or a video summarization tool) obtains feedback about a talk. The feedback is in the form of video segments (engaging and non-engaging parts of talk) and the predicted label. As the audio-visual information are correlated, visual information is also provided to user during a speech segment.

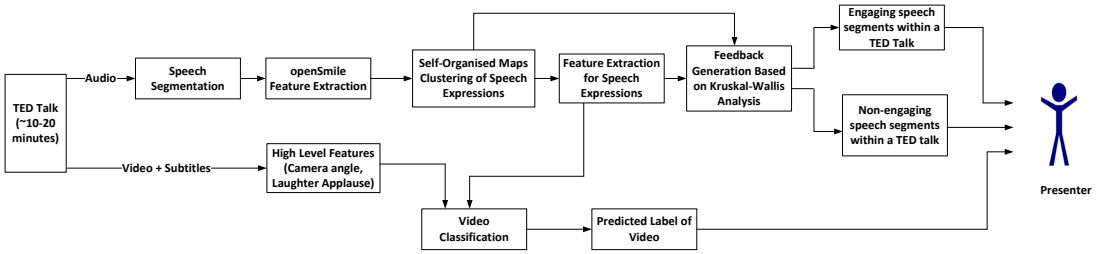


Figure 3.16: System Architecture.

Three experiments are performed using three different feature set for classification as described below:

Experiment One: It uses the high-level visual and paralinguistic features (camera angles, laughter, applause) extracted from video and subtitles.

Experiment Two: It uses the speech expressions features. However, features set is increased by also considering the duration of speech segments in each cluster along with the number of speech segments in each cluster. The speech expressions features are extracted with different cluster sizes (10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60): this helped us find the best cluster size for engagement detection.

Experiment Three: The previous two experiments' feature sets are fused.

Classification Methods

The classification is performed using Linear Discrimination Analysis (LDA) in 10-fold cross-validation setting. This classifier is employed in MATLAB[4] using the

[4]https://uk.mathworks.com/products/matlab.html – last verified Aug 2017

statistics and machine learning toolbox. LDA works by assuming that the feature sets of the classes to be discerned are drawn from different Gaussian distributions and adopting a pseudo-linear discriminant analysis (i.e. using the pseudo-inverse of the covariance matrix (Raudys & Duin, 1998)). There is no other classifier is used for comparison as the objective is to demonstrate the proposed features discrimination power not the classifier performance. The Deep neural network can not be employed as the number of instances (1338) are not sufficient for training.

Results and Discussion

The duration and number of speech segments in each cluster are used as a feature vector in detecting engagement, as defined earlier. The results are depicted in Table 3.5. All engagement levels are detected using the proposed feature vector above the blind guess baseline (50% A-weighted F-score (averaged harmonic mean of both classes) for engagement detection problem). The results show that the visual and paralinguistic features (Vis+Para) extracted from video and subtitles provide better results than speech expressions (SE: clusters) features for 7 out of 14 user ratings. The fusion of speech expression and visual+para features (Fusion: clusters) improve results for 11 out of 14 user ratings.

Haider, Salim, Luz, Conlan, and Campbell evaluated the relationship between high-level features (camera angles, pitch, laughter and applause) and user ratings (Haider et al., 2015). It only showed that high-level features are statistically different for user ratings. This current study, however, not only analyses the relationship between speech expressions and user rating but it also proposes a system to detect the engagement with a novel combination of speech expressions and high-level features. The proposed system also generates feedback for the

Table 3.5: A-weighted F-score (averaged harmonic mean of both classes (Yes and No)). Where Vis+Para means high-level visual and paralinguistic features, SE: Clusters means Speech Expressions and the corresponding number of clusters and Fusion: Clusters mean Fusion of Vis+Para and SE along with the corresponding number of clusters.

Rating	Vis+Para	SE: clusters	Fusion: clusters
Beautiful	55.28	60.58:15	61.70:30
Confusing	58.21	53.47:30	56.18:10
Courageous	60.33	58.08:20	61.25:10
Fascinating	52.65	54.02:55	58.04:45
Funny	70.96	61.61:35	71.85:10
Informative	59.08	61.24:40	64.09:40
Ingenious	57.42	56.67:30	57.35:55
Inspiring	54.85	53.05:55	56.13:60
Jaw-dropping	58.33	58.38:10	59.39:10
Longwinded	64.17	62.53:40	64.45:15
Obnoxious	48.87	52.14:45	54.25:45
OK	64.46	61.86:50	63.68:15
Persuasive	56.67	59.02:60	60.35:35
Unconvincing	56.4	56.86:20	58.11:10

presenter or viewers in the form of video segments. The clustered video segments can potentially be used as training material for presenters to advise about using certain speech expressions for a particular type of engagement with viewers. For example in Table 3.4 it shows that clusters 3 and 7 have significant p-value with 'Longwinded'. It may be used to guide a presenter to avoid speech expression of cluster 3 and utilisation of speech expression in cluster 7 to make sure that their presentations do not become 'Longwinded'. Similarly, from a viewers perspective, it is a recommender system that can predict the engaging talks using multi-modal features. It can guide a potential viewer to avoid segments that have a higher number of speech segments in that cluster. From the 3.15a, It is also observed that the cluster 3 and 2 have a lesser distance between them than cluster number 3 and 7. Due to that lesser distance, the probability of sounding similar can be

high for both clusters and the cluster number 2 instances may also be named as engaging one. The clustering approach may also support in video summarization and segmentation e.g. summarising all the Inspiring parts of a video etc.

3.4 Attitude Recognition of Video Bloggers

The video blogger's attitudes have an influence on the viewers (e.g. level of attention gain on-line i.e. number of views). So an automatic attitude recognition system can help potential video bloggers in providing feedback about their attitudes. In this study, the acoustic and visual features (body movements that are captured by low-level visual descriptors) are used to predict the attitudes annotated in the speech of video bloggers. The automatic detection of attitude can be helpful in a scenario where a machine has to automatically provide feedback to video bloggers about their performance in terms of the extent to which they manage to engage the audience by displaying certain attitudes. In this section, a novel automatic attitude recognition system is proposed using low level audio-visual descriptors that is able to recognise the six attitudes (Amusement-A, Enthusiasm-E, Friendliness-Fd, Frustration-Fr, Impatience-I and the added label for Neutral-N) using audio and visual features. However, the classifiers and feature sets are also evaluated for a three-class task: positive (Amusement-A, Enthusiasm- E, Friendliness-Fd), negative(Frustration-Fr, Impatience-I) and neutral.

3.4.1 Experimentation

This section describes the dataset used for automatic detection of attitude along with methodologies for acoustic and visual descriptor extraction and analysis of

principle components for attitude recognition.

Data Set

The video-blog dataset used in this study is the same used in (N. A. Madzlan et al., 2015) augmented with the annotation of hundred video segments with a neutral label. In total, it contains the 613 audio-visual segments (for each subject the number of segments is as follows: 34, 53, 54, 111, 46, 36, 93, 104, 34, 48) from around 250 different videos that are annotated for six different attitudes (Amusement-A, Enthusiasm-E, Friendliness-Fd, Frustration-Fr, Impatience-I and Neutral-N) as depicted in Table 3.6. The data annotation was performed by two annotators with an inter-coder agreement of 75% as reported in (N. A. Madzlan, Reverdy, Bonin, Cerrato, & Campbell, 2016). The data comes from 10 different native speakers of English. The duration of each video clips is around 1-3 seconds.

Table 3.6: Number of instances (video segments) for each attitude

Attitude	Instances
Amusement	100
Enthusiasm	107
Friendliness	101
Frustration	103
Impatience	102
Neutral	100

Feature Extraction

This study analysis acoustic and visual descriptors for the recognition of attitudes.

Acoustic Features The openSMILE (Eyben, Weninger, Groß, & Schuller, 2013a) tool kit is used to extract the acoustic features, this has been widely used for emo-

tion recognition (Liu et al., 2014). The acoustic feature set contains the MFCC, voice quality, fundamental frequency (F0), F0 envelope, LSP and intensity features along with their first and second order derivatives. However, many statistical functions are also applied to the features which resulted in-total of 950 feature for every speech segment.

Visual Features For visual features, Dense Histogram of Gradients (DHOG), Dense Histogram of Flows (DHOF) and Dense Motion Boundary Histograms (DMBH) are extracted that have been used to capture the movements of subjects for human action recognition in videos (Uijlings, Duta, Sangineto, & Sebe, 2015). The purpose of using these features is to capture gesture movements, particularly hand gestures as depicted in Figure 3.17.

Figure 3.17: An example of impatience attitude where the attitude is also reflected in the hand gestures

The block size chosen for each descriptor (e.g. DHOG) is 6 by 6 pixel by 6 frames. For aggregating the descriptor response a single frame (out of 6 frames) for HOG (frame 3), and 3 frames (frame 2, 4 and 6) for HOF and MBH is used. The motivation behind this aggregation comes from a video classification task, where the authors report the best aggregation size for a human action recognition

problem (Uijlings et al., 2015). As a result, 144 descriptors are extracted for each aggregated frame. Later, a Fisher vector representation of the visual descriptor (Vedaldi & Fulkerson, 2008; Perronnin, Sánchez, & Mensink, 2010) is generated using the two common clusters size 64 and 256 for Gaussian Mixture Model(GMM) (Chatfield, Lempitsky, Vedaldi, & Zisserman, 2011). As a result, 18432 (cluster size = 64) and 73728 (cluster size = 256) features are extracted for each visual segment of attitude. The Fisher vectors are $l2$ normalised.

Feature Analysis

A high dimensionality of features are extracted so that it is not possible to evaluate the significance of each feature. So the Principle Component Analysis (PCA) is performed on the high dimensional data and then the ANOVA test is performed on the first three components of the PCA and report the most significant component p-values for both modalities (audio, visual). The visual feature set ($l2$ normalised Fisher vector using GMM cluster size of 256) is analysed after applying PCA.

There are in total 6 different attitudes so in multiple comparison tests there are 15 possible combinations, and their p-values are reported in Table 3.7. In most cases, acoustic features account for the only statistically significant differences. However, for group E-I (comparison between Friendliness and Enthusiasm) the visual modality differences are statistically significant, while audio differences are not.

Table 3.7: Multiple comparisons test results for six-class problem

Groups	Audio (PC1)	Audio (PCA2)	Video (PCA2)
A-E	**2.0701e-08**	0.40376	0.99183
A-Fd	0.40677	0.99347	0.99999
A-Fs	0.98516	0.57615	1
A-I	0.98516	**0.019799**	**0.038078**
A-N	0.73227	0.96758	0.99725
E-Fd	**7.0677e-06**	0.14032	0.99769
E-Fs	**2.0676e-08**	**0.0047534**	0.98143
E-I	0.66379	0.78285	**0.004536**
E-N	**2.0676e-08**	**0.073095**	0.89597
Fd-Fs	0.10479	0.90017	0.99989
Fd-I	**2.2372e-08**	**0.003089**	**0.028179**
Fd-N	**0.013747**	0.99991	0.99158
Fs-I	**2.0676e-08**	**2.0907e-05**	**0.049163**
Fs-N	0.97814	0.96251	0.99921
I-N	**2.0676e-08**	**0.0010494**	0.13143

Classification

In this task, the 'python Scikit-learn' [5] is used for random forest classifier (Liaw & Wiener, 2002; Breiman, 2002) model training and testing in 5-fold cross validation settings. The classification is performed on 608 instances of attitude because for some of the segments (5 out of 613) it is not possible to extract visual features. As the number of instances is small compared to the number of features, random forest learning is employed, which is robust in such situation as compared to other methods (discriminant analysis, support vector machines and neural networks) (Breiman, 2002). 2500 number of estimators for each feature set (e.g. DHOG) are selected and for all visual features (DHOG, DHOF, DMBHx DMBHy) fusion there are 10000 estimators, while there are 12500 estimators (number of trees in the forest) for audio-visual fusion. The reason to choose a large number of trees

[5] http://scikit-learn.org/stable/ – last verified Aug 2017

in the forest is the very high dimensionality of data (Breiman, 2002).

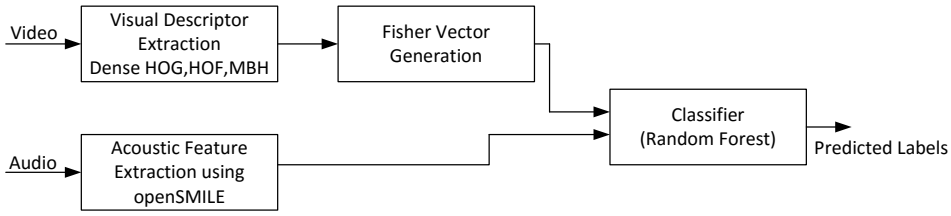


Figure 3.18: Attitude recognition process uses the feature fusion method

3.4.2 Results and Discussion

The results of the six-class problem are shown in Table 3.9 for each feature set and the detailed confusion matrices are reported in Table 3.8. From the results, it is observed that for the six-class problem all the features provide results above blind guess (16.67%) and DHOF provides better results than other visual descriptors, while the fusion of all the visual descriptors slightly increases the accuracy (almost 2%). The acoustic feature set provides the best results with an accuracy of 58.72% and the fusion of audio and visual features results in a small decrease in accuracy (almost 2%). From the confusion matrix reported in Table 3.8, it is observed that Amusement is difficult to detect as compared to other attitudes using the acoustic information, while Impatience is detected with low accuracy using the visual information. The fusion of both modalities results in a general decrease in accuracy but improves the accuracy of Impatience and Frustration. It is also observed that Frustration and Amusement are often confused which means that their audio-visual information appears similar to the classifier, hence the misclassification of amusement as frustration (24 for audio, 28 using video and 30 with fusion). Most of the Impatience instances (28 using audio and 38 using video) are

classified as Enthusiasm using audio and visual information, but here the fusion reduces this misclassification.

Table 3.8: Confusion Matrix for attitude Recognition. The visual analysis is performed using cluster size of 256 for GMM for fisher vector generation

Audio						
	A	E	Fd	Fs	I	N
A	**36**	8	1	24	12	19
E	4	**76**	2	4	17	4
Fd	4	8	**65**	7	2	10
Fs	12	5	1	**69**	6	10
I	12	28	0	4	**49**	9
N	17	6	2	9	4	**62**

Video						
	A	E	Fd	Fs	I	N
A	**12**	19	19	28	6	16
E	4	**38**	24	26	11	4
Fd	6	18	**50**	11	6	5
Fs	7	17	7	**52**	5	15
I	14	38	17	18	**8**	7
N	9	12	15	31	7	**26**

Fusion						
	A	E	Fd	Fs	I	N
A	**17**	11	1	30	16	25
E	1	**71**	4	5	22	4
Fd	2	11	**57**	9	7	10
Fs	6	8	0	**73**	3	13
I	3	17	2	6	**61**	13
N	6	10	2	13	5	**64**

Table 3.9: Accuracy (%) of classifier for six-class problem (blind guess (16.67%))and three-class problem (blind guess is 33.33%)

	6-Class Problem			3-Class Problem			
Features	majority Guess	GMM 64	GMM 256	majority Guess	GMM64	GMM256	Trees
DMBHx	16.67	25.00	28.29	49.84	51.15	**53.29**	2500
DMBHy	16.67	26.15	27.14	49.84	50.99	51.48	2500
DHOG	16.67	17.43	20.56	49.84	49.84	50.00	2500
DHOF	16.67	27.30	**28.78**	49.84	52.80	50.00	2500
Visual	16.67	25.02	**30.59**	49.84	50.66	51.64	10000
Audio	16.67	**58.72**		49.84	**63.98**		2500
Fusion	16.67	55.62	**56.41**	49.84	60.03	**58.06**	12500

Table 3.10: Number of instances (attitudes) along with classifier accuracy (six-class problem) in percentage for each subject

	Audio	Video	Fusion	A	E	Fd	Fs	I	N
S1	58.49	32.08	**60.38**	7	16	3	12	5	10
S2	60.38	33.96	**71.70**	1	8	2	13	19	10
S3	**79.41**	52.94	**79.41**	3	2	14	1	2	12
S4	**61.47**	33.03	55.05	26	25	10	25	14	9
S5	54.35	34.78	**58.70**	6	4	17	3	8	8
S6	**40.00**	14.29	37.14	4	2	10	6	3	10
S7	**64.13**	35.87	59.78	23	11	19	7	26	6
S8	**46.15**	26.92	44.23	21	22	12	27	9	13
S9	**61.76**	11.76	58.82	0	6	1	4	13	10
S10	**66.67**	22.92	50.00	10	11	8	5	2	12

The results of the three-class problem (depicted in Table 3.9 and Table 3.11) shows that the acoustic feature set performs better than visual features, and the

fusion causes a decrease in an overall accuracy. The visual features do not achieve better results (although the blind guess is 33.33%, the majority guess is almost 50%) that can be due to the data imbalanced nature for three class problem, because in the six-class problem the visual feature provides better results than baseline (where the blind guess (16.67%) and majority guess is almost the same due to the balanced nature of dataset). However, the audio features provide good results well above blind and majority guesses. One of the main reason of mis-classification might be that the inter-coder agreement between the two annotators of the data set is 75% reported by N. A. Madzlan et al. (N. A. Madzlan et al., 2015) and when a sub-corpus of the dataset is tested using the 20 subjects, the inter-coder agreement becomes far less with a k-value of 0.27 using weighted Fleiss Kappa (Fleiss, n.d.) as reported in (N. A. Madzlan et al., 2016). The results of the three-class problem are also contradicted with a previous study (N. A. Madzlan et al., 2015), one of the possible reason of this can be the introduction of the Neutral label in this study instead of using the Friendliness as Neutral (N. A. Madzlan et al., 2015). However, in the previous study, the imbalanced nature of the dataset (majority guess is 40%) does not affect the results as much as in this case.

Table 3.11: Confusion Matrix for attitude Recognition (three-class problem). The visual analysis is performed using cluster size of 64 for GMM for fisher vector generation

		Positive	Negative	Neutral
Audio	Positive	**253**	45	5
	Negative	92	**107**	6
	Neutral	66	5	**29**
Video	Positive	**298**	5	0
	Negative	195	**10**	0
	Neutral	94	6	**0**
Fusion	Positive	**265**	37	1
	Negative	116	**87**	2
	Neutral	81	6	**13**

The visual descriptors' poor performance in the three-class setting can be due to the fact that the gestures are correlated with a specific type of attitude (like Amusement and Frustration) instead of being related to the valence (positive and negative) of the attitudes. Although the results of the acoustic features are promising in both type of classification problem, they are tested in a mixture model setting where the training and test data may contain the same vlogger instances. The performance of the classifier may be reduced in speaker independent training settings and improved in speaker dependent settings. However, a detailed analysis of the predicted labels of six class problem shows that the fusion is able to increase accuracy for some subjects as depicted in Table 3.10. The dataset is balanced in terms of the number of attitudes, but it is not balanced for each subject, and that probably causes a decrease in accuracy for the fusion approach because the classifier trained on less instances of a particulate attitude for a video blogger. Moreover, the expression of attitudes is not in a natural interaction and probably some subjects are better in expressing their attitudes using both modalities (good actors) and some are not.

3.5 Political Debates: Data Collection and Synchronisation

A political debates training system (multimodal dialogue system with abilities to interact with humans in a natural way) could help humans in improving their presentation and negotiation skills through instructional advice. The development of a debate training system requires audio-visual data for training. The development

process is divided into different pilot studies. The goal of the initial pilot system (pre-pilot system) is to primarily observe users and give feedback on their interaction, regardless of whether the interaction is successful. The system uses audio and visual features and limits its interventions to inaction feedback (Helvert, Rosmalen, Börner, Petukhova, & Alexandersson, 2015) in the form of a green or red light, so as to minimise participant distraction during data collection. This limited form of feedback is meant to reflect the system's view on whether a participant is interacting in a "successful" way or not based on its analysis of audio and visual input features. Prosodic features, facial expressions and body gestures are used by the system to make a judgement about the participant's metacognitive skills. A metacognitive skill is defined as the ability of a participant in an interaction to understand, control and modify their own cognitive process. Such skills are believed to be useful in real life learning and training processes, and in debating skills in particular (Tumposky, 2004).

This study describes the data collection process of political debates, with participants of drawn from the Hellenic youth parliament [6] student cohort.

3.5.1 Related Corpora

This section discusses the possibility of using other available corpora for the purposes of the research outlined above, and their limitations.

IFA Dialogue Corpus

The IFA (Institute of Phonetic Sciences) dialogue corpus contains a collection of face-to-face dialogue videos with annotated labels. Even though the language is

[6] http://www.efivoi.gr/ — Last verified July 2018

Dutch the corpus gives examples of informal and friendly dialogue. This corpus could be useful to model friendly behaviour which can be used especially for training humans as call centre agents after annotating the corpus for meta-cognitive skills. There are in total 20 dialogue conversation videos, and each one lasts for around 15 minutes. There is no topic restriction imposed on the participants of the dialogue. The recording is performed with the two gen-locked JVC TK-C1480B analogue video cameras. The overall duration of the corpus is around 5 hours. To make the dialogue more useful and friendly, selection of participants is based on a subset of the following factors:

- Good friends
- Relatives
- Long-time colleagues

The corpus can be used and distributed under the GNU General Public Licence (an open source license) (van Son, Wesseling, Sanders, & van den Heuvel, 2008).

AMI Meeting Corpus

The AMI (Augmented Multi-party Interaction) meeting corpus (McCowan et al., 2005) consists of 100 hours of recordings. The corpus is multimodal since it includes several inputs: voice, video and writing. This corpus is annotated at different levels (Dialogue Acts, Topic Segmentation, Individual Actions, Person Location, Focus of Attention, AmiEmotion) and could be helpful to model the formal behaviour of a person in an interactive communicative situation. The language of the corpus is English and most of the participants are non native speakers. However, some of the

recordings are performed in different rooms with different acoustic properties. The AMI meeting corpus is released under a creative common attribution shareAlike license.

MIMLA Data

The MLA-14 data contains students' presentations (Ochoa, Worsley, Chiluiza, & Luz, 2014c). In total there are 441 oral presentations delivered by Spanish speaking students presenting projects about entrepreneurship ideas, literature reviews, research designs, software design etc. Recordings were placed in regular classroom settings and include multimodal data: speech, facial expressions and physical movements in video, skeletal data gathered from Kinect [7] for each individual, and presentations slides. In total, approximately 19 hours of multimodal data is recorded. In addition individual ratings for each presentation is included as well as a group grade related to the quality of the slides used when doing each presentation. Each presentation has a rating based on the following performance factors:

1. Structure and connection of ideas.
2. Presents relevant information with good pronunciation.
3. Maintains an adequate voice volume for the audience.
4. Language used in presentation according to audience.
5. Grammar of presentation slides.
6. Readability of presentation slides.

[7] https://developer.microsoft.com/en-us/windows/kinect – last verified Aug 2017

7. Impact of the visual design of the presentation slides.

8. Posture and body language.

9. Eye contact.

10. Self confidence and enthusiasm.

Limitations

While the above described corpora provide useful resources in terms of training and testing general models of dialogue interaction, none of them directly fits the scenario of political debates. The first two corpora (IFA and AMI) are relevant in terms of interactivity and multimodality, but lack the instructional element. The MIMLA corpus is also situated in an educational context, contains rich multimodal data, but lacks the dialogue and interactivity elements. These limitations motivated the data collection activity described below.

3.5.2 Data Collection Process

Several studies indicate that a presenter should speak with lively voice and make an eye contact with the audience. Moreover the presenter should stand straight, avoid crossing his legs, and use his hands, body and face to do gestures at the appropriate time (Stassen et al., 1993; Lamerton, 2001; Grandstaff, 2004; DeCoske & White, 2010). However, the face pose information is also helpful in increasing the speech recognition performance (Slaney, Stolcke, & Hakkani-Tür, 2014).

The political debates are also a form of public speaking situation which also require public speaking skills which are reflected in speech and body gestures of a

presenter. Hence, it is needed to gather the full skeleton/face data of participants along with audio and video recordings to models the presentation's skills. Since this setting requires the use of several independent cameras (Kinect and conventional video cameras), synchronisation issues need to be addressed. An example of such issues is the dropping of frames by the Kinect sensors during recording, which was addressed by duplicating some of the neighbouring frames. The motivation behind adding conventional cameras for recording is to obtain high quality frames for image analysis (emotion and effect recognition etc) in addition to Kinect's built-in skeleton tracking functionality.

This section describes the complete data collection process including recording settings, environment, room and equipments specification. Around 2 hours of audio-visual training material simulating the political debates is recorded. In total This section 11 sessions have been recorded and each session lasts around 10-15 minutes.

Recording Settings

The recording takes place in controlled settings and it includes a quite room, no windows behind participants and their faces are not in shadows. However, the participants are allowed to face their opponent and audience but restricted to remain in the field of view of Kinect during the debate. In a recording session, there are two students standing in-front of an audience and debating on a social issue (either smoking should be banned on public place or not). One student is in favour of smoking on public places and the other student is against it. Any of them can start the debate and after outlining their views (in 2-4 minutes), they listen to their opponent's intervention, taking turns as the debate proceeds. The overall

duration of a session is typically 10 to 15 minutes. A schematic representation of this recording set up along with recording hardware is shown in Figure 3.19.

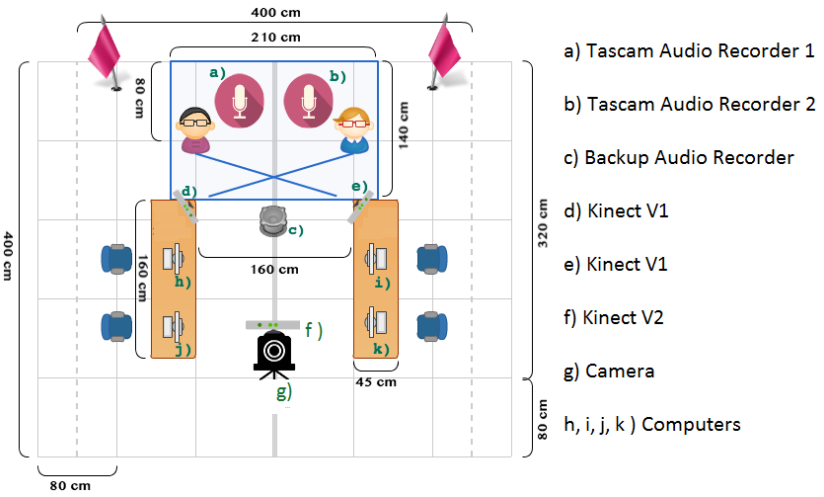


Figure 3.19: Recording Settings

Wizard of Oz Software

A simple WOZ system prototype has been implemented in Python[8] which consists of two programs:

RedGreenUser.py the user's interface, that is, a frame displaying two panels (a red and a green one) which light up according to the feedback sent to the participants by the wizard. The program starts as a server and listens for feedback from up to 10 concurrent connections, which makes it possible for multiple wizards to control the interface collaboratively.

RedGreenWOZ.py the wizard's control panel, through which the wizard chooses different categories of feedback to send to the participants' screen.

[8]http://www.python.org/ – last verified Aug 2017

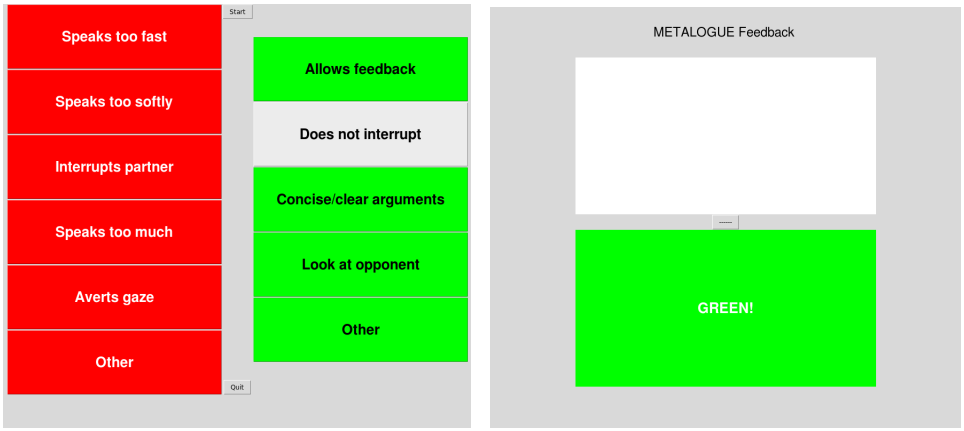


Figure 3.20: Wizard (left) and participant (right) user interfaces.

Although the feedback is categorised, the participants only sees unspecified red or green feedback on their screen. The wizard's and participant's user interfaces are shown in Figure 3.20. The types of "red" (negative) feedback the wizard can choose from are:

- a participant is speaking too fast,
- a participant is speaking too softly,
- a participant has inappropriately interrupted the other participant,
- a participant speaks too much,
- a participant averts gaze,
- other (a "catch all" category).

The types of "green" (positive) feedback the wizard can choose from are:

- a participant allows the other to provide feedback
- a participant does not interrupt the other's speech,

- a participant presents clear and concise arguments,
- a participant looks at the other while debating,
- other.

When any of these items is chosen by the wizard, RedGreenWOZ.py timestamps it and writes a log of this feedback to an XML file. In these trials two such files are recorded per session, one containing the actual WOZ feedback (as displayed on the participant's screen) and an additional file generated by a "silent wizard" who interacts with an instance of RedGreenWOZ which simply records the feedback without actually presenting it to the participants. This file can be useful, for instance, in assessing the level of agreement between the two wizards through comparison of the feedback given by each of the participants on the time line. For instance, a wizard might decide to give a "red" feedback when participant A interrupts participant B, while the other wizard might decide to give "green" feedback to participant B for allowing A to provide feedback.

The wizard log files will be synchronised with the other media streams gathered during the sessions. See deliverable for further details on the processing and storage of these data.

First sitting for Recording

In this sitting, three recording sessions (10-15 minutes each) have been recorded using a real time feedback light (red or green) which is simulated by a wizard (as described in section 3.5.2). Two Kinects sensors are used to track body skeleton and facial landmarks and their details are saved in XML files. A video camera is

also used to record the whole recording session besides Kinect sensors. A snapshot of the recording set-up is shown in Figure 3.21.

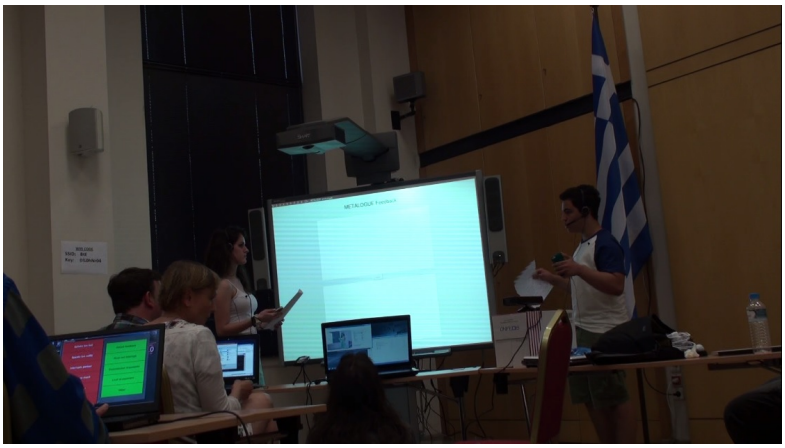


Figure 3.21: A snapshot of recording settings of first sitting recordings using WOZ software

Second Sitting for Recording

In this sitting, in total 10 sessions have been recorded. Three Kinect sensors (two Kinect V1 and one Kinect V2) have been used to track the movements of skeletons and facial features. To avoid the inaccuracies in tracking, the Kinects' fields of view should not overlap. However, in this case the proposed set-up have an overlap which might affect the tracking performance of Kinect. Two Kinect V1 sensors, each facing one participant as much as possible, is placed at a distance of 1.5-2 meters to the participants. Participants are facing each other and/or audience, and markers will be placed on the floor (movement space: 50cm x 50cm). For Kinect V1, the skeleton and face information are tracked and time stamped in real time and saved in XML files. Beside that Kinect V2 is also used to record the raw data using Kinect Studio 2.0 for off-line processing. The motivation behind

Figure 3.22: A snapshot showing Synchronised video Streams and Kinects Tracking of second sitting recordings (WOZ software is not used.)

using the Kinect V2 is that to create backup in case any Kinect V1 crashes or vice versa. In future, Kinect V2 raw data can also be used to extract the new feature (hand open or close etc.) The difference between previous sitting and this one is the removal of feedback tool (described in Section 3.5.2), introduction of Kinect V2 and placing marker on the floor instead of just telling the participants to be remain in a range.

Data Synchronisation

Synchronisation of video and audio streams is performed using Final Cut pro X and a snapshot of the synchronised session is shown in Figure 3.22. Moreover the details of each synchronised stream (e.g. time offsets) are imported in FCPXML files. The Kinect devices were started manually so these offset help us to synchronise the data with the other streams. To accomplish this objective, the tracked XML files are parsed using python and update their time stamps (by adding or subtracting the offset). Automatic speaker diarization is performed using the LIUM toolkit

Table 3.12: Speaker ID, their role (pro and against) and the location of speaker (left or right)

Setting	Session No.	Speaker ID (Left)	Role	Speaker ID (Right)	Role	Duration
Pre Pilot	1	S5	pro	S6	against	11:10
1st sitting	2	S6	against	S4	pro	09:50
	3	S4	against	S5	pro	07:07
	1	S0	pro	S1	against	10:25
	2	S2	pro	S3	against	11:39
	3	S0	pro	S2	against	09:43
	4	S4	pro	S1	against	12:05
	5	S5	pro	S0	against	13:26
Pre Pilot	6	S3	pro	S5	against	13:09
2nd sitting	7	S1	pro	S4	against	13:11
	8	S4	pro	S0	against	19:39
	9	S1	pro	S3	against	16:31
	10	S3	pro	S4	against	11:49
	11	S5	pro	S0	against	13:18

(Rouvier et al., 2013) and speech chunks are extracted using speaker diarization information.

Speaker Characteristics

The participants of debate were recruited by the HeP (Hellenic Parliament)[9]. The speaker are young school students from two different schools and were aged between 17 to 20 year. They are non native speakers (Greek) of English, know each other, have participated in HeP annual debating sessions and take part in data collection activities as volunteers. There are in total 6 participants, three females (S2, S3 and S5) and 3 males and their role (smoking should be banned on public places (smoking should be banned on public places (pro).) and the opposite (smoking shouldn't be banned on public places (against).) and duration of each session is depicted in Table 3.12.

[9]https://www.hellenicparliament.gr/en/

3.6 Conclusion

This chapter mainly presents public speaking training systems for three different kind of public speech and also describes a data collection activity of students' debates which can be used to train a machine to provide instructional advice to students for preparing them for debates. The conclusion of each study is described below.

3.6.1 Students Presentations

This study presents a system for exploiting audio-visual features for public speaking abilities detection and shedding light on how prosodic and visual features are related to the delivery of a presentation. The proposed system and empirical findings may be useful in the field of multimodal learning analytics which seeks to analyse different aspects of public presentations in order to understand the learning process and provide feedback to the trainee presenter. These techniques have been implemented as a component of a multimodal dialogue system intended to monitor the presentation performance of a public speaker, for example the EU METALOGUE project (Alexandersson et al., 2014). The results of this study are published in international conference (Haider, Cerrato, Campbell, & Luz, 2016).

3.6.2 TED Talks

The high level visual, paralinguistic and spoken expression feature have a relationship with user engagement. Camera views are not based on some random selection. But it is not possible to increase engagement level by just increasing or decreasing the number of camera views in a video. These views are depended on

the speaker's way of speaking, if he/she displays emotions or do gestures then a professional camera man/editor may take his/her close up shot and if he/she uses body language and slides then a camera man will focus through 'Distance Shots' and at the end if some content is really important then a camera man/editor just show the slides. Believing that the camera men/editors are highly professional, these views may help in building some feedback to the speaker e.g. they need to show more content (because their video has less number of 'person not on screen' camera shots), use of more body language (because their video has less number of 'Distance Shots') and facial expressions (because their video has less number of 'close-up shots'). Moreover, these features can also predict the engagement level of a talk. The proposed approach also demonstrates that characteristics of speech can be used to detect the engagement level of a talk. It is also a step towards generating feedback in the form of video chunks for presenters so that they will know the parts of videos which are engaging or not. It is also possible to create summary of a TED talk using the engaging parts of a talk. The results of this study are published in international conferences (Salim, Haider, Conlan, Luz, & Campbell, 2015; Haider et al., 2015; Haider, Salim, et al., 2017)

3.6.3 Attitude Recognition of Video Bloggers

This study have shown that acoustic and visual features can be used to detect the set of attitudes labelled in a corpus of vloggers with good accuracy. While the fusion of audio-visual descriptors does not improve accuracy in general, it improves accuracy of Frustration and Impatience detection in the six-class problem. It is also observed that the fusion causes an increase in accuracy for some subjects.

Based on the results of the three-class and six-class problems, it is concluded that the attitude recognition system can provide better results for balanced datasets (sixclass problem). The results of this study are published in international workshop (Haider, Cerrato, Luz, & Campbell, 2016).

3.6.4 Political Debates: Data Collection and Synchronisation

The data recorded with the setup and according to the procedures described in this study have been made available to the public using cloud server[10]. Approximately 3 hours of data have been recorded, and all recorded streams have been precisely synchronised and pre-processed for statistical learning. The data consists of audio, video and 3-dimensional skeletal movement information of the participants. The novel collected data will facilitate the development of a dialogue system which will exploit metacognitive reasoning in order to deliver feedback on the user's performance in debates and negotiations. This data collection activity is published in international conference (Petukhova et al., 2018) and workshops (Haider, Luz, & Campbell, 2017, 2016b).

[10]metalogue.scss.tcd.ie/owncloud/ – Last verified September 2017

Chapter 4

Cognitive Processing Components for Interactive Systems

4.1 Introduction

This chapter describes methods used to propose the systems which can improve the interaction abilities of machines. It presents novel systems which can automatically detect if someone speaks to machine or not using multi-sensor information. A cognitive state detection system is also proposed which can help in sensing the user experience with a machine translation system in the form of cognitive states (amusement, frustration, surprise and neutral). In the end, an active speaker detection system is proposed which can detect who is speaking. The validation of hypothesis, systems' evaluation and discussion are also the part of this chapter.

4.2 Speaking to a Machine or Not?

This section describes the novel methods for detection of On-Talk (Speaking to a machine) and Off-Talk (Speaking but not speaking to a machine).

4.2.1 Dataset

A data set was used which consists of recorded human dialogues mediated through a speech-to-speech machine translation system (Hayakawa, Campbell, & Luz, 2014). The participants communicated remotely through the system to solve a map task problem, where one participant (the instruction giver) has a complete map and the other (the instruction follower) has a map with missing information (Anderson et al., 1991). Three different types of talk were observed in this setting: 1) on-talk, where the speaker directed speech to the ASR for transmission to the other participant 2) self-speaking, where participants spoke to themselves (e.g. venting frustration at system component failure) producing utterances not intended for ASR or transmission, and 3) other-talk, where participants spoke directly to other people than their remote task partner (e.g. a colleague that happened to be in the same room). Both self-speaking and other-talk are regarded as off-talk in this study. The data used for the research described in this paper includes precisely synchronised audio and EEG signals, from the Interlingual Map Task (ILMT-s2s) corpus (Hayakawa, Luz, Cerrato, & Campbell, 2016).

The ILMT-s2s System In the recording setup of the system, two subjects are communicating with each other from two different locations (rooms) using the ILMT-s2s system. The ILMT-s2s machine translation system is developed

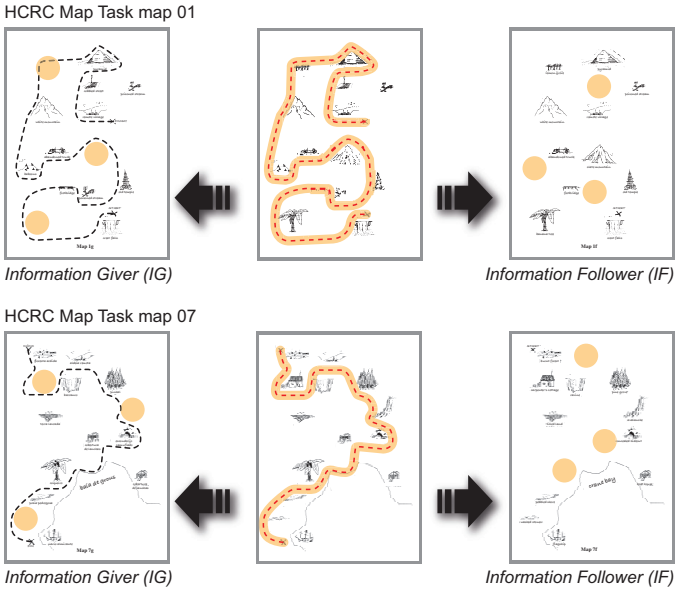


Figure 4.1: Maps, with differences highlighted

using off-the-shelf technologies as depicted in Figure 4.2. The subjects use a push to talk button to speak to the other person. However, they cannot hear each other, and the speech synthesis technology provides the output of ASR and MT to them as depicted in Figure 4.3 . Only one of the dialogue participants uses the physiological recording equipment (Hayakawa, Luz, et al., 2016) in any particular session. In total, there are 30 participants (15 English and 15 Purtogise speakers). However, half of them are equipped with the biosignal recording equipment, and the duration of dialogue is between 20 and 74 minutes. This study uses the datasets of 13 subjects (who are equipped with bio signals equipment). The number of on-off talks produced by all subjects along with the mean and standard deviation values is depicted in Table 4.1.

Audio Recordings Two audio and five video streams are part of the ILMT-s2s corpus. However, this study uses the audio recorded by the two video cameras not

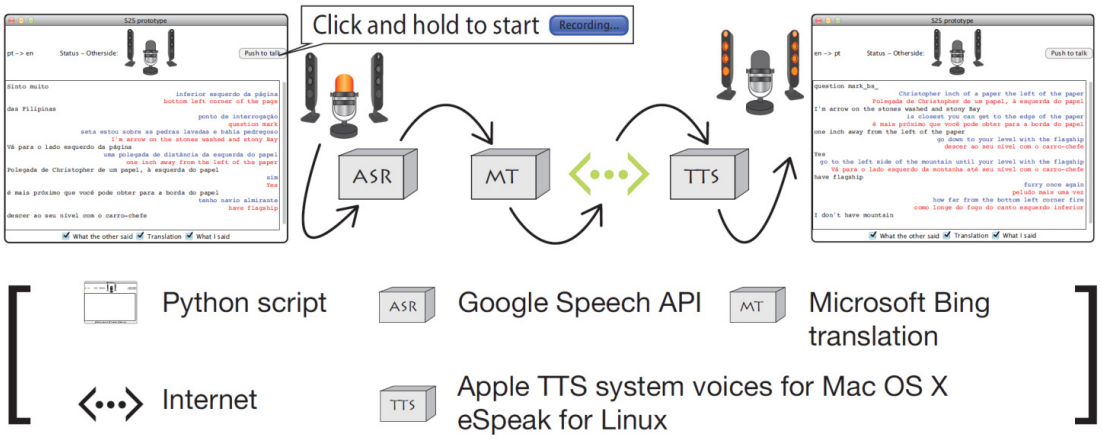


Figure 4.2: *ILMT-s2s System used to collect the data*
(Hayakawa, Luz, et al., 2016)

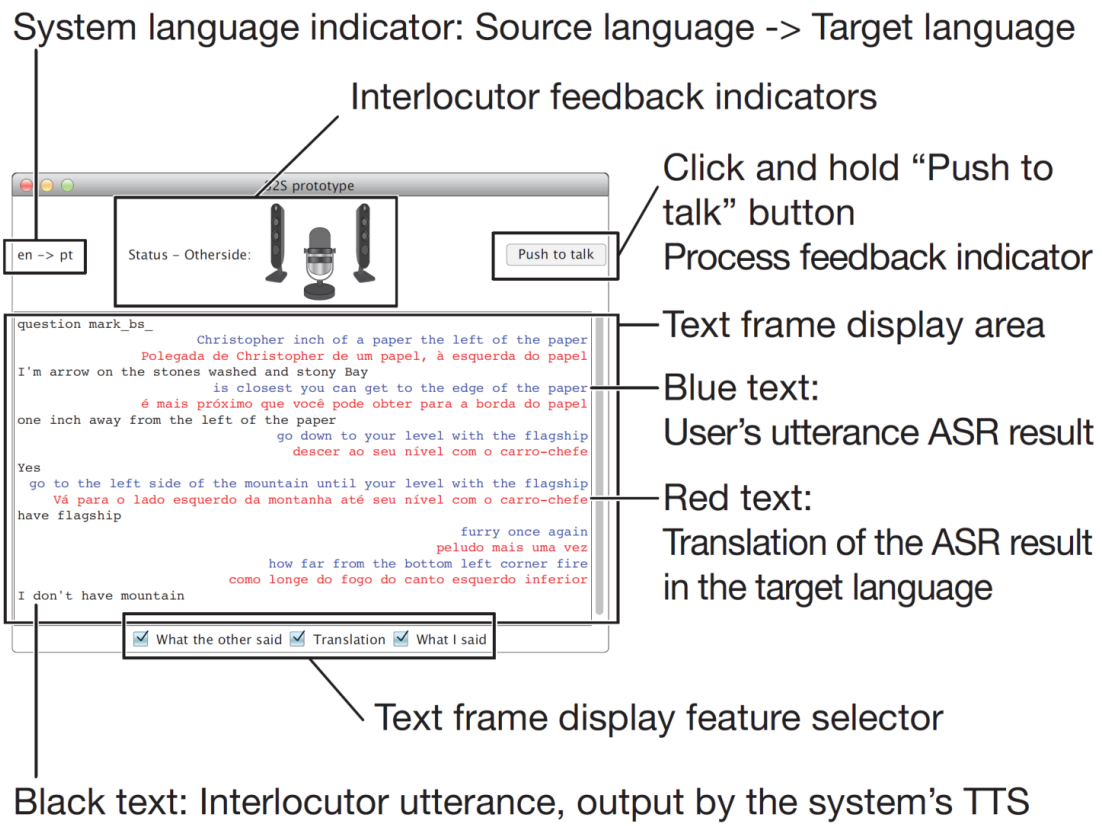


Figure 4.3: *User Interface of the ILMT-s2s System*
(Hayakawa, Luz, et al., 2016)

Table 4.1: Dataset description along with number of subjects their on-off talk instances with mean and standard deviation of duration (in seconds)

Subject	Self Talk			On Talk			other Talk		
	Instances	mean	Std.	Instances	mean	Std.	Instances	mean	Std.
S1	25	0.65	0.26	33	0.87	0.88	0	-	-
S2	55	0.77	0.70	100	2.1	7.8	4	0.81	0.54
S3	10	0.65	0.31	120	1.87	7.13	30	0.83	0.90
S4	5	0.74	0.24	98	1.65	7.83	0	-	-
S5	46	0.73	0.53	92	2.2	8.12	32	0.84	0.74
S6	60	2.35	10.01	103	1.21	1.32	57	0.89	0.80
S7	40	0.77	0.81	73	2.27	9.08	6	0.79	0.25
S8	2	0.89	0.26	201	1.52	5.56	0	-	-
S9	106	1.61	7.53	49	1.66	1.8	0	-	-
S10	10	0.83	0.28	57	0.76	0.70	0	-	-
S11	44	0.75	0.72	78	2.2	8.79	0	-	-
S12	13	0.72	0.29	34	0.79	0.86	12	0.75	0.38
S13	6	0.84	0.31	89	1.63	8.20	1	0.93	0

by the push to talk microphone.

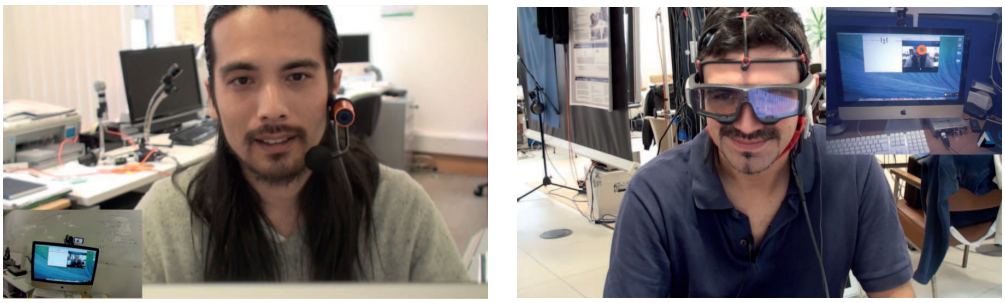


Figure 4.4: *Recording setup.*

EEG Recording The Electroencephalography (EEG) is recorded using the Mind Media B.V., NeXus-4. The EEG sensors are placed in the F4, C4, P4 (located at right hemisphere of the brain that is responsible for the control of speech prosody (Shapiro & Danly, 1985; Weintraub et al., 1981)) with a ground channel placed at A1 of the 10–20 location system as depicted in Figure 4.5. The sampling frequen-

cies for the EEG is 1,024 Hz.

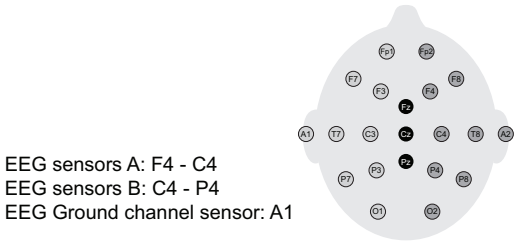


Figure 4.5: *10 – 20 system layout map*

4.2.2 On-Off Talk Detection

This study has evaluated the discrimination power of different modalities (audio and bio) and proposed novel system to predict the on-off talk.

For the following experiments, the start and end times of the *On-Talk*, *Off-Talk* label annotation were used to segment the synchronised audio and biosignal files. Two of the fifteen EEG recordings provided faulty readings and were excluded from the dataset. This resulted in 1,127 *On-Talk*, 554 *Off-Talk* (422 *Self* and 132 *Other*) utterance locations being used for this experiment. For the detection of *On-Talk* and *Off-Talk* audio and biosignals are extracted, and the potential use of these features is explored to identify *On-Talk* and *Off-Talk*.

Exp. 1: A 2-Class experiment which distinguish the difference between *On-Talk* and *Off-Talk*.

Exp. 2: A 3-Class experiment which distinguish the difference between *On-Talk*, *Off-Talk Self* and *Off-Talk Other*.

Feature Extraction

The following features were used for the classification task.

Audio features: For the classification task, the INTERSPEECH 2013 Computational Paralinguistic Challenge (ComParE) feature set (Schuller et al., 2013b) is used. This contains energy, spectral, cepstral (MFCC) and voicing related low-level descriptors, as well as other descriptors such as logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness. To ignore the most irrelevant acoustic feature, K Means clustering algorithm is employed. This divides the feature set into 9 clusters and of these only the cluster with highest number of features is selected for classification. As a result, the total number of acoustic features reduces from 6,373 to 6,356.

Biosignal features: For the biosignals (HR, SC and EEG), Shannon entropy, mean, standard deviation, median, mode, maximum value, minimum value, maximum ratio, minimum ratio, energy and power are calculated. This feature set is calculated for each biosignal and its first and second order derivative. In total 33 features for each biosignal are extracted. The EEG lower gamma (frequency band) signals from sensor A and B (10 – 20 system: F4–C4 and C4–P4) are considered in this initial study to evaluate the EEG discrimination power which leads to further studies reported in Section 4.2.3 and 4.2.4. The minimum ratio of an observation is measured by counting the number of instances which have a lower value compared to their preceding and following instance and then dividing it by the total number of instances in that observation. Similarly, the maximum ratio of an observation is measured by counting the number of instances which have a higher value compared to their preceding and following instance and then dividing it by the total number of instances in that observation.

Classification Method

The classification method was implemented in MATLAB[1] using Statistics and Machine Learning Toolbox and employed discriminant analysis in 10-fold cross validation experiments. The classification method works by assuming that the feature sets of the classes to be discerned are drawn from different Gaussian distributions and adopting a pseudo-linear discriminant analysis (i.e. using the pseudo-inverse of the covariance matrix (Raudys & Duin, 1998)).

Result: Detection of On-Talk & Off-Talk

The following results were obtained. See Table 4.2 and Figure 4.6 for details.

Table 4.2: *Discriminative Analysis Method Results – F-Score (%)*

Signal	Experiment 1		Experiment 2		
Class (Talk type):	On	Off	On	Off Self	Off Other
EEG	79.91	34.47	79.89	26.55	5.13
HR	80.41	31.55	80.44	21.68	8.54
SC	81.61	48.30	81.39	36.31	5.19
HR + SC	80.85	49.40	81.64	40.12	11.56
All Bio (EEG+HR+SC)	80.32	50.31	80.88	42.60	13.33
Audio	94.14	87.55	94.00	77.60	36.64
Audio + EEG	94.31	87.94	94.13	78.50	33.60
Audio + HR + SC	94.87	88.91	94.17	77.17	34.62
Audio + All Bio	94.09	87.47	94.38	79.08	38.13

The results of experiment 1 show that the acoustic and biological measures significantly contribute to the prediction of *On-Talk* and *Off-Talk*. The acoustic feature set provides the optimum performance with a maximum F scores of 94.14% for On-Talk and 87.55% for *Off-Talk*. Also, the SC feature set performs better than other biological features, but a fusion of the bio feature sets causes an increase

[1] http://uk.mathworks.com/products/matlab/

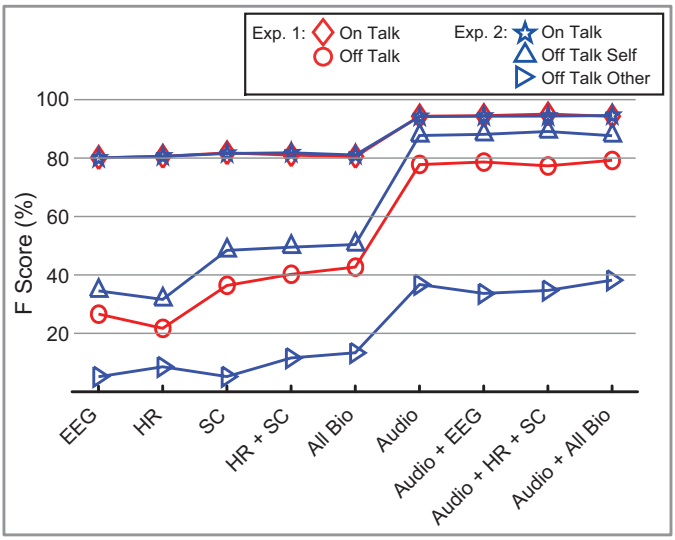


Figure 4.6: *Discriminative Analysis Method Results*

in prediction. However, a fusion of acoustic and bio features improves the performance in two cases, but has almost no effect as compared to audio feature alone when audio features are fused with all the bio features. From the results of experiment 2, it can be seen that the 3-Class results for *On-Talk* are almost the same as the 2-Class *On-Talk* results. Also results for *Off-Talk Other* are poor using bio features alone (max. 13.33%) but significantly improve when combined with the acoustic feature set (38.13%) — considering that the dataset is imbalanced, with less instances for *Off-Talk Other* (7.85%) these results can be regarded as quite good. The HR is found to have more prediction power as compared to EEG and SC and the fusion of biosignals improves the prediction. A decrease in *Off-Talk Other* results is observed when audio (36.64%) feature set is combined with EEG (33.60%) and with HR and SC (34.63%) feature sets. This might be due to the lower number of bio features since when they are fused all together (All Bio: HR, SC, and EEG) with an increase in the number of bio features, the highest F-Score

(38.13%) is obtained as expected.

Although the acoustic feature set performs best as compared to other signal sets, it is believed that there is still room for improvement from the biosignals since they currently use a limited number of features (only 33 features for each signal) and may contain some noise components (head movements of subjects etc).

4.2.3 Improving Response Time of On-Talk & Off-Talk Detection System

This section describes the proposed novel models for automatic detection of on and off-talk using prior to articulation EEG information which could decrease the response time of an interactive speech driven system in accepting or rejecting a speech utterance as depicted in Figure 4.7. This model employs electroencephalography (EEG) features collected prior to articulation. The alternative models are assessed that employ such features in isolation and in combination with prosodic feature for on- and off-talk detection. The EEG signal is recorded from the right hemisphere of the brain, which is the area responsible for the control of speech prosody (Shapiro & Danly, 1985; Weintraub et al., 1981; Ross & Mesulam, 1979).

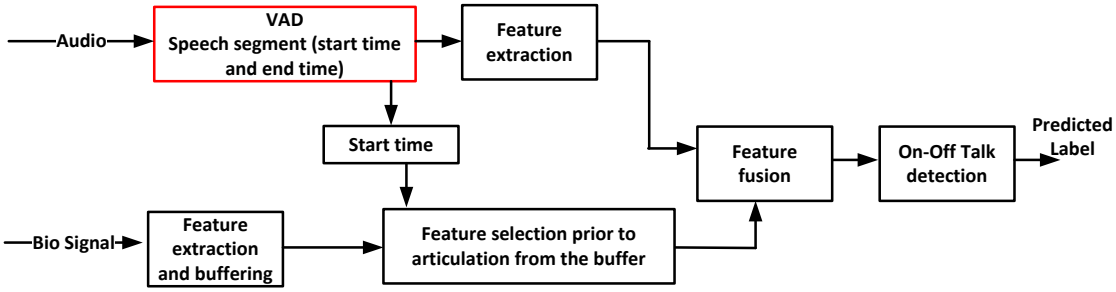


Figure 4.7: The system architecture where the system processes the EEG features prior to articulation as soon as it received 10 ms of audio.

Feature Extraction and Classification method

EEG Power Spectrum Feature Extraction (EEG Frame): The feature extraction is performed on the EEG signal two seconds before articulation. A frame rate of 250 ms is used for feature extraction. The first step is to take the Fourier transform of the EEG signal frame and then calculate the power spectrum. Later a frequency bin resolution of 5 Hz (from 0-40 Hz) is set that resulted in 8 frequency bins. The frequencies above $40Hz$ are ignored in this study, in line with clinical standards. It is noted that, while contrary to a common misconception the human skull does not filter out higher frequencies (Gotman, 2010), neural activity at such frequencies is harder to detect due attenuation caused by the skull's resistivity (Oostendorp, Delbeke, & Stegeman, 2000). In future work, It is intended to explore higher EEG frequency bands. For the current study, however, the power in each frequency bin bellow 40Hz, and the ratio and range of power between all eight frequency bins were calculated, resulting in 64 features per frame for each EEG sensor.

Prosodic Features: The mean and standard deviation value of sound intensity, loudness and fundamental frequency were extracted from the speech segments using the openSMILE toolkit (Schuller et al., 2013b).

Classification Methods: This study uses the Scikit-learn[2] implementation of the random forest (RF) classifier (Liaw & Wiener, 2002; Breiman, 2002) for model training and testing in a 10-fold cross validation setting. As the number of instances is close to the number of features in some of the cases, random forest (250 trees) learning is employed, which is robust in such situations as compared to other methods (discriminant analysis, support vector machines and neural net-

[2]http://scikit-learn.org/stable/ – last verified: May 2017.

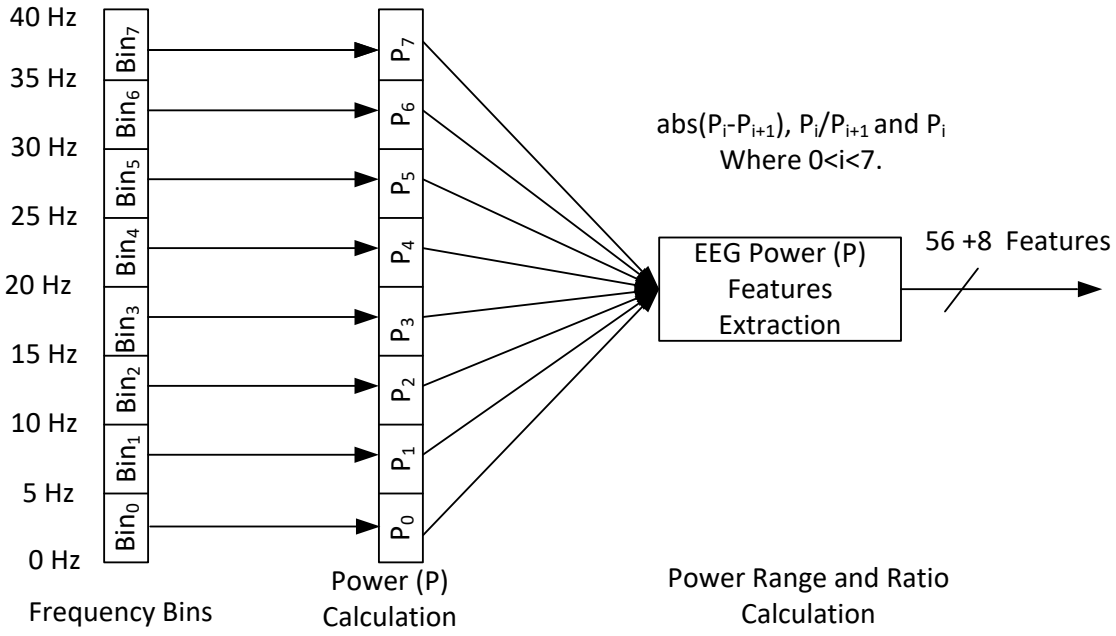


Figure 4.8: *Frame (250 ms) level feature Extraction on EEG Signal*

works) (Breiman, 2002). K-nearest neighbour (with K=3) is also used for comparison. The results are assessed using the A-weighted F_1-score statistic which is the average of F_1-score of both classes (on- and off-talk). The baseline is 50%.

Results and Discussion

The discrimination power of EEG signals two seconds prior to articulation with a frame size of 0.25 seconds on 0–40Hz frequency bands is evaluated. In total, the following three different experiments are conducted.

Experiment 1: This setting evaluates the discriminative power of eight frames (250 ms) of EEG (2 seconds before articulation) for on-off talk classification using the KNN and random forest classifiers. The EEG signal from both sensors is used in this experiment. The prediction power of each sensor is evaluated individually and in combination (feature fusion of Sensor A and Sensor B). The features of

all frames are then fused for classification. The results are depicted in Table 4.3. The best result (80.25%) is obtained using the frame level features of both EEG sensors that are extracted from two second prior to articulation.

Table 4.3: 10-fold cross validation Results (A-Weighted F-Score %) for each frame before articulation, and feature fusion of one second (4 frames) and two seconds (8 frames) before articulation.

EEG Features	Window (sec)	Sensor A (F4-C4)		Sensor B (C4-P4)		Fusion	
		KNN	RF	KNN	RF	KNN	RF
EEG Frame 1	0.00-0.25	53.90	72.09	52.26	63.27	53.12	79.40
EEG Frame 2	0.25-0.50	54.33	73.00	55.68	64.92	53.94	79.08
EEG Frame 3	0.50-0.75	51.24	70.99	52.61	63.86	56.17	79.21
EEG Frame 4	0.75-1.00	54.88	72.20	51.99	62.08	56.45	78.70
EEG Frame 5	1.00-1.25	53.65	71.24	54.23	64.67	54.20	79.05
EEG Frame 6	1.25-1.50	55.38	71.86	52.99	64.57	55.49	79.38
EEG Frame 7	1.50-1.75	55.66	72.75	55.52	66.36	57.57	79.21
EEG Frame 8	1.75-2.00	55.98	71.66	54.65	64.75	56.22	78.08
EEG Frame 1S	0.00-1.00	57.52	73.75	52.96	64.01	55.75	79.50
EEG Frame 2S	1.00-2.00	51.80	72.65	53.84	66.82	56.21	79.46
EEG Frame (1S+2S)	0.00-2.00	54.12	74.04	52.79	64.93	54.97	80.25

Experiment 2: The prosodic features are used for the classification task, and the results of the prosodic features are depicted in Table 4.4. The best result (81.83%) is obtained using the fusion of prosodic features (mean and standard deviation of loudness, intensity and fundamental frequency).

Experiment 3: The results of experiments 1 and 2 suggest that the EEG signal of all eight frames and their combinations are able to predict the on- and off-talks, well above the 50% baseline. It was also observed that the EEG signal fusion from both sensors also improves accuracy. Therefore, the acoustic information are fused with the EEG features of both sensors and two second before utterance, and performed inference. The fusion of EEG and prosodic features improves results as depicted in Table 4.4.

Discussion: The classification results show that EEG features prior to artic-

Table 4.4: 10-fold cross validation Results (A-Weighted F-Score %) with a baseline of 50%

Features	KNN	RF
Intensity	74.37	NaN
Loudness	75.89	75.24
Fundamental Frequency (fo)	67.49	68.63
Audio Fusion	68.29	*81.83*
Audio + EEG Frame 1	53.44	86.38
Audio + EEG Frame 2	54.87	86.72
Audio + EEG Frame 3	56.56	86.76
Audio + EEG Frame 4	56.48	86.39
Audio + EEG Frame (1S+2S)	54.93	85.95

Table 4.5: Confusion Matrix of the best results obtained from the three experiments

	Fusion		Audio		EEG	
	Off-Talk	On- Talk	Off-Talk	On- Talk	Off-Talk	On- Talk
Off-Talk	407	146	378	175	391	162
On- Talk	40	1086	82	1044	126	1000

ulation can predict on- and off-talk as depicted in Table 4.3. EEG signal from F4-C4 location (sensor A (74.04%)) provides better results than the C4-P4 location (sensor B (66.82%)). The predictive accuracies of both EEG Sensors are also compared using the mid-p-value McNemar test with a null hypothesis which is sensor A and sensor B have equal accuracy for predicting the target (on- and off-talk). The statistical test reject the null hypothesis ($p = 0.01$). The RF classifier provide better results than KNN using EEG features and the both classifiers predictive accuracies are also compared using the mid-p-value McNemar test with a null hypothesis which is the KNN and RF classifiers both have equal accuracy for predicting the target (on- and off- talk). The statistical test rejects the null hypothesis ($p = 3.2688e - 39$).

The EEG signal of 2 seconds prior to articulation is investigated. The most

discriminative time location in EEG signal for classification is frame 1 (0.000.25 seconds before utterance) and the fusion of 8 frames (2 second prior to articulation) yields an increase in performance. However, the fusion of both sensor features also improves the performance. The confusion matrix of the best results obtained from the three experiments are shown in the Table 4.5. The mid-p-value McNemar test is also conducted to compare the best results of the three experiments with a null hypothesis which is the audio, EEG and fusion features have equal accuracy for predicting the target (on- and off- talk). The test rejects the null hypothesis for 'EEG and fusion' ($p_{Exp.1-Exp3} = 2.7336e-12$), and 'audio and fusion' ($p_{Exp2-Exp3} = 4.4818e-07$) but unable to reject the null hypothesis for 'EEG and audio' ($p_{Exp2-Exp1} = 0.11$).

The basic prosodic features are also evaluated for the classification task that results in better results than EEG features, and the fusion of EEG and acoustic features improves the accuracy. However, the EEG system has a quick response time (RTEEG) as compared to the prosodic system (RTAudio) as depicted in Figure 4.9.

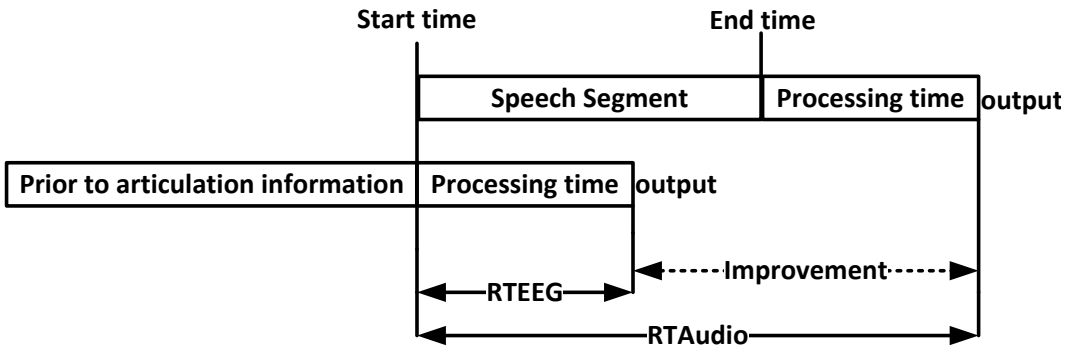


Figure 4.9: The baseline of the response time (RTAudio) and the proposed system response time (RTEEG)

To explore the mutual information of the best results of each experiment, the

Venn diagram is used as shown in Figure 4.8. The blue circle represents the labels (target), yellow circle represents the predicted labels using audio, green circle represents the predicted labels using EEG and the red circle represents the predicted labels using fusion of audio and EEG. From the Venn diagram, it is observed that there are 1207 instances which are correctly recognised by all three experiments (EEG, Audio and fusion of 'EEG and Audio'). However there are 75 instances (70 are off-talk and 5 are on-talks) which have not been recognised. Those instances belong to S1 (7 off-talk), S2 (2 on-talks), S3 (20 off-talk), S4 (1 on talk), S5 (25 off-talk), S6 (3 off-talk and 2 on-talk), S8 (2 off-talk), S9 (1 off-talk), S10 (6 off-talk) and S13 (6 off-talk). The fusion is able to recognise 11 instances (7 on-talk and 4 off-talk) correctly which have not been recognised by EEG and audio. Those instances belong to S4 (1 off-talk) , S5 (2 off-talk and 1 on-talk), S8 (1 off-talk and 2 on-talk), S10 (1 on-talk) and S11 (3 on-talk).

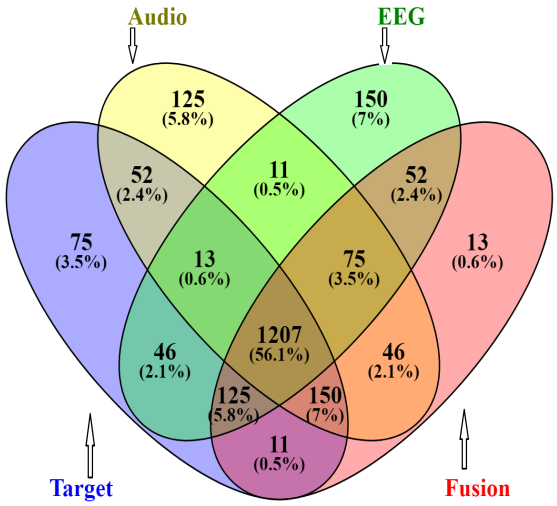


Figure 4.10: Venn Diagram showing the mutual information obtained from the best results of the three experiments.

The right hemisphere is responsible for the control of speech prosody, the results

obtained from the EEG signal may be due to the prosodic information processing in the brain. The audio features (81.83%) provide almost the same results as the EEG features (80.25%). This may be due to the fact that the intonation pattern of produced speech is defined before articulation, as suggested by previous studies (Bock, 1982; Dell, 1986; Garrett, 1975, 1988; Kempen, 1977; Kempen & Hoenkamp, 1987; Levelt, 1989). The loudness feature provides better results (75.24%) than other prosodic features, highlighting the importance of speech volume variations in distinguishing between on- and off-talk. This is consistent with the observation by Batliner et al. that users tend to interact with an ASR system as they would with a person who has a hearing impairment (Batliner, Hacker, & Nöth, 2009). Although the prosodic features provide slightly better results than EEG, the prosodic models may perform less well in a noisy environment as the data is collected in a controlled acoustic environment.

The failure of system components (e.g. ASR) may results in a kind of behaviour that is not as common as in human-human communication, as discussed above. Therefore, it might be assumed that a brain-computer interface (where there is no overt speech) might experience the same situations (the brain signal reading components fails) that results in a neural activity (off-thoughts) which should not be processed by the system. While the proposed models may also work in a covert speech situation (on- and off-thoughts), the brain has different activity patterns for overt and covert speech (Christoffels, Formisano, & Schiller, 2007; Pei et al., 2011), which may cause a decrease in accuracy of the proposed models for brain-computer interfaces.

4.2.4 On-Talk & Off-Talk Talk Detection using Wavelet Analysis

In this section, wavelet analysis is performed on the EEG signal to demonstrate the discrimination power of different EEG frequency bands. A novel automatic On-Off Talk detection system is also proposed using the EEG Signal.

EEG Signal Decomposition

The EEG signal (S) is decomposed into 11 components using the Discrete Wavelet Transform (DWT) using MATLAB,[3] where $S = d1 + d2 + d3 + ... + d10 + a10$ as shown in Figure 4.11 and Figure 4.12. The DWT helps us in evaluating the discrimination power of each component ($d1, d2$ etc) for *On-Talk Off-Talk* prediction.

A wavelet transform analysis of off-talk and on-talk samples is depicted in 4.12a and 4.12b, respectively, showing that the frequency band have different amplitudes and variations for both types of talks.

Feature Extraction

The following features are used for the classification task.

Audio features: The openSMILE (Eyben et al., 2013a) is used to extract the acoustic features that has been widely used for emotion recognition (Liu et al., 2014). The acoustic feature set contains the MFCC, voice quality, fundamental frequency (F0), F0 envelope, LSP and intensity features along with their first and second order derivatives. However, many statistical functions are also applied to the features which resulted in-total of 988 feature for every speech segment.

[3]https://uk.mathworks.com/products/matlab.html – last verified Aug 2017

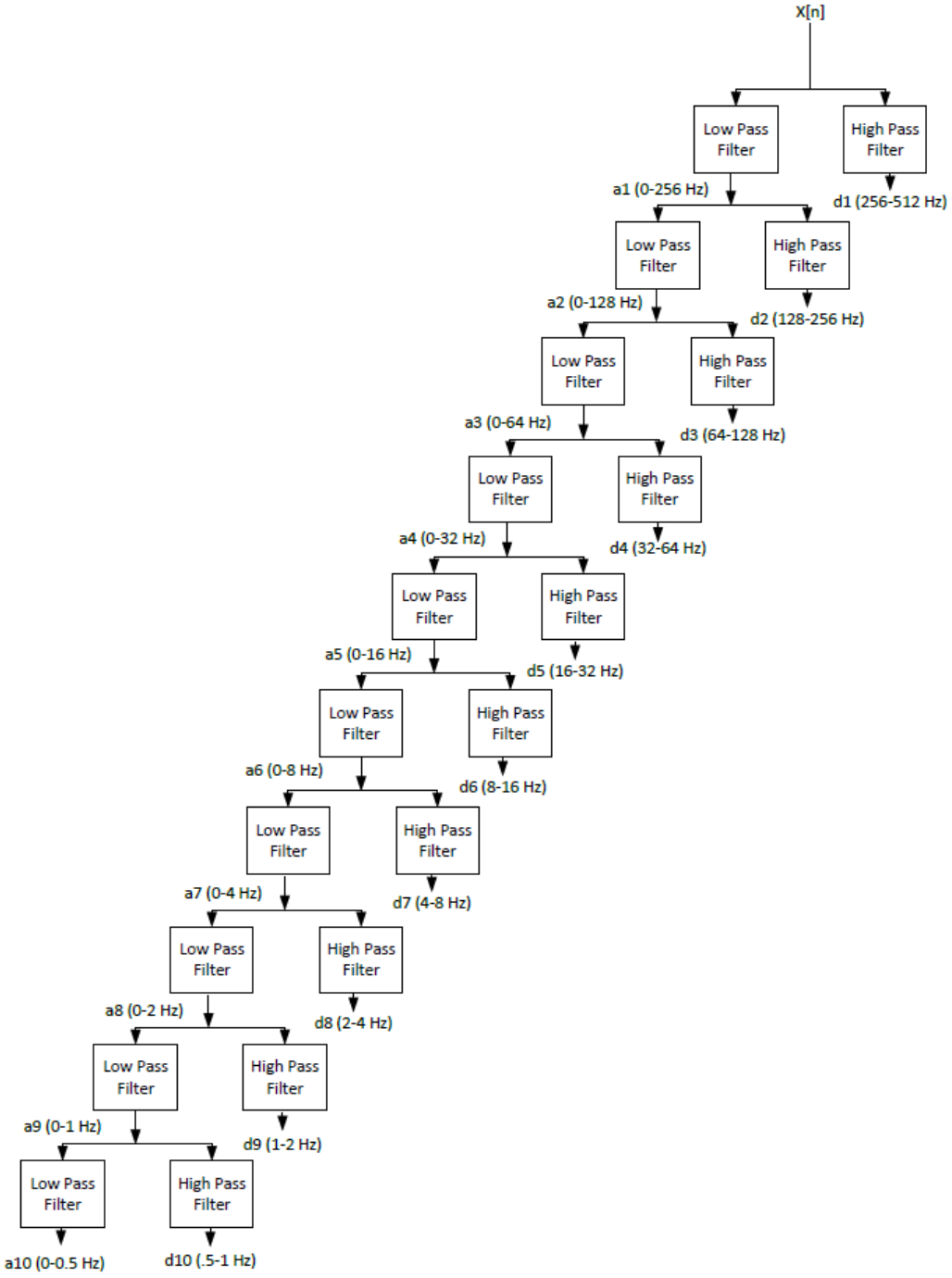
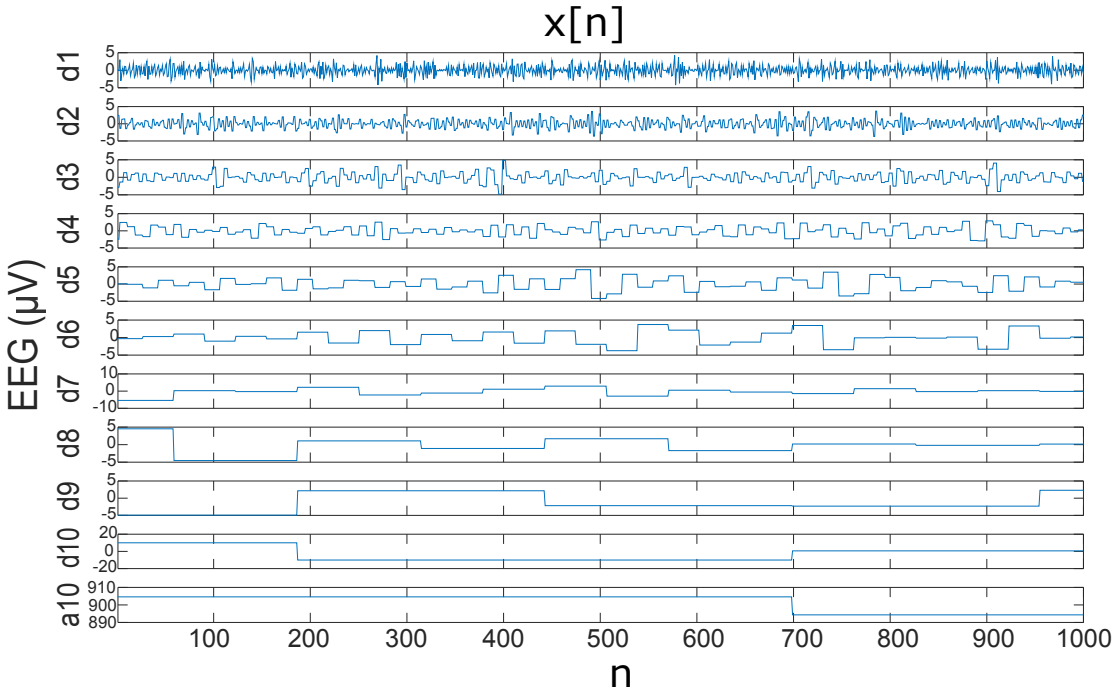
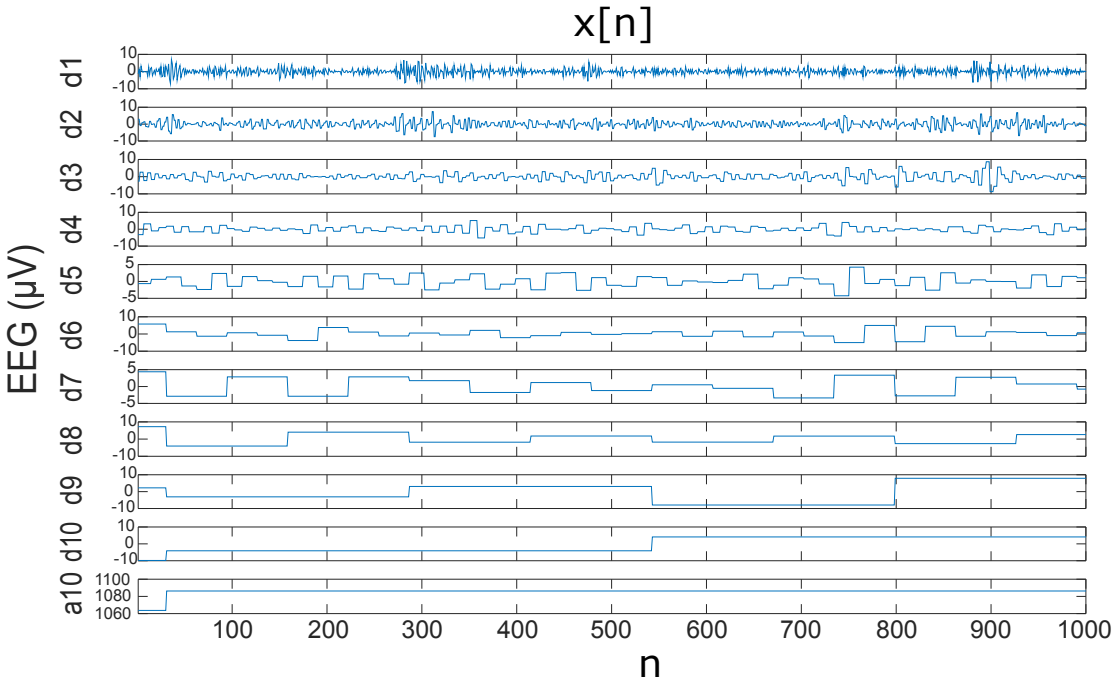


Figure 4.11: *Structure of the tenth level wavelet decomposition of EEG.*



(a) EEG epoch during off talk



(b) EEG epoch during on talk

Figure 4.12: A Wavelet decomposition of EEG signal (S) into 11 components $(d1, d2, d3, ..., d10, a10)$ where $S = d1d2 + d3 + ... + d10 + a10$

Physiological features: For each annotated label the Shannon Entropy, mean, std, mode, min, max, median, energy, power, minimum ratio and maximum ratio along with their first and second order derivatives are extracted. As a result, 33 features for HR and SC are obtained. For the EEG signals, there are 66 features for each component (e.g. d1, d2); 33 for sensor A and 33 for Sensor B. In total, there are 726 features (EEG) for each annotated label.

Hypothesis Testing

This study's working null hypothesis is as follows, H_o: The data (Shannon entropy of EEG Signal) of *On-Talk* and *Off-Talk* comes from the same distribution. In order to investigate this hypothesis, a Kruskal-Wallis test is performed, comparing the Shannon entropy of both kinds of utterances. For this purpose, an average of the Shannon entropy values from EEG sensor A (F4–C4 location) and B (P4–C4 location) are used. The motivation to chose Shannon entropy for statistical analysis is to demonstrate that the EEG signal recorded is not random during *On-Talk* and *Off-Talk* utterance which is indicated by high negative values of Shannon entropy and the difference between both groups is statistical significant (p) as depicted in Table 4.6.

The test results show that a significant difference exists between the *On-Talk* and *Off-Talk* utterances for all frequency bands. This may be due to the neural activity for the lower frequency bands due to the sensitivity of these bands to neural activity. However the higher frequency bands, that are more sensitive to muscle artefacts than lower bands, are also statistical different for *On-Talk* and *Off-Talk*.

Table 4.6: Kruskal-Wallis Test Results using Shannon Entropy.

Component	p	Off (s)	On (s)	Off (m)	On (m)
$d1$ (256–512 Hz)	0.00	$23e^5$	$71e^5$	$-13e^4$	$-95e^4$
$d2$ (128–256 Hz)	0.00	$92e^5$	$15e^6$	$-45e^4$	$-20e^5$
$d3$ (64–128 Hz)	0.00	$34e^6$	$43e^5$	$-15e^5$	$-68e^4$
$d4$ (32–64 Hz)	0.00	$44e^9$	$16e^5$	$-18e^8$	$-31e^4$
$d5$ (16–32 Hz)	0.00	$21e^9$	$64e^4$	$-90e^7$	$-13e^4$
$d6$ (8–16 Hz)	0.00	$12e^9$	$13e^5$	$-54e^7$	$-24e^4$
$d7$ (4–8 Hz)	0.00	$16e^8$	$27e^5$	$-69e^6$	$-51e^4$
$d8$ (2–4 Hz)	0.00	$20e^8$	$51e^5$	$-89e^6$	$-97e^4$
$d9$ (1–2 Hz)	0.00	$20e^8$	$90e^5$	$-87e^6$	$-17e^5$
$d10$ (0.5–1 Hz)	0.00	$25e^7$	$15e^6$	$-13e^6$	$-28e^5$
$a10$ (0–0.5 Hz)	0.00	$47e^{11}$	$29e^{12}$	$-28e^{11}$	$-81e^{11}$

Classification Methods

The results reported in section 4.2.4 suggest that the EEG values (Shannon entropy) may serve as good predictors of *Off-Talk* events. Therefore, this study investigated automatic detection of '*On-Talk* and *Off-Talk* utternaces' using four machine learning methods, namely Linear Discrimination Analysis (LDA), Nearest Neighbour (KNN with K=15), Decision Trees (DT) and Random Forest (RF). These classifiers are employed in MATLAB using the statistics and machine learning toolbox but the RF classifier is employed in python using scikit-learn library.[4] LDA works by assuming that the feature sets of the classes to be discerned are drawn from different Gaussian distributions and adopting a pseudo-linear discriminant analysis (i.e. using the pseudo-inverse of the covariance matrix (Raudys & Duin, 1998)). KNN and DT are non-parametric methods.

Results and Discussion

An experiment is conducted using, different EEG frequency bands, HR, SC and prosodic features. The results are assessed using the A-weighted F-score statistic

[4]http://scikit-learn.org/stable/ – last verified Aug 2017

(with the β parameter set to 1). In this setting, the A-weighted F-score is equivalent to the A-weighted harmonic mean of the precision and recall scores and the baseline is 50%.

The classification results of the 1,127 *On-Talk* and 554 *Off-Talk* utterances are reported in Table 4.8. Of the four classification methods, the results indicate that the Random Forest (RF) classifier provides the best results in all tested settings. The highest frequency band ($d1$) achieved an A-weighted F-score of 72.19%, and the second highest frequency band ($d2$) provides an A-weighted F-score of 68.77%. The reason why results are better using high EEG frequency bands is probably that these frequencies are reflecting speech related muscle artefacts in the recorded EEG signal, as explained in Section 2.2.2. The EEG frequencies ($> 15\,\text{Hz}$ and $< 40\,\text{Hz}$) contains muscle artefacts and neural activities, and able to detect the *On-Talk* and *Off-Talk*.

The lowest frequency band $a10$ produced the best results (74.80%) for the classification task which may be due to the fact that the right hemisphere of human brain is responsible for speech prosody and that prosodic information may be encoded in lower bands of the EEG signal because the lower bands $< 15\,\text{Hz}$ do not contain the muscle artefacts, as explained in section 2.2.2. The audio features set provides the best classification results (91.36%), and the fusion of audio and EEG features ($d1$) improves the performance slightly (92.08%). The HR and SC provide an A-weighted F-score of 60.60% and 68.59% receptively and their fusion improve results (69.47%). The fusion of skin conductance and EEG does not improve results (79.08%). The information in the lower components of EEG signals ($a1, a2$ and $a3$) is also evaluated, and observed that $a3$ (0–256 Hz) component provide the best result (79.47%), but the difference is very small

(79.18%) compared to $a3$ (0–64Hz). A Venn digram is used to explore the mutual information of the top three results which are obtained using $d1$, $a10$ and the Audio signal as depicted in Figure 4.13, and the confusion matrix of this figure is listed in Table 4.7.

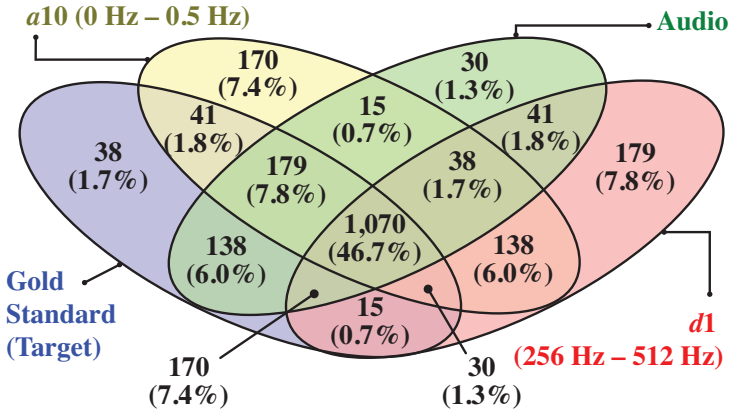


Figure 4.13: Mutual Information: Venn diagram of the results

Table 4.7: Confusion Matrix of the top three best results, showing classification of instances

	$a10$		$d1$		Audio	
	Off-T.	*On-T.*	*Off-T.*	*On-T.*	*Off-T.*	*On-T.*
Off-Talk	365	189	308	246	457	97
On-Talk	103	1,024	150	977	27	1,100

In Figure 4.13, the 'blue circle (Target)' represents the annotated labels, the 'yellow circle' represents the predicted labels by the features of $a10$ frequency band using the RF classifier, the 'green circle' represents the predicted labels by the acoustic features using the RF classifier, and finally the 'red circle' represents the predicted labels by the features of $a10$ frequency band using the RF classifier. From the Venn diagrams overlap, it is observed that there are 38 instances (8 *On-Talk* and 30 *Off-Talk*) which have not been recognised by any of the feature sets.

However there are 1,070 instances (875 *On-Talk* and 195 *Off-Talk*) which have been detected by all three feature sets. The EEG features provide less accurate results than audio features but are able to capture some information (41 (yellow circle: $a10$), 30 (overlap of yellow: $a10$ and red circles: $d1$) and 15 (red circle: $d1$) instances) which is not captured by the audio features as depicted in Figure 4.13.

A mid-p-value McNemar test is used to compare the results of $a10$, $d1$ and Audio features with a null hypothesis which is that $a10$, $d1$ and Audio features have equal accuracy for predicting the target (*On-Talk* – *Off-Talk* detection). The test rejects the null hypothesis for 'Audio and $a10$' ($p_{Audio-a10} = 1.67 \times 10^{-36}$), and 'Audio and $d1$' ($p_{Audio-d1} = 9.44 \times 10^{-52}$) but fails to reject the null hypothesis for '$a10$ and $d1$' ($p_{a10-d1} = 0.08$). High-frequency bands (> 40 Hz e.g., $d1$) provide good results, and due to the muscle activity they capture, it confirm that the *On-Talk* and *Off-Talk* utterances have a different muscle activity pattern. In addition, good results are also obtained from the $\{a10$ (0 Hz – 0.5 Hz)$\}$ band which has robustness against muscle activities, which indicates that *On-Talk* and *Off-Talk* utterances also have different neural activity patterns.

In a previous study, Hayakawa et al (Hayakawa, Haider, et al., 2016a) explored the EEG Gamma band along with SC, HR and acoustic features for the detection of *On-Talk* and *Off-Talk* and reported an A-weighted F-scores of 57.19% when using only the EEG Gamma band. Our results of the wavelet analysis of the EEG signals significantly improves the performance for *On-Talk* and *Off-Talk* detection up to 74.80%. The acoustic features provide the best results for *On-Talk* and *Off-Talk* detection in this study and in the results from Hayakawa et al (Hayakawa, Haider, et al., 2016a). However, the results from Hayakawa et al (Hayakawa, Haider, et al., 2016a) do not provide promising results using physiological signals

alone and used more acoustic features (6,371 acoustic features) than those used in the method (988 acoustic features) reported in this paper. In the previous study Hayakawa et al (Hayakawa, Haider, et al., 2016a) only present an idea of detecting *On-Talk* and *Off-Talk* using different modalities (e.g., EEG gamma band, audio) instead of demonstrating and evaluating the results in detail, which this study covers.

Table 4.8: 10-fold cross validation results (A-Weighted F-score%) for *On-Talk, Off-Talk* detection. (Baseline is 50%)

Results	LDA	KNN	DT	RF
$d1$ (256–512 Hz)	67.53	65.64	65.92	72.01
$d2$ (128–256 Hz)	65.59	65.87	62.24	68.77
$d3$ (64–128 Hz)	65.24	64.11	61.41	67.70
$d4$ (32–64 Hz)	59.10	56.30	59.49	63.11
$d5$ (16–32 Hz)	55.11	55.63	59.01	61.51
$d6$ (8–16 Hz)	57.55	56.11	59.05	62.93
$d7$ (4–8 Hz)	58.64	54.33	55.40	60.54
$d8$ (2–4 Hz)	58.36	51.45	59.19	61.86
$d9$ (1–2 Hz)	55.46	52.28	55.78	62.67
$d10$ (0.5–1 Hz)	56.02	44.64	55.37	60.20
$a10$ (0–0.5 Hz)	56.32	64.03	70.76	74.80
Components fusion of S (0-512)	57.83	63.99	68.01	77.16
$a1$ (0–256 Hz)	67.42	69.29	72.08	79.47
$a2$ (0–128 Hz)	64.28	68.51	71.46	79.13
$a3$ (0–64 Hz)	65.55	68.72	71.08	79.18
Audio	82.73	67.91	84.14	91.36
Audio+$d1$	82.63	65.98	83.62	92.08
Audio +$a1$	87.45	68.56	85.23	92.06
Audio +$a1 + d1$	88.72	68.21	86.56	92.06
$a1 + d1$	66.62	68.26	72.37	79.72
HR	57.68	57.09	57.34	60.60
SC	65.01	59.51	64.90	68.59
HR + SC	65.02	59.15	64.92	69.45
$a1 + d1$+ SC	68.76	67.95	72.20	79.08

4.3 Cognitive States Detection

This section describes the methods to propose a novel cognitive states (i.e. temporary psychological states which are annotated using facial expressions) detection

system for machine translation systems for sensing user experiences with the system in terms of cognitive states. However the cognitive states occurs in a Speech-to-Speech, Machine Translation mediated Map Task as described in Section 4.2.1 and the dataset is annotated for three cognitive states (amusement, surprise and frustration using facial expressions) by two annotators with an inter-coder agreement of above 60% which was calculated on one of the dialogues.

4.3.1 Features Extraction

In order to detect the cognitive states from the biosignal and the speech audio files, this study computes a joint feature set for the single modalities (speech vs physiological) for which a discriminative analysis pattern classifier is tested, and then compare the results of the recognition rates for separate and integrated modalities. The idea is to verify whether the information from the biosignal combined with the prosodic analysis can improve the results of the detection.

Physiological Features Extraction

Two physiological signals are used in cognitive state recognition: The heart rate (HR) from the BVP sensor and skin conductance measured from the SG/GSR sensor. The feature set contains the median, mean, standard deviation, minimum, maximum, minimum ratio, and maximum ratio of the data values. These features are also calculated for the 1st and 2nd order derivatives of physiological signals. This results in 21 features for SC and 21 for HR. However the HR signal, maximum ratio and median of 1st and 2nd derivative have a zero value for most of the observations. These zero values are removed from the physiological feature set

resulting in a total of 18 features for HR. The minimum ratio of an observation is measured by counting the number of instances which have a lower value compared to their preceding and following instance and then dividing it by the total number of instances in that observation. Similarly, the maximum ratio of an observation is measured by counting the number of instances which have a higher value compared to their preceding and following instance and then dividing it by the total number of instances in that observation.

Prosodic Feature Extraction

OpenSMILE (Eyben, Weninger, Groß, & Schuller, 2013b) is used to extract a prosodic feature-set from the high quality audio files that have been down sampled to 48 kHz, 16 bit. The feature set employed is the 2013 ComParE challenge (Schuller et al., 2013b) set. It comprises 6,373 features, including energy, spectral, cepstral (MFCC) and voicing related low-level descriptors, as well as other descriptors such as logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness.

Classification Method

Using the MATLAB statistical tool box, a 10-fold cross validation was performed with the discriminant analysis (DA) method for the classification of this balanced dataset. The method assumes that the feature-set of different classes follow different Gaussian distributions and follows the pseudolinear discriminative type.

4.3.2 Experimentation

The following 6 experiments are conducted to evaluate the performance of physiological and/or prosodic features for cognitive states recognition on a balanced dataset using different analysis windows as shown in Figure 4.14. Linear discrimination analysis is used to classify the states in the below experiments. However, the prosody analysis is performed on an utterance-by-utterance basis.

Exp. 1: The physiological feature vectors are calculated over the 372 annotated labels of cognitive states.

Exp. 2: 124 observations of neutral state are introduced to the data of Exp. 1 in order to compare the physiological characteristics of neutral states with the other three states.

Exp. 3: Our hypothesis is that the cognitive state starts developing after the participant has read the displayed ASR result of the utterance spoken to the ILMTs2s system. 249 observations of the biosignal are extracted starting from when the ASR result is displayed on the screen and ending at the following cognitive state label end-time. The physiological feature vectors over this duration are calculated.

Exp. 4: Our hypothesis is that the cognitive state affect the participant's speech even after the labelled cognitive state has ended (that is when the cognitive state is no longer observable). 390 utterances spoken to the ILMT-s2s system are selected that occurred after the labelled cognitive state and calculated the prosodic feature vectors for these utterances.

Exp. 5: It is investigated whether the cognitive state will affect the participant's physiological characteristics even after the cognitive state is no longer ob-

servable (after the labelled section has ended). The physiological feature vectors are calculated for the 222 observations over an extended window starting when the labelled section starts, but ending when the following speech utterance ends.

Exp. 6: The physiological features from the fifth experiment are combined with the prosodic feature of the utterances that followed the cognitive state (on the time-line) of the ILMT-s2s system.

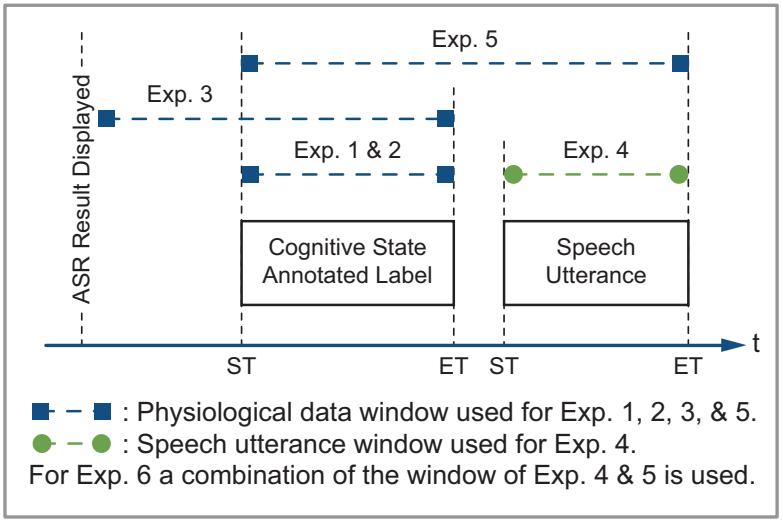


Figure 4.14: *Analysis window explanation.*

4.3.3 Results and Discussions

Table 4.9 and Figure 4.15 show the recognition rates of the classification. The average rates of classification for physiological data (combination of SC and HR) in the 1st experiment is 61.29% and improves in the 2nd experiment to 70.77%. However, in the 3rd experiment, the average percentages of classification accuracy using HR, SC and a combination of both physiological signals are reduced to 47.79%, 49.00% and 52.61% respectively, which contradicts our hypothesis. The

Table 4.9: *Classifier accuracies for each class & overall average.*

Features		A (%)	F (%)	S (%)	N (%)	Ave. (%)
Exp. 1:	HR	70.97	52.42	40.32	-	54.57
	SC	62.90	59.68	61.29	-	61.29
	HR + SC	68.55	62.10	53.23	-	61.29
Exp. 2:	HR	71.77	56.45	46.77	100.00	68.75
	SC	62.10	55.65	35.48	70.16	55.85
	HR + SC	68.55	64.52	50.00	100.00	70.77
Exp. 3:	HR	62.65	40.96	39.76	-	47.79
	SC	54.22	49.40	43.37	-	49.00
	HR + SC	60.24	51.81	45.78	-	52.61
Exp. 4:	Prosody	56.15	43.85	50.77	-	50.26
Exp. 5:	HR + SC	70.27	51.38	45.95	-	55.86
Exp. 6:	Prosody + Bio	75.68	67.57	60.81	-	68.02

decrease in accuracies compared to 1st experiment shows that the cognitive state takes some time to develop instead of developing right after the ASR display. In the 4th experiment, the preceding utterances were analysed to detect the cognitive state that follows the utterance and the average result of classification was 50.26%. It is observed that the cognitive states have an impact on the preceding utterances. In the 5th experiment, the physiological data under the extended window provides an average accuracy of 55.86% and finally, in the 6th experiment, it is observed that the combination of physiological and prosodic features gives the optimum accuracy of 68.02%. At the end, the results also show that physiological data, in all cases except SC in the 1st experiment, provides better accuracy for amusement (e.g. 70.97%) than frustration (e.g. 52.42%) or surprise (e.g. 40.32%). Contrary to this, the prosodic data provides a better accuracy for surprise (50.77%) than

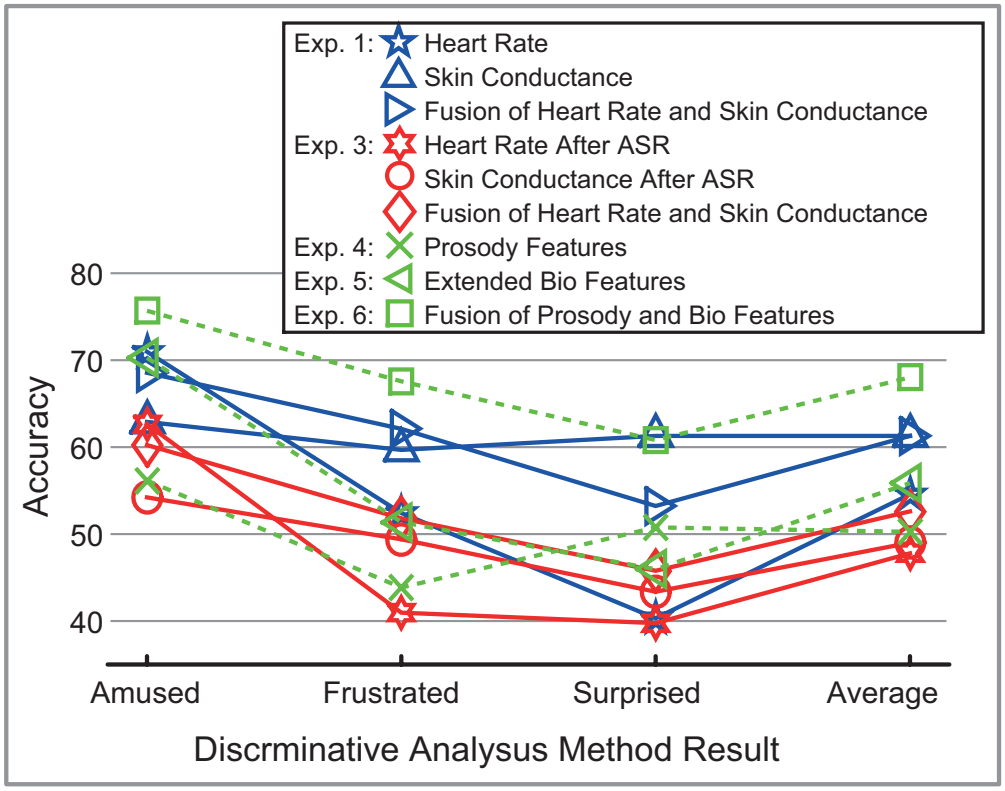


Figure 4.15: *The figure shows the classifier average accuracies.*

frustration (43.85%). This proves that the prosodic and physiological characteristics of amusement vary more than those of frustration and surprise. This also indicates that the physiological characteristics of frustration (e.g. 62.10%) differ more than the prosodic characteristics (43.85%).

The results here reported validate the hypothesis that the performance of bimodal cognitive state recognition gives better results compared to unimodal recognition: the bimodal approach (combination of prosodic and physiological characteristics provides 68.02% accuracy) gives an improvement of almost 7 percent compared to the performance of the physiological signals (61.29%). The results of physiological signals using neutral labels show that the characteristics of neutral

labels are clearly separable from the three cognitive states, which resulted in good classification of these labels. Moreover, these results also show that the HR of neutral state is completely different from the three cognitive states. In addition it is observed that HR has more correlation with amusement than SC, while, SC is more correlated with frustration and surprise than HR. Using the "extended window" results in an increase of accuracy for amusement but a decrease in frustration and surprise accuracies.

4.4 Active Speaker Detection

This section describes the methodology and experimentation to propose a novel active speaker detection system using visual prosody information for human-machine multi-party dialogue.

4.4.1 Dataset

An audio-visual dataset (Haider & Al Moubayed, 2012; Haider, Luz, & Campbell, 2016a) was collected as follow. Four participants (3 males and 1 female) converse with the "machine", but they are not allowed to speak with each other directly. They are free to gesture, display emotions, etc. so long as this does not change their location. The machine perspective is simulated by a fifth person (S0) in a separate location, using video conferencing software, and his face is displayed on the computer screen visible to the other four. The motivation for using full facial information is to simulate a humanoid robot/avatar that can handle and generate social signals and behaviours. The recording session consisted of two parts. In part one, the subjects (in-front of the camera) ask questions from the machine

(S0) one by one through video conferencing, and the machine (S0) answers them. In the second part, the machine (S0) asks the questions from the other participants. Some sample questions are as follow:

1. Where is the nearest train station?

2. How can I reach to the football ground?

3. Where can I find a place for lunch?

The fifth person (So) makes the interaction seem more natural, as full facial information is present to the participants. All subjects make use of gestures, body movements and display emotions. However, they are not allowed to change their positions. The recording equipment and set-up are shown in Figure 4.16. A high-definition JVC video camera was used for recording the session. It recorded the video with a frame rate of 25 fps, and the duration of the video (dialogue) is 21 minutes. The distance between speakers and machine interface (S0) was approximately 2 meters. The segments of speech for all participants are annotated using ELAN annotation software as shown in Figure 4.17 (Brugman, Russel, & Nijmegen, 2004). Each speaker produced an average of approximately 70 speech segments. Details of speech/non-speech frames are depicted in Table 4.10.

Table 4.10: *Speech/non-speech frames and their data distribution.*

Subject	non-speech frame	speech frame
S1	28442 (89.47%)	3355 (10.55%)
S2	29301 (92.49%)	2380 (7.51%)
S3	29697 (93.25%)	2148 (6.75%)
S4	28276 (98.29%)	492 (1.71%)
Total	115716 (93.25%)	8375 (6.75%)

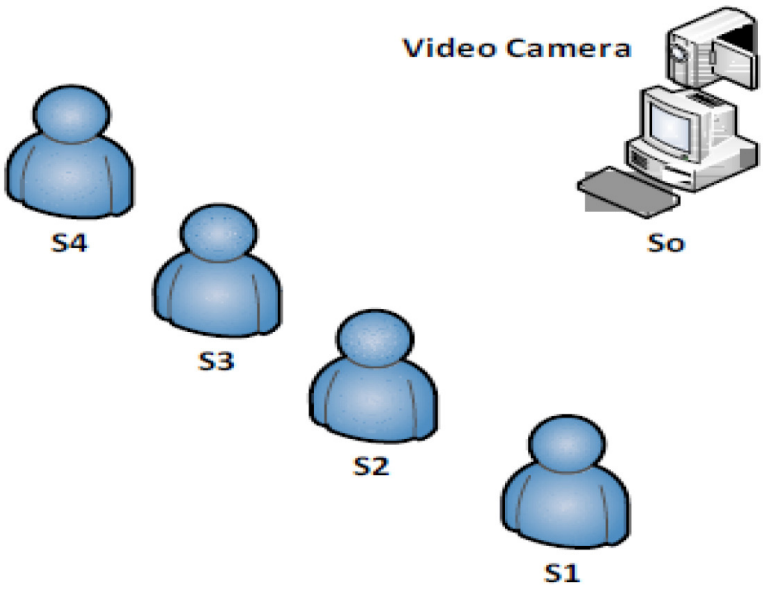


Figure 4.16: *Recording setup.*

4.4.2 Active Speaker Detection using Visual Prosody Information

This section presents a novel system for active speaker detection using the head and lip movements during speech articulation.

Feature Extraction

The FaceAPI SDK (Machines, 2009) is used for tracking of facial landmarks and head coordinates for every speaker. FaceAPI is a commercially available software (a product of Seeing Machines) capable of tracking head pose and lip location as well as the location of jaw, eyebrows and eyes. Features used in this study are the lips inner height, outer height and width (in meters) calculated by the position (x, y and z in meters) of face landmark ID numbers 101, 104, 202, 206,

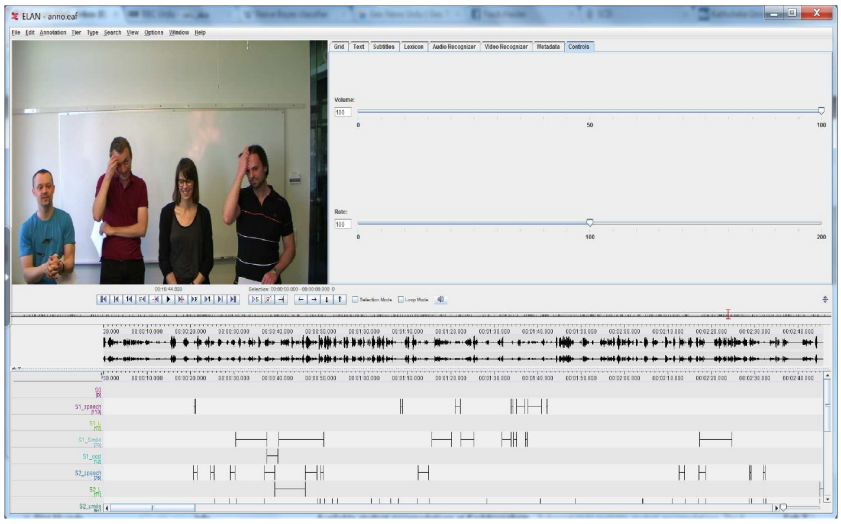


Figure 4.17: *Video annotation by ELAN (A snapshot showing the annotation interface on the recorded video.)*

200 and 204 as shown in Figure 4.19, and the head rotation along x, y and z-axis (in radian) as shown in Figure 4.18. To calculate the feature set, average rates of change of lip (inner height, outer height and width) and head coordinates (x,y and z coordinates) are calculated. Then, a window with a length of 25 frames (1 second) with 96% overlap (24 frames) is applied. After that, the standard deviation and mean values for each window are calculated. The resulting dataset has four features (2 for lip and 2 for head movements for each frame) for each speaker and number of instances for speech and non-speech are reported in Table 4.10.

Statistical Analysis

From the ANOVA (Analysis of Variance) test results, it is found that head movements groups means (speech/non-speech) are significantly different for speech frames ($p = 0.00$), and lip movements exhibit the same behaviour ($p = 0.00$), as shown in Table 4.11. Based on these results, in the following section a voice

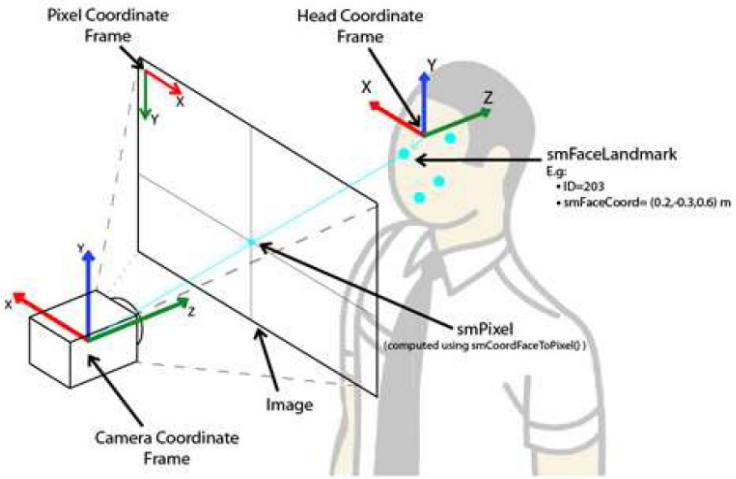


Figure 4.18: The face traking API coordinate frames (Machines, 2009)

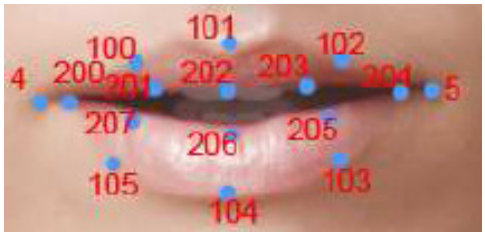


Figure 4.19: The face tracking API lip tracking points (Machines, 2009)

activity detection (VAD) system is proposed, and then it is extended to a speaker detection system.

Table 4.11: *Statistical significance test (ANOVA test) results for head and lip movements. The mean and standard deviation (Std.) values for NS (non-speech) and S (speech) are also reported*

	Head ($p = 4.4554e-11$)		Lip ($p = 3.2916e-04$)	
	NS (Head)	S (Head)	NS (Lip)	S (Lip)
mean	-1.5215e-05	2.4867e-04	4.3311e-07	-5.1023e-06
Std	0.0036	0.0029	1.3590e-04	1.4040e-04

Classification Method

The classification is performed using four different methods namely Linear Discrimination Analysis (LDA), Naïve Bayes (NB) classifier, Nearest Neighbour (KNN with K=1) and Decision Trees (DT). These classifiers are employed in MATLAB[5] using the statistics and machine learning toolbox. LDA works by assuming that the feature sets of the classes to be discerned are drawn from different Gaussian distributions and adopting a pseudo-linear discriminant analysis (i.e. using the pseudo-inverse of the covariance matrix (Raudys & Duin, 1998)). The NB classifier also assumes a normal distribution for the feature set. KNN and DT are non-parametric methods.

Results

The section reports results on active speaker and speech activity detection using the head and lip movements as features. VAD is performed in three different model training settings, namely: speaker dependent, speaker independent and hybrid and the F-Score (harmonic mean) is reported to compare the results of different feature sets and classifiers. In the case of LDA and NB classifiers, the F-Score for 'speech frames detection' contains NaN (not a number) values which are due to the zero value of precision or recall. So, it is found that the LDA and NB classifier performs poorly (NaN (Not a Number) values of F-Score due to zero value of precision or recall) in our case, and their results are not reported here. It can be due to the fact that the data is highly imbalance.

[5]http://uk.mathworks.com/products/matlab/

Speaker Dependent VAD

In this setting, a 5-fold cross validation is performed on the data of each subject separately, and the results are reported in Table 4.12. The results show that the 1-NN provides the best F-Score for speech detection (68.40% for S3) and head movements are better than lip movements.

Table 4.12: *Speaker Dependent VAD Results (F-Score %) for Non-Speech(NS) and Speech(S) using 1-Nearest Neighbour (1NN) and Decision Tree (DT) classifiers. The 5-fold cross validation is performed on each subject data separately.*

Subject	Classifier	Lip		Head		Fusion	
		NS	S	NS	S	NS	S
S1	1NN	90.10	16.61	**91.82**	**30.96**	***95.60***	***62.40***
	DT	90.81	15.43	91.81	25.82	93.63	43.28
S2	1NN	93.15	19.53	**94.23**	**28.25**	***97.30***	***67.22***
	DT	93.87	16.41	94.50	25.81	95.83	47.02
S3	1NN	94.30	20.88	**94.67**	**25.41**	***97.75***	***68.40***
	DT	94.71	19.32	94.66	22.12	96.42	48.54
S4	1NN	98.46	9.35	**98.91**	**33.30**	***99.36***	***63.70***
	DT	98.52	4.55	98.83	24.08	99.03	43.55

Hybrid VAD

In this setting, all the speech and non-speech data is concatenated and 5-fold cross validation is performed. The results are reported in Table 4.13. 1-NN provides the best F-Score for speech detection (83.90%) and head movements are better than lip movements.

Table 4.13: *Hybrid Method Results (F-Score %) for Non-Speech (NS) and Speech (S) using 1-Nearest Neighbour (1NN) and Decision Tree (DT) classifiers.*

Classifier	Lip		Head		Fusion	
	NS	S	NS	S	NS	S
1NN	95.00	30.13	97.23	61.49	***98.83***	***83.90***
DT	95.36	31.97	**97.30**	**61.66**	97.18	60.25

Speaker Independent VAD

In this setting, 4-fold cross validation (leave one subject out from training) is performed, and the results of all four fold are reported in Table 4.14. The models were trained on data from three participants, while testing was performed on data from the fourth person, exclusively. The results show that the lip movements give better F-score than head movements and fusion of them do not improve results.

Table 4.14: *Speaker Independent VAD Results (F-Score %) for Non-Speech (NS) and Speech (S) using 1-Nearest Neighbour (1NN) and Decision Tree (DT) classifiers.*

Test Subject	Classifier	Lip		Head		Fusion	
		NS	S	NS	S	NS	S
S1	1NN	91.21	**9.47**	91.49	7.76	91.21	***9.07***
	DT	91.86	8.73	92.06	7.27	91.62	8.90
S2	1NN	93.50	**8.22**	92.91	7.67	93.33	***8.55***
	DT	94.21	7.51	93.58	8.03	93.52	8.85
S3	1NN	94.24	***9.51***	94.23	6.14	94.77	**9.09**
	DT	94.60	7.36	94.76	5.39	94.48	9.07
S4	1NN	94.60	3.42	95.15	3.28	95.06	2.95
	DT	95.65	**3.63**	95.60	2.73	95.25	***3.92***

Speaker Detection

For the speaker detection problem, those frames are selected that have both features (both lip and head movement have been tracked by FaceAPI) for all participants and ignore those frames where two subject speak simultaneously. As a result, the data contains 21301 frames of silence and 2408, 1886, 1900, 411 frames of speech for S1, S2, S3 and S4 respectively. The 5-fold cross validation results are depicted in Table 4.15.

Table 4.15: *Speaker Detection Results (F-Score % of each class) for Lip and Head movements using 1-Nearest Neighbour (1NN) and Decision Tree (DT) classifiers.*

Feature	Classifier	Silence	S1	S2	S3	S4
Lip	1NN	97.86	91.91	**92.94**	**93.27**	88.89
	DT	89.89	57.37	59.98	60.24	55.43
Head	1NN	**97.98**	**93.59**	92.81	91.95	**92.23**
	DT	93.70	75.28	75.89	77.29	66.03
Fusion	1NN	***98.18***	***93.68***	***93.55***	***93.43***	***92.77***
	DT	94.23	76.92	76.96	76.52	73.54

Discussion

The results show that head movements provide better result than lip movements in speaker dependent and hybrid VAD and fusion of both features improves the accuracy as depicted in Table 4.12 and Table 4.13. However, in speaker independent setting lip movements have better results than head movements but the difference between them is quite small as depicted in Table 4.14. The poor F-scores obtained for speech detection in speaker independent settings are mainly due to data imbalance (as reported in Table 4.10), the small number of subjects in the training data set, and the absence of speaker-specific features. However, in the other settings the proposed approach performed quite well over blind guess (50%). The speaker dependent setting provides less accurate results than hybrid settings. This indicates a strong relationship between head and lip movements of different subjects. The hybrid model provides the best results for speech detection (F-score = 83.90%). This also suggests that this approach will scale well as more subjects' data is used. If the training data set contains lip and head movements of the several people (say, 50 subjects) then the likelihood of finding a similar relationship among the multiple speakers and the test subject is increased, possibly leading to better accuracy. In the speaker detection problem (blind guess = 20%), head movements also provide better results than lip movement. Fusion improves accuracy, but only

slightly. As regards classification methods, LDA and NB classifier provides poor results as compared to decision trees and nearest neighbour. The reasons for the poor performance of LDA and NB are the highly imbalanced nature of the data and the poor fit between the high-level features that are used in this study and the underlying assumptions of these models. In (Wang & Schmid, 2013), the authors use low-level visual descriptor to capture upper body movements including head movements for active speaker detection, but they do not explicitly explore head movements. Furthermore, the scenario of the dataset is more formal (students and jury) than this study. In a previous study, lip movements are also considered for detecting speaker and speech with some success. However, that study was conducted on a balanced dataset (Haider & Al Moubayed, 2012).

4.4.3 Improving Response Time of Active Speaker Detection

In Human-Human interaction, it is observed that the listener turns their gaze towards the speakers around 30-–80% of the time (Kendon, 1967). Hence, from the social robotics perspective, it is useful to detect the active speaker as soon as possible to enable the robot to turn gaze/head towards the speaker to show that it is attending to the speaker. In particular, it is useful if one may anticipate who the next active speaker will be, in order to accelerate this process.

This study continues the author's past work (Haider & Al Moubayed, 2012; Haider, Luz, & Campbell, 2016a) which demonstrated the use of lip and head movements during speech articulation for active speaker detection but did not assess the discriminative power of visual prosody data captured just before and/or

after articulation. This study proposes methods for detection of active speakers through use of visual prosody information one second before/after speech articulation and also evaluate the visual prosody information of the first second of the speech utterance. The system architecture is depicted in Figure 4.20. The system processes visual information before articulation from the memory buffer as soon as Voice Activity Detection (VAD) detects 10 ms of voice activity. The proposed methods are a step towards decreasing the response time of a robot in generating multimodal attention towards the user in situated interactions and experimental findings help in understanding the discrimination power of visual prosody for those regions (one second before/after the articulation). To the author's best knowledge, it is the first automatic active speaker detection system with input from one camera which uses the visual information particularly head movements before articulation. Moreover, it is the first study which demonstrates the discrimination power of visual prosody (one second before and after articulation) for active speaker detection.

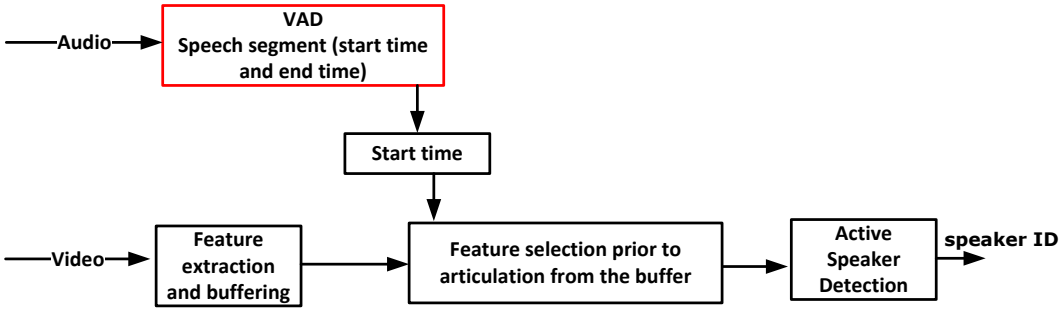


Figure 4.20: The proposed system architecture for active speaker detection.

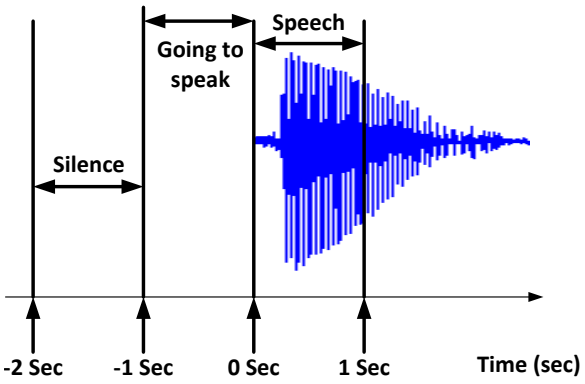


Figure 4.21: Regions of interest for 'Going to Speak', Silence and Speech

Pearson Correlation Test

In line with evidence of low variation in electrical potential up to one second before the speech articulation in the human brain (McAdam & Whitaker, 1971) It is hypothesised that visual prosody also shows some characteristics (e.g. subject is going to start moving his or her lips for articulation) that are manifest during the one second before articulation. To validate this assumption, Pearson correlation test is performed using a subset of the corpora. For the test, 20 speech segments are considered for each subject and extracted features (mean value of the rate of change in lip and head movements) from the regions shown in Figure 4.21. There are 20 instances of Silence Region (SilR), Going To Speak Region (GTSR) and Speech Region (SR) for each subject. SR is defined as the first second of a speech utterance, and the GTSR is defined as a time window of one second before SR. The SilR is defined as a time window of one second before GTSR. All these regions are concatenated as depicted in Figure 4.21. The results of Pearson correlation test along with the null hypothesis are described below. This test helps us in finding the correlation between visual prosody of SR, GTSR and SilR.

Hϕ^1: There is no correlation between the visual prosody of SilR and SR data.

The Pearson correlation failed to rejected this null hypothesis at the $p_{\text{GTSR-SilR}} < 0.05$ significance level in all cases, as depicted in Table 4.16.

Hϕ^2: There is no correlation between the visual prosody of SR and GTSR data.

For this hypothesis, the Pearson correlation test rejected the null hypothesis ($p_{\text{GTSR-SR}} < 0.05$) in 4 out of 6 cases as depicted in Table 4.16. Only for S1 head and S2 lip data, the test was unable to reject the null hypothesis ($p_{\text{GTSR-SR}} > 0.05$).

Hϕ^3: There is no correlation between the visual prosody of GTSR and SilR data.

For Hϕ^3, the Pearson Correlation test rejected the null hypothesis ($p_{\text{GTSR-SilR}} < 0.05$) in 2 out of 6 cases as depicted in Table 4.16. S3 lip and S1 head data showed statistically significant correlation ($p_{\text{GTSR-SilR}} < 0.05$).

Table 4.16: *Pearson Correlation test results (statistical significance (p) and correlation coefficient (r)) for Silence Region (SilR), Speech Region (SR) and Going To Speak Region (GTSR).*

Feature	Subject	SilR-SR		GTSR-SR		GTSR-SilR	
		$r_{\text{SilR-SR}}$	$p_{\text{SilR-SR}}$	$r_{\text{GTSR-SR}}$	$p_{\text{GTSR-SR}}$	$r_{\text{GTSR-SilR}}$	$p_{\text{GTSR-SilR}}$
Lip	S1	0.164	0.489	**0.668**	**0.001**	0.297	0.204
	S2	0.065	0.786	-0.089	0.709	0.067	0.781
	S3	0.211	0.372	**0.626**	**0.003**	**0.627**	**0.003**
Head	S1	0.078	0.743	0.238	0.313	**0.687**	**0.001**
	S2	0.443	0.051	**0.656**	**0.002**	-0.151	0.525
	S3	-0.070	0.769	**0.791**	**0.000**	0.116	0.625

From these correlation tests, it is concluded that: 1) the GTSR is highly correlated with the SR and this correlation is statistically significant ($p < 0.05$) in 4 out of 6 cases. It suggests that every speaker has some visually detectable means (i.e. head and/or lip movements) of communicating their intention to speak; 2) the GTSR is less correlated with the SilR than SR, and this correlation is statis-

tically significant ($p < 0.05$) in 2 out of 6 cases; 3) the data shown no significant correlation between SR and SilR.

Experimentation

Given that SR information seems better correlated with GTSR than SilR, feature sets are created that reflected this fact, and trained models for an automatic active speaker detection system using classification methods. We have performed three experiments using three different feature sets for classification as described below:

Experiment One: In this experiment, The features are extracted one second before articulation (GTSR, see Figure 4.21).

Experiment Two: In this experiment, The features are extracted one second after articulation. (SR, see Figure 4.21).

Experiment Three: In this experiment, The features have been fused from the previous two experiments.

Classification Methods

The classification is performed using four different methods, namely Linear Discriminant Analysis (LDA), Naïve Bayes (NB), Nearest Neighbour (KNN with K=1) and Decision Trees (DT). These classifiers are employed in MATLAB[6] using the statistics and machine learning toolbox in the 10-fold cross-validation setting. LDA works by assuming that the feature sets of the classes to be discerned are drawn from different Gaussian distributions and adopting a pseudo-linear discriminant analysis (i.e. using the pseudo-inverse of the covariance matrix (Raudys & Duin, 1998)). The NB model also assumes a kernel distribution for the feature

[6]http://uk.mathworks.com/products/matlab/ – Last verified June 2017

set. KNN and DT are non-parametric, and non-linear classification methods.

Results and Discussion

The classification is performed on subset of the dataset, Where each subject has 21 instances. In this case, blind guess and majority guess are the same (33.33%), and it is set as a baseline for the classification task. The results of experiment one are shown in Table 4.17. It is observed that the lip movements (42.86%) provide better results than the head (38.10%) and fusion of lip and head movements (47.62%) improves the performance. The LDA classifier provides the best results. However, the fusion of lip and head movements does not improve accuracy for NB and KNN, and overall lip movements provide better results than head movements. This probably reflect the fact that these models are more prone to being misled by irrelevant features than LDA, as is well known of KNN, for instance.

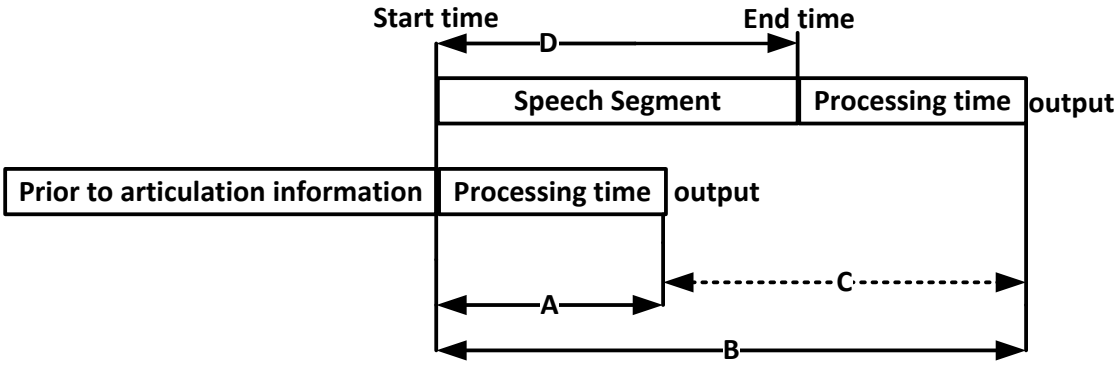


Figure 4.22: Highlighted the proposed system response time (A), baseline response time (B), improvement in response time (C) and duration (couple of seconds) of Speech segment (D). The output is the predicted label and processing time is the time taken by a machine's processor for classification purpose.

The results of experiment two are shown in Table 4.18. It is observed that the lip movements (52.38%) provide better results than the head (46.03%) and fusion of lip and head movements (49.21%) cause a slight increase in accuracy. The LDA

Table 4.17: *Accuracy (%) for experiment one (10-fold cross-validation): facial features one second before articulation.*

Feature	Baseline	KNN	DT	NB	LDA
Head	33.33	31.75	28.57	25.40	38.10
Lip	33.33	36.51	38.1	30.16	42.86
Fusion	33.33	28.57	41.27	30.16	**47.62**

classifier provides the best results. The fusion of lip and head movements do not improve accuracy.

Table 4.18: *Accuracy (%) for experiment two (10-fold cross-validation): features one second after articulation.*

Feature	Baseline	KNN	DT	NB	LDA
Head	33.33	39.68	46.03	42.86	44.44
Lip	33.33	42.86	44.44	42.86	**52.38**
Fusion	33.33	36.51	46.03	42.86	49.21

The results of experiment three are depicted in Table 4.19. It is observed that the lip movements provide better results than head movements using LDA, DT and KNN classifiers. The fusion of lip and head movements improves accuracy for LDA.

Table 4.19: *Accuracy (%) experiment Three (10-fold cross-validation): fused features.*

Feature	Baseline	KNN	DT	NB	LDA
Head	33.33	30.16	41.27	37.10	39.68
Lip	33.33	33.33	50.79	36.51	52.38
Fusion	33.33	30.16	42.86	33.87	**55.56**

From the above three experiments results, it is observed that the visual prosody one second before articulation provide good results for active speaker detection. This can be due to the fact that the subjects start moving their lips before articulation of speech.

We use a Venn diagram to visualise the range of classification overlaps of the best performing classifier (LDA) for each experiment. In Figure 4.23, the red circle (Target) represents the annotated labels, the yellow circle (Exp.2) represents the predicted labels by the lip movements in experiment two, the blue circle (Exp.1) represents the predicted labels by the fusion of head and lip movement in experiment one, and finally the green circle (Exp.3) represents the predicted labels by the fusion of head and lip movements in experiment three. It is observed that the yellow and green circles have the highest overlap (38 out of 63), and that both these circles have an overlap of 35 samples with the red circle (Target). It is also observed that there are 12 instances (4 of S1, 6 of S2 and 2 of S3) which have no overlap with any circle. There are 16 instances (2 of S1, 7 of S2 and 7 of S3) which have been detected by all the three experiments as depicted in Figure 4.23. We also compare the predictive accuracies of our three best results using the mid-p-value McNemar test (testcholdout[7]) with a null hypothesis that predicted labels of Exp.1, Exp.2 and Exp.3 have equal accuracy for predicting the target. The statistical test was unable to reject the null hypothesis ($p_{\text{Exp.1-Exp.2}} = 0.58$, $p_{\text{Exp.1-Exp.3}} = 0.29$ and $p_{\text{Exp.2-Exp.3}} = 0.66$) and shows that although GTSR provides less accurate results than SR and fusion of both regions but the difference is not statistically significant. Hence demonstrating that the GTSR and SR regions have similar characteristics for active speaker detection. In previous studies, the speech/non-speech frames were distinguish and the active speaker is classified with good accuracy using visual prosody information during speech articulation. However, the proposed methods were not developed for a quick response time, and the main objective was to evaluate the lip and head movements as discriminative fea-

[7]https://uk.mathworks.com/help/stats/testcholdout.html (Aug 2017)

tures for active speaker detection in human-machine multi-party setting (Haider & Al Moubayed, 2012; Haider, Luz, & Campbell, 2016a).

Most studies (Chakravarty, Zegers, Tuytelaars, & Van hamme, 2016; Chakravarty et al., 2015) to date process the full speech segment acoustic and visual information ('D' region as depicted in Figure 4.22) and then assign each utterance a speaker label that added a latency and decrease the response time of a machine that may results in turning the gaze and head of a robot to the active speaker in its view after the subject finished speaking ('B' region as depicted in Figure 4.22). In a previous study, The head and lip movements of the 'D' region were evaluated for speaker detection and observed an average of 71.29% accuracy using lip movements on the same dataset used in this study (Haider & Al Moubayed, 2012). While the accuracy for the 'D' region is better than GTSR and SR regions for speaker detection, the former will generate a multi-modal output for a robot only after processing the full speech segment with a duration of some seconds (depending on the speech segment length, which is typically couple of second to around 20 seconds, plus processing time). The latter will generate the output after 0-1 second (plus processing time) of speech articulation. The current and previous study (Haider & Al Moubayed, 2012) both require an input from audio-VAD. The strength of the current study is its focus on quick response time ('A' region as depicted in Figure 4.22) which can increase the naturalness of a machine in a human-machine multi-party interaction. In another study (Haider, Luz, & Campbell, 2016a), a visual active speaker detection system is proposed at frame level which do not need an input from audio-VAD to operate and detect the speaker at video segment level. This involved processing of consecutive 25 frames (1 second of video segment) with an overlap of 24 frames with neighbouring video segment,

hence detecting who is speaking in each video frame, instead of speech segment level (detected by audio-VAD) using lip and head movements with GTSR treated as SilR (Haider, Luz, & Campbell, 2016a). While we observed high accuracy ($> 90\%$) in classifying video segments of active speaker, that study did not explicitly demonstrate the discrimination power of speech frames (video segments) one second before/after articulation, which this study covers. Based on the experimental findings (GTSR is more correlated with SR than SilR), we recommend that GTSR should be treated as SR instead of SilR for the development of visual active speaker detection systems for noisy environments.

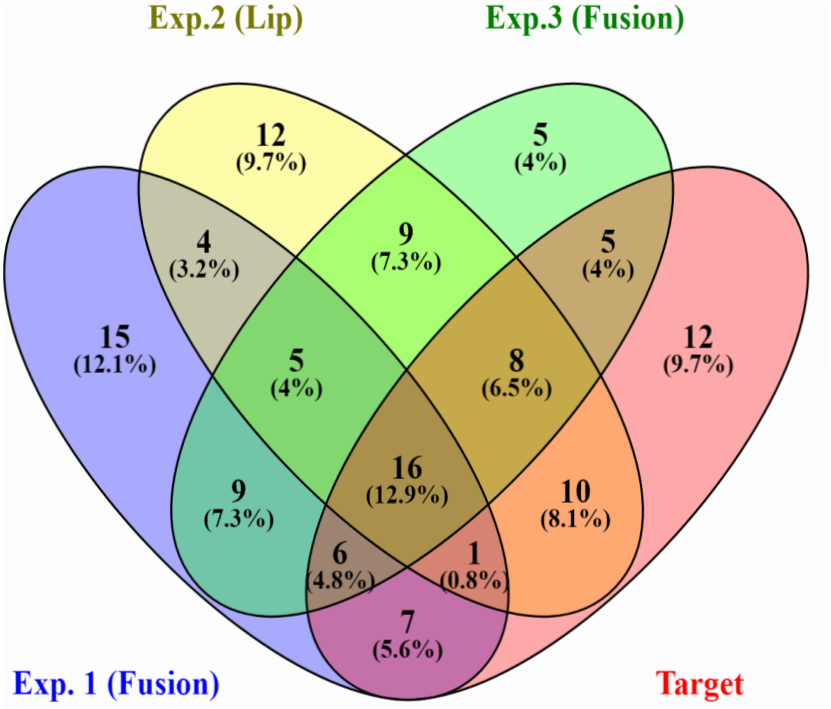


Figure 4.23: *Venn Diagram of the best results of three experiments and annotated labels (Target).*

4.5 Conclusion

This chapter mainly presents cognitive processing components for public speaking training system which can help a machine to interact and manage multiple trainees. Without these components, a multimodal spoken dialogue system is not able to manage multiple trainees at a time. It is also unable to handle spoken utterances which are directed towards it and unable to deploy certain repair strategies (e.g., switching between ASR models) if humans' are facing trouble in interacting with a machine due to systems components failure. The conclusion of each study is described below.

4.5.1 On-Talk & Off-Talk Detection

EEG features are able to predict the 'on- and off-talk' maximum accuracy of 80.25%. For on-/off-talk detection, prosodic information provides slightly better accuracy (81.83%) than EEG (80.25%) and fusion of prosodic and EEG information results in an improvement in performance (86.76%). However, prosodic features extracted from the whole utterance add latency because the system needs to wait until the utterance is finished, while EEG features (prior to articulation) do not have this problem. Moreover, a system based on prosodic features will typically perform poorly in noisy settings. In future work, it is planned to investigate higher frequency bands, as well as detection of on-talk using the EEG signal from the mid-line and left hemisphere of the brain to encode the phonological and motor area activation information.

The results of this study also indicate that the EEG frequency bands contribute significantly towards the detection of On-Talk and Off-Talk utterances, and sta-

tistical significance tests suggest that EEG features (during speech articulation) are statistically different for these two types of utterances. The higher frequencies contain significant information which could be further explored for emotion recognition. It is also found that the muscle activity (talk related muscles artefacts) may influence the results positively. A possible direction for future work is to explore facial movements during On-Talk and Off-Talk to further corroborate these assumptions (muscle activities). It is also worth exploring, how different uncontrolled acoustic environments affects the prediction power of a classifier (that is trained using data of controlled environment) for On-Talk and Off-Talk detection. Some of the results of this study are published in international conference (Hayakawa, Haider, Luz, Cerrato, & Campbell, 2016b; Haider, Akira, Luz, Carl, & Campbell, 2018).

4.5.2 Cognitive State Detection

The results of study indicate that (a) the association between the cognitive state and the biosignals does not seem to persist until the next sentence is uttered, as suggested by the poor state detection performance in time windows that include following utterance, and (b) that features of the speech signal can be used to complement biosignals in detecting cognitive states in time windows that include the following utterance. Extending the window to the end of the utterance following the cognitive state yields poor detection on biosignals alone, but improves considerably if features of the speech signal are added, thus showing the potential usefulness of speech features as a biosignal. The results of this study are published in international conference (Akira et al., 2015)

4.5.3 Active Speaker Detection

The results show that head and lip movements (during speech articulation) are significantly correlated with one's speech and can be used in detecting speech and speaker. The results also show that the 'going to speak' region (one second before speech articulation) contains a significant amount of information about who holds the floor in a dialogue. The visual prosody features extracted from this region provide less accurate results than the 'speech' region (one second after articulation) for the classification task but the difference is not statistically significant. The fusion of features from both regions improves performance for linear discrimination analysis and decision tree. Possible future work is to evaluate the low-level visual descriptors (e.g. histogram of the gradient) extracted from 'going to speak' and 'speech' region for active speaker detection and its fusion with the audio features. Another possible future work is to detect the 'going to speak region' by using visual information only instead of relying on 10 ms of speech utterance. The results of this study are published in international conference (Haider, Luz, & Campbell, 2016a) (Haider, Luz, Vogel, & Campbell, 2018)

Chapter 5

Conclusion and Future Work

This thesis presents multiple novel models and systems which can help in developing a multimodal multiparty spoken dialogue system for public speaking training. This thesis proposes novel models and systems which are mainly focused on two parts. The first part deals with the public speaking abilities and the social signal and behaviour cues that represent the public speaking abilities. In the context of this thesis, prosody, face, and body gestures information are used to make inference about a presentation. However, the thesis does not focus on the actual verbal content spoken because the objective of this thesis is to provide models and systems for the training of non-verbal aspects of public speaking. The first part of thesis considers four different types of public speaking situations (TED talks, students' presentation, video blogs and political debates in parliament) and provides novel models in recognizing some public speaking abilities in those contexts. However, there are many public speaking situations (e.g., journalist speech, teachers lectures, politician speech to public and job interviews) where automatic systems are needed for training humans which can be the future work of the the-

sis. The student's debates dataset is collected but is not used to train machines to recognize the political debates skills. One of the possible future work is to use the debates dataset to train models for debates training skills. The second part of the thesis focuses on three different types of challenges (on-off talk, active speaker detection, and cognitive state detection) which a multiparty spoken dialogue system may face while training humans for public speaking. It could help a machine in deploying certain repair strategies in case some of the machine components (e.g., ASR failure, self-speaking) fail, sense user experience and manage multiple users (trainees) at a time. However, there are many other humans' behaviours (like the use of body gestures while interacting with machines) which can be the possible future work of the thesis.

Bibliography

Abdel Rahman, R., van Turennout, M., & Levelt, W. J. (2003). Phonological encoding is not contingent on semantic feature retrieval: an electrophysiological study on object naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 850–860. doi: 10.1037/0278-7393.29.5.850

Adeli, H., Ghosh-Dastidar, S., & Dadmehr, N. (2007). A Wavelet-Chaos Methodology for Analysis of EEGs and EEG Subbands to Detect Seizure and Epilepsy. *Biomedical Engineering, IEEE Transactions on*, *54*(2), 205–211.

Akira, H., Haider, F., Cerrato, L., Campbell, N., & Luz, S. (2015). Detection of cognitive states and their correlation to speech recognition performance in speech-to-speech machine translation systems. In *Sixteenth annual conference of the international speech communication association* (pp. 2539–2543).

Albers, M. J., & Mazur, M. B. (2014). *Content and complexity: information design in technical communication*. Routledge.

Alexandersson, J., Aretoulaki, M., Campbell, N., Gardner, M., Girenko, A., Klakow, D., ... others (2014). Metalogue: A multiperspective multimodal dialogue system with metacognitive abilities for highly adaptive and flexible dialogue

management. In *Intelligent environments (ie), 2014 international conference on* (pp. 365–368).

Allwood, J., & Henrichsen, P. J. (2013). Predicting the attitude flow in dialogue based on multi-modal speech cues. In *Nealt proceedings. northern european association for language and technology; 4th nordic symposium on multimodal communication; november 15-16; gothenburg; sweden* (pp. 47–53).

Al Moubayed, S., Beskow, J., Skantze, G., & Granström, B. (2012). Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive behavioural systems* (pp. 114–130). Springer.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., ... Weinert, R. (1991, October). The hcrc map task corpus. *Language and Speech*, *34*(4), 351–366. Retrieved 2014-03-21, from `http://las.sagepub.com/content/34/4/351` doi: 10.1177/002383099103400404

Anwar, A., Salama, G. I., & Abdelhalim, M. B. (2013). Video Classification And Retrieval Using Arabic Closed Caption. In *Icit 2013 the 6th international conference on information technology video.*

Attfield, S., Piwowarski, B., & Kazai, G. (2011). Towards a science of user engagement (Position Paper). In *Wsdm workshop on user modelling for web applications.* Hong Kong.

Balomenos, T., Raouzaiou, A., Karpouzis, K., Kollias, S., & Cowie, R. (2003). An introduction to emotionally rich man-machine intelligent systems. In *Proceedings of the 3rd European Symposium on Intelligent Technologies, Hybrid Systems and their Implementation on Smart Adaptive Systems (EUNITE'03).*

Baron-Cohen, S., Ring, H. A., Wheelwright, S., Bullmore, E. T., Brammer, M. J., Simmons, A., & Williams, S. C. (1999). Social intelligence in the normal and autistic brain: an fmri study. *European Journal of Neuroscience*, *11*(6), 1891–1898.

Batliner, A., Hacker, C., & Nöth, E. (2006). To talk or not to talk with a computer: On-Talk vs. Off-Talk. In *How People Talk to Computers, Robots, and Other Artificial Communication Partners* (pp. 79–100). Hansewissenschaftskolleg, Delmenhorst, Germany: SFB/TR 8 Spatial Cognition.

Batliner, A., Hacker, C., & Nöth, E. (2009). To talk or not to talk with a computer: Taking into account the user's focus of attention. *Journal on multimodal user interfaces*, *2*(3–4), 171–186.

Benini, S., Migliorati, P., & Leonardi, R. (2010). Statistical Skimming of Feature Films. *International Journal of Digital Multimedia Broadcasting*, *2010*, 1–11. Retrieved from `http://www.hindawi.com/journals/ijdmb/2010/709161/` doi: 10.1155/2010/709161

Bernstein, L. E., Tucker, P. E., & Demorest, M. E. (2000). Speech perception without hearing. *Perception & Psychophysics*, *62*(2), 233–252.

Biel, J.-I., Aran, O., & Gatica-Perez, D. (2011). You are known by how you vlog: Personality impressions and nonverbal behavior in youtube. In *Icwsm* (pp. 446–449).

Biel, J.-I., & Gatica-Perez, D. (2011). Vlogsense: Conversational behavior and social attention in youtube. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *7*(1), 33.

Biel, J.-I., Teijeiro-Mosquera, L., & Gatica-Perez, D. (2012). Facetube: predicting personality from facial expressions of emotion in online conversational video. In *Proceedings of the 14th acm international conference on multimodal interaction* (pp. 53–56).

Blank, S. C., Scott, S. K., Murphy, K., Warburton, E., & Wise, R. J. (2002). Speech production: Wernicke, Broca and beyond. *Brain*, *125*(8), 1829–1838. doi: 10.1093/brain/awf191

Bock, J. K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*, *89*(1), 1–47. doi: 10.1037/0033-295x.89.1.1

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Breazeal, C. (2003). Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, *59*(1), 119–155.

Breiman, L. (2002). Manual on setting up, using, and understanding random forests v3. 1. *Statistics Department University of California Berkeley, CA, USA*, *1*.

Brezeale, D., & Cook, D. J. (2009). Learning video preferences using visual features and closed captions. *IEEE Multimedia*, *16*(3), 39–47. doi: 10.1109/MMUL.2009.51

Brugman, H., Russel, A., & Nijmegen, X. (2004). Annotating Multi-media/Multimodal Resources with ELAN. In *Lrec.*

Burgoon, J. K., Jensen, M. L., Meservy, T. O., Kruse, J., & Nunamaker, J. (2005). Augmenting human identification of emotional states in video. In *Intelligence Analysis Conference, McClean, VA.*

Campbell, N. (2008). Multimodal processing of discourse information; the effect of synchrony. In *2008 second international symposium on universal communication* (pp. 12–15).

Cech, J., Mittal, R., Deleforge, A., Sanchez-Riera, J., Alameda-Pineda, X., & Horaud, R. (2013). Active-speaker detection and localization with microphones and cameras embedded into a robotic head. In *2013 13th ieee-ras international conference on humanoid robots (humanoids)* (pp. 203–210).

Chakravarty, P., Mirzaei, S., Tuytelaars, T., et al. (2015). Who's speaking?: Audio-supervised classification of active speakers in video. In *Proceedings of the 2015 acm on international conference on multimodal interaction* (pp. 87–90).

Chakravarty, P., Zegers, J., Tuytelaars, T., & Van hamme, H. (2016). Active speaker detection with audio-visual co-training. In *Proceedings of the 18th acm international conference on multimodal interaction* (pp. 312–316). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/2993148.2993172` doi: 10.1145/2993148.2993172

Chatfield, K., Lempitsky, V. S., Vedaldi, A., & Zisserman, A. (2011). The devil is in the details: an evaluation of recent feature encoding methods. In *Bmvc* (Vol. 2, p. 8).

Chen, F., De Vleeschouwer, C., & Cavallaro, A. (2014). Resource allocation for personalized video summarization. *IEEE Transactions on Multimedia*, *16*(2), 455–469. doi: 10.1109/TMM.2013.2291967

Chen, L., Leong, C. W., Feng, G., & Lee, C. M. (2014). Using multimodal cues to analyze MLA'14 oral presentation quality corpus: Presentation delivery and slides quality. In *Proceedings of the 2014 acm workshop on multimodal learning analytics workshop and grand challenge* (pp. 45–52). New York, NY, USA: ACM. doi: 10.1145/2666633.2666640

Christian, A. D., & Avery, B. L. (1998). Digital smart kiosk project. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 155–162).

Christoffels, I. K., Formisano, E., & Schiller, N. O. (2007). Neural correlates of verbal feedback processing: an fMRI study employing overt speech. *Human Brain Mapping*, *28*(9), 868–879. doi: 10.1002/hbm.20315

Criswell, E. (2010). *Cram's introduction to surface electromyography.* Sudbury, Massachusetts, USA: Jones & Bartlett Publishers.

Curtis, K., Jones, G. J., & Campbell, N. (2016). Speaker impact on audience comprehension for academic presentations. In *Proceedings of the 18th acm international conference on multimodal interaction* (pp. 129–136).

Cutler, R., & Davis, L. (2000). Look who's talking: Speaker detection using video and audio correlation. In *Multimedia and expo, 2000. icme 2000. 2000 ieee international conference on* (Vol. 3, pp. 1589–1592).

DeCoske, M. A., & White, S. J. (2010, August). Public speaking revisited: delivery, structure, and style. *American journal of health-system pharmacy : AJHP : official journal of the American Society of Health-System Pharmacists*, *67*(15), 1225—1227. Retrieved from `http://www.ajhp.org/cgi/content/full/67/15/1225` doi: 10.2146/ajhp090508

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321. doi: 10.1037/0033-295x.93.3.283

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21.

De Moraes, J. A., Rilliard, A., de Oliveira Mota, B. A., & Shochi, T. (2010). Multimodal perception and production of attitudinal meaning in brazilian portuguese. *Proc. Speech Prosody, paper*, *340*.

Dobrian, F., Awan, A., Joseph, D., Ganjam, A., Zhan, J., Sel, V., ... Zhang, H. (2013). Understanding the Impact of Video Quality on User Engagement. In *Communications of the acm* (pp. 91–99). doi: 10.1145/2428556.2428577

D O'Donnell, R., Berkhout, J., & Adey, W. R. (1974). Contamination of scalp EEG spectrum during contraction of cranio-facial muscles. *Electroencephalography and Clinical Neurophysiology*, *37*(2), 145–151.

Dong, A., & Li, H. (2008). Ontology-driven annotation and access of presentation video data. *Estudios de Economía Aplicada*.

Duncan-Johnson, C. C., & Kopell, B. S. (1981). The stroop effect: brain potentials localize the source of interference. *Science*, *214*(4523), 938–940. doi: 10.1126/science.7302571

Echeverría, V., Avendaño, A., Chiluiza, K., Vásquez, A., & Ochoa, X. (2014). Presentation skills estimation based on video and kinect data analysis. In *Proceedings of the 2014 acm workshop on multimodal learning analytics workshop and grand challenge* (pp. 53–60). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/2666633.2666641` doi: 10.1145/2666633.2666641

Edlund, J., Gustafson, J., Heldner, M., & Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. *Speech communication*, *50*(8), 630–645.

Ekman, P., & Friesen, W. V. (1981). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Nonverbal communication, interaction, and gesture*, 57–106.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*(3), 572–587.

Evangelopoulos, G., Zlatintsi, A., Potamianos, A., Maragos, P., Rapantzikos, K., Skoumas, G., & Avrithis, Y. (2013). Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, *15*(7), 1553–1568. doi: 10.1109/TMM.2013.2267205

Eyben, F., Weninger, F., Groß, F., & Schuller, B. (2013a). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st acm international conference on multimedia* (pp. 835–838).

Eyben, F., Weninger, F., Groß, F., & Schuller, B. (2013b). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st acm international conference on multimedia* (pp. 835–838).

Eyben, F., Wöllmer, M., & Schuller, B. (2009). Openear—introducing the munich open-source emotion and affect recognition toolkit. In *Affective computing and intelligent interaction and workshops, 2009. acii 2009. 3rd international conference on* (pp. 1–6).

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th acm international conference on multimedia* (pp. 1459–1462).

Fleiss, J. (n.d.). Statistical methods for rates and proportions. john wiley & sons; new york: 1981. *The measurement of interrater agreement*, 212–36.

Friendly, M. (2002). Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, *56*, 316–324.

Garrett, M. F. (1975). The Analysis of Sentence Production. *Psychology of Learning and Motivation*, *9*, 133–177. doi: 10.1016/S0079-7421(08)60270-4

Garrett, M. F. (1988). Processes in language production. In *Language: Psychological and biological aspects* (pp. 69–96). Cambridge University Press.

Gerl, F., & Herbig, T. (2008, October 10). *Speaker recognition system.* Google Patents. (US Patent App. 12/249,089)

Goleman, D., & Boyatzis, R. (2008). Social intelligence and the biology of leadership. *Harvard Business Review*, *86*(9), 74–81.

Gotman, J. (2010, February). High frequency oscillations: The new EEG frontier? *Epilepsia*, *51*(0 1), 63–65. Retrieved 2017-08-30, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786932/ doi: 10.1111/j.1528-1167.2009.02449.x

Graf, H. P., Cosatto, E., Strom, V., & Huang, F. J. (2002). Visual prosody: Facial movements accompanying speech. In *Automatic face and gesture recognition, 2002. proceedings. fifth ieee international conference on* (pp. 396–401).

Grandstaff, D. (2004). *Speaking as a professional: Enhance your therapy or coaching practice through presentations, workshops, and seminars.* W.W. Norton & Company. Retrieved from https://books.google.ie/books?id=UvmrZdAmNcYC

Gunes, H., & Piccardi, M. (2005). Fusing face and body display for bi-modal emotion recognition: Single frame analysis and multi-frame post integration. In *Affective Computing and Intelligent Interaction* (pp. 102–111). Springer.

Guo, P. J., Kim, J., & Rubin, R. (2014). How Video Production Affects Student Engagement : An Empirical Study of MOOC Videos. In *L@s 2014 - proceedings of the 1st acm conference on learning at scale* (pp. 41–50). doi: 10.1145/2556325.2566239

Haesen, M., Meskens, J., Luyten, K., Coninx, K., Becker, J. H., Tuytelaars, T., ... Moens, M.-F. (2011, May). Finding a needle in a haystack: an interactive video archive explorer for professional video searchers. *Multimedia Tools and Applications*, *63*(2), 331–356. Retrieved from http://link.springer.com/10.1007/s11042-011-0809-y doi: 10.1007/s11042-011-0809-y

Haider, F., Akira, H., Luz, S., Carl, V., & Campbell, N. a. (2018). On-talk and off-talk detection: A discrete wavelet transform analysis of electroencephalogram. In *2018 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 960–964). Calgary, Canada.

Haider, F., & Al Moubayed, S. (2012). Towards speaker detection using lips movements for humanmachine multiparty dialogue. *FONETIK 2012*.

Haider, F., Cerrato, L., Campbell, N., & Luz, S. (2016). Presentation quality assessment using acoustic information and hand movements. In *Proceeding of 41st ieee international conference on acoustics, speech and signal processing (icassp).*

Haider, F., Cerrato, L. S., Luz, S., & Campbell, N. (2016). Attitude recognition of video bloggers using audio-visual descriptors. In *Proceedings of the workshop on multimodal analyses enabling artificial agents in human-machine interaction* (pp. 38–42). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/3011263.3011270` doi: 10.1145/3011263.3011270

Haider, F., Luz, S., & Campbell, N. (2016a). Active speaker detection in human machine multiparty dialogue using visual prosody information. In *Proceedings of the ieee global conference on signal and information processing (globalsip)* (pp. 1207–1211). Washington, D.C., USA.

Haider, F., Luz, S., & Campbell, N. (2016b). Metalogue: Data collection using a real time feedback tool for non verbal presentation skills training. In *Proceedings of the lrec 2016 workshop just talking – casual talk among humans and machines* (p. 13-15). Portoroz, Slovenia.

Haider, F., Luz, S., & Campbell, N. (2017). Data collection and synchronisation: Towards a multiperspective multimodal dialogue system with metacognitive abilities. In K. Jokinen & G. Wilcock (Eds.), *Dialogues with social robots: Enablements, analyses, and evaluation* (pp. 245–256). Singapore: Springer Singapore. Retrieved from https://doi.org/10.1007/978-981-10-2585-3_19 doi: 10.1007/978-981-10-2585-3_19

Haider, F., Luz, S., Vogel, C., & Campbell, N. (2018). Improving response time of active speaker detection using visual prosody information prior to articulation. In *INTERSPEECH.* Hyderabad, India.

Haider, F., Salim, F. A., Luz, S., Conlan, O., & Campbell, N. (2015). High level visual and paralinguistic features extraction and their correlation with user engagement. In *2015 ieee international symposium on signal processing and information technology (isspit)* (pp. 326–331).

Haider, F., Salim, F. A., Luz, S., Vogel, C., Conlan, O., & Campbell, N. (2017). Visual, laughter, applause and spoken expression features for predicting engagement within ted talks. In *Proc. interspeech 2017* (pp. 2381–2385). Retrieved from http://dx.doi.org/10.21437/Interspeech.2017-1633 doi: 10.21437/Interspeech.2017-1633

Han, J., Gilmartin, E., & Campbell, N. (2013). Herme, yet another interactive conversational robot. In *Affective computing and intelligent interaction (acii), 2013 humaine association conference on* (pp. 711–712).

Hashimoto, T., Hiramatsu, S., Tsuji, T., & Kobayashi, H. (2007). Realization and evaluation of realistic nod with receptionist robot saya. In *Ro-man 2007-the*

16th ieee international symposium on robot and human interactive communication (pp. 326–331).

Hashimoto, T., Kato, N., & Kobayashi, H. (2011). Development of educational system with the android robot saya and evaluation. *Int J Adv Robotic Sy*, *8*(3), 51–61.

Hayakawa, A., Campbell, N., & Luz, S. (2014). Interlingual Map Task Corpus Collection. In *Proceedings of INTERSPEECH'14: the 15th Annual Conference of the International Speech Communication Association* (pp. 189–191). Singapore: ISCA.

Hayakawa, A., Haider, F., Cerrato, L., Campbell, N., & Luz, S. (2015). Detection of Cognitive States and Their Correlation to Speech Recognition Performance in Speech-to-Speech Machine Translation Systems. In *Proceedings of INTERSPEECH'15: the 16th Annual Conference of the International Speech Communication Association* (pp. 2539–2543). Dresden, Germany: ISCA.

Hayakawa, A., Haider, F., Luz, S., Cerrato, L., & Campbell, N. (2016a). Talking to a system and oneself: A study from a Speech-to-Speech, Machine Translation mediated Map Task. In *Proceedings of Speech Prosody 2016 (SP8)* (pp. 776–780). Boston, Massachusetts, USA: ISCA.

Hayakawa, A., Haider, F., Luz, S., Cerrato, L., & Campbell, N. (2016b). Talking to a system and oneself: A study from a Speech-to-Speech, Machine Translation mediated Map Task. In *Proceedings of Speech Prosody 2016 (SP8)* (pp. 776–780). Boston, Massachusetts, USA: ISCA.

Hayakawa, A., Luz, S., Cerrato, L., & Campbell, N. (2016, May). The ILMT-s2s Corpus — A Multimodal Interlingual Map Task Corpus. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 605–612). Paris, France: European Language Resources Association (ELRA).

Helvert, J. V., Rosmalen, P. V., Börner, D., Petukhova, V., & Alexandersson, J. (2015). Observing, coaching and reflecting: A multi-modal natural language-based dialogue system in a learning context. In *Workshop proceedings of the 11th international conference on intelligent environments* (Vol. 19, pp. 220–227).

Helvert, J. V., Rosmalen, P. V., Börner, D., Petukhova, V., & Alexandersson, J. (2015). Observing, Coaching and Reflecting: A Multi-modal Natural Language-based Dialogue System in a Learning Context. In *Workshop Proceedings of the 11th International Conference on Intelligent Environments* (Vol. 19, pp. 220–227). IOS Press.

Hincks, R. (2005). Measuring liveliness in presentation speech. In *Interspeech* (pp. 765–768).

Hjalmarsson, A. (2010). Human interaction as a model for spoken dialogue system behaviour.

Hudlicka, E. (2003). To feel or not to feel: The role of affect in human–computer interaction. *International Journal of Human-Computer Studies*, *59*(1), 1–32.

Ishii, R., Kumano, S., & Otsuka, K. (2016). Analyzing mouth-opening transition pattern for predicting next speaker in multi-party meetings. In *Proceedings of the 18th acm international conference on multimodal interaction* (pp. 209–216).

Jasper, H. H. (1958). The ten twenty electrode system of the international federation. *Electroencephalography and clinical Neurophysiology*, *10*, 371–375.

Kapoor, A., & Picard, R. W. (2005). Multimodal affect recognition in learning environments. In *Proceedings of the 13th annual ACM international conference on Multimedia* (pp. 677–682).

Kempen, G. (1977). Conceptualizing and formulating in sentence production. In *Sentence production: Developments in research and theory* (pp. 259–274). Erlbaum.

Kempen, G., & Hoenkamp, E. (1987). An incremental procedural grammar for sentence formulation. *Cognitive Science*, *11*(2), 201–258. doi: 10.1207/s15516709cog1102_5

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta psychologica*, *26*, 22–63.

Krajewski, J., Batliner, A., & Kessel, S. (2010). Comparing multiple classifiers for speech-based detection of self-confidence-a pilot study. In *Pattern recognition (icpr), 2010 20th international conference on* (pp. 3716–3719).

Kumar, S., Narayan, Y., & Amell, T. (2003). Power spectra of sternocleidomastoids, splenius capitis, and upper trapezius in oblique exertions. *The Spine Journal*, *3*(5), 339–350.

Lamerton, J. (2001). *Public speaking. everything you need to know.* Harpercollins Publishers Ltd.

Lee, C. M., & Narayanan, S. (2005). Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, *13*(2), 293–303. doi: 10.1109/TSA.2004.838534

Levelt, W. J. (1989). *Speaking: From intention to articulation*. MIT press.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R news*, *2*(3), 18–22.

Lienhart, R., Kuranov, A., & Pisarevsky, V. (2003). Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In *Proceedings of the 25th dagm pattern recognition symposium* (pp. 297–304). Retrieved from `http://www.springerlink.com/index/WYP0XR0WMTMYY4RK.pdf$\delimiter"026E30F$nhttp://link.springer.com/chapter/10.1007/978-3-540-45243-0_39` doi: 10.1007/978-3-540-45243-0_39

Liotti, M., Woldorff, M. G., Perez, R., & Mayberg, H. S. (2000). An ERP study of the temporal course of the Stroop color-word interference effect. *Neuropsychologia*, *38*(5), 701–711. doi: 10.1016/S0028-3932(99)00106-2

Liu, M., Wang, R., Li, S., Shan, S., Huang, Z., & Chen, X. (2014). Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In *Proceedings of the 16th international conference on multimodal interaction* (pp. 494–501).

Lucas, S. E. (2008). *The art of public speaking 11th edition*. McGraw-Hill.

Luz, S. (2012). The non-verbal structure of patient case discussions in multidisciplinary medical team meetings. *ACM Transactions on Information Systems*, *30*(3), 17:1–17:24.

Luzardo, G., Guamán, B., Chiluiza, K., Castells, J., & Ochoa, X. (2014). Estimation of presentations skills based on slides and audio features. In *Proceedings of the 2014 acm workshop on multimodal learning analytics workshop and grand challenge* (pp. 37–44). New York, NY, USA: ACM. doi: 10.1145/2666633.2666639

Machines, S. (2009). Faceapi. *URL: http://www.seeingmachines.com*.

Madzlan, N., Han, J., Bonin, F., & Campbell, N. (2014). Towards automatic recognition of attitudes: Prosodic analysis of video blogs. *Speech Prosody, Dublin, Ireland*, 91–94.

Madzlan, N. A., Han, J. G., Bonin, F., & Campbell, N. (2014). Automatic recognition of attitudes in video blogs-prosodic and visual feature analysis. In *Interspeech* (pp. 1826–1830).

Madzlan, N. A., Huang, Y., & Campbell, N. (2015). Automatic classification and prediction of attitudes: Audio-visual analysis of video blogs. In *International conference on speech and computer* (pp. 96–104).

Madzlan, N. A., Reverdy, J., Bonin, F., Cerrato, L., & Campbell, N. (2016). Annotation and multimodal perception of attitudes: A study on video blogs. In *Proceedings from the 3rd european symposium on multimodal communication, dublin, september 17-18, 2015* (pp. 50–54).

Martin, J.-C., Niewiadomski, R., Devillers, L., Buisine, S., & Pelachaud, C. (2006). Multimodal complex emotions: Gesture expressivity and blended facial expressions. *International Journal of Humanoid Robotics*, *3*(03), 269–291.

Matejka, M., Kazzer, P., Seehausen, M., Bajbouj, M., Klann-Delius, G., Gisela, Menninghaus, W., ... Prehn, K. (2013). Talking about Emotion: Prosody and Skin Conductance Indicate Emotion Regulation. *Frontiers in Psychology*, *4*, 260. doi: 10.3389/fpsyg.2013.00260

McAdam, D. W., & Whitaker, H. A. (1971). Language production: Electroencephalographic localization in the normal human brain. *Science*, *172*(3982), 499–502.

McCowan, I., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., & Wellner, P. (2005). The AMI meeting corpus. In *In: Proceedings measuring behavior 2005, 5th international conference on methods and techniques in behavioral research. l.p.j.j. noldus, f. grieco, l.w.s. loijens and p.h. zimmerman (eds.), wageningen: Noldus information technology.*

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* University of Chicago Press.

McNeill, D. (2008). *Gesture and thought.* University of Chicago Press.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception. *Psychological Science*, *15*(2), 133–137.

O'Brien, H. L., & Toms, E. G. (2008). What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, *59*(6), 938–955.

O'Brien, H. L., & Toms, E. G. (2013). Examining the generalizability of the User Engagement Scale (UES) in exploratory search. *Information Processing and Management*, *49*(5), 1092–1107. Retrieved from `http://dx.doi.org/10.1016/j.ipm.2012.08.005` doi: 10.1016/j.ipm.2012.08.005

Ocak, H. (2009). Automatic detection of epileptic seizures in EEG using discrete wavelet transform and approximate entropy. *Expert Systems with Applications*, *36*(2), 2027–2036.

Ochoa, X., Worsley, M., Chiluiza, K., & Luz, S. (2014a). Mla'14: Third multimodal learning analytics workshop and grand challenges. In *Proceedings of the 16th international conference on multimodal interaction* (pp. 531–532). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/2663204.2668318` doi: 10.1145/2663204.2668318

Ochoa, X., Worsley, M., Chiluiza, K., & Luz, S. (2014b). Mla'14: Third multimodal learning analytics workshop and grand challenges. In *Proceedings of the 16th international conference on multimodal interaction* (pp. 531–532).

Ochoa, X., Worsley, M., Chiluiza, K., & Luz, S. (2014c). Mla'14: Third multimodal learning analytics workshop and grand challenges. In *Proceedings*

of the 16th international conference on multimodal interaction (pp. 531–532). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2663204.2668318 doi: 10.1145/2663204.2668318

Oliveira, M. (2002). The role of pause occurrence and pause duration in the signaling of narrative structure. In *Advances in natural language processing* (Vol. 2389, pp. 43–51). Springer.

Oostendorp, T. F., Delbeke, J., & Stegeman, D. F. (2000, November). The conductivity of the human skull: results of in vivo and in vitro measurements. *IEEE transactions on bio-medical engineering*, *47*(11), 1487–1492. doi: 10.1109/TBME.2000.880100

Oppermann, D., Schiel, F., Steininger, S., & Beringer, N. (2001). Off-Talk — a Problem for Human-Machine-Interaction? In *Proceedings of EUROSPEECH 2001 Scandinavia: the 7th European Conference on Speech Communication and Technologyand the 2nd INTERSPEECH Event* (pp. 2197–2200). Aalborg, Denmark: ISCA.

Pantic, M., & Rothkrantz, L. J. (2003). Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, *91*(9), 1370–1390.

Pavlović, V., Garg, A., Rehg, J. M., & Huang, T. S. (2000). Multimodal speaker detection using error feedback dynamic bayesian networks. In *Computer vision and pattern recognition, 2000. proceedings. ieee conference on* (Vol. 2, pp. 34–41).

Pei, X., Leuthardt, E. C., Gaona, C. M., Brunner, P., Wolpaw, J. R., & Schalk, G. (2011). Spatiotemporal dynamics of electrocorticographic high gamma activity

during overt and covert word repetition. *NeuroImage*, *54*(4), 2960–2972. doi: 10.1016/j.neuroimage.2010.10.029

Pentland, A. (2007, July). Social signal processing [exploratory dsp]. *IEEE Signal Processing Magazine*, *24*(4), 108-111. doi: 10.1109/MSP.2007.4286569

Perronnin, F., Sánchez, J., & Mensink, T. (2010). Improving the fisher kernel for large-scale image classification. , 143–156.

Petrantonakis, P. C., & Hadjileontiadis, L. J. (2010). Emotion Recognition From EEG Using Higher Order Crossings. *Information Technology in Biomedicine, IEEE Transactions on*, *14*(2), 186–197.

Petukhova, V., Malchanau, A., Oualil, Y., Klakow, D., Luz, S., Haider, F., ... Alexandersson, J. (2018, May 7-12, 2018). The Metalogue Debate Trainee Corpus: Data Collection and Annotations. In N. C. C. chair) et al. (Eds.), *Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018).* Miyazaki, Japan: European Language Resources Association (ELRA).

Picard, R. W. (2000). *Affective computing.* MIT press.

Porbadnigk, A., Wester, M., & Jan-p Calliess, T. S. (2009). EEG-based Speech Recognition – Impact of Temporal Effects. In *Biosignals 2009 - proceedings of the international conference on bio-inspired systems and signal processing* (pp. 376—381). Porto, Portugal: INSTICC Press.

Raudys, S., & Duin, R. P. W. (1998). Expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix. *Pattern Recognition Letters*, *19*(5-6), 385–392.

Ray, W., & Cole, H. (1985). EEG Alpha Activity Reflects Attentional Demands, and Beta Activity Reflects Emotional and Cognitive Processes. *Science*, *228*(4700), 750–752. doi: 10.1126/science.3992243

Reynolds, D. A. (2002). An overview of automatic speaker recognition technology. In *2002 ieee international conference on acoustics, speech, and signal processing* (Vol. 4, pp. 4072–4075). Orlando, Florida, USA: IEEE. doi: 10.1109/ICASSP .2002.5745552

Ross, E. D., & Mesulam, M.-M. (1979). Dominant language functions of the right hemisphere?: Prosody and emotional gesturing. *Archives of Neurology*, *36*(3), 144–148. doi: 10.1001/archneur.1979.00500390062006

Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T., & Meignier, S. (2013). *An open-source state-of-the-art toolbox for broadcast news diarization* (Tech. Rep.). Idiap.

Salim, F. A., Haider, F., Conlan, O., Luz, S., & Campbell, N. (2015). Analyzing multimodality of video for user engagement assessment. In *Proceedings of the 2015 acm on international conference on multimodal interaction* (pp. 287–290).

Sansen, H., Torres, M. I., Chollet, G., Glackin, C., Petrovska-Delacretaz, D., Boudy, J., ... Schlögl, S. (2016). The roberta ironside project: A dialog capable humanoid personal assistant in a wheelchair for dependent persons. In *2016 2nd international conference on advanced technologies for signal and image processing (atsip)* (pp. 381–386).

Schmitt, B. M., Münte, T. F., & Kutas, M. (2000). Electrophysiological estimates of the time course of semantic and phonological encoding during implicit picture naming. *Psychophysiology*, *37*(4), 473–484. doi: 10.1111/1469-8986.3740473

Schmitt, B. M., Schiltz, K., Zaake, W., Kutas, M., & Münte, T. F. (2001). An electrophysiological analysis of the time course of conceptual and syntactic encoding during tacit picture naming. *Journal of Cognitive Neuroscience*, *13*(4), 510–522. doi: 10.1162/08989290152001925

Schneider, A., & Luz, S. (2011). Speaker alignment in synthesised, machine translated communication. In *International workshop on spoken language translation* (pp. 254–260). San Francisco, California, USA.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., ... others (2013a). The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., ... others (2013b). The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In *Proceedings of INTERSPEECH'13: the 14th Annual Conference of the International Speech Communication Association* (pp. 148–152). Lyon, France: ISCA.

Shapiro, B. E., & Danly, M. (1985). The role of the right hemisphere in the control of speech prosody in propositional and affective contexts. *Brain and Language*, *25*(1), 19–36. doi: 10.1016/0093-934X(85)90118-X

Sinclair, J. (1995). *The collins cobuild english language dictionary*. HarperCollins, London.

Slaney, M., Stolcke, A., & Hakkani-Tür, D. (2014). The relation of eye gaze and face pose: Potential impact on speech recognition. In *Proceedings of the 16th international conference on multimodal interaction* (pp. 144–147).

Sousa, R., & Ferreira, A. (2008, Sept). Evaluation of existing harmonic-to-noise ratio methods for voice assessment. In *Signal processing algorithms, architectures, arrangements, and applications (spa), 2008* (p. 73-78).

Stassen, H., et al. (1993). Speaking behavior and voice sound characteristics in depressive patients during recovery. *Journal of Psychiatric Research*, *27*(3), 289–307.

Stern, W. (1921). *Die differentielle psychologie: in ihren methodischen grundlagen*. JA Barth.

Sternberg, R. J. (1990). *Metaphors of mind: Conceptions of the nature of intelligence*. Cambridge University Press.

Stone, J. L. (1991). Paul Broca and the first craniotomy based on cerebral localization. *Journal of Neurosurgery*, *75*(1), 154–159. doi: 10.3171/jns.1991.75.1.0154

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, *26*(2), 212–215.

Takeuchi, S., Hashiba, T., Tamura, S., & Hayamizu, S. (2009). Voice activity detection based on fusion of audio and visual information. *Proceedings of the Auditory-Visual Speech Processing. AVSP (Norwich, UK)*.

Tan, S., Bu, J., Qin, X., Chen, C., & Cai, D. (2014, March). Cross domain recommendation based on multi-type media fusion. *Neurocomputing*, *127*, 124–134. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0925231213009260` doi: 10.1016/j.neucom.2013.08.034

Traunmüller, H., & Eriksson, A. (1995). The perceptual evaluation of f0 excursions in speech as evidenced in liveliness estimations. *The Journal of the Acoustical Society of America*, *97*(3), 1905–1915.

Tumposky, N. R. (2004). The debate debate. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *78*(2), 52–56.

Uijlings, J., Duta, I., Sangineto, E., & Sebe, N. (2015). Video classification with densely extracted hog/hof/mbh features: an evaluation of the accuracy/computational efficiency trade-off. *International Journal of Multimedia Information Retrieval*, *4*(1), 33–44.

van Son, R., Wesseling, W., Sanders, E., & van den Heuvel, H. (2008, may). The IFADV corpus: a free dialog video corpus. In *Proceedings of the sixth international conference on language resources and evaluation (lrec'08).* Marrakech, Morocco: European Language Resources Association (ELRA). (http://www.lrec-conf.org/proceedings/lrec2008/)

van Turennout, M., Hagoort, P., & Brown, C. M. (1997). Electrophysiological evidence on the time course of semantic and phonological processes in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 787–806. doi: 10.1037/0278-7393.23.4.787

Vedaldi, A., & Fulkerson, B. (2008). *VLFeat: An open and portable library of computer vision algorithms.* `http://www.vlfeat.org/`.

Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and vision computing*, *27*(12), 1743–1759.

Viola, P., Jones, M. J., & Snow, D. (2005). Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, *63*(2), 153–161.

Vogel, C., & Mamani Sanchez, L. (2012). Epistemic signals and emoticons affect kudos. In P. Baranyi (Ed.), *3rd ieee conference on cognitive infocommunications* (p. 517-522).

Von Borstel, A. I., Esquivel, J. Z., & Meyer, P. L. (2015, March 24). *Voice activity detection technologies, systems and methods employing the same.* Google Patents. (US Patent App. 14/666,525)

Wang, H., & Schmid, C. (2013). Action recognition with improved trajectories. In *Proceedings of the ieee international conference on computer vision* (pp. 3551–3558).

Wechsler, D. (1958). The measurement and appraisal of adult intelligence.

Weintraub, S., Mesulam, M.-M., & Kramer, L. (1981). Disturbances in prosody: A right-hemisphere contribution to language. *Archives of Neurology*, *38*(12), 742–744. doi: 10.1001/archneur.1981.00510120042004

Wernicke, S. (2010). *Lies, damned lies and statistics (about tedtalks)* [video]. http://go.ted.com/bDrm.

Whitaker, H. A. (1970). *A model for neurolinguistics*. University of Rochester.

Yumoto, E., Gould, W. J., & Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *The Journal of the Acoustical Society of America*, *71*(6), 1544–1550.

Zanna, M. P., & Rempel, J. K. (1988). *Attitudes: A new look at an old concept. s. 315-334 in: Bar-tal, d./kruglanski, aw (hrsg.), the social psychology of knowledge.* Cambridge: Cambridge University Press.

Zhang, C., Yin, P., Rui, Y., Cutler, R., Viola, P., Sun, X., ... Zhang, Z. (2008). Boosting-based multimodal speaker detection for distributed meeting videos. *Multimedia, IEEE Transactions on*, *10*(8), 1541–1552.

Zue, V., Seneff, S., Polifroni, J., Phillips, M., Pao, C., Goddeau, D., ... Brill, E. (1994). Pegasus: A spoken language interface for on-line air travel planning. In *Proceedings of the workshop on human language technology* (pp. 201–206).