Dysarthric Speech Emotion Classification

Lubna A. Alhinti

Supervisors: Dr Heidi Christensen and Dr Stuart Cunningham

Department of Computer Science
The University of Sheffield

This dissertation is submitted for the degree of
Doctor of Philosophy

Speech and Hearing Research Group July 2021

I would like to dedicate this thesis to
My husband Tamim - the source of my strength, inspiration, and happiness
The memory of my father Ahmad - who is always on my mind, forever in my heart
My mother Noura - who I can never thank enough
My children Hamad and Wafa - my reasons to keep going

Declaration

I hereby declare that I am the sole author of this thesis. The contents of this thesis are my original work and have not been submitted for any other degree or any other university. Some parts of the work presented in Chapters 3, 4, 5, 7 and 8 have been published in a conference proceedings, and journals as given below:

- Lubna Alhinti, Heidi Christensen, & Stuart Cunningham, "An exploratory survey questionnaire to understand what emotions are important and difficult to communicate for people with dysarthria and their methodology of communicating", International Journal of Psychological and Behavioral Sciences, 14(7), 187-191., 2020.
- Lubna Alhinti, Heidi Christensen, & Stuart Cunningham, "Acoustic Differences in Emotional Speech of People with Dysarthria", Speech communication, 126, 44-60., 2021.
- Lubna Alhinti, Stuart Cunningham, & Heidi Christensen, "Recognising Emotions in Dysarthric Speech Using Typical Speech Data", in Proceedings of Interspeech 2020.

Lubna A. Alhinti
July 2021

Acknowledgements

Undertaking this PhD has been a life-changing experience for me and it would not have been possible to do without the invaluable support I received from many people in both my personal and professional lives.

First and foremost, I would like express my deepest gratitude to my supervisors, Dr Heidi Christensen and Dr Stuart Cunningham for their endless support, guidance, and patience. They have both encouraged and supported my ideas. Words cannot adequately express my appreciation for your outstanding mentorship, kindness, and advice, which made me a much better student and researcher. I would also like to extend my gratitude to my panel members, Professor Roger K. Moore and Dr Dawn Walker for their helpful comments and suggestions regarding my work through each panel meeting. It has been an honour and privilege to work with this exceptional supervisory team. My deepest appreciation to Mr Simon Judge (Senior Clinical Scientist, University of Sheffield, Rehabilitation and AT Group. Barnsley Assistive Technology Team) for his valuable comments in the early stages of this research. I would also like to thank my examiners, Dr Stefan Goetze and Professor Gwen Van Nuffelen for the valuable comments and an intellectually stimulating and enjoyable viva.

My sincere thanks go to my colleagues in the Speech and Hearing Lab who have been always a continuous source of knowledge and friendship, especially Asif Jalal, Bahman Mirheidari, Dalia Attas, Eidah Alzahrani, Erfan Loweimi, Fatimah Alzahrani, Gerardo Roa Dabike, Hector Ramirez, Hussain Yusufali, Jack Deadman, Jisi Zhang, Lucy Skidmore, Madina Hasan, Mashael AlSaleh, Megan Thomas, Najwa Alghamdi Rabab Algadhy, Sadeen Alharbi, Yilin Pan, Zehai Tu, and Zhengjun Yue.

I also wish to express my appreciation to all of the participants for freely giving their time and effort and enabled this research to be possible.

I am hugely grateful to my family for always being there for me and for all their shown encouragement and love. My dear mother Noura, I can never thank you enough. I feel privileged to have been raised by you. Thank you for your endless love and prayers.

أمّي وإن طالَ الحديث بها فلا شِعرٌ يُوَفّيها ولا الأقلامُ

My beloved husband Tamim, I find it difficult to express my appreciation because it is so boundless. Your unconditional and genuine love, support and understanding underpinned my persistence in this journey and made the completion of this thesis possible. It means so much knowing that you are always there, whether they are times of happiness or times of sadness. Thank you for always being my source of comfort.

My prince Hamad and my princess Wafa, thank you for being such a bundle of joy and laughter. Seeing your eyes and hearing your giggles after a long day of work made a big difference.

I also feel incredibly blessed for the love and support that I continuously receive from my brothers Riyadh, Fahad, Khaled, and Nayef, sisters Najla and Ruba, sisters-in-law Fawziyah, Dhay, and Nouf, and brothers-in-law Moawia and Majed. My aunt Hessah, I have been deeply touched by your kindness. I owe special thanks to my nephews and nieces Bodour, Hadeel, Maha, Rakan, Hamad, Bader, Noura, Ahmad, Alanoud, Aljawharah, Riyadh, and Meshal. To all of you, thank you all for all your loving actions, warm welcomes whenever I am home, and incredible sense of humour.

My second family, who I am wholeheartedly blessed to have in my life. My father and mother-in-low Hamad and Wafa, thank you for being a source of constant and unconditional love and support. My sisters and brothers-in-law Khaled, Layla, Leena, Lubna, Hend, Fahad, and Noura, my gratitude to you is infinite and endless. Thank you for all your kindness, loving hearts, and warm welcomes.

Special thanks to my uncles Abdulrahman and Abdullah. Thank you for your kindness, generosity, and presence in my life over the years. Thank you my wonderful aunts Nawal and Jawaher and all my cousins, especially, Reem, Ruba, and Najla for always being there and for giving me so much love and support.

Thank you from the bottom of my heart to all my friends for their support and love and for keeping me strong throughout this journey.

I have been blessed with a very loving and supportive family and friends, who are a priceless gift.

Finally, I thankfully acknowledge the financial funding for this work from King Saud University and the Saudi Ministry of Education, to which I am grateful.

Abstract

Emotions play critical role in our lives. Communicating emotion is essential in building and maintaining relationships. Misunderstanding them or being unable to express them clearly may lead to problems in communication. People communicate their emotional state not just with the words they use, but also in how they say them. Changes in the rate of speech, energy and pitch all help to convey emotional states like 'angry', 'sad', and 'happy'.

People with dysarthria, the most common speech disorder, have reduced articulatory and phonatory control. This can affect the intelligibility of their speech. However, producing less intelligible speech may not be the only problem affecting their communication; having dysarthria may make it hard to convey emotions in their speech in a way that can be perceived and understood clearly by listeners. Recent research shows some promise on automatically recognising the *verbal* part of dysarthric speech. However, we know very little about the ability of people with dysarthria to convey their emotional state through *nonverbal* cues. This thesis investigates the ability of people with dysarthria, caused by cerebral palsy and Parkinson's disease, to communicate emotions in their speech, and the feasibility to automatically recognise these emotions. Recognising emotions from speech is by itself a challenging problem. In the case of disordered speech, this may exacerbate the problem more as the speakers often have less control of the signifying features.

A survey was designed and distributed to achieve a better understanding of different aspects related to emotion communication by people with dysarthria. A parallel multimodal, dysarthric and typical emotional speech database, which is a first of its kind, was collected and will be made publicly available. The ability of people with dysarthria to make systematic changes to their speech to convey their emotional state is investigated through analysing a set of potential acoustic features which are subsequently compared to those made by typical speakers. Their ability is also assessed perceptually and human listening performance on the collected database is reported. Two main approaches investigating the ability of automatically classifying emotions in dysarthric speech are followed: using models trained on dysarthric (speaker-dependent, matched) and typical (speaker-independent, unmatched) speech. The results of these investigations show it is possible to automatically recognise the emotional state of a speaker with dysarthria with a high degree of accuracy for some speakers.

The work in this thesis shows that despite some speakers with dysarthria having a more limited articulatory and prosodic control, they can make systematic changes in their speech that help in the communication of their emotions. These changes are shown to be successfully perceived by human listeners as well as by automatic emotion recognition models. These findings demonstrate the potential for improved, more expressive voice input communication aids.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

The quality of a person's life is affected by multiple factors. One of the fundamental ones is their ability to communicate in a meaningful way. People need to communicate to express their feelings and needs, share their thoughts, ask questions, socialise, etc. The spoken language is one of the most common ways to communicate and it plays a critical role in defining who we are. However, not everybody can communicate well using natural speech (Beukelman et al., 2005). For example, people with speech disorders may lose their ability to produce intelligible and audible speech sounds. As a result, people with speech disorders may also suffer from low self-esteem and may be hindered from achieving their goals in education, employment, and life in general (Beukelman et al., 2005; Major and O'brien, 2005; Walshe and Miller, 2011).

There are many types of speech disorders. The most common acquired speech disorder is dysarthria (Duffy, 2013; Walshe and Miller, 2011). It is defined as a neurological disorder that affects different aspects of speech production caused by weakness in the muscles responsible for speaking, miscoordination or inaccuracy of articulatory movements, or irregularity in the tone, steadiness, or speed. Dysarthric speech has been characterized prosodically as having monoloudness, monopitch, impaired ranges of F0, vocal intensity, and rate (Duffy, 2013). These prosodic impairments can significantly affect the intelligibility of the speech (De Bodt et al., 2002; Miller and Bachrach, 2017).

Human to human communication can be viewed, simply, as the process of producing and receiving messages. Messages are formulated using different signs and codes that are in turn interpreted by the receiver (Steinberg, 1995). The ability to communicate effectively relies on a number of aspects that are not only limited to the intelligibility of the spoken words. In particular, nonverbal cues play a critical role in the correct comprehension of the delivered message. For example, a sentence may have different meanings when it is spoken with different voice tones (Training, 2012). Thus, relying on communication with missing or

ambiguous nonverbal cues may lead to problems as people cannot express all their feelings using words alone. Nonverbal information conveys part of a person's feelings and emotions. Combining verbal and nonverbal communication results in a better perception of a speaker's feelings and emotions (Calero, 2005). To avoid any confusion of what is meant by verbal and nonverbal communication in the context of this thesis, from here after, verbal communication is referred to the spoken words while nonverbal communication is referred to how these words are spoken (voice characteristics).

People with dysarthria face many barriers to communicating effectively, and it is widely acknowledged that their lives can be impacted negatively (Martens et al., 2011, 2013; Pell et al., 2006). The impaction heavily depends on the individual, and on the severity of their dysarthria. In an exploration study of speakers' experience living with acquired chronic dysarthria, six dimensions were illustrated where dysarthria has a negative influence on their lives (Walshe and Miller, 2011). One of these dimensions is that having dysarthria changes people's way of communicating. These changes were reflected in different communication aspects including, but not limited to, the speaker's style and communication behavior, and the capacity to put feelings into their voices. Having less intelligible and monotonous voice may increase the potential of being socially withdrawn. In addition, having dysarthria can result in a number of negative feelings such as embarrassment, lack of confidence, and frustration as have been reported by the people with dysarthria who participated in the study (Walshe and Miller, 2011).

With the emergence of augmentative and alternative communication (AAC) technology, people with speech disabilities have been given a way to support their communication. AAC methods do not necessarily involve high and complex technology. It could be any method that is used to supplement someone's speech ranging all the way from pen and paper to more complex electronic communication aids. AAC users have social roles and therefore have desires, demands and expectations in their social participation that they would like to fulfill (Fried-Oken et al., 2012). However, current AAC technologies do have challenges in addressing all the communication interactions and needs. One of the most reported difficulties is their slow rate of communication (Mcnaughton and Bryen, 2007). Another is their lack of recognising and producing context and nonverbal information is one of the main challenges. The limitations of the current AAC technology also negatively affect social interactions and relationships as its design and function does not take into account how human factors affect human-to-human communication in different conditions and scenarios (Higginbotham et al., 2007). The authors in (Mcnaughton and Bryen, 2007) and (Higginbotham et al., 2007) have listed a number of recommendations for the development of an AAC technology that allows

full participation in society. Below is a summary of the main recommendations listed by the authors. AAC technology should:

- Allow its users to express their own thoughts and beliefs by providing quick and easy access to a wide range of vocabulary.
- Provide context-related vocabulary suitable for the ongoing communication activity.
- Provide a variety of input and output options to support different circumstances. For example, providing a screen and a built-in printer as two alternative output methods would be helpful in situations where display screens may not be viewable such as being under direct sunlight.
- Accurately recognise disordered speech and resolve the issues related to intelligibility and noise.
- Be usable in different environmental challenges and seating positions.
- Have an appealing design.
- Construct pragmatically appropriate utterances for speech output as a way to increase the emotional context of the technology. For example the single phrase "oh.", "oh?", and "oh!".
- Facilitate the access of distance communication technologies.
- Verify its user identity.
- Be connected to a different tools such as calendars and support access to different applications and internet functions.
- Be aware of its user's place, time, people and objects surrounding, know what the person wants to say and when to say it, i.e., understanding context.

In addition to the above points, an AAC device should be able to reflect the emotional state of its user. One of the recent advances in the AAC technology is the use of voice-driven AAC. The voice-input-voice-output communication aid (VIVOCA) which is a communication aid that recognises disordered speech and reproduces it in a synthesized voice is a form of voice driven communication aid. It helps not only in retaining the responsiveness, speed, and naturalness of the speech communication, but also opens new doors into recognising nonverbal information (Hawley et al., 2006, 2013; Therapy Box). Adding expressiveness to the synthesised voice will enhance the users' communication experience and social

relationship. Many people with dysarthria show a strong preference for using their own voice (also known as *residual voice*) when they communicate as it is the natural means of communication (Beukelman et al., 2007).

In this thesis, the ability of people with dysarthria due to cerebral palsy and Parkinson's disease (PD) to communicate emotions through suprasegmental and prosodic features and the degree to which these could be recognised automatically, as would be required by a VIVOCA system, will be investigated. The terms suprasegmental features and prosodic features are usually used interchangeably although they can have different meanings in classical metrical composition and Phonology. In this thesis these two terms are used interchangeably referring to the various features that reflect phonological properties of speech.

1.1 Motivation

There has been much effort concentrated on finding ways to automatically recognise the verbal part of dysarthric speech. Developing dysarthric automatic speech recognition (ASR) is considered to be a challenging task due to the intra- and inter-speaker variability in dysarthric speech and the difficulty of obtaining suitable data (Christensen et al., 2012, 2014; Darley et al., 1969a; Hawley et al., 2007; Ma et al., 2010; Wilson, 2000; Xiong et al., 2020; Yue et al., 2020b). Researchers have been working on developing and improving ASR systems and employing different techniques to overcome the challenges such as using transfer learning, adaptation techniques, speaker-dependent models, and data augmentation techniques (Christensen et al., 2012, 2013; Geng et al., 2020; Keshet, 2018; Mengistu and Rudzicz, 2011; Shor et al., 2019; Takashima et al., 2020; Xiong et al., 2018, 2020; Yue et al., 2020a).

The ability to recognise emotion is an important component in social interaction. People can express their emotions through a number of different modalities including speech, voice characteristics, facial expressions, and gestures. Automatic speech emotion recognition (SER) gained a lot of interest due to its increasingly important role in many fields including assistive technology and healthcare. There is a huge body of work done by researchers to automatically recognise emotions from human *typical* speech including recognising discrete emotions, recognising positive and negative emotions, and recognising nonverbal sounds such as cries and laughter (Dissanayake et al., 2020; Huang et al., 2020, 2019; Neumann and Vu, 2019; Zhou et al., 2020).

Emotions, to the best of our knowledge, have never been investigated in dysarthric speech caused by cerebral palsy and very little in dysarthic speech caused by PD. Pell et al. (2006) found that listeners faced a lot of difficulties recognising most of the emotions produced by

English speakers with dysarthria caused by PD in comparision to typical speakers. Although Martens et al. (2011) found no significant difference in recognising emotions by listeners between two groups of Dutch speakers where emotion identification were above random for both groups, it was shown that it was a difficult task regardless of the type of speech. On the other hand, a number of studies investigated the prosodic and phonatory features of dysarthric vocalisation. Despite having speech which is less intelligible, many studies show that even with the limited phonological and prosodic dimensions, many people with dysarthria have enough control to signal prosodic contrast on different tasks. For example, several studies investigated the ability of people with dysarthria caused by either cerebral palsy or PD to signal question-statement contrast in different languages (Liu et al., 2019; Ma et al., 2010; Ma and Hoffmann, 2010; Martens et al., 2011; Patel, 2002a, 2003; Pell et al., 2006). Other studies investigated the prosodic and acoustic characteristics such as the fundamental frequency (F0), intensity, and speech rate of speakers with dysarthria in comparison to typical speakers (Canter, 1963; Ghio et al., 2014; Gu et al., 2017; Hammen and Yorkston, 1996; Illes et al., 1988; J. Holmes et al., 2000; Patel, 2002b; Rusz et al., 2011). The ability to perform boundary marking by people with dysarthria caused by PD and contrastive stress by people with dysarthria caused by cerebral palsy and PD was examined (Martens et al., 2011; Patel and Campellone, 2009; Pell et al., 2006). In addition, the ability of people with dysarthria caused by PD to produce a perceptually detectable accent was also investigated and their strategy was compared to typical speakers (Ramos et al., 2020). Although speakers with dysarthria can differ from speaker with typical speech in their prosodic characteristics, most studies showed that they were able to perform different productive prosodic skills similar to healthy control speakers. High variability among speakers was observed in some of these studies such as their ability to vocally mark questions. These studies show that they may have enough control to convey emotions and show intentions in their speech which opens up new horizons for improving communication aids.

The overarching aim of this research is to improve AAC technology that is used by people with dysarthria, in particular to improve nonverbal emotion communication in voice-input AAC. Figure 1.1 shows a hypothesised speech-driven AAC device that would be ideal to achieve. As can be seen from the figure, the AAC device is composed of multiple subsystems. The dysarthric ASR is responsible for processing the linguistic part of the dysarthric speech and turns it into text. The dysarthric SER is responsible for processing the paralinguistic part of the dysarthric speech and recognising the emotional aspects of the speech. The context awareness ability is responsible for understanding the context of the communication situation including recognising who is present, what is the topic being discussed, and the place where the communication taking place. This will help in improving the output of the AAC. The

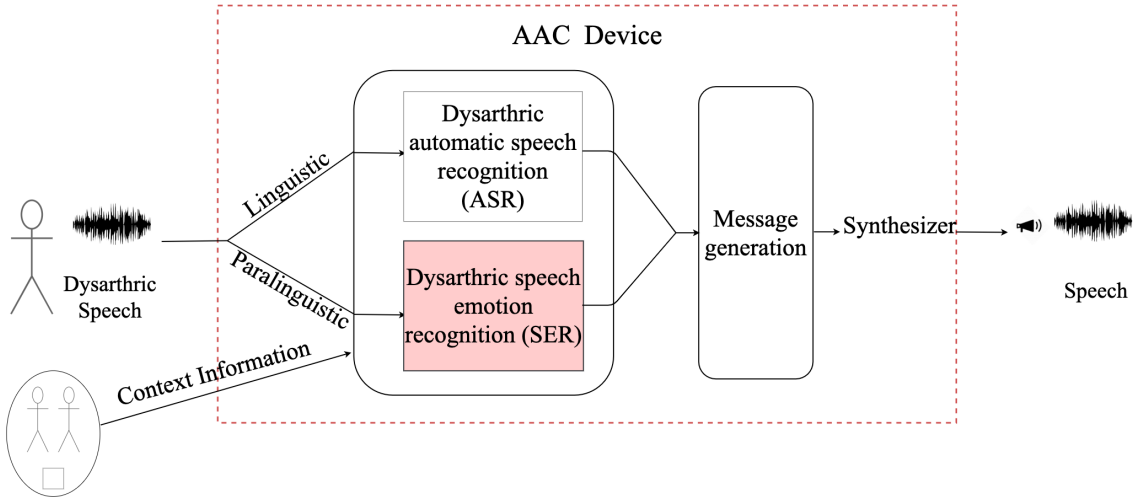


Fig. 1.1 Hypothesised speech-driven AAC device.

message generation is responsible for combining the output of the dysarthric ASR, dysarthric SER, and the context based information to regenerate the message. The output of the message generation subsystem will be fed into a synthesizer that is responsible for producing the expressive audible message.

This requires studying in more depth the ability of people with dysarthria due to cereberal palsy and PD to convey emotions in their speech in the first place and, if they were able, how to automatically recognise the emotional state of a speaker. Thus, this will be the focus of this thesis as highlighted in red in Figure 1.1.

1.2 Thesis aim

Many studies have focused on understanding the articulation errors in dysarthric speech and finding ways to automatically recognise their speech. In addition, a number of studies have investigated the prosodic control ability of speakers with dysarthria in different tasks. However, their ability to convey emotions in their speech through suprasegmental and prosodic features remains unexplored for speakers with dysarthria caused by cerebral palsy and needs more explorations for speakers with dysarthria caused by PD. This work is motivated by a long term goal to improve voice-input communication aids used by people with dysarthria in a way that makes it more sensitive to specific cues in the vocalization signal produced by the speaker with dysarthria and hence act more according to the speaker's intention. Therefore, this research will investigate:

- what emotions people with dysarthria are able to express in their speech;

- how people with different levels of dysarthria convey emotions in their speech, and how consistent a certain emotion is expressed among speakers and within the speaker him/herself;
- how their acoustic signalling of emotions might differ to that used by typical speakers;
- how well these emotions, if, at all, are recognised by human listeners;
- how these emotions can be automatically classified.

Answering the above questions is the first step towards the main goal of this research, which is **building an automatic dysarthric speech emotion classification model**. It is essential to first establish a good understanding of the above questions. In order to establish such an understanding, a database with a number of speakers is needed. Currently, there are no available dysarthric speech emotional databases.

By developing an automatic dysarthric speech emotion classification model and incorporating it in voice-input communication aids, it is hoped that the usefulness of these devices for people with dysarthria will be improved and their communication efficiency will be enhanced.

1.3 Contributions

The novel contributions of this thesis includes the following:

- **Investigating what emotions are important but difficult to communicate for people with dysarthria and their strategies for overcoming this.** To achieve a better understanding of the problem and to help in defining the scope of the research, an online survey was designed and distributed. Establishing a good understanding of the survey findings is very important before starting the process of collecting data and developing an automatic dysarthric speech emotion classification. This study was presented as a talk in
  - Workshop on Speech Processing for voice, speech, and hearing Disorders (WSPD), 2018, India.

  and published as:
  - L. Alhinti, H. Christensen, & S. Cunningham, "An exploratory survey questionnaire to understand what emotions are important and difficult to communicate for

people with dysarthria and their methodology of communicating", International Journal of Psychological and Behavioral Sciences, 14(7), 187-191., 2020.

- **Establishing a parallel multimodal emotional database of dysarthric speech and typical speech.** In automatic emotion classification for typical speech, researchers tend to use or record speech or multimodal databases to develop automatic emotion classification systems. The few available databases of dysarthric speech including the TORGO Database (Rudzicz et al., 2012), the Nemours database (Menendez-Pidal et al., 1996), and the Dutch dysarthric speech database (Yilmaz et al., 2016) are not emotional databases and therefore cannot serve the purpose of analysing emotions in dysarthric speech nor developing automatic emotion classification techniques. A database containing audio recordings of multiple people is therefore essential for this research because people with dysarthria may express emotions differently and therefore, results from a single speaker cannot be generalised. Thus, a proper database, called the Dysarthric Expressed Emotion Database (DEED), was collected. Both kinds of speech, typical and dysarthric, were recorded from a number of speakers in the same recording studio using the same settings. This allows a fair comparison and analysis to be made between the two types of speech. The database will be made publicly available for research purposes in the near future.

- **Analysing the acoustic differences in emotional speech of people with dysarthria using the collected database.** In order to investigate the ability of people with dysarthria caused by cerebral palsy and PD to make systematic changes to their speech to convey their emotional state, an acoustic analysis using some relevant features that are correlated with different vocal emotion expressions was conducted. The analysis was also carried out on the typical speech and the results were compared to see how the changes made by people with dysarthria are similar to those made by speakers with typical speech. The results show that people with dysarthria are able to make consistent and reliable changes to convey their emotions. The changes to the studied features appear similar to those of typical speakers, despite speakers with dysarthria having a more limited phonetic and prosodic control. This study was published as:

  - L. Alhinti, H. Christensen, & S. Cunningham "Acoustic Differences in Emotional Speech of People with Dysarthria", Speech Communication, 126, 44-60., 2021.

- **Assessing how well human listeners perceive emotions communicated by speakers with dysarthric speech.** This study aimed to evaluate the collected database subjectively. The evaluation included evaluating all speakers with dysarthria and part

of the typical speakers including all the 7 emotions. Although, the overall classification performance on typical speech was generally better than on dysarthric speech, participants in this study were able to classify emotions spoken by speakers with dysarthria even for the speaker who has severe dysarthria and highly unintelligible speech. These results indicated that speakers with dysarthria in this study were able to communicate different emotions effectively.

- **Developing an automatic dysarthric speech emotion classification model.** A model that can automatically classify emotions from dysarthric speech was developed using a speaker-dependent approach. The model was tested on all speakers with dysarthria in the collected database and the results of classifying 7 and 4 classes of emotions using categorical and dimensional approaches were presented. The model was also tested on the typical speech part of the database in order to allow a direct comparison of the classification results between the two types of speech. The results were very encouraging. Part of this study was published as:
  - L. Alhinti, S. Cunningham, & H. Christensen, "Recognising Emotions in Dysarthric Speech Using Typical Speech Data", in Proceedings of INTERSPEECH 2020.
- **Investigating approaches to automatically classifying emotions from dysarthric speech using models trained on typical speech data.** Due to the difficulty of collecting large databases of emotional dysarthric speech, which is needed when using deep learning techniques that may improve the model's performance, this study investigated to what extent it is feasible to automatically classify emotions from dysarthric speech using models trained on typical speech. Thus, a model that can automatically classify emotions from dysarthric speech using not only a speaker-independent approach but also trained on typical speech data rather than dysarthric speech was developed. The results were very encouraging. Part of this study was published as:
  - L. Alhinti, S. Cunningham, & H. Christensen, "Recognising Emotions in Dysarthric Speech Using Typical Speech Data", in Proceedings of Interspeech 2020.

1.4 Thesis structure

The remainder of this thesis is presented in Chapters 2 to 9. The content of these chapters can be summarised as follows:

- **Chapter 2: Background** This chapter presents a literature review covering human communication in general and emotion communication in particular. It also discusses

the notion of basic emotions and illustrates the applications of emotion classification in AAC. The chapter ends with defining dysarthria and discussing the nonverbal information reported in the literature as being present in dysarthric speech.

- **Chapter 3: Towards the Understanding of Communicating Emotions for People with Dysarthria** This chapter outlines the design, distribution, and results of the survey that is designed to achieve a better understanding of different aspects related to emotion communication by people with dysarthria.

- **Chapter 4: Disordered Speech Emotional Database Collection** This chapter starts with discussing the issues related to the development of emotional databases in general. Then, the chapter presents the collection of a parallel multimodal emotional database of dysarthric speech and typical speech, covering the corpus design, technical information related to the data capture, description of the data files, and data accessibility information.

- **Chapter 5: Acoustic Differences in Emotional Speech of People with Dysarthria** This chapter investigates the ability of people with dysarthria caused by cerebral palsy and PD to communicate emotions in their speech and to what extent they are similar to typical speakers in terms of the changes happening to the acoustic features. This is based on an extensive acoustic analysis of the collected database and the application of statistical models. The chapter illustrates the features used, methodology adopted, and results obtained from the analysis followed by an interpretation and discussion.

- **Chapter 6: Subjective Evaluation of DEED** This chapter presents the design and procedure followed in the evaluation of the collected database subjectively. It presents the results that show how well human listeners can recognise emotions in dysarthric speech. The evaluation process also includes evaluating a subset of the typical speech part in the collected database. The chapter ends with a discussion of the obtained results.

- **Chapter 7: Towards the Automatic Recognition of Emotion in Dysarthric Speech** This chapter starts with reviewing the popular speech emotion classification techniques. The development of an automatic dysarthric and typical speech emotion classification system is then presented. The chapter presents the baseline results of classifying 7 emotions and 4 emotions on the collected database using different classification approaches covering the details of the features and classifiers used when developing the model. The chapter ends with a discussion that includes a comparison between

the results obtained from the automatic emotion classification system and the ones obtained from the subjective evaluation presented in Chapter 5.

- **Chapter 8: Automatic Dysarthric Speech Emotion Recognition Using Models Trained on Typical Speech** This chapter presents the development of an automatic dysarthric speech emotion classification system that is trained on typical speech data covering the extracted features, classifiers, and classification approaches used. Classification results of classifying 4 classes of emotions, and pairs of emotions are presented and discussed.
- **Chapter 9: Conclusion** The final chapter contains the conclusion of the thesis and illustrates potential directions for future work.

Chapter 2

Background

2.1 Human communication

Communication has a vital role in our lives and a significant effect in almost every activity we do. Effective communication is central to spreading knowledge and building and maintaining relationships. Communication and exchanging information can be done through verbal and nonverbal forms. Keyton (2011) has defined communication as "The process of transmitting information and common understanding from one person to another". From this definition, we can conclude that understanding the message is an essential part of the communication process.

The model of a communication process provided by Shannon and Weaver (1949) presents the communication as a simple linear process. It is one of the most influential models of communication. Figure 2.1 shows Shannon and Weaver's schematic diagram of a communication system.

According to Shannon-Weaver's model, the five essential components of communication systems are (Shannon and Weaver, 1949):

- An information source: which produces the message or sequence of messages.
- A transmitter: which encodes the message into suitable signals that could be sent through the channel.
- A channel: which is the medium that is responsible for carrying the signals from the transmitter to the receiver.
- A receiver: which decodes the signals to rebuilt the message.
- The destination: is where the message is intended to arrive.

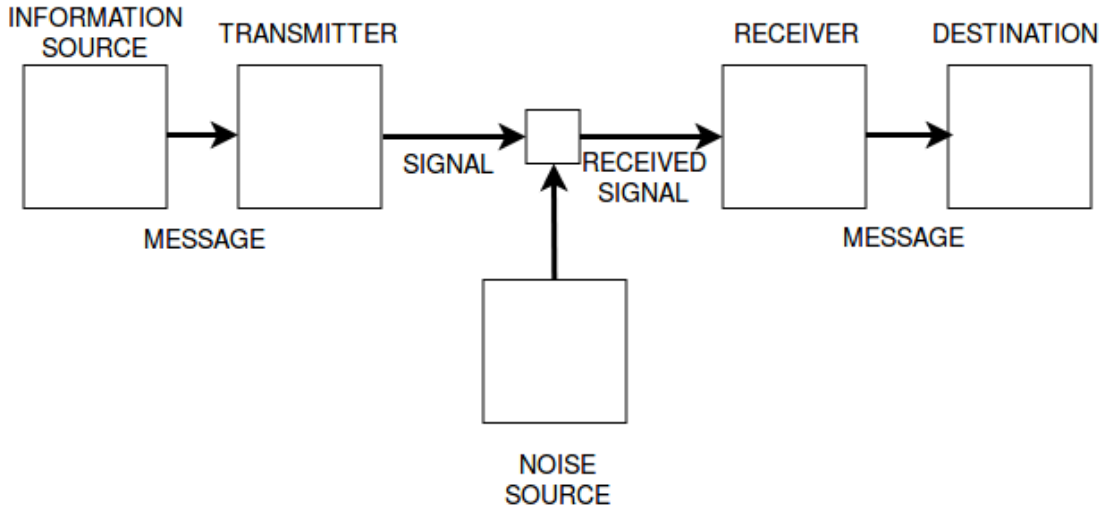


Fig. 2.1 Shannon-Weaver's model of the communication process (Shannon and Weaver, 1949).

Shannon and Weaver's model can be seen as an abstract level of the communication process and sets the common ground elements of communication and their relationship to one another. Obviously, what happens in real life communication is more complex than what is presented by this model. In most scenarios, there is no single source and single destination, both source and destination play both roles. Also, communication is not a simple linear process, it is composed of sending and receiving multiple signals in parallel (Foulger, 2004).

The transactional model of communication defined by Verderber (1990) presents all the different elements that play role in the communication process. Figure 2.2 shows Verderber's transactional model which illustrates the activation and engagement of both participants in the communication process rather than presenting them as taking turns. The two circles represents the participants. The message is in the center of each circle and can be communicated verbally or nonverbally. The message is surrounded by different factors that influence the meaning of the communicated message. As can be seen the participant's values, culture, background, occupation, sex, feelings, knowledge, and attitude all contribute to the meaning of the communicated message. The bar in the middle represents the medium where the message transmission and feedback are shown as a continuous process. The surrounding area shows the different kinds of noise that may disrupt the communication process. These noises includes semantic, internal, and external noises. Having a speech disorder, such as dysarthira, poses a challenge in Verderber's transactional model as it can affect the message encoding process. Depending on the dysarthria severity, the message encoding process could

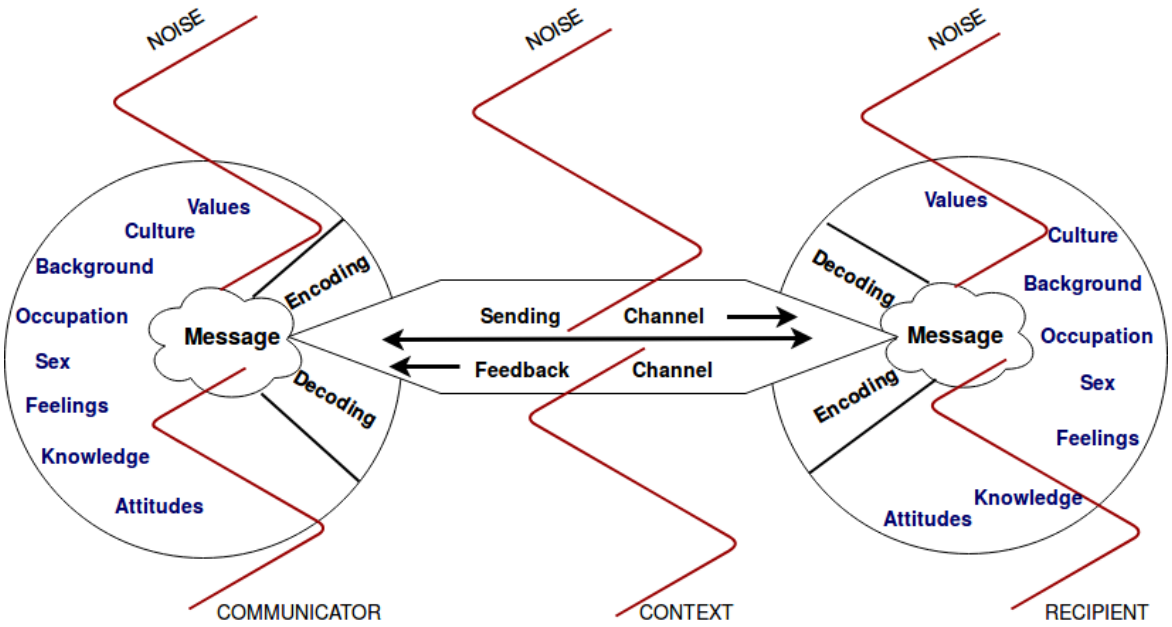


Fig. 2.2 Transactional model of communication (Verderber, 1990).

be affected by the sender's inability to put proper meaning to the formulated message all the way to the inability to produce intelligible message at all.

2.1.1 Noise in communication

The purpose of communication is to send a message and the goal is to achieve an understanding of this message. A successful and effective communication is determined by the understanding of the communicated message. One more component in Shannon-Weaver's model is the presence of noise during the process of transmitting and receiving the signals. Any interference with the transmitted message is considered a noise (Lunenburg, 2010). Therefore, noise is not necessarily associated with technical issues, it includes any barrier to an effective message delivery. Theses barriers lead to miscommunication or ineffective communication. Miscommunication may cause minor effects such as workflow delays, uncertainty that may lead to conflicts or it may result in very serious problems such as death especially when it happens in medical or airplane cockpit situations.

2.1.2 Types of communication barriers

There are different types of communication noise that place barriers to effective communication. Lunenburg (2010) has described three types of communication noise:

- *Physical barriers*: which include any physical distraction that interfere with the communicated message. Walls, doors, distance, drop-in visitors, and background visitors are all examples of physical barriers.
- *Semantic barriers*: are related to the chosen words and the meaning of those words. A word may hold many meanings to different people and therefore, a message may be interpreted differently by the receiver according to his or her own understanding of the meaning of that word. Technical language also has a part in raising a semantic barrier. The use of jargon may block the understanding of the sent message if the receiver is unfamiliar with the used terminology.
- *Psychosocial barriers*: which relate to the individual differences among others. Field of experience, filtering, and psychological distance are significant concepts related to the psychosocial barriers. The understanding of the sent message becomes difficult when the sender and receiver come from different backgrounds, perceptions, and expectations. The influence of the receiver's emotions, needs, and interests on what is heard or seen, which is known as filtering, affect the real meaning of the sent message. Also, the gap between the sender and receiver known as the psychological distance affect the effective delivery of the message.

A very important barrier to add to the above is the disability barrier that may affect person's speaking, hearing, vision or cognition ability. As discussed in Section 2.1, people with dysarthria for example, due to their disability, may lack the production of common acknowledged cues for communicating emotions in their speech, which stands as a barrier to an effective communication. Having physical barriers in addition to a disability barrier may make the communication even harder. Poor lighting, for example may make it more difficult for a listener to lip read a speaker with dysarthria. Background noise may also negatively affect the speaker's intelligibility. For those who rely on an AAC device in their communication, Figure 2.3 represents the same transactional model in Figure 2.2 but with adding the AAC device as part of the communication process. The use of current AAC devices add to the list of barriers to an effective communication. In addition, the attitudes of people interacting with speakers with speech disoreder, such as dysarhtria, may serve as a significant barrier to an effective communication (Connaghan et al., 2020; Howe, 2008). Unfavourable and inaccurate attitudes and impressions may lead to negative stereotypes, which hinder opportunities in relationships, employment, and education (Major and O'brien, 2005). These negative stereotypes has been found for people with different communication disorders including stuttering, voice disorders, hearing impairment, and dysarthria (Boyle, 2017; Dickson et al., 2008; Freeman, 2018; Jaywant and D PELL, 2010; Nagle et al., 2015).

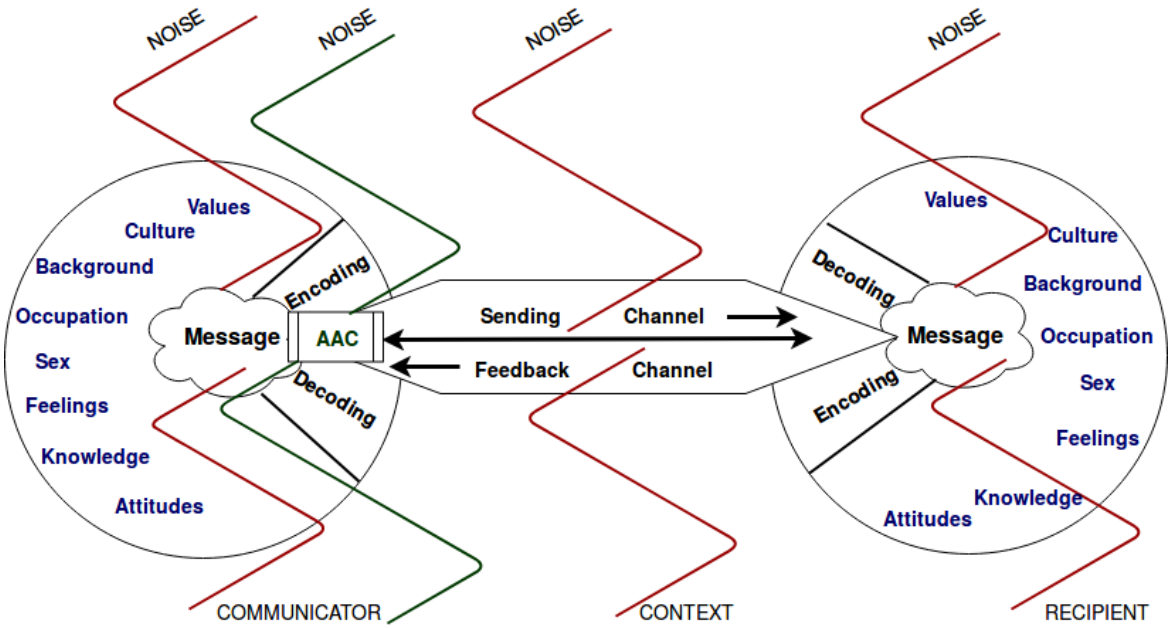


Fig. 2.3 Transactional model of communication with AAC ((Verderber, 1990) with modification).

2.2 Disordered speech

As discussed in Section 2.1, effective communication is the essence of a successful human interaction. Communication in its different modalities is based on four elements: sending a message also known as encoding, transition medium, receiving a message or decoding, and feedback. The way we communicate and the signals we send through communication have an impact on how other people respond and feel. This impact may be either negative or positive depending on the style of communication used.

Speech is the main and most powerful way of communication as it carries information beyond the spoken words. Examples of other information that are carried out by speech include: speaker's emotional state, age, health condition, paralinguistic information such as pitch and tone, and speaker's identity. Effective oral communication relies on a number of factors. The clarity of speech is one of the main aspects of effective oral communication. The pitch, speed, and volume all have an impact on the effectiveness of the oral communication.

The speech production, however, may be affected by different speech disorders. There are many different causes of speech disorders, some of the causes are known while others are still unknown. Some of the known causes includes neurological disorders, physical impairments, hearing loss, brain injury, and intellectual disability. Dysarthria, aphasia, apraxia of speech, cluttering, stuttering, and mispronunciation are all examples of speech disorders that affect the effectiveness of oral communication.

This section talks in more depth about dysarthria and the acoustic features of dysarthric speech.

2.2.1 Dysarthria

Motor speech disorders (MSDs) are speech disorders that occur as a result of damage in the nervous system (Duffy, 2013). The most common acquired speech disorder is Dysarthria which is one of the MSDs (Duffy, 2013; Walshe and Miller, 2011). According to Duffy (2013), dysarthria can be defined as a neurological disorder that affects different aspects of the speech production caused by weakness in the muscles responsible for speaking, miscoordination or inaccuracy of articulatory movements, or irregularity in the tone, steadiness, or speed. There are different types of dysarthria where each type reflects a different abnormality. Darley et al. (1969b) outline the following five different types of dysarthria in their study of seven neurological disorder groups:

Spastic dysarthria A result of an impairment in the upper motor neuron system that cause a lesion in the pyramidal and extrapyramidal systems known as pseudobulbar palsy. Along other disorders, this condition causes spastic dysarthria. The main prominent characteristics of spastic dysarthria ordered by their prominancy are: imprecise consonants, monopitch, reduced stress, harsh voice, monoloudness, low pitch, slow rate, hypernasality, strained-strangled voice, and phrases short. With imprecise consonants being the more prominent characteristic in this type.

Hypokinetic dysarthria A result of an impairment in the extrapyramidal system that cause movement reduction. This type of dysarthria is found in people who have Parkinson's disease (PD). The main prominent characteristics of hypokinetic dysarthria ordered by their prominancy are: monopitch, reduces stress, monoloudness, imprecise consonants, inappropriate silence, short rushes, harsh voice, and continuous breathy voice. With monopitch being the more prominent characteristic in this type. The severity level of monopitch and monoloudness in this type is greater than the above mentioned types.

Hyperkinetic dysarthrias A result of an impairment in the extrapyramidal system that cause increase in the movement. This type of dysarthria is found in people who have dystonia and chorea. The main prominent characteristics of hyperkinetic dysarthria ordered by their prominancy are: imprecise consonants, vowels distorted, harsh voice, irregular articulatory breakdown, strained, strangled voice, monopitch, and monoloudness. With imprecise consonants being the more prominent characteristic in this type.

Flaccid dysarthria A result of an impairment in the lower motor neuron system or peripheral nervous system. This type of dysarthria is found in people who have bulbar palsy. The main prominent characteristics of flaccid dysarthria ordered by their prominancy are: hypernasality, imprecise consonants, continuous breathiness of voice, and monopitch, with hypernasality being the more prominent characteristic in this type.

Mixed dysarthria A result of progressive degeneration of the upper and lower neuron system. This type of dysarthria is found in people who have amyotrophic lateral sclerosis. As the name implies, this type of dysarthria have mixed characteristics of flaccid and spastic dysarthria. The main prominent characteristics of mixed dysarthria ordered by their prominancy are: imprecise consonants, hypernasality, harsh voice, slow rate, monopitch, phrases short, vowels distorted, low pitch, monoloudness, excess and equal stress, and intervals prolonged. With imprecise consonants being the more prominent characteristic in this type and have more severity than it has in flaccid and spastic dysarthria.

Ataxic dysarthria A result of a damage in the cerebellum that cause movement inaccuracy. This type of dysarthria is found in people who have cerebellar disorders. The main prominent characteristics of ataxic dysarthria ordered by their prominancy are: excess and equal stress, irregular articulatory breakdown, vowels distorted, harsh voice, and imprecise consonants. With excess and equal stress being the more prominent characteristic in this type.

Although that most of the characteristics listed above overlap among different types of dysarthria, they vary in their level of severity. Different types of dysarthria can be distinguished. Each one of them sounds differently depending on its speech and voice dimensions (Darley et al., 1969b).

2.2.2 Human communication for people with dysarthria

As can be seen from the above, dysarthria interfere with articulation, respiration, phonation, and resonance. Therefore, dysarthria changes the way people communicate. According to Walshe and Miller (2011) and Hartelius et al. (2008) findings from their exploratory study of speakers' experiences of living with dysarthria, dysarthria changes the speaker's style, communication behavior and the capacity to put feelings into their voices. The tendency to make conversations short, hesitation to participate in 'small talk', avoidance of deep discussions, let other people speak for them, avoidance of situations were they are expected

to talk, and reluctant to talk in cases were they think they might not be understood are all examples of how dysarthria changes the speaker's style and communication behavior.

It was also found that some people with dysarthria rely on some strategies to help them be better understood. These startegies include drinking water, trying to speak distinctly, trying to slow down their speech rate, practicing words and sentences, providing alphabet and topic cues, writing some words down, and word/message modification (Hustad et al., 2003; Walshe and Miller, 2011).

In terms of how easy and difficult a communication situation is for a speaker with dysarthria, it was found that some speakers with dysarthria found it easier to communicate with family members, however, it was not the case for everyone. Talking to strangers, though, was generally agreed to be the hardest (Hartelius et al., 2008; Walshe and Miller, 2011). There was also no agreement on the level of difficulty related to the size of the audience, while some people with dysarthria found it easier to communicate in a one-to-one situations, others found it harder or feel no difference. Emotions were also reported to have an effect on the difficulty of communication, where people with dysarthria often indicated that they find it more difficult to speak when being 'angry' or 'sad' (Hartelius et al., 2008). Although the severity of dysarthria can affect the perception of communication difficulties, it was found that it was highly subjective (Hartelius et al., 2008; Walshe and Miller, 2011; Yorkston et al., 1994).

Nevertheless, people with dysarthria show a strong preference for using their own voice (also known as residual voice) when they communicate as opposed to using other AAC aids (Beukelman et al., 2007). However, having limited phonological and prosody dimensions may not only result in producing unintelligible speech, but it may make it hard to convey emotions in the speech in a way that can be captured and understood clearly and easily by recipients. This may increase the potential of them being socially withdrawn.

2.2.3 Paralinguistic information in dysarthric speech

The literature includes a number of studies that compares acoustic differences between dysarthric and typical speech (Ansel and Kent, 1992; Bunton and Weismer, 2001; Connaghan and Patel, 2017; Darley et al., 1975; Kempler and Van Lancker, 2002; Le Dorze et al., 1994; Liss and Weismer, 1994; Martens et al., 2011; Patel, 2004; Pell et al., 2006; Platt et al., 1980; Rosen et al., 2006). Despite having speech which is less intelligible, many studies show that even with the limited phonological and prosodic dimensions, many people with dysarthria have enough control to signal prosodic contrast on different tasks. Below is a highlight of some of the conducted studies that investigated the paralinguistic precisely acoustic, prosodic

and phonatory features of dysarthric vocalization on speakers with spastic and hypokinetic dysarthria caused by cerebral palsy and PD, repectively along with their main findings.

Studies on speakers with spastic dysarthria

Patel (2002a) conducted an experiment with eight speakers with severe spastic dysarthria caused by cerebral palsy to investigate the use of pitch contour and syllable duration for phrase-level productions and whether there exist a vocal control to signal a linguistic contrast. The stimulus list contained ten unique phrases that were composed of three syllables where all the phrases were produced five times as questions and five times as statements. This resulted in having 100 recordings for each speaker. Forty-eight normal-hearing English monolingual speakers who were unfamiliar with dysarthric speech and do not know the stimulus materials were selected as listeners to classify the production of statements and questions. Using prosodic cues only, the listeners were able to achieve 87% classification accuracy. In the same study, fundamental frequency (F0) and duration cues were systematically removed to determine their importance in classifying dysarthric utterances. The effect of removing the pitch contour reduced listeners' accuracy scores by 32% while removing the durational cues reduced the accuracy by 8%. This implied that syllable duration is less informative than F0 contour to the listeners. In a following study, Patel (2003) replicated the above experiment with eight healthy controls to investigate the strategies used by the speakers with dysarthria due to cerebral palsy to signal the question-statement contrast. Precisely to find out what are the acoustic cues they use to do so and whether they are using different strategies to signal the contrast compared to the healthy controls speakers. To understand the acoustic consistencies among speakers with dysarthria and the listener perception judgments, logistic regression analyses were used. It was found that the average of the fundamental frequency of the first syllable, the duration of the second and third syllable, the peak and slope of the fundamental frequency of the third syllable and the average and slope intensity of the third syllable were significant predictors of contrasting questions and statements by dysarthric speakers. The analysis showed that the third syllable differs mainly in questions versus statements in the dysarthric speech. The results indicated that F0, duration, and intensity were all used by speakers with dysarthria to signal the question-statement contrast while F0 and duration cues were primarily used by health control speakers.

Patel (2002b) also examined the ability of people with severe dysarthria caused by cerebral palsy to convey information using duration and pitch cues. The study was conducted on a group of eight speakers with severe dysarthria caused by cerebral palsy. The results indicated that during sustained vowel production, all the eight speakers had a consistent control over duration while producing the vowel /a/. They were all able to produce at least

three different levels of vowel duration. Their ability to control the pitch, however, varied. All speakers were able to produce at least two different levels of F0.

In addition, the ability to perform contrastive stress by people with dysarthria caused by cerebral palsy was examined. Patel (2004) examined the ability of three speakers with severe dysarthria caused by cerebral palsy to produce contrastive stress placed on different positions of three, four-word phrases. Their ability was compared to three healthy control speakers. Based on the performed acoustic analysis, the author found that all speakers from both groups successfully managed to produce contrastive stress at all the different positions in the utterance. Although both groups relied on increased levels of F0, duration, and intensity while marking the stressed word, it was found that speakers with dysarthria depended more on intensity. In a follow-up study, the ability of twelve listeners, who were unfamiliar with dysarthric speech, to identify the stress position within an utterance was examined. The listeners achieved a high stress identification accuracy of 78%-97% in the dysarthric production. It was found that to identify stress, listeners tended to rely on F0 cues more than duration and intensity (Patel and Watkins, 2007). In a more in depth investigation of the study presented in (Patel, 2004), Patel and Campellone (2009) assessed the ability of twelve speakers with different levels of dysarthria caused by cerebral palsy to produce contrastive stress and compared it to twelve healthy control speakers using acoustic and perceptual experiments. It was found that although speakers with dysarthria had reduced ranges of F0 and intensity, both groups used F0, duration, and intensity cues to mark contrastive stress. However, speakers with dysarthria relied much more on duration. Listeners who were unfamiliar with dysarthric speech achieved high stress identification accuracy in both groups, ($> 93\%$) on dysarthric speech and ($> 97\%$) on typical speech.

Studies on speakers with hypokinetic dysarthria

Using prosodic characteristics, the ability of people with hypokinetic dysarthria caused by PD to signal question-statement contrast was also investigated in Mandarin, German, and Cantonese languages and compared to healthy control speakers (Liu et al., 2019; Ma et al., 2010; Ma and Hoffmann, 2010). Ma and Hoffmann (2010) investigated the ability of twenty-four German speakers with mild or moderate hypokinetic dysarthria caused by PD to produce imperatives, questions, and statements compared to twelve typical speakers. Their intonation contrast ability was measured using an acoustic analysis of F0 statistics, speech rate, and intensity range and envelop. Although, the results showed significant differences in some of these features, speakers with PD were able to mark contrasts between intonations similar to typical speakers. Using acoustic and perceptual methods, the ability of fourteen Cantonese speakers with mild, mild-to-moderate and moderate hypokinetic dysarthria caused

by PD to signal question-statement contrast was investigated (Ma et al., 2010). Twenty normal hearing listeners who are native Cantonese speakers and had limited familiarity with dysarthric speech identified the stimuli as either questions or statements. In addition, F0, duration, and intensity variations were analysed. The results showed that statements were identified with high accuracy by listeners while less accuracy was obtained in questions identification. High variability in the ability of speakers with PD to signal question-statement contrast were observed. Similar acoustic cues were found to be used by speakers who were able to signal the contrast in comparison to typical speakers. Another study investigated the ability of twenty native Mandarin speakers with PD to signal question-statement contrast compared to twenty typical speakers (Liu et al., 2019). An analysis of F0, duration, and intensity was conducted. The results indicated the inability of speakers with PD to signal the contrast.

In addition to the ability of people with dysarthria caused by PD to signal question-statement contrast, their ability to perform phonemic and contrastive stress, and express six different emotions were investigated and compared to healthy control speakers (Pell et al., 2006). The study was performed on twenty-one English speakers with dysarthria with high intelligible speech and twelve healthy control speakers. Twenty listeners who are native English speakers participated in the four perceptual identification tasks. For the phonemic and contrastive stress identification tasks, listeners performed significantly worse in speakers with dysarthria. Although listeners were able to identify statements produced by speakers with dysarthria, it was not the case when identifying questions where they faced much more trouble. In terms of emotion identification task, it was reported that speakers with dysarthria often sounded as if they were 'sad'. In comparison to the identification on healthy control speakers, 'sad' was the only emotion that was recognised with similar accuracy. Much more difficulty was reported when recognising other emotional expressions produced by speakers with dysarthria. It was found that 'anger', 'happy', 'disgust', and 'surprised' were frequently perceived as 'neutral'. Similarly, Martens et al. (2011) assessed the ability of people with dysarthria caused by PD when performing four prosodic functions: boundary marking, contrastive stress (focus), intonation (questions vs statements), and emotion expression (anger, happy, sad, and neutral). The study included samples collected from eighteen Dutch-speaking people with dysarthria with different levels of severity and eighteen healthy control speakers using two different approaches: reading and imitation. Three professional listeners experienced with dysarthric speech participated in the evaluation of the speech material. It was found that speakers with dysarthria performed significantly worse than healthy control speakers on imitation when performing boundary marking, contrastive stress, and intonation tasks. It was also found that on the contrastive stress task, speakers with moderate and severe

dysarthria performed significantly worse on imitation compared to reading. In contrast to the results reported in (Pell et al., 2006), there was no significant difference found between the group of speakers when expressing emotions. In general, for both groups, the identification results were relatively high (78.6% – 98.5%) for all the prosodic functions except for emotion expression where the results ranged were (47.7% - 63.6%).

The ability of people with dysarthria caused by PD to produce perceptually detectable accent in Dutch was also investigated perceptually and objectively and their strategy was compared to typical speakers (Ramos et al., 2020). Three expert listeners perceptually judged speech samples from fifty speakers with dysarthria with different severities and thirty healthy control speakers. An acoustic analysis of features related to F0, duration, and intensity was also performed. Using linear discriminant analysis (LDA), an automatic classification and statistical analysis was performed. High classification accuracy of accented versus unaccented syllables was reported for both groups of speakers. Although most features were used by both groups to produce accent, a significant difference was found between the two groups of speakers. The study showed that speakers with dysarthria still have residual control of F0. It was also found that they tended to rely on variations of intensity and/or duration as a compensatory mechanism for their lack to control F0.

The prosodic and acoustic characteristics such as F0, speech rate, and intensity of people with dysrathria caused by PD were also investigated and compared to healthy control speakers (Canter, 1963; Ghio et al., 2014; Hammen and Yorkston, 1996; Illes et al., 1988; J. Holmes et al., 2000; Rusz et al., 2011). There is a lack of consistency in some of the reported results as some studies show differences in some of these features between the two groups while other reported no differences. For example, Canter (1963), Illes et al. (1988) and Ma and Hoffmann (2010) showed that speakers with PD had high average of F0 in comparison to typical speakers while Ghio et al. (2014) and Gu et al. (2017) reported no difference between the two groups in the average of F0. In terms of intensity, similar levels of intensity was found by Canter (1963) while lower levels of intensity was found in speakers with PD in comparison to typical speakers by Illes et al. (1988) and J. Holmes et al. (2000). Similarly, inconclusive findings were reported with regards to speech rate. Hammen and Yorkston (1996) found that speakers with PD had faster speech rate than typical speakers, Canter (1963), Gu et al. (2017), and De Letter et al. (2006) found no differences in speech rate between the two groups, and Rusz et al. (2011) found that speakers with PD had slower speech rate in comparison to typical speakers. This may indicate the complexity of PD and the high variability in the acoustic characteristics of people with dysarthria. It is important, however, to highlight that there are some methodological and language differences in these studies, which could have an effect on some of these inconsistent findings. Although most of

these findings were based on performing acoustic analysis, there are few where perceptual judgment were included.

Overall findings discussion

It is important to note that the majority of these studies were investigating the linguistic prosodic skills of people with dysarthria such as questions vs statements rather than their ability to express emotions. Also, it is important to note that some of these studies were carried out on small samples, in different languages that have different characteristics (for example, tone and non tone languages), and/or on speakers with different severity levels of dysarthria. Thus, any attempt of generalizing the findings is difficult, however, there are several consistent points. First, the majority of these studies show that even with having different prosodic characteristic to typical speakers, speakers with dysarthria caused by cerebral palsy and PD were able to signal most prosodic functions. Second, high inter-speaker variability among speakers with dysarthria was observed in some of these studies related to signaling the question-statement contrast. Third, there is a lack of consistency in some of the reported results on the prosodic and acoustic characteristics between speakers with dysarthria and typical speakers. Nevertheless, these studies show that people with dysarthria may have enough control of prosodic and phonatory features that allows them to communicate emotions, convey intentions, and obtain listeners' attention. This potential control opens many new doors of investigations related to the kinds of paralinguistic information people with dysarthria can communicate and their consistency of doing so, i.e., the intra- and inter-speakers variability when communicating specific information.

2.3 Communication of emotions

Feelings and emotions play critical and very important role in our lives. Without these emotions, none of the events that we experience would have any value or meaning to us. Therefore, it has been a topic of interest to many researchers and psychologists for a long time. In fact, it is one of the most confused and debatable topics. There is no common agreement of the definition of emotion in the scientific and research world. According to Plutchik (2001), emotion has been given more than ninety definition in the twentieth century. This may be a result of the many different disciplines that are involved. A very simple and useful definition of emotion is the one that is defined by Levenson (1994) as:

> "Emotions are short-lived psychological-physiological phenomena that represent efficient modes of adaptation to changing environmental demands."

A central and a widely accepted assumption in a social functional approach to emotion is that human beings are social by nature (Keltner and Haidt, 1999). Sharing emotions are assumed to be very important in forming and maintaining relations as well as restoring and strengthening social bonds that leads to a beneficial relationships (Keltner and Kring, 1998). In this Section, we will look into more details on the functional role of emotions and the modalities of expressing emotions.

2.3.1 Emotions' functional role

Darwin (1872) in his book "The Expression of The Emotions in Man and Animals", states that emotional expression carries information about the internal state of the individual. Understanding these emotional expressions and the information that they convey helps in coordinating social interactions. In this Section, we will draw the attention to the functional role of emotion in three different levels, namely, intrapersonal, interpersonal, and the cultural functions of emotion.

Intrapersonal functions of emotion

A primary function of emotion is to provide information (Clore, 1994a). This internal information affects a number of cognitive functions such as judgement, reasoning, and decision making. It also helps in deciding the urgency and significance of an event, and re-prioritising processes. Emotions play a critical role in determining human development, personality, well being, and can affect decision making. (Clore, 1994a,b; Fox et al., 2018).

Emotions occured to solve specific adaptation problems and demands (Ekman, 1992a; Tooby and Cosmides, 2008). The emotions that we feel play a central role in dealing with the immediate challenges and threats that we may face. To illustrate this adaptation, let us consider the human body. A balanced and healthy body is a one that maintains its hormones and homeostatic levels. However, sometimes to deal with a certain situation or threat, the body needs to get itself to an unbalanced state. Some of our emotions are the trigger to this needed body state shift. For example, the emotion of fear leads to a number of abnormalities in our bodies such as a rise in the heart beat rate. These abnormalities prepare the body for an escape response. Likewise,the emotion of disgust helps in developing a disease-avoidance behaviour. For example, the body will reject spoiled food by vomiting it out which is the opposite to the normal way where the body ingest food (Levenson, 1999; Tooby and Cosmides, 2008). On the other hand, positive emotions such as happiness, amusement and contentment will help the body to return to its balanced state after the occurrence and effect

of a negative emotion. Thus, these kinds of emotions provide means for soothing (Levenson, 1999).

Interpersonal functions of emotion

Emotions are considered as essential elements in our everyday social interaction with others. Greetings rituals, caregiving, and discourse are all forms of social interaction that involves one or more emotions. The emotion information that we send affect others' responses, actions and feelings. Keltner and Haidt (1999) present the three main points that were discussed by the theorists in this field about the interpersonal functions of emotion. First, the communication of emotion helps the receiver understands the sender's intention and emotion and also reveal information about surrounding objects and events and therefore leads to a coordinated interaction. That is, it serves as a window to understand the sender's emotional state. Second, expressing certain emotions triggers reciprocal emotions responses in others. Distress, for example, evoke the response of sympathy in others. Third, communicating emotions regulate others' social interaction and behaviour by encouragement or discouragement.

Cultural functions of emotion

There are many different and complex definitions of culture, each of which is defined with a certain perspective. For this research, a relative definition would be the one that is defined by Matsumoto (1996) as:

> "The set of attitudes, values, beliefs, and behaviors shared by a group of people, but different for each individual, communicated from one generation to the next."

Thus, the culture provides its individuals a kind of information system and social order to follow which helps in setting rules, boundaries, and regulations to insure smooth and healthy social interactions. An observer may infer and learn specific culture norms and values when observing emotion expressions in socialization practices. This learning process helps the individual to fit and act flawlessly within this culture (Hareli et al., 2013). The culture also plays a critical role in defining what, when, and where emotions could be expressed and who can express these emotions. For example, in some cultures, women are expected to be emotionally spontaneous inside the house while they are expected to have their emotion regulated in public (Gordon, 1990).

Many studies found that gender has an effect on the level of emotion expression. Women tend to express emotions more than men do with the exception of the 'anger' emotion (Kashdan et al., 2009).

2.3.2 Modalities of expressing emotions

There are a number of different modalities and channels people use to express their emotions. Emotions can be communicated through verbal, nonverbal, and vocal channels. People may use a single channel or multiple channels at the same time when communicating emotion.

Verbal communication of emotions

Spoken language is one of the main and direct way to describe emotions. People tend to share their emotional experiences by talking about it to other people. The reasons and benefits behind this differ from case to another. Seeking for emotion recovery and gaining social support are two examples of the reasons why people talk about their emotions explicitly (Kashdan et al., 2009; Zech and Rimé, 2005).

Nonverbal communication of emotions

Emotions can be communicated nonverbally through different modalities. Everyday, people tend to receive and send a lot of emotional signals through nonverbal behaviours. Facial expression is one of the main modalities in emotion communication and gets the attention of many researchers. It carries the greatest nonverbally communicated amount of information and yet it is universal (Matsumoto et al., 2013). Therefore, facial expressions could be interpreted no matter what language the person is speaking or what culture and backgrounds he/she comes from. Body posture reflects the person's emotional state. The way people sit, move, and walk all encode emotional information that could be decoded by others (Lhommet and Marsella, 2014). Gestures also convey different emotional information that is communicated consciously or unconsciously. Self scratching, playing with the hair, interlocking fingers, and hiding the face are examples of gestures that communicate certain emotions (Lhommet and Marsella, 2014). Gaze is one of the powerful nonverbal behaviours that is used to reveal an emotional intensity level (Kimble and Olszewski, 1980). Emotions can also be communicated through touch. Thompson and Hampton (2011) showed in their study that universal and prosocial emotions were successfully communicated through touch by both strangers and romantic couples. However, self-focused emotions of envy and pride were only successfully communicated by romantic couples.

People with dysarthria vary in their facial expressions, gestural and body movement abilities depending on the etiology and severity of dysarthria. For example, a reduction in facial expressiveness ability, also known as *facial masking*, is commonly observed in people with PD. This result in them being consistently interpreted by others as 'depressed', 'anxious', 'sad', and/or 'suspicious', which negatively affect future interaction with others and

the ability to communicate successfully (Katsikitis and Pilowsky, 1991; Pentland et al., 1987; Pitcairn et al., 1990a,b; Smith et al., 1996; Spielman et al., 2003). In terms of gesticulations, a number of studies showed the positive influence of using gestures on the intelligibility of limited number of speakers with dysarthria and mild motor impairments (Garcia and Cannito, 1996a; Garcia and Dagenais, 1998; Garcia and Cannito, 1996b). However, many people with dysarthria have reduced or affected motor capabilities which decrease or affect the use of gestures (Duffy, 2013).

Vocal communication of emotion

The human voice is a powerful and complicated mean of communication. Its power comes from the fact that it transmits many signals and information beyond the spoken words. Biological, psychological, social and emotional status, age, sex, and weight are examples of the information that could be transmitted through the voice (Karpf, 2007). Pitch, sound pressure, timber, and tone are the main characteristics of the human voice (Dasgupta, 2017). Emotions affect the voices we speak by changing the speech pattern and tonal quality. For example, the use of shrill or high pitched voices may relate to a scared or panicked emotional state; the use of long pauses and slow rate of speaking may indicate a pensive emotional state; and the use of lower intensity may indicate a sad or shame emotional state (Dasgupta, 2017; Sauter et al., 2010).

Nonverbal vocalizations such as laughing, crying, screaming, and the deliberate use of moaning, yawning, coughing and snoring convey different emotional information as well (Trouvain and Truong, 2012).

More details on how emotions affect different acoustic characteristics is presented and discussed in Chapter 5.

2.4 Emotion classification

Emotions share a set of characteristics such as short duration, unbidden occurrence, and automatic appraisal, while at the same time each emotion has a set of unique and distinctive characteristics (Ekman, 1992a). The two main approaches in understanding emotions are: evolutionary psychology and social construction (Prinz, 2004). Both have their supporters and their powerful evidences. The evolutionary psychology approach is based on the belief that emotions are adaptive. Evidences provided by the evolutionary psychologist in favor of their approach includes the universality of certain emotions and that the adaptive responses of emotions are associated with biological adaptation, ancient brain structures and nervous systems. The social construction approach is based on the belief that emotions are socially

constructed. Constructionists believe in the critical role of cognition in emotion and reject the idea that they are related to bodily states. The variations of emotions across cultures is the strongest evidence used by constructionists to support their argument (Prinz, 2004). In this Section, we will talk about basic emotions, emotion classification approaches, and the applications of automatic emotion classification in augmentative and alternative communication (AAC).

2.4.1 Basic emotions

There is a lack of agreement among researchers on the number of basic emotions. This disagreement is not just on what emotions are considered basic and what are considered composite; it is also on the definition of the word "basic" and its many possible meanings. The main problem is that there is no generally acceptable definition of "basicness" and yet, the issue that some emotions are more basic than others (Averill, 1994; Ortony and Turner, 1990).

Paul Ekman is one of the theorist who has contributed greatly to the notion of basic emotion theory, the number of basic emotions, the characteristics of basic emotions, and the expression and physiology of emotion (Ekman, 1992a,c, 2003). Ekman and Friesen (1971); Ekman et al. (1969) and Ekman (1971) carried a number of studies in different cultures and proved the universality of facial expressions of certain emotions. In all of these studies, six emotions were considered. The experiments were conducted by presenting photographs of facial expression of six different emotions: happiness, surprise, sadness, anger, fear, and disgust to five literate cultures. These six basic emotions become one of the widely accepted candidates for basic emotions. Izard (1971) also conducted similar experiment to judge the universality of facial expressions but used a set of eight pairs of emotion words. These pairs were enjoyment-joy, interest-excitement, distress-anguish, anger-rage, disgust-contempt, surprise-startle, fear-terror, and shame-humiliation. Results indicated strong cultural agreement in identifying these facial expressions.

A recent research in the University of Glasgow conducted by Jack et al. (2014) suggested that there are only four basic expressions of emotions instead of six. These four categories of emotions are happy, sad, fear/surprise, and disgust/anger. They found that fear and surprise expressions look very much alike to the observers. The same goes for disgust and anger. The researchers suggested that the differences between fear and surprise and between disgust and anger arise later to serve the purpose of social interaction needs rather than survival needs.

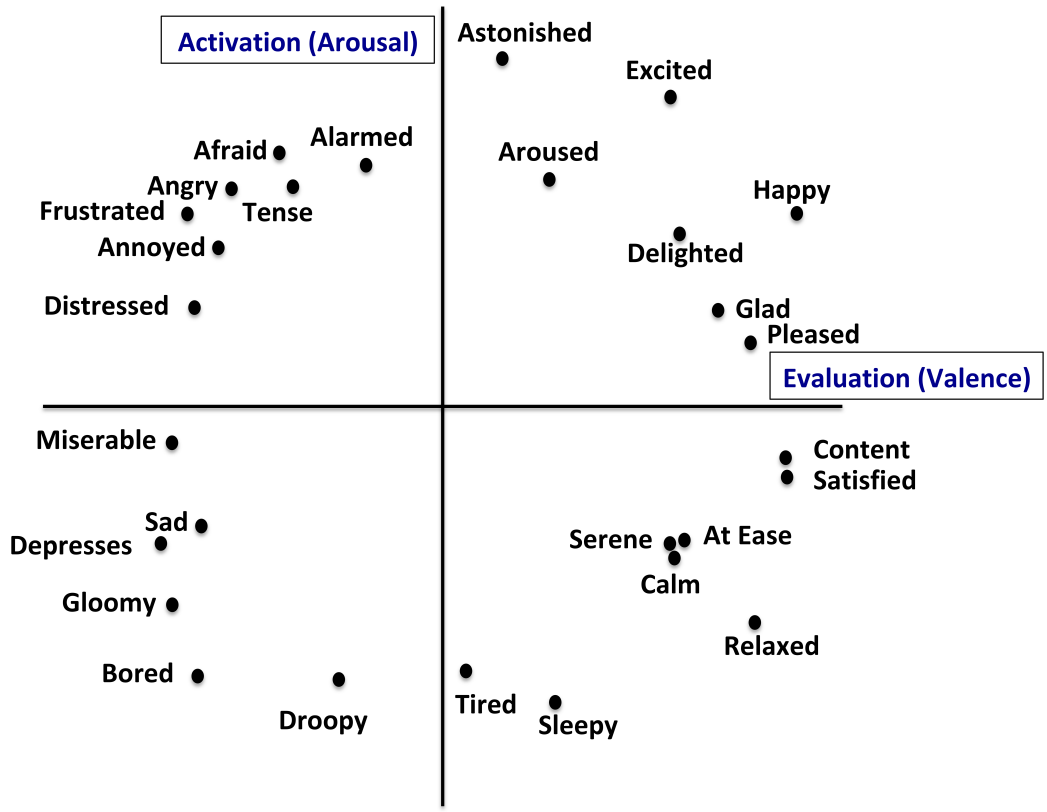


Fig. 2.4 Scaling of 28 emotional states on the arousal-valence space (Russell (1980)).

2.4.2 Emotion classification approaches

Based on the many given definitions of emotions, two ways to conceptualise emotions have become common. It is either done by using a discrete, also known as categorical or dimensional approach.

In the discrete approach, emotions are identified using a small number of basic and primary emotions. Usually six basic emotions are used: happiness, surprise, sadness, anger, fear, and disgust which are usually experienced for a short period of time. (Ekman, 1971; Ekman and Friesen, 1971; Ekman et al., 1969; Johnson-Laird and Oatley, 1992). People tend to use this approach when describing observed emotions in their daily life. Thus, it is justifiable, intuitive, and matches people experience to use this labeling scheme. However, relying on only these six basic emotions categories will result on the inability to describe the more complex emotions which also occur in everyday communication (Akçay and Oğuz, 2020; Zeng et al., 2008).

In the dimensional approach, emotions are identified using a small number of independent latent dimensions such as arousal, valence, and dominance. These dimensions are considered to be generic and definitive to describe emotional states. One of the most used dimensional models is a two dimensional model that uses valence and arousal or sometimes pleasant-unpleasant and arousal-sleepiness dimensions (Abelson and Sermat, 1962; Russell, 1980; Russell and Bullock, 1985). According to Russell's (1980) model, emotions are not discrete but rather systematically interrelated and could be represented in a two dimensional spatial model. Figure 2.4 presents the scaling of 28 emotional states in the arousal-valence space. The two dimensions upon which the emotions vary are the pleasure-displeasure (horizontal) dimension and arousal-sleep (vertical) dimension. The valence dimension represents how positive or negative a felt emotion is ranging from pleasant to unpleasant while the arousal dimension represents how strong a felt emotion is ranging from frantic excitement to boredom or sleepiness. In a three dimensional model, a dominance dimension is added which represents the ability of the speaker to handle a situation, that is the speaker's strength and power. It is useful when distinguishing anger from fear for example, where they could only be distinguished using this dimension (Grimm et al., 2007). Although using the two dimensional approach is aimed at reflecting the main aspects of emotions, there are several disadvantages of this approach. Mainly, the loss of information as a result of projecting high dimensional emotions into a two dimensional space may make some emotions identical or indistinguishable such as anger and fear. Also, using this approach is not intuitive (Akçay and Oğuz, 2020; Zeng et al., 2008).

Since categorical and dimensional approaches have each their own advantages and disadvantages and they both can be useful in enhancing the functionality of AAC technology, both approaches are adopted in the development of automatic recognition of emotions in dysarthric speech models that this thesis investigates with more emphasis on the categorical approach.

2.4.3 Applications of automatic emotion recognition in augmentative and alternative communication (AAC)

The automatic recognition of emotions is becoming an interesting and important field for human-computer interaction (HCI), AAC, and assistive technology. Emotions are very complex constructs with substantial variations among individuals in expressing and experiencing different emotions. This makes the process of automating naturalistic human emotion recognition a very challenging task. The popularity of automatic emotion recognition rises from the prospect of the substantial applications that can be developed. The existence of

different modalities and channels individuals may use to express their emotions, as discussed in Section 2.3.2, opens many doors for many recognition methods to be developed. Depending on the purpose of the developed application, different sensors needs to be incorporated to be able to capture and recognise emotions from the different channels that individual use to express emotions. These sensors or recognisers might be unimodal in a way that it captures emotions from a single channel only or multimodal systems in a way that it captures emotions from multiple channels. Facial expression recognisers, speech recogniser, biosignals recognisers, motion recognisers, and linguistic recognisers are among the most popular emotion recognisers used to detect and recognise emotions. There are many potential applications of automatic emotion recognition in the field of AAC and assistive technology. The below examples provide some insight into these potential applications.

Facial expression recognisers

Emotional facial expression recognisers may be helpful for people who have difficulties in recognising facial expressions. These difficulties can be a consequence of comprehension, perceptive, or sensorial problems. People with vision disabilities, for example, may benefit from knowing the emotional state from the facial expressions of the interlocutor by producing the correct response such as comforting the interlocutor when he/she is sad. The need of such applications rises especially in situations where other verbal and nonverbal cues are unavailable. On the other hand, people who have autistic or Asperger syndrome do not only have problems in understanding facial expressions of other people but also do not know how to respond to these expressions (El Kaliouby and Robinson, 2005; Garay et al., 2006). Therefore, a technology that identifies the emotional state of the interlocutor and sends recommended reactions to the disabled user would be helpful in supporting effective communication. The Emotional Hearing Aid proposed by El Kaliouby and Robinson (2005) is an example of such an application.

Speech recogniser

Emotional speech recognisers use different voice parameters such as prosody features to infer the emotional state of the speaker. Deaf people, for example, may benefit from emotional speech recognisers to know the emotional state of the speaker as a way to support effective communication especially in the situations where nonverbal cues are unavailable. An example of such a situation is the case of using text-telephone technology (TTY) which is widely used by deaf people. In these situations, a useful application would be an emotional speech recogniser that will help deaf people to understand the emotional state of the interlocutor

(Garay et al., 2006). In addition, speech recognition technologies used in call centers to help in interpreting the emotional state of callers, can be adapted to be used by people with hearing impairment to support effective communication (Garay et al., 2006). The literature includes many studies on automatic speech emotion recognition (SER) that investigated the performance of different algorithms and techniques. Chapter 7, Section 7.2 reviews some of these studies.

Biosignals recognisers

Biosignals cary information about the emotional state of the person. It has the advantage of being able to be measured by non-invasive sensors as well as being free from social masking (Van Den Broek et al., 2009). Biosignals sensors measure factors such as the person's blood volume pulse (BVP), heart rate, fluctuations of skin temperature, and skin conductance. People with speech impairment can benefit from combining biosignal sensors applications with telephones in order to provide to or obtain from the person they are having conversation with emotional information that cannot be obtained be speech (Garay et al., 2006). Another example of a prospect application of using biosignals recognisers is a teaching tool that help people with autism to understand the emotional state of the interlocutor and help them build this skill progressively (Garay et al., 2006).

As speech is the most preferred means of communication with rich acoustic information and with the increased interest in and development of automatic SER models, this thesis focuses on automatically recognising emotions from dysarthric speech. Looking into automatically recognising emotions from other modalities such as facial expressions and gestures will be the focus of future investigations. Thus, the collected database presented in Chapter 4 contains both audio and video data.

2.5 Summary

This chapter highlighted the importance of human communication and discussed the barriers to effective communication including physical, semantic, psychosocial, and disability barriers. The second section of this chapter identified dysarthria and reviewed its five different types. The prosodic and acoustic characteristics of speakers with spastic and hypokinetic dysarthria and their ability to perform different prosodic functions were then discussed. The vital role of emotions and emotion communication were highlighted followed by the different modalities used by humans to communicate emotions. The notion of basic emotions and emotion classification approaches were then presented. An insight of the potential applications of automatic emotion classification systems in AAC was provided.

The potential control of prosodic and phonatory features found in speakers with dysarthria caused by cerebral palsy and PD discussed in Section 2.2.3 encourages investigations related to other kinds of paralinguistic information people with dysarthria can communicate and their consistency of doing so. The focus of this thesis is to investigate their ability to communicate emotions in English through suprasegmental and prosodic features and if they are able, then to what extent is it feasible to automatically classify emotions.

Since little is known about the ability of people with dysarthria to communicate emotions through their speech, it is important to explore the topic from the speakers' point of view in terms of the difficulty and importance of communicating emotions and the methodology used to do so as a pre-study before starting to collect the data and automating the emotion classification process. This is explored and discussed in the next chapter.

Chapter 3

Towards the Understanding of Communicating Emotions for People with Dysarthria

The content of this chapter has been published in the International Journal of Psychological and Behavioral Sciences 2020 (Alhinti et al., 2020a).

3.1 Introduction

People have been looking into different ways of applying expressiveness to synthetic speech. The study of the 17 ways to say "yes" revealed four perspectives of the voice tone: emotional state, conversational intent, social context, and vocal qualities (Pullin and Hennig, 2015). Communicating emotions is part of the expressiveness that can be added to the AAC devices. There are several possible input channels that can be used to communicate emotions using AAC devices. For example, the use of emotion words from the AAC vocabulary list such as 'happy', 'sad', etc., or the use of visual emotional symbols. Out of the many possible ways, it would be interesting to be able to communicate emotions directly using voice-input-voice-output communication aid (VIVOCA), especially as the literature shows in Chapter 2 that many AAC users prefer to use their residual voices (Hawley et al., 2006, 2013). Figure 1.1 presented in Chapter 1 illustrates a high level description of the hypothesized dysarthric speech driven AAC device that would be ideal to achieve in which the AAC device would be able to recognize and interpret its user's disordered speech along with their emotional state and then deliver the message with the effect of the detected emotion in a clear synthesized voice.

Since a little is known about the ability of people with dysarthria caused by cerebral palsy to communicate emotions, this chapter focuses on understanding more about the nonverbal communication ability of people with dysarthria from the speakers' point of view in terms of the difficulty, importance, and methodology used to communicate their emotions as a preliminary study before attempting to automate the emotion recognition process in dysarthric speech. Therefore, a survey was designed to address the above points. The rest of this chapter is structured as follows. Section 3.2 describes the methodology followed in conducting this study. Section 3.3 presents the survey's results and discusion of the findings. Finally, a conclusion is presented in Section 3.4.

3.2 Methodology

This research received ethical approval from the ethical review panel of the Department of Computer Science at the University of Sheffield as the first stage of several stages towards the final aim, which is automatic recognition of emotion in dysarthric speech. The aim of this stage of the research is to achieve a better understanding of how people with dysarthria communicate emotions. Therefore, a survey was designed to address the following questions:

- How difficult it is for people with dysarthria to get their emotions across?
- What are the emotions that are important to them to get across?
- What are the ways that they tend to use to get their emotions across?
- Is there a difference in the way emotions are communicated to familiar and unfamiliar people?

Knowing the answers to the above questions will help in defining the scope of the research. It will also help in identifying the generalisability of this research area among people with dysarthria. The survey was distributed using special email lists, flyers, and personal approach that targeted participants who have dysarthria caused by cerebral palsy within the United Kingdom. The following section discusses and analyzes the main findings of the survey.

3.3 Survey results

The survey contains a total of 27 questions. Closed questions, open-ended questions and rank order questions were included to get the maximum information out of this survey. To follow a logical flow of the questions, the survey is arranged into three sections. The first section is

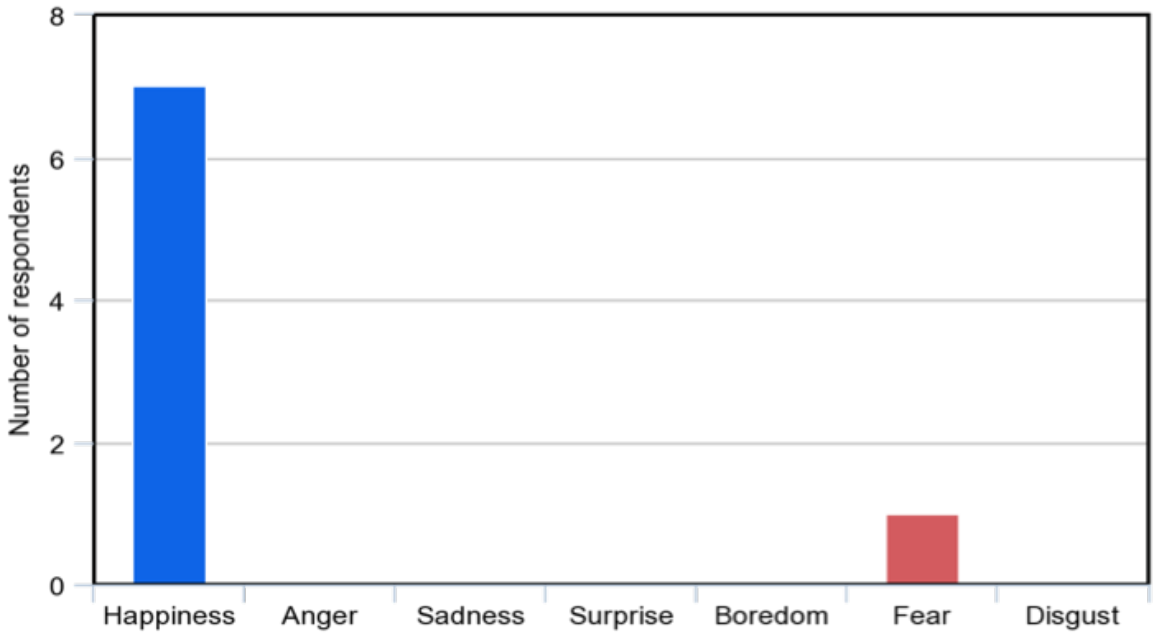


Fig. 3.1 Survey result of the most useful emotion to try to communicate in social life settings for people with dysarthria.

related to the use of a communication aid. The second section, which is the main section, is related to questions about emotions. The third and final section is related to demographic information. The full list of the survey questions and their results can be found in Appendix A. Below, we will discuss the findings of the main questions in the survey from eight native English respondents – six male, one female, and one participant who preferred not to say. Five of the respondents have severe dysarthria and three of them have moderate dysarthria caused by cerebral palsy.

All but one of the respondents are users of one form or another of a voice output communication aid. The non-communication aid user respondent indicated their preference of using their residual speech over communication aids as it is a faster means of communication. This is a typical preference from this group of people, however, the survey shows that people with dysarthria can face difficulties when communicating with familiar people, if they are not using their communication aid, but that this problem is exacerbated when they are communicating with unfamiliar people.

Given a list of seven different emotions (happiness, sadness, anger, surprise, boredom, disgust, and fear), respondents were asked about what emotion do they feel is the most useful to try to communicate in their social life. As can be seen from Figure 3.1, all but one of the respondents chose 'happiness', with the remaining respondent choosing 'fear'. One of the respondents justified the importance of communicating 'happiness' in social life settings by

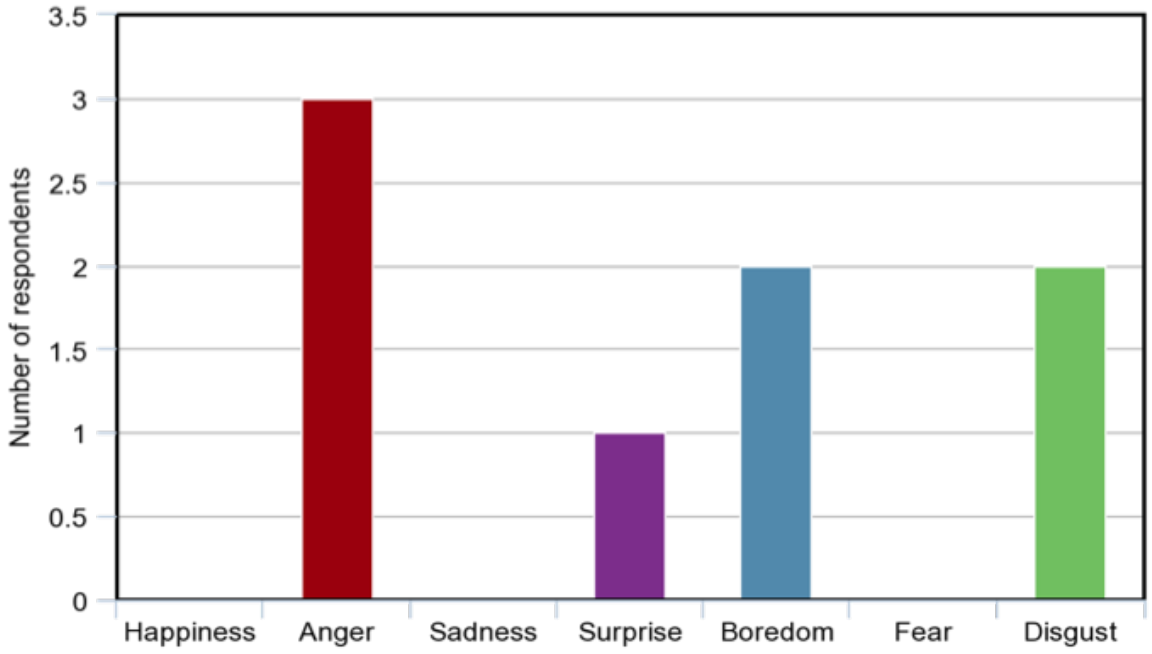


Fig. 3.2 Survey result of the most difficult emotion to try to communicate to familiar people.

the following: "People need to know that I am happy with them so they want to come back and be with me". 'Happiness' was also chosen by the majority of the respondents as the most important emotion they feel they want to communicate in everyday life. From the set of questions that focus on addressing the difficulty of communicating emotions, the following question was asked: "What emotion do you feel is the most difficult for you to communicate to familiar people?". 'Anger' was chosen by almost half of the respondents. 'Surprise', 'boredom', and 'disgust' were chosen by the other respondents. This question results are presented in Figure 3.2. When communicating with unfamiliar people, respondents' answers vary. However, 'anger' and 'boredom' were the most chosen emotions among the others. These emotions, in typical speech, are perhaps characterised by being more subtle (boredom) or easily confusable (anger/surprise) compared to e.g., happiness (Lugger and Yang, 2007; Yacoub, 2003). Looking into the channels that people with dysarthria tend to use when communicating emotions to familiar people, the following question was asked: "How do you communicate your emotions to familiar people?". As can be seen from Figure 3.3, the majority of the respondents indicated their use of facial expressions and/or speech. The use of gestures, and eye gaze were also indicated by some respondents. There was little difference in respondent's answers to this question regardless of whether they are communicating with familiar or unfamiliar people.

In a ranking order question, respondents were asked to rank a set of emotions according to their importance to them in terms of being able to communicate them successfully "For

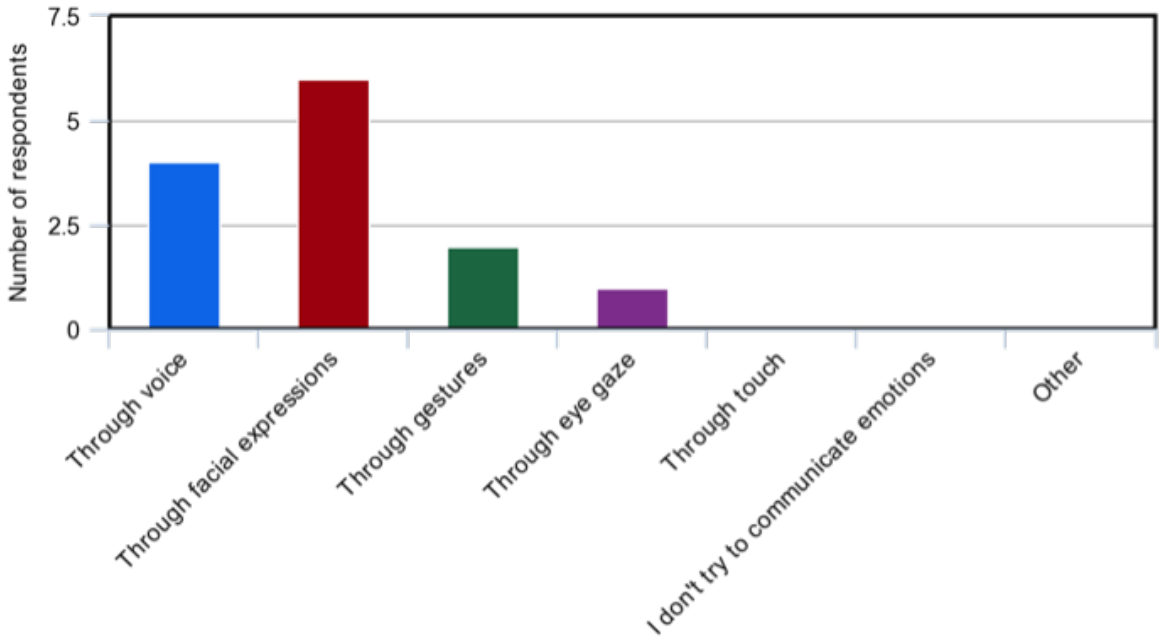


Fig. 3.3 Survey result of the channels that people with dysarthria tend to use when communicating emotions to familiar people.

1 being the most important and 7 being the least important, please number the following emotions according to their importance to you in terms of being able to communicate them successfully.". Figure 3.4 presents the results obtained from this question where the average ranking of each answer choice is displayed. The results show an indication of the importance of successfully communicating 'happiness', 'anger', and 'sadness'.

Despite the relatively small number of responses the survey has, the aim of this survey is to provide an insight and understanding of some related aspects to communicating emotions for people with dysarthria rather than give final conclusions at this stage. It can be inferred, however, how complex the problem is as many factors have already been indicated to playing a critical role in the way emotions are communicated. The nature of the person and the severity of their speech disorder are some of the main factors that may influence the way emotions are communicated.

3.4 Conclusion

The survey was kept live for more than a year trying to collect as much responses as possible to help in enabling the generalisability of the findings beyond the relatively small number of responses it currently has. However, even with the survey being online, the short time commitment needed from the participants to complete it, and the many different ways used

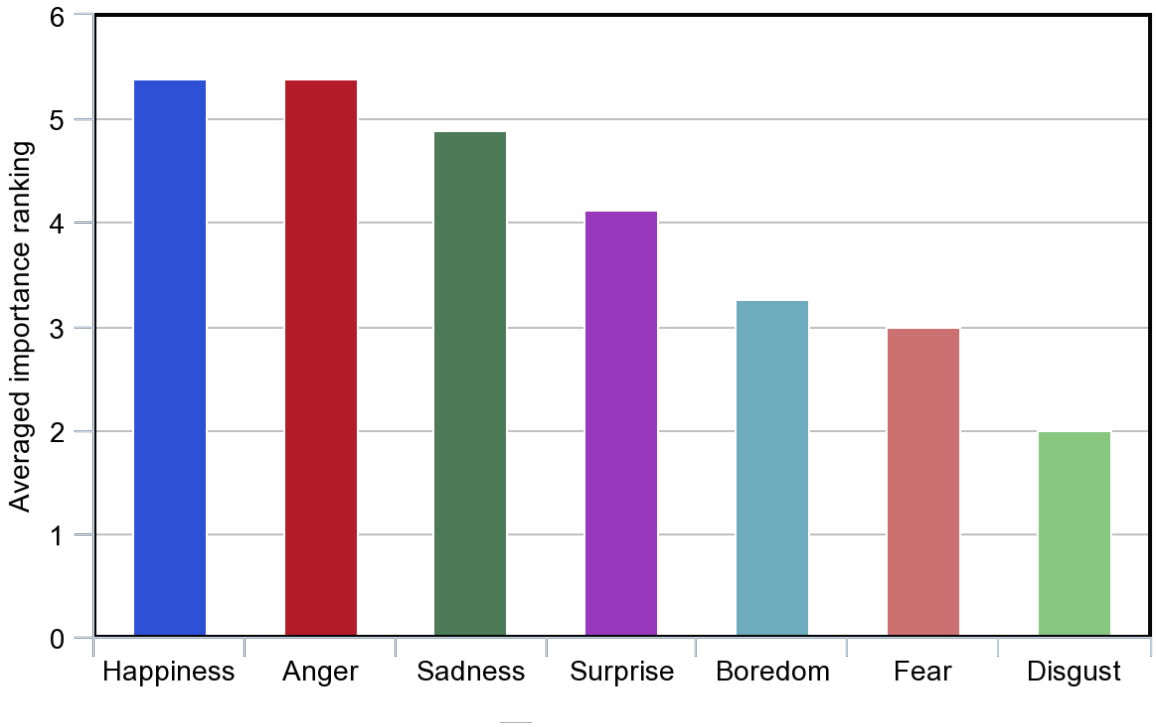


Fig. 3.4 Survey result of the most important emotions to communicate for people with dysarthria.

to distribute the survey across targeted people, it was very challenging to get people to participate. Beside that, the survey shows that people with dysarthria find difficulties when communicating emotions with familiar people; however, the problem is exacerbated when they are communicating with unfamiliar people. Therefore perhaps a VIVOCA that could assist with this could be beneficial. This new field of research will need a lot of understanding of how people with dysarthria communicate their emotions, how this may be encoded and thereby automatically extracted via prosodic and acoustic features, and how consistent a certain emotion is expressed among speakers and within the speaker him/herself (inter- and intra- speaker variability). All of these questions and probably more need to be addressed before we will be able to fully automate the process of identifying emotions in dysarthric speech and adding this information to the output of the AAC.

In order to investigate the ability of people with dysarthria to communicate emotions in their speech and the feasibility of automatically recognising these emotions, a database of dysarthric emotional speech needs to be recorded. The next chapter will present all the details about the database recordings in terms of the design and development.

Chapter 4

Disordered Speech Emotional Database Collection

Part of the content of this chapter has been published in Speech Communication Journal (Alhinti et al., 2021).

4.1 Introduction

People with dysarthria show a strong preference for using their residual voices when they communicate as it is the natural mean of communication (Beukelman et al., 2007). This was also confirmed by the respondents of our survey presented in Chapter 3. Therefore, voice-input-voice-output communication aids (VIVOCAs) would be one of the preferred means of communication especially for those with mild to moderate dysarthria (Hawley et al., 2006, 2013). People with dysarthria do not have any problems understanding other people's speech and emotions. Their problem is mainly about producing intelligible speech (Miller and Bachrach, 2017). Having such limited phonological and prosodic dimensions may result in making it hard for them to convey emotions in their speech in a way that can be captured and understood clearly and easily by recipients. This is because their way of conveying emotions can be different from that of typical speakers. Having said that, it is important to highlight the main points towards achieving such a goal. We need to understand how people with dysarthria communicate emotions through their voices, how different emotional states can be detected and classified automatically, and how to produce emotional synthetic speech that best reflect and adapt to the emotional state of the user. To address the above questions, it is essential to have a speech emotional database of both types of speech: dysarthric speech and typical speech (i.e., a parallel database), and hence this database was developed.
This chapter presents the design and development of the Dysarthric Expressed Emotion

Database (DEED) which is a first of its kind. Section 4.2 describes the corpus design and development. Section 4.3 contains the discussion and conclusion.

4.2 Corpus design

DEED is an audio-visual British English database of emotional speech that contains both typical and dysarthric speech. It is designed for the purpose of:

- Analysing the features used by people with dysarthria when communicating different emotions (see Chapter 5).
- Comparing these features with that used by people with typical speech when communicating the same emotions (see Chapter 5).
- Developing automatic emotion classification model for the dysarthric speech (see Chapter 7 and Chapter 8).
- Applying the findings in the development of emotional speech synthesis for the dysarthric speech in a voice-input-voice-output communication aids (VIVOCAs) (Hawley et al., 2006, 2013).

A controlled approach has been adopted for the design and development of this database. The below subsections will discuss the adopted approach. The recording approach of DEED has been ethically approved by the University of Sheffield, UK. Before any recording session, a written consent form has been obtained from every participant.

4.2.1 Scope

Whereas there exist a few databases on dysarthric speech such as the TORGO Database (Rudzicz et al., 2012), the Nemours database (Menendez-Pidal et al., 1996), and the Dutch dysarthric speech database (Yilmaz et al., 2016), they are not emotional databases and therefore cannot serve the purpose of analyzing emotions in dysarthric speech nor developing automatic emotion classification techniques. To the best of our knowledge, DEED is the first database of its kind. It is not only that it contains a multimodal (audio and video) emotional dysarthric speech but it also contains emotional typical speech. Both kinds of speech were recorded in the same recording studio using the same settings. This allows a fair comparison and analysis to be made between the two types of speech. Table 4.1 includes the details of DEED audio recordings. Figures 4.1 and 4.2 show the tree diagram of the design of the DEED typical speech and dysarthric speech corpus parts, respectively. The numbers written

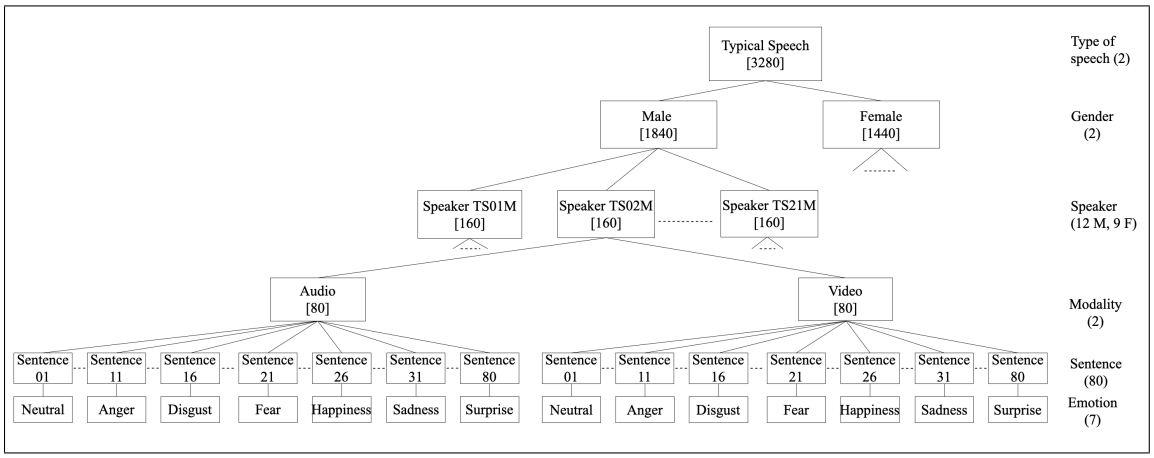


Fig. 4.1 Tree diagram of the design of the DEED typical speech corpus part.

Typical speech				
Number of audio recordings (utterances)	1680 (80 utterances per speaker)			
Number of speakers	21 speakers (12 male and 9 female)			
Total length of recordings	1.268 hours			
Average length of utterance in seconds	2.178			
Dysarthric speech				
Speakers	DS01F	DS02F	DS04F	DS03M
Number of audio recordings (utterances)	80	80	80	80
Total length of recordings in minutes	15.393	4.008	3.805	3.599
Average length of utterances in seconds	11.545	3.006	2.854	2.699

Table 4.1 The details of DEED audio recordings.

between square brackets indicate the total number of files for that level. All speakers have been audio and video recorded, except one of the male typical speakers who was only audio recorded. As can bee seen, the size of the DEED-Typical speech part makes it suitable for many machine learning approaches. Although the main reason for developing this database is for dysarthric emotion classification purposes, it can be used for training ASR for dysarthric and typical speech.

4.2.2 Selection of methodology

The fact that emotional states are caused by many factors is the reason behind the difficulty of collecting samples of people under a particular emotional state (Douglas-Cowie et al., 2003). Over the past decades, there has been considerable debate over the type of methodology that should be used. The studies on vocal emotional expressions falls into one of the following

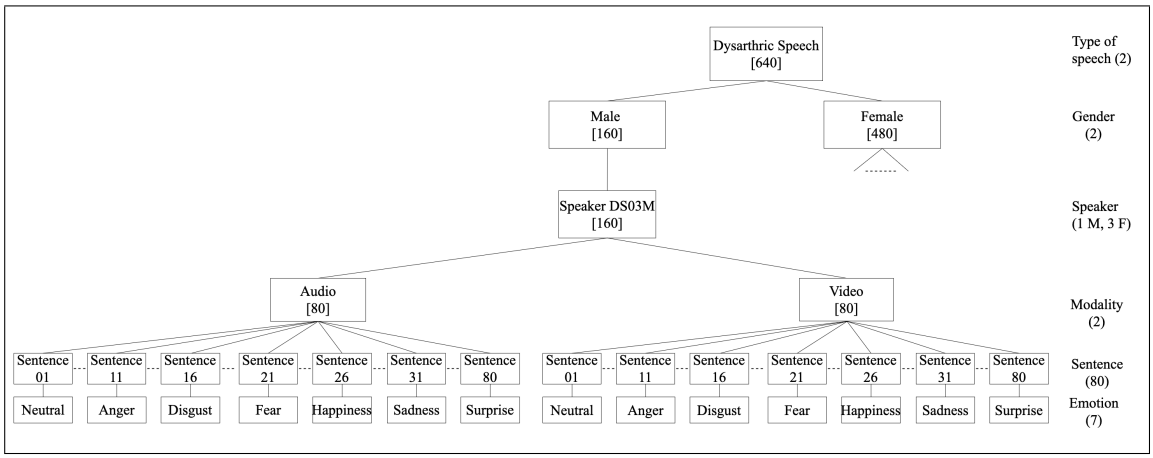


Fig. 4.2 Tree diagram of the design of the DEED dysarthric speech corpus part.

three main paradigms: natural vocal expressions, induced (elicited) vocal expressions, and acted (simulated) vocal expressions.

Natural vocal expressions databases Natural vocal expressions are those emotional expressions that were recorded during different states of naturally emotional situations. An example of these kinds of databases includes those who were recorded for pilots in a dangerous flight situations. Another example is recordings from different shows on TV and news where journalists report events that elicit emotions. Although this could be seen as the ideal research paradigm because of its high ecological validity, there are a number of methodology issues that are related to these kinds of recordings:

- These recordings are usually very brief and are taken from single or limited number of speakers.
- Mostly, suffers from poor recording quality.
- The task of determining the underlying emotion can be challenging.
- Mostly, protected by copyright laws and privacy policies which prevent them from being shared and distributed widely.
- The task of processing the data is very challenging due to the lack of control on the different aspects that are related to the recording settings such as: the environment and the background noise where the recorded data is taking place, the position of the microphones and cameras, the content, and the conveyed emotion (Busso et al., 2008; Scherer, 2003).

Examples of corpora that fall under this category are the Belfast natural database (Douglas-Cowie et al., 2000), the Geneva airport lost luggage study (Scherer and Ceschi, 1997), and the databases in (Chung, 2000) and in (France et al., 2000).

Induced (elicited) vocal expressions databases Eliciting emotional states in a speaker and recording his/her speech is another way followed by some researcher to study the effect of emotions on the voice. This intermediate paradigm falls between the natural and acted paradigms. A number of techniques have been used to induce specific emotions. For instance, Cowie and Douglas-Cowie (1996), Iriondo et al. (2000), Livingstone and Russo (2018), and Amir et al. (2000) ask the speakers to recall previous events and experiences where they have felt certain emotional states known as self-induction, presenting emotional materials such as pictures, films, and stories that trigger emotions (Martin et al., 2006; Scherer et al., 1985; Tolkmitt and Scherer, 1986), ask the speakers to perform selected emotional scripts or improvised hypothetical scenarios that are designed to trigger certain emotions (Busso et al., 2008), and solving specific problems under different induced levels of stress and/or time pressure (Fernandez and Picard, 2003; Karlsson et al., 2000). One of the advantages of this methodology that makes it favoured by experimental psychologists is the degree of control it offers which leads to having more consistent speech samples. However, this methodology contains a number of drawbacks, including relatively weak effects are often produced from following these procedures and following the same procedure on different people does not necessarily mean that the same emotional effect will be produced by all of them (Scherer, 2003).

Since this paradigm needs to be aided by additional resources, it cannot be seen as that much different from the acted paradigm (Douglas-Cowie et al., 2003).

Acted (simulated) vocal expressions databases This strategy is the most preferred strategy in the field of collecting emotional speech databases (Scherer, 2003). Speakers who are sometimes professional actors, lay actors, or non actors are asked to produce emotional verbal expressions that are usually based on standard verbal content (Banse and Scherer, 1996; Burkhardt et al., 2005; Cosmides, 1983; Fairbanks and Pronovost, 1939; Jackson and Haq, 2011; Kaiser, 1962; Scherer et al., 1972, 1973; Whiteside, 1999). There have been some raised doubts about this methodology that falls mainly under two points. First, this methodology may produce more intense or exaggerated emotional expressions when compared to the ones that result from induced and natural methodologies. Second, actors often tend to overemphasize powerful and obvious cues while they miss more subtle cues that help in differentiating discrete emotions

in natural expressions (Scherer, 1986). It could be argued, however, that with the existence of social constraints over emotional expressions and unconscious tendencies toward self-presentation, all the public expressions could be seen to some extent as "portrayals" (Scherer, 2003). In addition, since the listeners recognise reliably the emotional states from the acted speech, a reflection of part of the normal expressions at least could be assumed (Scherer, 2003). This strategy is still being followed and used by researchers in the field and the main advantage of this strategy is the full control that it provides which result in:

- Having a high quality recordings that enable later speech processing and analysis.
- Having utterances with unambiguous emotional states.
- Allowing comparisons to be made among emotions and speakers as the same utterances have been recorded by all speakers.
- The ability to recruit reasonable number of speakers over targeted group to act all kinds of emotions that are under the study to enable generalisation (Staroniewicz and Majewski, 2009).

As can be seen each paradigm has its advantages and disadvantages and what really can be seen as the determinant factor is the goal of the research. Naturalness may not be the optimal methodology of some research goals. In addition, more attention should be paid to the strategies followed when selecting the acted and elicited methodologies to insure that the results are adequate reflection of reality. As part of these strategies, natural database should be used as a comparison and a way to help the development of acted and elicited databases (Douglas-Cowie et al., 2003). Also, since studying emotion is a multidisciplinary area with many variables to consider, a single corpus may not be sufficient to address all the open questions but a set of databases that comply with the core requirements and standards would probably do (Busso et al., 2008).

If we look again into these three methodologies, then it would be clear that on one hand we have the natural databases. On the other hand, we have the acted databases. It is unclear, however, where would the elicited speech databases stand. Would it be something in between, more towards the acted speech databases, or more towards the natural speech databases? Would the technique used in the elicitation process affect this categorisation? This uncertainty is probably a result of the lack of clear definition of what is considered as an elicitation technique and what is not. For example, employing the self-induction technique by asking the speakers to remember a situation from the past where he/she felt a specific emotion and give the speaker time to put him/herself into that specific emotion before the start of the recordings, would presumably be very similar to what some speakers implicitly do

in the acted approach. Speakers in the acted approach cannot just start recording emotional speech directly, there must be some kind of internal eliciting approach they follow that helps them produce the required emotional state even if they have not been explicitly directed to a specific approach. This leads to the following two questions being raised: is there purely acted speech and are the acted and elicited approaches two distinct approaches or are they somehow part of each other?

This grey area may lead to a confusion in the databases categorization process. The Berlin database of emotional speech (Burkhardt et al., 2005) for example, has been categorized as an acted emotional speech database, even though self-induction approach was adopted as their database recording methodology. While at the same time, an emotional speech corpus in Hebrew (Amir et al., 2000), and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone and Russo, 2018), has been categorized as semi-natural (induced) speech emotional database, although, a self-induction approach was adopted in the database recording. Another example, is the SAVEE database that has been categorized as acted emotional speech although the emotion stimuli, including photos and short video clips, have been presented to the speakers before recording the emotional utterances (Jackson and Haq, 2011). This methodology could be seen by many as induced methodology rather than acted.

There is a need for researchers in the emotion and psychology field to come up with a clear definition and set boundaries that helps in distinguishing between these two approaches.

One of the very important questions is which methodology should be adopted to develop DEED to insure that it is resourceful, reliable, and valid? Since there are no available emotional databases on dysarthric speech, there are no means of practical comparisons on the methodologies that are best to follow. Natural methodologies will not be appropriate in forming this database due to the many problems related to natural database recordings discussed above. More layers of difficulties are added to these problems giving the fact that the recordings here are of atypical speech. For example, the task of determining the underlying emotion would be much more challenging for a dysarthric speech than it would be for a typical speech. This is in itself one of the research questions. Since this database is the first of its kind, having such an ambiguity in determining the expressed emotions would not help in constructing a reliable database. Also, practical consideration such as wheelchairs that are used by some people with dysarthria add more layers of difficulties to the problem. In fact, having a natural database of dysarthric emotional speech may make the distribution of the database to the research community difficult or impossible due to ethical and privacy issues. Therefore, a choice between the acted and elicited methodologies should be made. According to the discussion made above, two main points have been highlighted:

- Although many emotional databases have been categorized as acted, when looking at the way they were recorded, it is clear that one way or another of elicited methodologies have been used, despite it have been miscategorised as acted.
- Is there really pure acted methodology? Even if no explicit elicitation techniques have been used such as presenting emotional stimuli, it cannot be guaranteed that at the very least self-induction techniques have not been used by the speakers even if they have not been asked to do so explicitly. In fact, one could argue that without self-induction, the acting would be quite poor.

As a result, one can see that the acted and elicited methodologies are very closely related and in practice they can rarely be thought of as two distinct methodologies.

Therefore, based on the nature of the speech disorder (dysarthria), a strong argument in favour of the elicited methodology can be made. The main issue that is in disfavour of the acted and some elicited speech methodologies, is the concern of producing more intense or exaggerated emotional expressions when compared to the natural methodology. However, in the case of disordered speech in general and dysarthric speech in particular, this is not considered as an issue. This is because people with speech disorder tend to exaggerate the way they speak and convey emotions anyway in an attempt to try to get their messages across. Therefore, the gap between the natural and elicited way is not that big. With the power of control offered by the elicited methodology, adopting the elicited approach in the development of dysarthric emotional speech database looks highly valid and thus it was selected as the methodology of recording the DEED.

4.2.3 Selection of speakers

There were three groups of participants in this study: speakers with dysarthria associated with cerebral palsy, speakers with dysarthria associated with Parkinson's disease (PD), and speakers with typical speech. All participants were recruited using advertising emails sent to special email lists, flyers, and word of mouth in the area of Sheffield, UK. The inclusion criteria for all of the three groups were that the participant must be a native British English speaker, over the age of 18, and have no known cognitive problems and no known literacy difficulties. None of the participants were professional actors. Informed consent was obtained from all participants.

Speakers with dysarthria				
Type of dysarthria	**Speaker**	**Gender**	**Age**	**Dysarthria severity [*] / time diagnosed**
Spastic dysarthria (cerebral palsy)	DS01F	Female	65 years	Severe / from birth
Hypokinetic dysarthria (PD)	DS02F	Female	71 years	Mild / 10 years
	DS04F	Female	68 years	Mild-to-moderate / 10 years
	DS03M	Male	66 years	Moderate / 9 years

Speakers with typical speech				
Gender	**Number of Speakers**	**Age**		
		Mean	**SD**	**Range**
Female	9	34.00	13.26	20-56
Male	12	35.67	16.81	19-70

[*] The dysarthria severity levels indicated in the table are informal judgments by the authors.

Table 4.2 Speaker's description.

Speakers with dysarthric speech

Two groups of participants with dysarthria were included in this study. The first group contained 1 female speaker with severe dysarthria associated with cerebral palsy. The second group contained 2 female speakers and 1 male speaker with dysarthria associated with PD. Recordings of speakers with PD were taken while they were under the anti-Parkinsonian medications effect. Table 4.2 lists the details of the speakers and their dysarthria severity levels.

Speakers with typical speech

Twenty-one speakers with typical speech were included in this study, 12 male and 9 female. Table 4.2 lists the details of the speakers of this group. All twenty-one speakers have been audio and video recorded except one male speaker who was only audio recorded.

4.2.4 Selection of emotions

With respect to the set of emotions captured, the widely adopted approach is to capture a small set of 'basic' emotions (Douglas-Cowie et al., 2003). Most of the discrete emotion models are taken from Darwin's "The Expression of Emotion in Man and Animals" (Darwin, 1872). This discrete emotion pattern approach has been popularized by scholars in this field: Tomkins, Ekman and Izard (Ekman, 1971, 1980, 1992b; Ekman et al., 1987; Izard, 1971, 1994; Levenson et al., 1992; Tomkins, 1962, 1963; Van Bezooijen et al., 1983). The Geneva

airport lost luggage study (Scherer and Ceschi, 1997), the Danish Emotional Speech (DES) database (Engberg et al., 1997), the Berlin database of emotional speech (Burkhardt et al., 2005), the eNTERFACE'05 audio-visual emotion database (Martin et al., 2006), the SAVEE database (Jackson and Haq, 2011), and the RAVDESS database (Livingstone and Russo, 2018), to name a few examples of databases that adopted this approach in the development of their databases. Two other approaches have emerged as a result of the debate on the range of emotions covered. The first approach is to cover a larger set of emotions and sometimes distinguish between different forms of an emotion. Banse and Scherer (1996) is an example of a study that followed that approach where cold and hot anger have been differentiated. The second approach is to cover a narrower set of emotions and therefore study it in depth. The most prominent databases that follow this approach are stress oriented. Fernandez and Picard (2003) and Tolkmitt and Scherer (1986) are examples of studies that followed this approach where different levels of stress have been investigated.

Ververidis and Kotropoulos (2003) found in their review that the most common recorded emotions ordered in decreasing order are: anger, sadness, happiness, fear, disgust, surprise, boredom and joy. Also, despite the number of different emotions, Ververidis and Kotropoulos (2006) found in their review of sixty-four emotional speech databases, that the majority have limited their recordings to five or six emotions. It is important in this aspect to consider that emotional life is strongly influenced by culture and are subject to social rules, display rules, and feeling rules (Boiger and Mesquita, 2012; Scherer and Ceschi, 2000). Also, moderate emotional states are more expressed during daily communication rather than full blown basic emotions (Douglas-Cowie et al., 2003). Therefore, if the developed technology is to be used for everyday life, it is important to consider the set of emotions that occurs frequently rather than studying emotions that occur rarely in everyday situations (Scherer, 2003).

All of the above approaches have their defenders; however, since the debate remains unsettled, the six basic or primary emotions: happiness, sadness, anger, surprise, fear, and disgust has been selected for the development of the DEED. Neutral state also has been added as a baseline conditions. Another reason of this selection of emotions is that these sets has been widely adopted in most existing databases (Jackson and Haq, 2011; Livingstone and Russo, 2018; Martin et al., 2006; Mazurski and Bond, 1993; Tottenham et al., 2009).

4.2.5 Stimuli

The text material is a subset of the material used in the SAVEE database (Jackson and Haq, 2011). Long sentences were excluded from the adopted set of sentences, as it might have been difficult for some people with dysarthria to be able to speak them. It consisted of 10 TIMIT sentences per emotion: 3 common, 2 emotion-specific and 5 generic sentences that

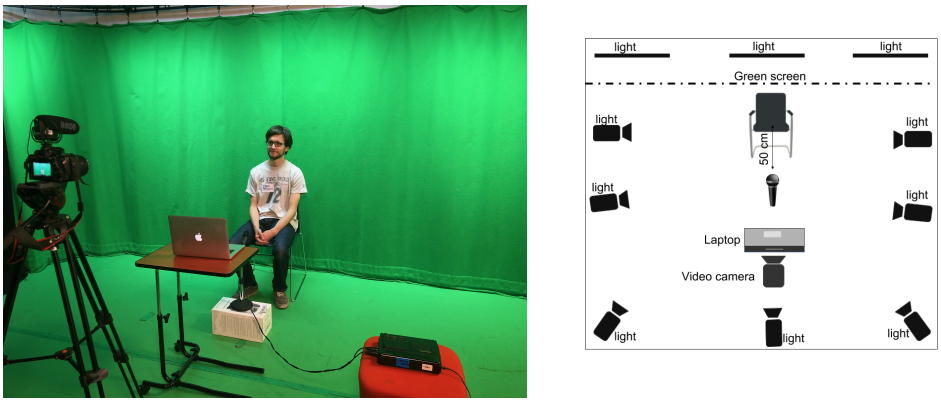
were different for each emotion. The 3 common and 2 x 6 = 12 emotion-specific sentences were recorded as neutral in addition to 2 neutral sentences and 3 generic sentences. This gives a total of 20 neutral sentences. Therefore, a total of 80 utterances per speaker is recorded. The SAVEE database has a total of 120 sentences per speaker (Jackson and Haq, 2011). A list of DEED sentences for anger, disgust, fear, happiness, sadness, surprise and neutral emotions can be found in Appendix B.

The sentences were divided into fourteen blocks, were each block contained 5 sentences from the same emotion, except for the neutral state where the block contained 10 sentences. Each recorded set began with a neutral block followed by one block from each emotion, giving a total of two rounds. This division procedure was applied to help in avoiding bias caused by speakers' fatigue and to ensure that speakers, mainly those who have dysarthria and could not record the whole set of sentences, he/she would at least been through recording one round which includes a subset of the sentences that covers all set of emotions.

The stimuli presentation consisted of three main stages: task presentation, emotive video presentation, and sentence presentation. In the task presentation stage, the emotional state that the speaker should perform next is presented as a text on a screen in front of the speaker for around 2 seconds. The emotive video presentation stage consisted of playing a short emotive video clip to help the speaker elicit the target emotional state. Finally, in the sentence presentation stage, each sentence within the current block of sentences is presented on the screen individually.

4.2.6 Emotion elicitation approach

In order to elicit specific emotions in speakers, emotion stimuli has been chosen as the eliciting technique. Very short video clips of emotion stimuli are presented in order to elicit specific emotions. The video clips that have been used are adopted from those used when recording the SAVEE database (Jackson and Haq, 2011). These video clips were taken from popular movies and television series. In addition to that, speakers were told that they can use Stanislavski's emotional memory techniques where they can remember the details of a situation with the same emotion if they think it will help them to put themselves in a particular emotional state (Stanislavsky et al., 1936). This follows standard protocols for recording such databases (Burkhardt et al., 2005; Jackson and Haq, 2011; Livingstone and Russo, 2018). Speakers were given all the time they needed to put themselves into a specific emotional state. They have been also told that they can repeat a sentence as many times as they want until they feel satisfied with their performance. Speakers have been explicitly instructed to provide genuine expressions of emotions as they would do in typical everyday



(a) (b)

Fig. 4.3 Recording studio physical setup.

scenarios. No instructions or guidance were given in how a particular emotion should be expressed.

4.2.7 Data capture

The database recordings took place in a professional recording studio at the University of Sheffield over several months. Figure 4.3 shows the data capture physical setup. All speakers were sitting all the time during the recordings facing the camera. A green screen cloth has been used as the background. The microphone was placed 50 cm from the speaker. A break is given after finishing one round of the sentences blocks, if needed, although speakers were permitted to ask for a break anytime they feel they needed to. The duration of the break was totally determined by the speaker.

Technical information

Speakers were recorded individually with Canon EOS 80D DSLR Kit with 18-135mm IS STM Lens. Video files were saved in MP4 format. The camera was placed around 1.5 meters from the speaker and zoomed as needed to only fit the speaker's face, upper half of his\her body, and the green screen cloth. For audio recording, Marantz PMD 670 recorder has been used. Audio files were saved in wave format. The microphone was placed approximately 50 cm from the speaker. Table 4.3 lists the camera and the recorder settings used in all the recording sessions. Speakers were illuminated using six ceiling mounted studio lighting rig. The lightning levels were adjusted to provide full spectrum lighting and to reduce facial and body shadows. The prompting material was displayed on a 13 inch MacBook Air and placed on a table 1 meter from the speaker. Each speaker audio file

Camera Settings	
ISO	500
Frame rate	25 fps
Shutter speed	25
Aperture	5.6
Recorder Settings	
Recording levels	4
Sampling rate	16000Hz
Audio channel	Mono

Table 4.3 Camera and recorder settings.

was exported from the recorder and sentence segmentation was performed using Audacity software (http://audacity.sourceforge.net).

4.2.8 Description of DEED files

The DEED contains 1680 recordings from 12 male and 9 female speakers with typical speech and 320 recordings from 1 male and 3 female speakers with dysarthric speech. All speakers produced 80 spoken utterances.

DEED filename convention

A unique filename has been given to each file in the DEED. Each filename consists of 9 alphanumeric identifiers (e.g., An01TS02M). Each 2-4 part characters represent a different information. Positions 1-2 represent the emotion. Each emotion has been identified using the first 2 characters as follows: Ne:Neutral, An:Anger, Di:Disgust, Fe:Fear, Ha:Happiness, Sa:Sadness, and Su:Surprise. Positions 3-4 represent the sentence number. The DEED has 80 sentences, therefore, sentences are identified using numbers from 01 to 80. Positions 5-8 represent the speaker ID, where TS identifies a speaker with typical speech and DS identifies a speaker with dysarthric speech. The following two digits number represent the speaker number. Position 9 represents the gender of the speaker where M indicates a male speaker and F indicates a female speaker.

Download and accessibility

Since one of the aims of developing the DEED is to provide the research community with a validated parallel database of typical and dysarthric emotional speech, it will be available

free of charge subject to non-commercial use. Instructions on how to access and download the database will be shared soon.

4.3 Discussion and conclusion

This chapter described the development and design of DEED. DEED has a number of important features that we believe that scientists and clinicians will be interested in. It is a parallel database of typical and dysarthric emotional speech which allows direct comparison and analysis to be made between those two types of speech. To the best of our knowledge, there are no available database with this parallelism feature of typical and dysarthric emotional speech nor dysarthric emotional speech database. It is a multimodal database which allows the integration and analysis of multiple information from different channels, audio and video. DEED also considered to be relatively large in its scope in terms of number of speakers and number of sentences. All design methodologies has been selected carefully to best serve the purpose of this database. The database is going to be freely available for the research community.

In order to determine the adequacy and quality of the DEED recordings, a subjective evaluation is performed. The methodology and the results are presented and discussed in details in Chapter 6. Also, an automatic speech emotion recognition (SER) baseline system is developed for both parts of the DEED, dysarthric and typical parts. The development and results are presented and discussed in Chapter 7. The next chapter will investigate the ability of speakers with dysarthria to control some acoustic features while communicating emotions.

Chapter 5

Acoustic Differences in Emotional Speech of People with Dysarthria

The content of this chapter has been published in the in Speech Communication Journal (Alhinti et al., 2021).

5.1 Introduction

People with speech disorders, such as dysarthria, can find it difficult to communicate with unfamiliar conversation partners. There is evidence that people can quickly adapt to speech from an unfamiliar person with a speech disorder, but often people can still find such speech difficult to understand (Borrie et al., 2012, 2017). We know that some of this difficulty comes from the fact that people with a speech disorder are sometimes unable to produce speech sounds accurately, or use typical intonation. People with dysarthria can struggle to be understood in conversation, not only because the intelligibility of their words is affected, but also because their paralinguistic information can be limited.

This chapter investigates to what extent people with dysarthria caused by cerebral palsy and Parkinson's disease (PD) are able to communicate emotions and how their acoustic signalling of emotions might differ to that used by typical speakers.

It has been shown that emotions have a direct effect on vocal production (Scherer et al., 1980). Different emotions are usually indexed by specific acoustic characteristics. Physiological changes that result from being in a particular emotional state affect the phonation, respiration, and articulation in a way that creates specific acoustic characteristics for each emotion (Banse and Scherer, 1996; Scherer, 2003). Although the exact details of which acoustic parameters are affected and what changes on these acoustic parameters occur while being in a certain emotion are still not very clear, there are however, some features that

usually change in emotional speech such as the fundamental frequency (F0) and the energy (Banse and Scherer, 1996). The analysis of the F0 (minimum, maximum, range, mean, etc.) has been included in most of the research on emotional speech. F0 statistics are one of the features that correlate with emotional vocal expressions. Higher F0 is usually associated with high-arousal emotions such as 'angry' and 'happy', while lower F0 is more associated with low-arousal emotions such as 'sad' (Breitenstein et al., 2001; Guo et al., 2016; Johnstone and Scherer, 2000). Also, 'happy' and 'angry' are found to be associated with a very wide range of F0 values compared to neutral speech while 'sad' is found to be associated with a less wide range of F0 values (Guo et al., 2016; Murray and Arnott, 1993). Research has also shown that higher energy is usually associated with high-arousal emotions such as 'angry' and 'happy', while lower energy is more associated with low-arousal emotions such as 'sad' (Johnstone and Scherer, 2000). Scherer (2003) and Johnstone and Scherer (2000) have presented more detailed results on the effect of different emotions on selected acoustic parameters. All of these studies were conducted on typical speech.

As have been discussed in Chapter 2 Section 2.2.3, comparing acoustic differences between speakers with dysarthria caused by either cerebral palsy or PD and typical (healthy control) speakers when performing different prosodic tasks have been investigated before. Few studies as well *perceptually* assessed the ability of speakers with dysarthria caused by PD to express emotions through speech and compared it to those with typical speech. The results of these studies did not tally with each other (see Chapter 2 Section 2.2.3 for more details on these studies). Despite having speech which is less intelligible, many studies show that even with the limited phonological and prosodic dimensions, many people with dysarthria have enough control to signal prosodic contrast on different tasks.

Since the ability to express emotions through speech, to the best of our knowledge, has never been investigated in dysarthric speech caused by cerebral palsy and only perceptually assessed in dysarthric speech caused by PD, this chapter investigates which acoustic characteristics people with dysarthria use to signal the different emotions, and if these are different to typical speakers. The purpose of this study is to answer the following questions. First, can people with dysarthria due to cerebral palsy or PD make systematic changes to their speech to convey their emotional state? Second, if they are able to make such changes are these similar to those made by speakers with typical speech?

The rest of this chapter is structured as follows. Section 5.2 describes the adopted methodology. Section 5.3 presents the results. A discussion of the results is presented in Section 5.4. Finally, Section 5.5 contains the conclusion.

Speakers with dysarthria				
Type of dysarthria	**Speaker**	**Gender**	**Age**	**Dysarthria severity [*] / time diagnosed**
Spastic dysarthria (cerebral palsy)	DS01F	Female	65 years	Severe / from birth
Hypokinetic dysarthria (PD)	DS02F	Female	71 years	Mild / 10 years
	DS04F	Female	68 years	Mild-to-moderate / 10 years
	DS03M	Male	66 years	Moderate / 9 years

Speakers with typical speech				
Gender	**Number of Speakers**	**Age**		
		Mean	**SD**	**Range**
Female	9	34.00	13.26	20-56
Male	12	35.67	16.81	19-70
Close in age female	1	56.00	-	65
Close in age male	2	66.00	5.66	62-70

[*] The dysarthria severity levels indicated in the table are informal judgments by the authors.

Table 5.1 Speaker's description including typical speakers close in age.

5.2 Methodology

5.2.1 Data

This study was carried out on the dysarthric and typical speech parts of the Dysarthric Expressed Emotion Database (DEED). All the details about the database in terms of the speakers, emotions, and recording settings can be found in Chapter 4. Table 5.1 presents the details of the speakers. It is a repetition of Table 4.2 with additional information on typical speakers who are close in age to the speakers with dysarthria.

5.2.2 Selection of emotions

In this study, a subset of the basic emotions recorded in DEED has been included, namely, 'angry', 'happy', and 'sad'. 'Neutral' state has also been included as a baseline condition. The selection of these emotions was guided by several points:

- Given that this is a first of its kind study, starting with a smaller non-overlapping set can provide the base for a more focused initial exploration of the problem. In particular, this can allow us to answer the main question of whether or not some people with dysarthria can convey emotions in their speech.

- Based on the survey conducted in order to understand emotion communication by people with dysarthria presented in Chapter 3, 'anger, 'happiness' and 'sadness' were chosen by people with dysarthria as the most important emotions in terms of being able to communicate them successfully (Alhinti et al., 2020a).
- This set of emotions are widely adopted in the literature when performing acoustic analysis on typical speech (Davletcharova et al., 2015; Kumbhakarn and Sathe-Pathak, 2015; Yildirim et al., 2004).
- 'Neutral' was included in this study to be able to compare how different emotions affect speech compared to the neutral state.

5.2.3 Acoustic analyses

A total of 50 utterances per speaker were included in the analysis, with each emotion consisting of 10 utterances, except for the 'neutral' which has 20 utterances. The acoustic features investigated in this analysis are: RMS energy, F0, speech rate, jitter, shimmer, and harmonic to noise ratio (HNR). All the acoustic features were extracted using the Praat tool (Boersma and Weenink, 2019), except for the RMS energy, which was extracted using Librosa, a python package for music and audio analysis (McFee et al., 2015). Default settings were chosen for all parameters unless specified otherwise. The choice of these features was guided by several points: 1) these features are among the most important and relevant features that show correlations with different vocal emotions expressions (Kim et al., 2013; Kumbhakarn and Sathe-Pathak, 2015; Laukka et al., 2005; Schuller et al., 2005; Toivanen et al., 2006; Yildirim et al., 2004), 2) all or part of these features are widely adopted in the literature with success for tasks related to analyzing the acoustic characteristics of emotional speech (Kim et al., 2013; Kumbhakarn and Sathe-Pathak, 2015; Schuller et al., 2005; Toivanen et al., 2006; Yildirim et al., 2004), and 3) all or part of these features have been included in some standardized sets developed for related tasks (Eyben et al., 2016; Schuller et al., 2009, 2013). Given that the purpose of this analysis is to mainly see whether the groups under study have enough control to communicate emotions objectively through their voices or not, and to see how different their way is, compared to the typical speech control group, it is sufficient to start with a minimal set of potential acoustic features.

RMS energy

The root mean square energy is a common way to calculate the energy in a speech signal. It is calculated as the square root of the average sum of the squares of the amplitude of the

signal samples. Research showed that high energy is usually associated with high-arousal emotions such as 'angry' and 'happy', while low energy is more associated with low-arousal emotions such as 'sad' (Johnstone and Scherer, 2000). The RMS energy of each utterance was computed using the utterance spectrogram with the following settings: 25 ms frame size and 10 ms overlap.

Fundamental frequency

Pitch is one of the most important perceptual features of sound that mainly depends on a sound's frequency and F0 (Plack et al., 2014). The analysis of the F0 including minimum, maximum, range, mean, has been included on most of the research on emotional speech. F0 statistics are one of the most important features that correlate with emotional vocal expressions. Higher and wider range of F0 is usually associated with high- arousal emotions such as 'angry' and 'happy' compared to neutral speech while lower and less wider range of F0 is more associated with low-arousal emotions such as 'sad' (Breitenstein et al., 2001; Guo et al., 2016; Johnstone and Scherer, 2000; Murray and Arnott, 1993). In this study, the F0 contour and related F0 statistics were computed using the autocorrelation method through the To Pitch command in Praat with the following pitch range settings: from 60 to 500 Hz. The two statistics that have been analyzed under this feature are the F0 mean and range. For each utterance, the range of F0 was calculated by subtracting the minimum F0 from the maximum F0 values.

Speech rate

Speech rate is determined by the number of syllables spoken per time unit. It is an important feature that has been used in different tasks such as determining fluency in second language learning and determining the speaker's emotional states. Research showed that speech rate has correlation with vocal arousal (Harrigan et al., 2008). The experiment reported by Breitenstein et al. (2001) showed that slow speech rate is associated with 'sad' emotion while fast speech rate is associated with 'angry' and 'happy'.

Speech rate per utterance was calculated using a Praat script where the syllable boundaries are estimated using energy-based syllable-nuclei detection method (De Jong and Wempe, 2009).

Jitter

In periodic signals, jitter shows how the signal deviates from its true periodicity. It is a measure of the fundamental frequency variations from cycle to cycle. There are several types

of jitter measurements. In this analysis, the jitter local absolute (known as jitta) was chosen. It is the average absolute difference between consecutive periods represented in seconds and was computed using the Get jitter (local, absolute) command in Praat.

Shimmer

In periodic signals, shimmer shows the cycle to cycle variations of amplitude. There are also several types of shimmer measurements. In this analysis the shimmer local (dB) was chosen. It represents the difference in peak to peak amplitude in decibels. It was computed using the Get Shimmer (local_dB) command in Praat.

Research has shown that jitter and shimmer are important features in emotion classification (Harrigan et al., 2008; Hossain and Naznin, 2018; Li et al., 2007). Whiteside (1998) found that high jitter and shimmer are associated with high-arousal emotions, such as 'angry', while low levels are associated with low-arousal emotions such as 'sad'.

Harmonics-to-noise ratio (HNR)

HNR is a measure of the additive noise in the voice signal. It is a useful feature to measure the breathiness and roughness (hoarseness) of a voice (Krom, 1995). Research showed that HNR has higher values in negative emotions such as 'anger' compared to the 'neutral' state (Alter et al., 1999). HNR values were computed using Praat To Harmonicity (cc) command with minimum pitch set to 60 Hz.

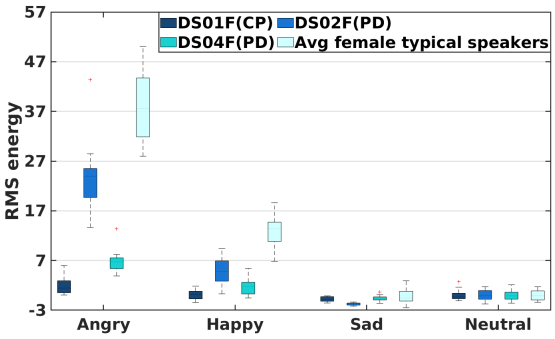
5.3 Results

Acoustic analysis was performed on all 200 utterances (50 utterances x 4 speakers) produced by speakers with dysarthria and on all 1050 utterances (50 utterances x 21 speakers) produced by speakers with typical speech. For each feature, a linear multi-level model was used to analyse the data. The feature being analysed was the response variable. The fixed factors were the type of speech, (hereinafter referred to as condition; typical speech (TS), dysarthric speech associated with cerebral palsy (CP), dysarthric speech associated with Parkinson's disease (PD)), gender (female, male), and emotion ('angry', 'happy', 'sad', 'neutral'). The interaction between these fixed factors was computed, with speaker identity and sentence as random factors. Using estimated marginal means, pairwise comparison for the main effects and their interaction were conducted on each feature where the p values were adjusted using the Bonferroni method. Since F0 and RMS energy are known to differ between male and female speakers (Biemans, 2000; Chen et al., 2020a; Izadi et al., 2012; Mendoza et al., 1996;

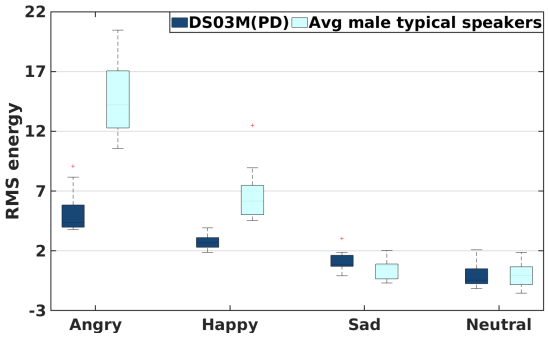
Teixeira and Fernandes, 2014), the analyses of the related features were done separately in both plots and statistical models where gender was added in the interaction terms. As normal aging can affect some acoustic characteristics of speakers, we also compared speakers with dysarthria to closely age-matched speakers with typical speech by plotting the results.

5.3.1 RMS energy

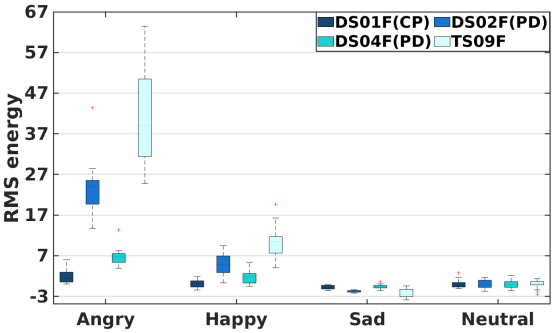
Figure 5.1 shows the boxplot of the RMS energy for female and male speakers after standardization. The RMS energy was standardized using the average energy of the 'neutral' state of each speaker/ group of speakers. The standardization was done to remove any effect of recording differences that can occur such as possible distance differences between speakers and the microphone. Figures 5.1a and 5.1b, show that female speakers with dysarthria have lower RMS energy compared to the average for female typical speakers in all of the three emotions. This is also the case for the male speaker with dysarthria in 'angry' and 'happy' emotions. Although there is a difference in the range of energy produced by speakers with dysarthria compared to typical speakers, all speakers seemed to have the ability to vary energy when trying to communicate different emotional states. Table 5.2 illustrates the pairwise comparison for the main effects of condition (TS, CP, PD), gender (female, male), and emotions ('angry', 'happy', 'sad', 'neutral') on RMS energy corrected using Bonferroni adjustment. The table indicates that the main effect of condition reflects a significant difference ($p < 0.001$) between TS and CP and a significant difference ($p < 0.01$) between TS and PD, while the difference between CP and PD is not significant. In addition, the difference between females and males is not significant. The differences between all pairs of emotions are significant except between 'neutral'/'sad'. A significant difference ($p < 0.001$) between 'neutral'/'angry', 'angry'/'happy', and 'angry'/'sad', ($p < 0.01$) between 'neutral'/'happy', and a significant difference ($p < 0.05$) between the pairwise comparison for the interaction effect of gender, condition, and emotion on RMS energy corrected using Bonferroni adjustment. The Table shows a significant difference ($p < 0.001$) between all pairs of emotions except between 'neutral'/'sad' for both female and male typical speakers. In addition, a significant difference ($p < 0.001$) between 'neutral'/'angry','angry'/'happy', and 'angry'/ 'sad' for female speakers with PD. Although there are differences between the means in the other two groups (female speaker with CP and male speaker with PD), having only one speaker in each group means differences would have to be large to be considered significant. However, overall observations can still be made for these groups. 'Angry' has the highest mean estimates of RMS energy for all groups.



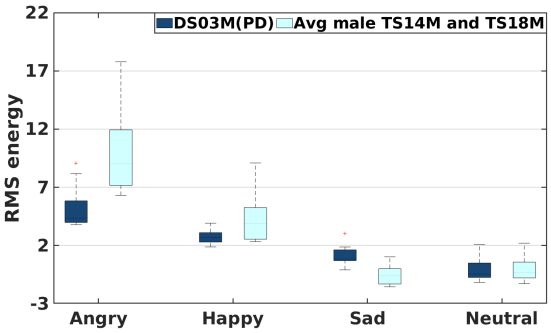
(a) RMS energy of female speakers with dysarthria and the average female typical speakers.



(b) RMS energy of a male speaker with dysarthria and the average male typical speakers.



(c) RMS energy of female speakers with dysarthria and a close in age female typical speaker.



(d) RMS energy of a male speaker with dysarthria and the average close in age male typical speakers.

Fig. 5.1 Boxplot of the RMS energy of female and male speakers.

Fixed Factor	(i)	(j)	Mean Difference (i-j)	SE	df	p value	95% Confidence Interval for Difference	
							Lower Bound	Upper Bound
Condition	TS	CP	6.045	1.335	1250.000	***	2.845	9.245
		PD	3.014	0. 848	1250.000	**	0.981	5.048
	CP	PD	-3.030	1.529	1250.000		-6.694	0.634
Gender	F	M	1.507	0.874	1250.000		-0.208	3.222
Emotions	N	A	-11.648	1.121	1250.000	***	-14.609	-8.686
		H	-4.041	1.121	1250.000	**	-7.002	-1.079
		S	-0.185	1.121	1250.000		-3.146	2.776
	A	H	7.607	1.294	1250.000	***	4.187	11.026
		S	11.648	1.294	1250.000	***	8.043	14.882
	H	S	3.856	0.808	1250.000	*	0.436	7.275

Interaction effects				Mean Difference (i-j)	SE	df	p value	95% Confidence Interval for Difference	
Gender	Condition	Emotion (i)	Emotion (j)					Lower Bound	Upper Bound
F	TS	N	A	-25.682	1.138	1250.000	***	-28.689	-22.676
			H	-8.928	1.138	1250.000	***	-11.935	-5.922
			S	-1.112	1.138	1250.000		-4.118	1.895
		A	H	16.754	1.314	1250.000	***	13.282	20.226
			S	24.571	1.314	1250.000	***	21.009	28.042
		H	S	7.817	1.314	1250.000	***	4.345	11.289
	CP	N	A	-1.966	3.414	1250.000		-10.986	7.054
			H	-0.184	3.414	1250.000		-9.204	8.836
			S	0.796	3.414	1250.000		-8.224	9.816
		A	H	1.782	3.942	1250.000		-8.634	12.197
			S	2.762	3.942	1250.000		-7.654	13.178
		H	S	0.980	3.942	1250.000		-9.435	11.396
	PD	N	A	-15.631	2.414	1250.000	***	-22.009	-9.253
			H	-3.306	2.414	1250.000		-9.684	3.073
			S	1.159	2.414	1250.000		-5.219	7.538
		A	H	12.325	2.787	1250.000	***	4.960	19.690
			S	16.790	2.787	1250.000	***	9.425	24.155
		H	S	4.465	2.787	1250.000		-2.900	11.830
M	TS	N	A	-9.671	0.985	1250.000	***	-12.275	-7.067
			H	-5.012	0.985	1250.000	***	-7.616	-2.408
			S	-0.659	0.985	1250.000		-3.263	1.945
		A	H	4.659	1.138	1250.000	***	1.652	7.666
			S	9.012	1.138	1250.000	***	6.006	12.019
		H	S	4.353	1.138	1250.000	**	1.347	7.360
	PD	N	A	-5.288	3.414	1250.000		-14.308	3.732
			H	-2.773	3.414	1250.000		-11.793	6.247
			S	-1.110	3.414	1250.000		-10.130	7.910
		A	H	2.515	3.942	1250.000		-7.901	12.930
			S	4.178	3.942	1250.000		-6.238	14.594
		H	S	1.663	3.942	1250.000		-8.752	12.079

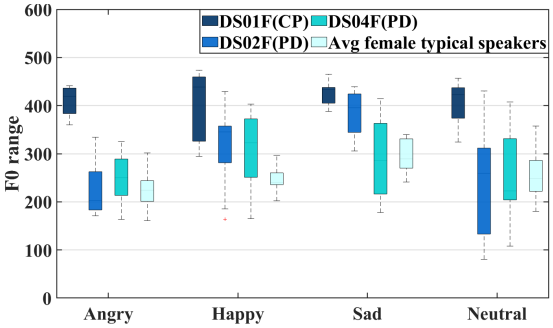
Table 5.2 Pairwise comparison of the estimated marginal means on RMS energy of the main effects and interaction effect of (gender*condition*emotion) using multilevel modeling. F/Female, M/Male, A/Anger, H/Happy, S/Sad, and N/ Neutral. (Where * = $p<0.05$, ** = $p<0.01$, *** = $p<0.001$).

5.3.2 Fundamental frequency

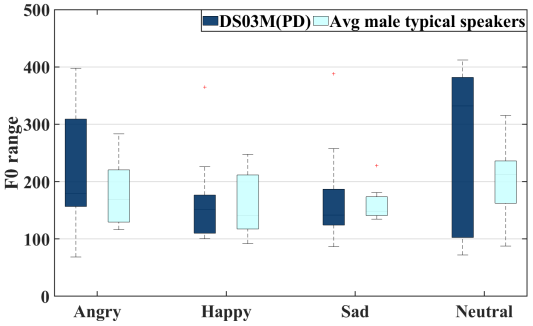
Two statistics of F0 were analysed, the F0 range and mean F0. Figures 5.2a, 5.2b, 5.2c, and 5.2d show the boxplot of F0 range and mean F0 per emotion for female and male speakers, respectively. Figures 3e and 3f show the boxplot of mean F0 per emotion for female and male speakers, respectively, in comparison to close in age typical speakers. Figures 5.3a and 5.3b show the F0 range of each utterance for the female speakers in the 'anger' and 'neutral' emotions, respectively.

Based on the conducted pairwise comparison for the main effects of condition, gender, and emotions on F0 range shown in Table 5.3, the following is observed: there is a significant difference ($p < 0.001$) between all conditions with higher mean estimates of F0 range for CP. The difference between females and males is significant ($p < 0.001$) with higher mean estimates for females. There is no significant difference between any pair of emotions. The interactions of gender, condition, and emotion on F0 range corrected using Bonferroni adjustment shown in Table 5.3, the difference between 'angry'/'sad' and 'happy'/'sad' is significant ($p < 0.01$) for female typical speakers. A significant difference ($p< 0.05$) between 'neutral'/'sad' and 'angry'/'sad' for female speakers with PD. The difference between 'neutral'/'happy' and 'neutral'/'sad' is significant ($p < 0.01$) for male typical speakers. There is no significant difference detected between any pair of emotions for the female speaker with CP and the male speaker with PD.

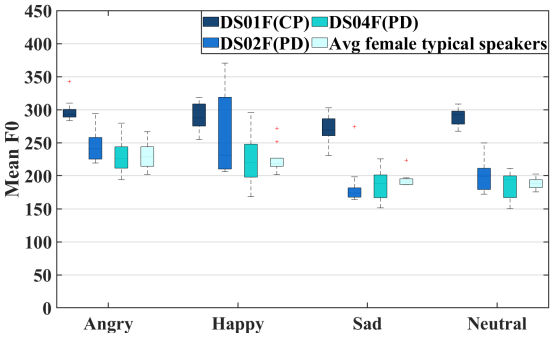
The pairwise comparison for the main effects of condition, gender, and emotions on mean F0 shown in Table 5.4 indicates a significant difference ($p < 0.001$) between TS and CP and between CP and PD with higher mean estimates of mean F0 for CP. The difference between females and males is significant ($p < 0.001$) with females having the highest marginal mean estimates. The differences between all pairs of emotions are significant ($p < 0.001$) except between ('neutral' and 'sad') and ('angry' and 'happy'). 'Angry' has the highest marginal mean estimates of F0 mean. Based on the conducted pairwise comparison for the interaction effect of gender, condition, and emotion on mean F0 presented in Table 4, the following is observed: there is a significant difference ($p < 0.001$) between all pairs of emotions except between 'neutral'/'sad' and 'angry'/'happy' for both female and male typical speakers and female speakers with PD. There is no significant difference detected between any pair of emotions for the female speaker with CP and the male speaker with PD except between 'neutral'/'angry' where a significant difference ($p < 0.001$) is found for the male speaker with PD.



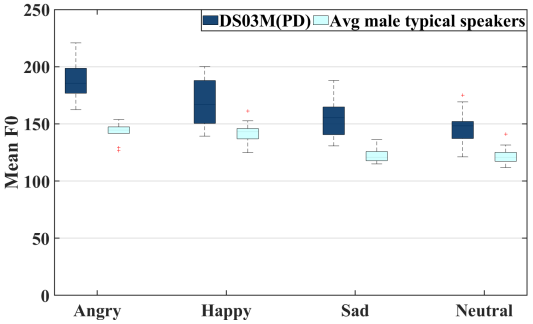
(a) F0 range of female speakers with dysarthria and the average female typical speakers.



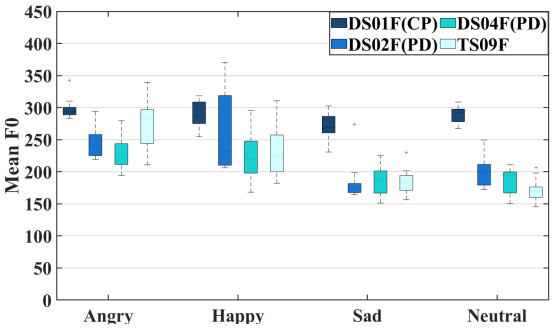
(b) F0 range of a male speaker with dysarthria and the average male typical speakers.



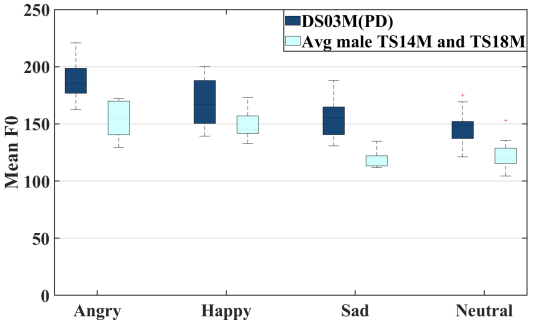
(c) Mean F0 of female speakers with dysarthria and the average female typical speakers.



(d) Mean F0 of a male speaker with dysarthria and the average male typical speakers.

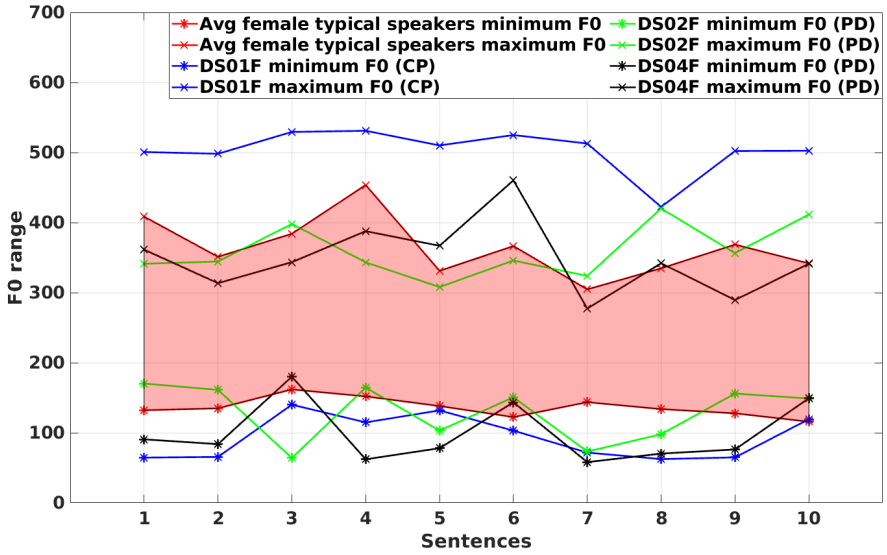


(e) Mean F0 of female speakers with dysarthria and a close in age female typical speaker.

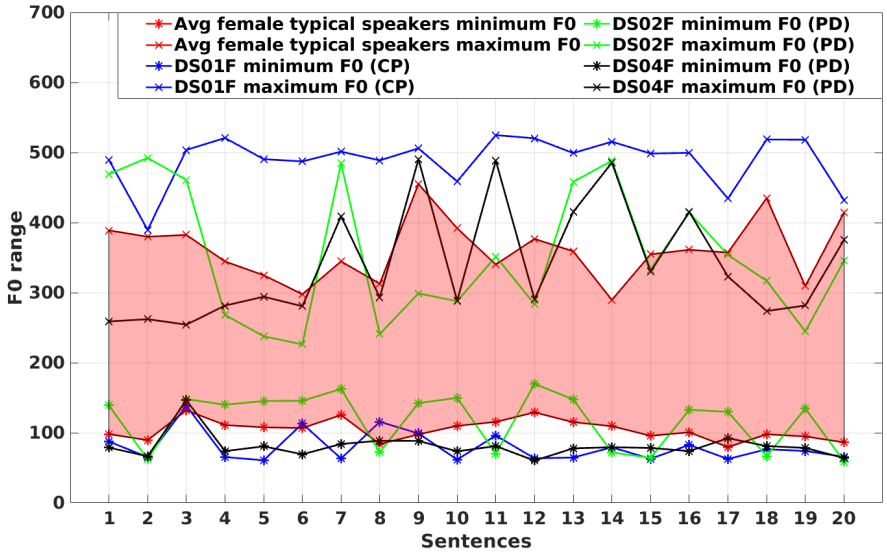


(f) Mean F0 of a male speaker with dysarthria and the average close in age male typical speakers.

Fig. 5.2 F0 range and mean of female and male speakers with dysarthria and typical speakers.



(a) F0 range of female speakers with dysarthria and the average female typical speakers in angry emotion.



(b) F0 range of female speakers with dysarthria and the average female typical speakers in neutral emotion.

Fig. 5.3 F0 range of female speakers in (a) anger and (b) neutral emotions.

Fixed Factor	(i)	(j)	Mean Difference (i-j)	SE	df	p value	95% Confidence Interval for Difference	
							Lower Bound	Upper Bound
Condition	TS	CP	-227.535	19.036	1192.910	***	-273.170	-181.900
		PD	-64.824	13.783	849.565	***	-97.885	-31.762
	CP	PD	162.711	20.546	1248.621	***	113.415	212.007
Gender	F	M	134.722	11.803	1249.624	***	111.566	157.878
Emotions	N	A	22.622	15.989	1245.913		-19.630	64.873
		H	18.830	15.4730	1249.972		-22.056	59.717
		S	-2.532	15.153	1248.843		-42.574	37.511
	A	H	-3.791	17.464	1248.579		-49.940	42.358
		S	-25.153	17.740	1249.938		-72.031	21.724
	H	S	-21.362	17.464	1248.579		-67.511	24.786

Interaction effects				Mean Difference (i-j)	SE	df	p value	95% Confidence Interval for Difference	
Gender	Condition	Emotion (i)	Emotion (j)					Lower Bound	Upper Bound
F	TS	N	A	32.267	16.207	1246.284		-10.559	75.093
			H	12.334	15.697	1249.986		-29.136	53.823
			S	-37.587	15.383	1248.815		-78.236	3.061
		A	H	-19.923	17.729	1248.556		-66.773	26.927
			S	-69.854	18.001	1249.918	**	-117.422	-22.286
		H	S	-49.931	17.729	1248.556	**	-96.781	-3.081
	CP	N	A	4.155	46.144	1248.803		-117.778	126.088
			H	7.455	45.967	1248.236		-114.012	128.922
			S	-15.221	45.861	1247.837		-136.407	105.965
		A	H	3.300	52.944	1247.799		-136.605	143.204
			S	-19.376	53.036	1248.103		-159.523	120.770
		H	S	-22.676	52.944	1247.799		-162.580	117.228
	PD	N	A	11.008	32.853	1249.537		-75.804	97.819
			H	-62.628	32.604	1248.701		-148.783	32.528
			S	-90.019	23.454	1247.977	*	-175.777	-4.260
		A	H	-73.635	37.459	1247.903		-172.620	25.349
			S	-101.027	37.588	1248.471	*	-200.353	-1.700
		H	S	-27.391	37.459	1247.903		-126.375	71.593
M	TS	N	A	26.912	14.294	1241.557		-10.861	64.685
			H	48.121	13.714	1249.605	**	11.882	84.360
			S	44.689	13.353	1249.102	**	9.405	79.974
		A	H	21.209	15.381	1248.796		-19.434	61.851
			S	17.777	15.693	1249.989		-23.691	59.246
		H	S	-3.431	15.381	1248.796		-44.074	37.211
	PD	N	A	38.766	46.144	1248.803		-83.167	160.700
			H	88.861	45.967	1248.236		-32.606	210.328
			S	58.478	45.861	1247.837		-35.707	206.664
		A	H	50.094	52.944	1247.799		-89.810	189.999
			S	46.712	53.036	1248.103		-93.434	186.858
		H	S	-3.382	52.944	1247.799		-143.287	136.522

Table 5.3 Pairwise comparison of the estimated marginal means on F0 range of the main effects and interaction effect of (gender*condition*emotion) using multilevel modeling. F/Female, M/Male, A/Anger, H/Happy, S/Sad, and N/ Neutral. (Where * = $p<0.05$, ** = $p<0.01$, *** = $p<0.001$).

Fixed Factor	(i)	(j)	Mean Difference (i-j)	SE	df	p value	95% Confidence Interval for Difference	
							Lower Bound	Upper Bound
Condition	TS	CP	-106.251	4.29	1237.413	***	-84.94	-64.633
		PD	-6.645	3.12	1111.983		-14.124	0.835
	CP	PD	99.606	4.624	1248.349	***	88.521	110.961
Gender	F	M	86.55	2.655	1249.043	***	81.343	91.758
Emotions	N	A	-31.906	3.598	1249.867	***	-41.415	-22.398
		H	-25.688	3.481	1249.643	***	-34.885	-16.491
		S	-1.227	3.408	1248.455		-7.778	10.232
	A	H	6.218	3.927	1248.301		-4.159	16.595
		S	33.133	3.99	1249.337	***	22.589	43.677
	H	S	26.915	3.927	1248.301	***	16.538	37.293

Interaction effects				Mean Difference (i-j)	SE	df	p value	95% Confidence Interval for Difference	
Gender	Condition	Emotion (i)	Emotion (j)					Lower Bound	Upper Bound
F	TS	N	A	-41.767	3.647	1249.9	***	-51.405	-32.13
			H	-36.942	3.531	1249.612	***	-46.273	-27.612
			S	-5.882	3.459	1248.438		-15.023	3.259
		A	H	4.825	3.987	1248.288		-5.71	15.36
			S	35.885	4.049	1249.306	***	25.186	46.584
		H	S	31.06	3.987	1248.288	***	20.525	41.595
	CP	N	A	-9.317	10.377	1248.43		-36.736	18.103
			H	1.295	10.336	1248.111		-26.018	28.609
			S	18.398	10.312	1247.902		-8.851	45.647
		A	H	10.612	11.905	1247.883		-20.846	42.07
			S	27.715	11.926	1248.04		-3.799	59.228
		H	S	17.103	11.905	1247.883		-14.355	48.56
	PD	N	A	-45.319	7.388	1248.932	***	-64.843	-25.796
			H	-48.718	7.332	1248.371	***	-68.092	-29.343
			S	6.79	7.297	1247.974		-12.493	26.074
		A	H	-3.398	8.423	1247.936		-25.655	18.859
			S	52.11	8.452	1248.24	***	29.775	74.445
		H	S	55.508	8.423	1247.936	***	33.251	77.765
M	TS	N	A	-21.038	3.218	1249.317	***	-29.54	-12.535
			H	-20.775	3.085	1249.871	***	-28.928	-12.622
			S	-1.49	3.003	1248.616		-9.425	6.444
		A	H	0.263	3.459	1248.427		-8.876	9.403
			S	19.547	3.53	1249.602	***	10.219	28.875
		H	S	19.248	3.459	1248.427	***	10.145	28.424
	PD	N	A	-42.089	10.377	1248.43	***	-69.509	-14.67
			H	-23.302	10.336	1248.111		-50.615	4.012
			S	-11.681	10.312	1247.902		-38.93	15.568
		A	H	18.788	11.905	1247.883		-12.67	50.246
			S	30.409	11.926	1248.04		-1.105	61.922
		H	S	11.621	11.905	1247.883		-19.837	43.079

Table 5.4 Pairwise comparison of the estimated marginal means on mean F0 of the main effects and interaction effect of (gender*condition*emotion) using multilevel modeling. F/Female, M/Male, A/Anger, H/Happy, S/Sad, and N/ Neutral. (Where * = p<0.05, ** = p<0.01, *** = p<0.001).

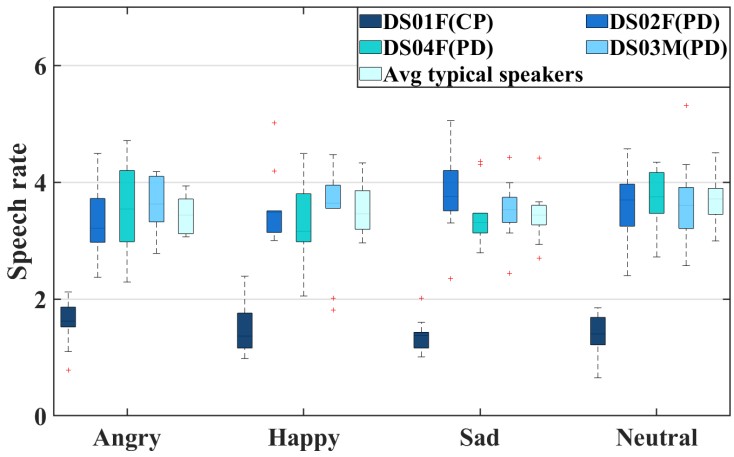


Fig. 5.4 Speech rate of speakers with dysarthria and the average typical speakers.

5.3.3 Speech rate

Figure 5.4 shows the boxplot of the speech rate for each emotion. The speech rate for speaker DS01F who has severe dysarthria caused by cerebral palsy is much slower than other speakers with dysarthria caused by Parkinson's disease and typical speakers. This is actually expected due to the nature and severity level of speaker DS01F as she takes much longer time to articulate. The pairwise comparison for the main effects of condition, gender, and emotions on speech rate presented in Table 5.5, indicates a significant difference ($p < 0.001$) between TS and CP and between CP and PD with CP having the lowest marginal mean estimates. The difference between females and males is significant ($p < 0.01$) with higher marginal mean estimates for comparison for the interaction effect of condition and emotions on speech rate presented in Table 5, indicates a significant difference ($p < 0.01$) between 'neutral'/'angry' and ($p < 0.001$) between 'neutral'/'sad' for the typical speakers group. There is no significant difference detected between any pair of emotions for the other groups.

5.3.4 Jitter

Figures 5.5a and 5.5b show the boxplot of the jitter local absolute feature of each emotion for female and male speakers, respectively. Figures 5.5c and 5.5d show the values of the jitter local absolute feature of each emotion between speakers with dysarthria and close in age typical speakers.

The pairwise comparison for the main effects of condition, gender, and emotions on jitter local absolute presented in Table 5.6, indicates the significant main effect of condition

Fixed Factor	(i)	(j)	Mean Difference (i-j)	SE	df	p value	95% Confidence Interval for Difference	
							Lower Bound	Upper Bound
Condition	TS	CP	2.249	0.112	1082.441	***	1.979	2.518
		PD	0.149	0.076	357.635		-0.033	0.332
	CP	PD	-2.099	0.12	1249.811	***	-2.387	-1.812
Gender	F	M	0.132	0.042	1207.415	**	0.05	0.214
Emotions	N	A	0.048	0.11	1225.386		-0.242	0.339
		H	0.073	0.108	1247.902		-0.211	0.358
		S	0.107	0.106	1249.267		-0.173	0.387
	A	H	0.025	0.122	1248.858		-0.298	0.348
		S	0.058	0.124	1249.552		-0.268	0.385
	H	S	0.033	0.122	1248.858		-0.29	0.356

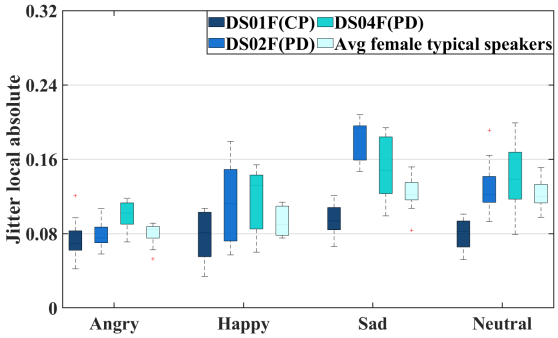
Interaction effects				Mean Difference (i-j)	SE	df	p value	95% Confidence Interval for Difference	
Gender	Condition	Emotion (i)	Emotion (j)					Lower Bound	Upper Bound
All	TS	N	A	0.219	0.66	1024.132	**	0.043	0.395
			H	0.119	0.62	1194.727		-0.046	0.284
			S	0.242	0.6	1249.214	***	0.84	0.399
		A	H	-0.1	0.069	1249.949		-0.281	0.081
			S	0.022	0.071	1221.308		-0.081	0.21
		H	S	0.123	0.69	1249.949		-0.059	0.304
	CP	N	A	-0.206	0.271	1249.679		-0.922	0.511
			H	-0.075	0.27	1248.628		-0.789	0.639
			S	0.024	0.27	1247.632		-0.688	0.737
		A	H	0.131	0.311	1247.529		-0.692	0.953
			S	0.23	0.312	1248.316		-0.593	1.054
		H	S	0.1	0.311	1247.529		-0.723	0.922
	PD	N	A	0.132	0.159	1247.614		-0.287	0.551
			H	0.176	0.157	1249.959		-0.238	0.591
			S	0.054	0.156	1248.335		-0.358	0.466
		A	H	0.044	0.18	1248.071		-0.431	0.52
			S	-0.078	0.181	1249.67		-0.555	0.4
		H	S	-0.122	0.18	1248.071		-0.597	0.353

Table 5.5 Pairwise comparison of the estimated marginal means on speech rate of the main effects and interaction effect of (condition*emotion) using multilevel modeling. A/Anger, H/Happy, S/Sad, and N/ Neutral. (Where * = p<0.05, ** = p<0.01, *** = p<0.001).

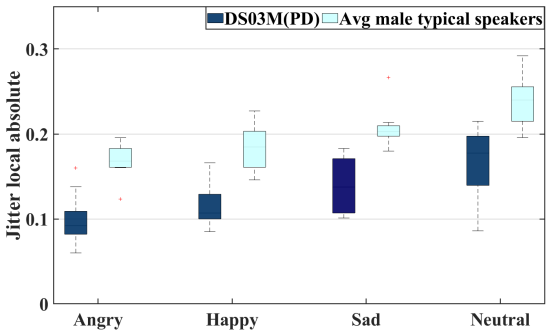
reflects a significant difference ($p < 0.001$) between TS and CP and between CP and PD and ($p < 0.05$) between TS and PD. CP has the lowest marginal mean estimates. The difference between females and males is significant ($p < 0.001$) with males having the highest marginal mean estimates. The differences between all pairs of emotions are significant ($p < 0.001$) except between ('neutral' and 'sad') and between ('angry' and 'happy'). Based on the pairwise comparison for the interaction effect of gender, condition, and emotion on jitter local absolute shown in Table 5.6, the following is observed: the difference is significant ($p < 0.001$) between all pairs of emotions for the female typical speakers group except between 'neutral'/'sad' and between 'angry'/'happy'. For the female speakers with PD, a significant difference ($p < 0.001$) is found between 'angry'/'sad', ($p < 0.01$) between 'neutral'/'angry' and 'happy'/'sad', and ($p<0.05$) between 'neutral'/'sad'. The differences between all pairs of emotions for the male typical speakers are significant except between 'angry'/'happy', with ($p<0.001$) for all the other pairs except between 'happy'/'sad' where ($p<0.01$). A significant difference ($p<0.01$) between 'neutral'/'angry' and ($p<0.05$) between 'neutral'/'happy' is found for the male speaker with PD. There is no significant difference found between any pair of emotions for the female speaker with CP.

5.3.5 Shimmer

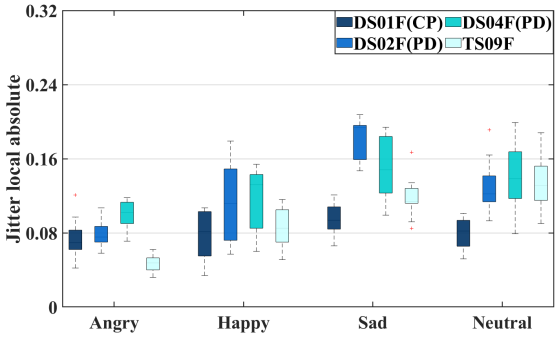
Figures 5.6a and 5.6b show the boxplot of the shimmer local feature in dB of each emotion for female and male speakers, respectively. The pairwise comparison for the main effects of condition, gender, and emotions on shimmer local shown in Table 5.7 indicates no significant difference of the main effect condition. The difference between females and males is significant ($p < 0.05$) with males having the highest marginal mean estimates. The differences between all pairs of emotions are significant except between 'neutral'/'sad' and between 'angry'/'happy', where ($p < 0.001$) for all the other pairs except between 'happy'/'sad' where ($p < 0.01$). From the pairwise comparison for the interaction effect of gender, condition, and emotion on shimmer local illustrated in Table 5.7 the following is observed: the differences between all pairs of emotions are significant ($p < 0.001$) for the female typical speakers except between 'neutral'/'sad' and between 'angry'/'happy'. A significant difference ($p<0.05$) between 'angry'/'sad' and between 'happy'/'sad' for the female speakers with PD. For the male typical speakers, a significant difference ($p < 0.001$) is found between 'neutral'/ 'angry', 'neutral'/'happy', and 'neutral'/'sad'. There is no significant difference found between any pair of emotions for the female speaker with CP and the male speaker with PD.



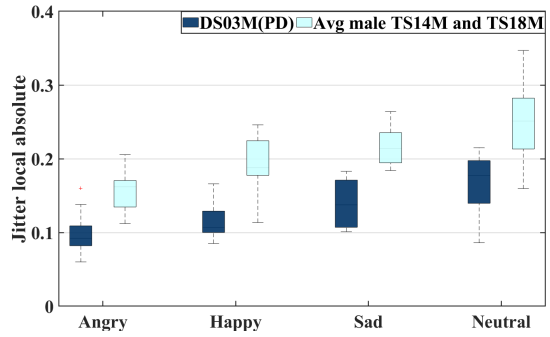
(a) Jitter values of female speakers with dysarthria and the average female typical speakers.



(b) Jitter values of a male speaker with dysarthria and the average male typical speakers.

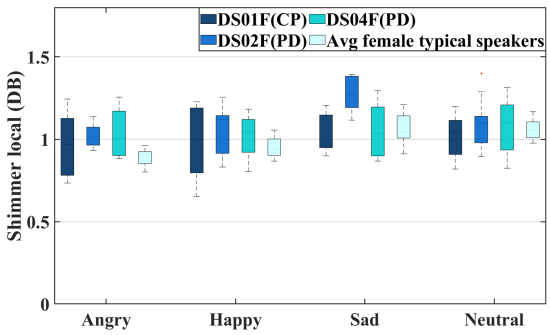


(c) Jitter values of female speakers with dysarthria and a close in age female typical speaker.

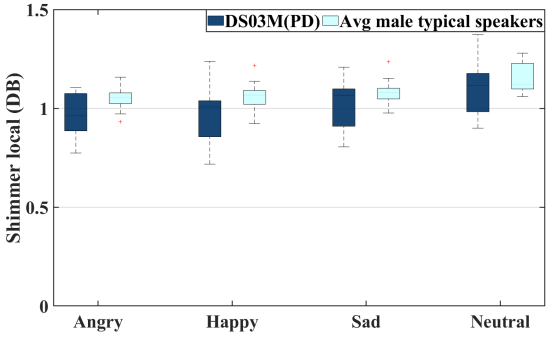


(d) Jitter values of a male speaker with dysarthria and the average close in age male typical speakers.

Fig. 5.5 Jitter values of (a) female and (b) male speakers.



(a) Shimmer values of female speakers with dysarthria and the average female typical speakers.



(b) Shimmer values of a male speaker with dysarthria and the average male typical speakers.

Fig. 5.6 Shimmer values in DB of (a) female and (b) male speakers

Fixed Factor	(i)	(j)	Mean Difference (i-j)	SE	df	p value	95% Confidence Interval for Difference	
							Lower Bound	Upper Bound
Condition	TS	CP	0.065	0.008	729.724	***	0.046	0.084
		PD	0.016	0.006	165.417	*	0.002	0.029
	CP	PD	-0.049	0.009	1249.518	***	-0.07	-0.029
Gender	F	M	-0.059	0.005	1248.585	***	-0.069	-0.050
Emotions	N	A	0.044	0.007	1081.311	***	0.026	0.061
		H	0.030	0.006	1218.991	***	0.013	0.047
		S	0.000	0.006	1249.988		-0.016	0.017
	A	H	-0.014	0.007	1249.839		-0.033	0.006
		S	-0.043	0.007	1235.86	***	-0.061	-0.024
	H	S	-0.030	0.007	1249.839	***	-0.049	-0.011

Interaction effects				Mean Difference (i-j)	SE	df	p value	95% Confidence Interval for Difference	
Gender	Condition	Emotion (i)	Emotion (j)					Lower Bound	Upper Bound
F	TS	N	A	0.040	0.007	1090.147	***	0.023	0.058
			H	0.028	0.006	1221.245	***	0.011	0.045
			S	-0.002	0.006	1249.998		-0.018	0.015
		A	H	-0.012	0.007	1249.802		-0.032	0.007
			S	-0.042	0.007	1237.039	***	-0.061	-0.022
		H	S	-0.03	0.007	1249.802	***	-0.049	-0.010
	CP	N	A	0.004	0.019	1250.000		-0.047	0.054
			H	0.002	0.019	1248.988		-0.048	0.053
			S	-0.015	0.019	1247.326		-0.065	0.036
		A	H	-0.001	0.022	1247.138		-0.059	0.056
			S	-0.018	0.022	1248.506		-0.076	0.040
		H	S	-0.017	0.022	1247.138		-0.075	0.041
	PD	N	A	0.044	0.014	1246.532	**	0.008	0.080
			H	0.019	0.013	1249.971		-0.017	0.055
			S	-0.037	0.013	1247.977	*	-0.072	-0.001
		A	H	-0.025	0.015	1247.643		-0.066	0.016
			S	-0.081	0.016	1249.64	***	-0.122	-0.040
		H	S	-0.056	0.015	1247.643	**	-0.097	-0.015
M	TS	N	A	0.065	0.006	992.912	***	0.050	0.081
			H	0.051	0.006	1193.446	***	0.036	0.066
			S	0.029	0.006	1249.538	***	0.014	0.044
		A	H	-0.014	0.006	1250.000		-0.031	0.003
			S	-0.036	0.006	1221.968	***	-0.053	-0.019
		H	S	-0.022	0.006	1250.000	**	-0.039	-0.005
	PD	N	A	0.065	0.019	1250.000	**	0.015	0.116
			H	0.051	0.019	1248.988	*	0.001	0.101
			S	0.026	0.019	1247.326		-0.025	0.076
		A	H	-0.015	0.022	1247.138		-0.072	0.043
			S	-0.04	0.022	1248.506		-0.098	0.018
		H	S	-0.025	0.022	1247.138		-0.083	0.033

Table 5.6 Pairwise comparison of the estimated marginal means on jitter local absolute of the main effects and interaction effect of (gender*condition*emotion) using multilevel modeling. F/Female, M/Male, A/Anger, H/Happy, S/Sad, and N/ Neutral. (Where * = p<0.05, ** = p<0.01, *** = p<0.001).

Fixed Factor	(i)	(j)	Mean Difference (i-j)	SE	df	p value	95% Confidence Interval for Difference	
							Lower Bound	Upper Bound
Condition	TS	CP	0.039	0.022	1250.000		-0.014	0.092
		PD	-0.011	0.014	1250.000		-0.044	0.023
	CP	PD	-0.050	0.025	1250.000		-0.110	0.011
Gender	F	M	-0.032	0.015	1250.000	*	-0.061	-0.003
Emotions	N	A	0.103	0.019	1250.000	***	0.054	0.152
		H	0.082	0.019	1250.000	***	0.033	0.132
		S	0.013	0.019	1250.000		-0.036	0.063
	A	H	-0.021	0.022	1250.000		-0.078	0.036
		S	-0.090	0.022	1250.000	***	-0.146	-0.033
	H	S	-0.069	0.022	1250.000	**	-0.126	-0.012

Interaction effects				Mean Difference (i-j)	SE	df	p value	95% Confidence Interval for Difference	
Gender	Condition	Emotion (i)	Emotion (j)					Lower Bound	Upper Bound
F	TS	N	A	0.165	0.019	1250.000	***	0.115	0.215
			H	0.110	0.019	1250.000	***	0.060	0.160
			S	0.004	0.019	1250.000		-0.046	0.054
		A	H	-0.054	0.022	1250.000		-0.112	0.003
			S	-0.161	0.022	1250.000	***	-0.219	-0.103
		H	S	-0.106	0.022	1250.000	***	-0.164	-0.049
	CP	N	A	0.050	0.057	1250.000		-0.101	0.200
			H	0.034	0.057	1250.000		-0.116	0.184
			S	-0.001	0.057	1250.000		-0.151	0.149
		A	H	-0.015	0.066	1250.000		-0.189	0.158
			S	-0.015	0.066	1250.000		-0.224	0.123
		H	S	-0.035	0.066	1250.000		-0.209	0.138
	PD	N	A	0.057	0.040	1250.000		-0.049	0.163
			H	0.045	0.040	1250.000		-0.610	0.151
			S	-0.085	0.040	1250.000		-0.191	0.021
		A	H	-0.012	0.046	1250.000		-0.135	0.110
			S	-0.142	0.046	1250.000	*	-0.264	-0.019
		H	S	-0.129	0.046	1250.000	*	-0.252	-0.007
M	TS	N	A	0.104	0.016	1250.000	***	0.060	0.147
			H	0.097	0.016	1250.000	***	0.054	0.140
			S	0.070	0.016	1250.000	***	0.027	0.114
		A	H	-0.006	0.019	1250.000		-0.056	0.044
			S	-0.033	0.019	1250.000		-0.083	0.017
		H	S	-0.027	0.019	1250.000		-0.077	0.023
	PD	N	A	0.140	0.057	1250.000		-0.010	0.290
			H	0.125	0.057	1250.000		-0.072	0.275
			S	0.078	0.057	1250.000		-0.072	0.228
		A	H	-0.015	0.066	1250.000		-0.188	0.159
			S	-0.062	0.066	1250.000		-0.235	0.112
		H	S	-0.047	0.066	1250.000		-0.220	0.126

Table 5.7 Pairwise comparison of the estimated marginal means on shimmer local of the main effects and interaction effect of (gender*condition*emotion) using multilevel modeling. F/Female, M/Male, A/Anger, H/Happy, S/Sad, and N/ Neutral. (Where * = $p<0.05$, ** = $p<0.01$, *** = $p<0.001$).

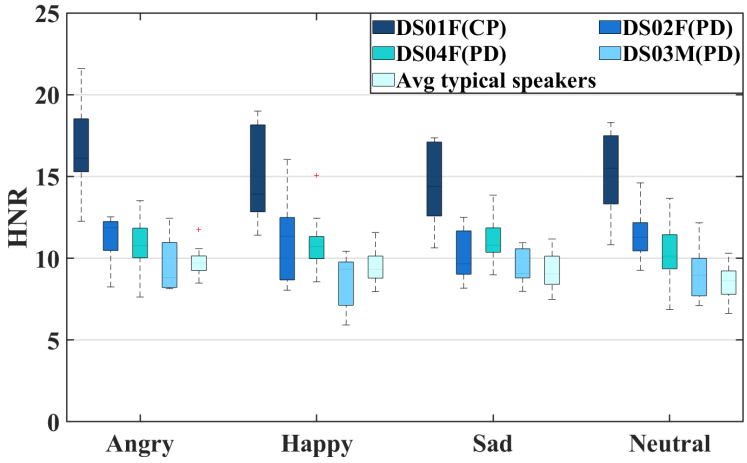


Fig. 5.7 HNR of speakers with dysarthria and the average typical speakers.

5.3.6 HNR

From Figure 5.7, it can be seen that speaker DS01F has higher levels of HNR compared to the other speakers with dysarthria caused by PD and the average typical speakers. From Table 5.8 we can see that based on the pairwise comparison for the main effects of condition, gender, and emotions on HNR, the significant main effect of condition reflects a significant difference ($p < 0.001$) between TS and CP and between CP and PD, but not between TS and PD. CP has the highest marginal mean estimates. The difference between females and males is significant ($p < 0.001$) with females having the highest marginal mean estimates. The differences between all pairs of emotions are not significant except between 'neutral'/'angry' and angry'/'sad' with ($p < 0.01$). From the pairwise comparison of the interaction effect of condition and emotion on HNR illustrated in Table 5.8, the following is observed: for typical speakers, the difference is significant ($p < 0.001$) between 'neutral'/'angry', 'neutral'/'happy', and 'neutral'/'sad' and ($p < 0.01$) between 'angry'/'sad' but not between 'neutral'/'sad' and 'happy'/'sad'. There is no significant difference between any pair of emotions for the speaker with CP and the speakers with PD, except between 'angry'/'sad' with ($p<0.05$) for the former.

5.4 Discussion

The results show that some people with dysarthria, even severe dysarthria, are able to control some aspects of the suprasegmental and prosodic features of their speech to communicate emotions.

Fixed Factor	(i)	(j)	Mean Difference (i-j)	SE	df	p value	95% Confidence Interval for Difference	
							Lower Bound	Upper Bound
Condition	TS	CP	-4.941	0.294	1243.882	***	-5.647	-4.236
		PD	-0.069	0.203	1108.300		-0.555	0.417
	CP	PD	4.872	0.312	1248.657	***	4.123	5.621
Gender	F	M	0.882	0.109	1249.583	***	0.668	1.096
Emotions	N	A	-0.998	0.287	1249.996	**	-1.757	-0.239
		H	-0.267	0.280	1249.338		-1.008	-0.474
		S	-0.046	0.276	1248.303		-0.684	-0.776
	A	H	0.731	0.319	1248.183		-0.111	1.573
		S	1.044	0.322	1249.042	**	0.193	1.896
	H	S	0.313	0.319	1248.183		-0.528	1.155

Interaction effects				Mean Difference (i-j)	SE	df	p value	95% Confidence Interval for Difference	
Gender	Condition	Emotion (i)	Emotion (j)					Lower Bound	Upper Bound
All	TS	N	A	-1.275	0.174	1244.435	***	-1.736	-0.814
			H	-1.070	0.163	1249.764	***	-1.501	-0.639
			S	-0.646	0.156	1249.13	***	-1.057	-0.235
		A	H	0.205	0.179	1248.848		-0.268	0.678
			S	0.629	0.185	1249.999	**	0.139	1.118
		H	S	0.424	0.179	1248.848		-0.049	0.897
	CP	N	A	-1.469	0.707	1248.459		-3.336	0.398
			H	0.244	0.704	1248.123		-1.615	2.104
			S	0.776	0.702	1247.903		-1.079	2.631
		A	H	1.713	0.811	1247.882		-0.429	3.855
			S	2.245	0.812	1248.049	*	0.099	4.390
		H	S	0.532	0.811	1247.882		-1.610	2.674
	PD	N	A	-0.251	0.414	1249.371		-1.344	0.842
			H	0.024	0.409	1248.645		-1.056	1.104
			S	0.009	0.406	1248.053		-1.064	1.081
		A	H	0.275	0.469	1247.994		-0.963	1.513
			S	0.260	0.471	1248.455		-0.985	1.504
		H	S	-0.015	0.469	1247.994		-1.254	1.223

Table 5.8 Pairwise comparison of the estimated marginal means on HNR of the main effects and interaction effect of (condition*emotion) using multilevel modeling. A/Anger, H/Happy, S/Sad, and N/ Neutral. (Where * = p<0.05, ** = p<0.01, *** = p<0.001).

Although no strong conclusions can be made and significant difference between emotions and groups can be difficult to observe due to the very limited number of speakers in some of the groups (1 speaker in some of them), we can still make observations. It is also important to note that the observations made for dysarthric speech may have high variance due to low number of available speakers in DEED. The changes to these features appear similar to those of typical speakers, despite speakers with dysarthria having a more limited articulatory and prosodic control. It is likely that these systematic changes help in the communication of the emotions.

One of the features used to distinguish emotion is the RMS energy. Typical speakers vary it while expressing the emotions investigated in this study, except between 'neutral'/'sad' utterances. Female speakers with PD also managed to vary it significantly when expressing some of the emotions. Similar to typical speakers, all speakers with dysarthria produced higher RMS energy when communicating high-arousal emotions such as 'angry' and 'happy' compared to low-arousal emotions such as 'sad' and 'neutral' [as can be seen from Figure 5.1 and from the statistical model presented in Table 5.2]. Despite that, the differences between some high-arousal emotions and low-arousal emotions were not marked as significant for some of the groups due to the limited number of observation. However, the differences between the marginal means are still observed and with more data, significancy might be confirmed. This aligns with the findings reported in the literature on typical speech (Johnstone and Scherer, 2000). There is a significant difference in the RMS energy between the group of speakers with dysarthria caused by cerebral palsy and the typical speakers, and between the group of speakers with dysarthria caused by PD and the group of typical speakers.

The range of F0 does not appear to be a strong distinguishing feature. A significant difference was found between the three groups of speakers in the range of F0.

The mean F0 is an important feature in distinguishing emotions. The mean F0 can be used to distinguish between high-arousal and low-arousal emotions as the difference differs significantly [as can be seen from Table 5.4]. This significant difference was also observed within each group with more than one speaker from the results of the interaction effect. This is also consistent with the findings in the literature for speakers with typical speech, where high F0 is usually associated with high-arousal emotions and low F0 is more associated with low-arousal emotions (Breitenstein et al., 2001; Guo et al., 2016; Johnstone and Scherer, 2000). From the effect of condition, a significant difference was found in the mean F0 between the speaker with dysarthria caused by cerebral palsy and the other two groups of speakers.

From our analysis, it is observed that the speech rate is not vary useful in distinguishing between pairs of emotions as the differences between pairs of emotion were not statistically

significant except between 'neutral'/'angry' and 'neutral'/'sad' for typical speakers. A significant difference was also found in the speech rate between the speaker with dysarthria caused by cerebral palsy and the other two groups of speakers.

The jitter local absolute is also a useful feature in distinguishing pairs of emotions. It can be used to distinguish between high-arousal and low-arousal emotions. The differences in the marginal means between these pairs of emotions were marked significantly within each group with more than one speaker. It is also observed that differences in the mean values of the jitter between the three groups of speakers: typical speakers, speakers with dysarthria caused by cerebral palsy and speakers with dysarthria caused by PD, were statistically significant.

Shimmer can be used to distinguish some pairs of emotions. As can be seen from the results of the interaction effect shown in Table 5.7, these pairs vary among groups. For example, while shimmer can be used for both female typical speakers and female speakers with dysarthria caused by PD to differentiate between 'anger'/'sad' and between 'happy'/'sad', where the mean difference between these emotions were found to be significant, in addition to other pairs of emotions for the group of female typical speakers, it can be used to distinguish neutral from all the other three emotions for the group of male typical speakers. There is no significant difference found for the effect of condition.

For typical speakers, the HNR feature appear to be sufficient to distinguish between pairs of emotions in our data except between 'angry'/'happy' and 'happy'/'sad'. No other significant difference was found in the HNR for the other groups except between 'angry'/'sad' for the speaker with severe dysrthria caused by cerebral palsy. From the effect of condition, a significant difference was found in the HNR between the speaker with dysarthria caused by cerebral palsy and the other two groups of speakers.

In this analysis, speaker DS01F, who has severe dysarthria due to cerebral palsy, has either higher values such as in F0 range and HNR or lower values such as in speech rate than the other female speakers with dysarthria caused by PD and the average female speakers with typical speech. This difference may be due to speaker DS01F having severe dysarthria, in contrast to the other speakers who have mild dysarthria. It is also observed that the characteristics of speakers with dysarthria caused by PD differ in some of the cases from those with typical speech. In addition, there is inter-speaker variation observed between the speakers with dysarthria caused by PD. This inter-speaker variation complies with the findings reported in the literature (Liu et al., 2019; Ma et al., 2010).

As the aim of this study is to know whether speakers with dysarthria have the ability to control some acoustic features while communicating emotions, some potential features have been analysed and the effect of a number of factors (condition, gender, and emotions) and their interaction on each feature has been investigated.

5.5 Conclusion

In this chapter, the ability of people with dysarthria, caused by cerebral palsy and PD, are able to communicate emotions in their speech has been investigated. A set of acoustic features of the two types of dysarthria under this study has been compared to those of typical speech. An analysis of the effect of different factors on each feature has been carried out. Although the conducted analysis has the limitations of having been carried out on a limited number of speakers with dysarthria and using a limited number of sentences, it does, however, show that these people may have enough control to communicate intentions, gain attention, and convey emotions. This level of control of articulatory and prosodic features may not only help to train listeners to better recognise the emotions of speakers with dysarthria, but also to improve communication aids in a way that makes it more sensitive to specific cues in the vocalization signal produced by the speaker with dysarthria and act according to the speaker's intention.

Classifying emotions from speech is by itself a challenging problem (El Ayadi et al., 2011). In the case of having disordered speech, this may be a more difficult classification problem as the speakers often have less control of the signifying features. The analysis presented in this chapter demonstrates the existence of significant differences between emotions in some of the investigated features. Yet, it is still unclear whether this level of difference is enough for people to accurately perceive these emotions. Assessing the ability of people with dysarthria to express emotions perceptually will be the focus of the next chapter.

Chapter 6

Subjective Evaluation of DEED

6.1 Introduction

DEED is a parallel database of typical and dysarthric emotional speech, the design and development were discussed in Chapter 4. This database will enable the investigation of automatically classifying emotions in dysarthric speech. However, it is important before that, to evaluate the DEED recordings subjectively. Obtaining the human performance on collected database is an approach that is followed in emotional typical speech. This will help in determining the task difficulty level for humans. It will also provide a benchmark for automatic emotion recognition models.

This chapter presents all the details of the subjective evaluation performed on DEED. Section 6.2 describes the evaluation methodology. Section 6.3 presents the results. A discussion of the results is presented in Section 6.4. Finally, Section 6.5 includes the conclusion.

6.2 Evaluation Methodology

The subjective evaluation approach of DEED has been ethically approved by the University of Sheffield, UK. Before any experiment, a written consent form has been obtained from every participant.

Participants

Twenty two normal hearing participants who are native speakers of British English or have lived in the UK for at least 1 year were recruited. Table 6.1 presents the participants' details. None of the participants were familiar with any of the speakers with dysarthria in DEED,

Demographic	Value			
Age	**Mean**	**SD**	**Range**	
	32.69	12.06	18-59	
Gender	**Female**		14 (63.64%)	
	Male		8 (36.36%)	
English proficiency	**Native**	**Lived in the UK for more than 5 years**	**Lived in the UK for 3-5 years**	**Lived in the UK for 1-2 years**
	9 (40.91%)	3 (13.64%)	5 (22.73%)	5 (22.73%)
Familiarity with the dysarthric speech	**Extremely familiar**	**Somewhat familiar**	**Slightly familiar**	**Not familiar at all**
	1 (4.55%)	5 (22.73%)	1 (4.55%)	15 (68.18%)

Table 6.1 Characteristics of participants in evaluation.

except one participant who were somewhat familiar with speaker DS01F, where they have met a few times in the past.

Stimuli, apparatus, and procedure

The evaluated stimuli consisted of the audio part of DEED. The selected stimuli included all the recordings of the speakers with dysarthria in addition to the recordings of 8 typical speakers who were randomly selected from the DEED-typical speech part (five female and three male). More female typical speakers were chosen in the evaluation as DEED-dysarthric speech part contains more female speakers than male. The randomly chosen typical speakers were: TS09F, TS13F, TS16F, TS17F, TS20F, TS06M, TS111M, and TS18M. Therefore, the stimuli consisted of a total of 960 audio recordings of emotional speech. Although, the main aim is to evaluate the dysarthric speech part of DEED, it was important to include recordings of typical speakers as a baseline assessment measure of the participants' evaluation level and to see where this database stands in comparison to previously published emotional typical speech databases.

The evaluation process began with a face-to-face approach, where participants were invited to the University of Sheffield and seated in a quiet room in front of a 13-inch MacBook Air laptop. Participants listened to the stimuli using a pair of headphones. However, due to the Coronavirus disease (COVID-19) pandemic and the lockdown imposed to stop the spread of the virus, it was not possible to carry on the evaluation using the same approach. Therefore, an online approach was proposed and adopted. Given the task in hand, the selection of the platform to carry out the evaluation was very critical as the audio files needs to be kept with

no distortion or modification while streamed. After comparing several platforms, the Zoom videoconferencing platform was selected (Zoom Video Communications Inc, 2016). Zoom allows a lot of flexibility in the audio settings to fit different needs. To make sure that the audio is heard by the recipients as it is without any modifications, these advanced settings were set:

- Disable automatically adjust audio volume.
- Enable original sound from microphone: This will turn off audio enhancements such as echo cancellation and noise suppression. This is a very important feature for audio streaming.
- Disable suppress persistence background noise.
- Disable suppress intermittent background noise.

People who teach vocals and music rely on these advanced settings as well (https://www.thenakedvocalist.com/zoom-for-singing-teachers/, https://www.makingmusic.org.uk/resource/zoom-online-rehearsals-vocal). In total, ten participants evaluated the data using the face-to-face approach, while the other twelve participated using the online approach.

Balanced evaluation sets were created, All the 7 emotions were included in the evaluation. For speakers with dysarthria, each speaker's recordings were divided into two equally sets in terms of the number of recordings resulting in having 40 recordings (utterances) per set. While for typical speakers, each speaker' recordings were divided into five equal sets in terms of the number of recordings resulting in having 16 recordings (utterances) per set.

The evaluation was carried out at utterance level. Each participant was presented with a chosen set of 288 stimuli as follows: 1 set (40 utterances) from each speaker with dysarthria recordings and 1 set (16 utterances) from each typical speaker's recordings except one participant who had evaluated the whole set of stimuli (960 utterances) and another participant who had evaluated the whole dysarthric stimuli (Set 1 + Set 2 for each speaker with dysarthria) in addition to one set from each typical speaker's recordings. Each participant was presented with a set from each speaker separately. Within each set, there is a training set and an evaluation set. To remove the systematic bias from the responses of the participants, the order of utterance in each set was randomised such that each participant was presented with each speaker's utterances in an order that is different from other participants. For example: participant 1 could be presented first with utterance 11 followed by utterance 45, and so on, from speaker DS01F while participant 2 was presented first with utterance 02 followed by utterance 72, and so on from the same speaker. Figure 6.1 illustrates the division of the data.

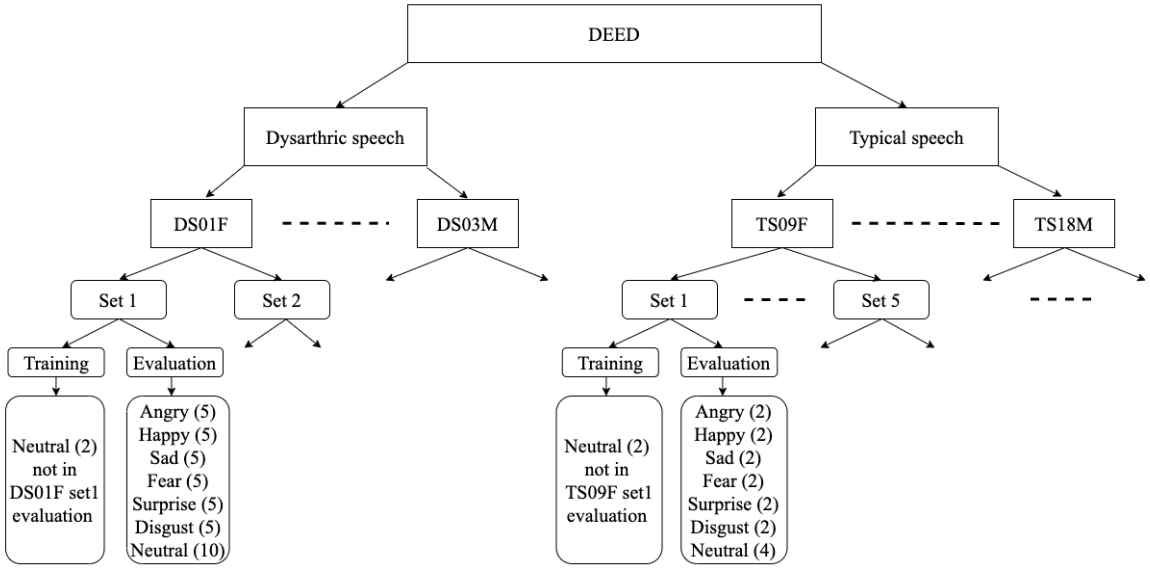
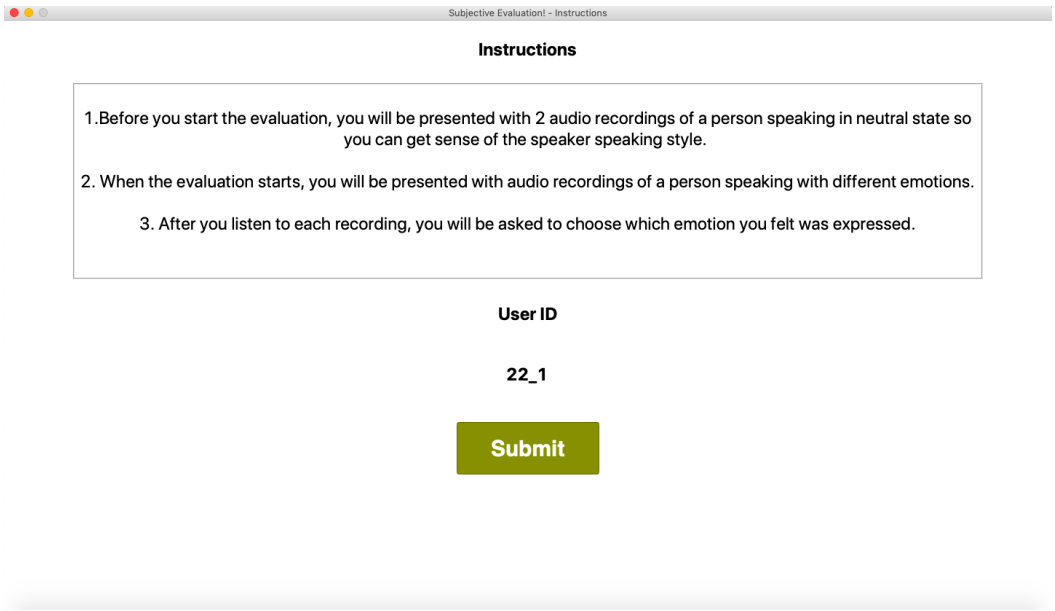


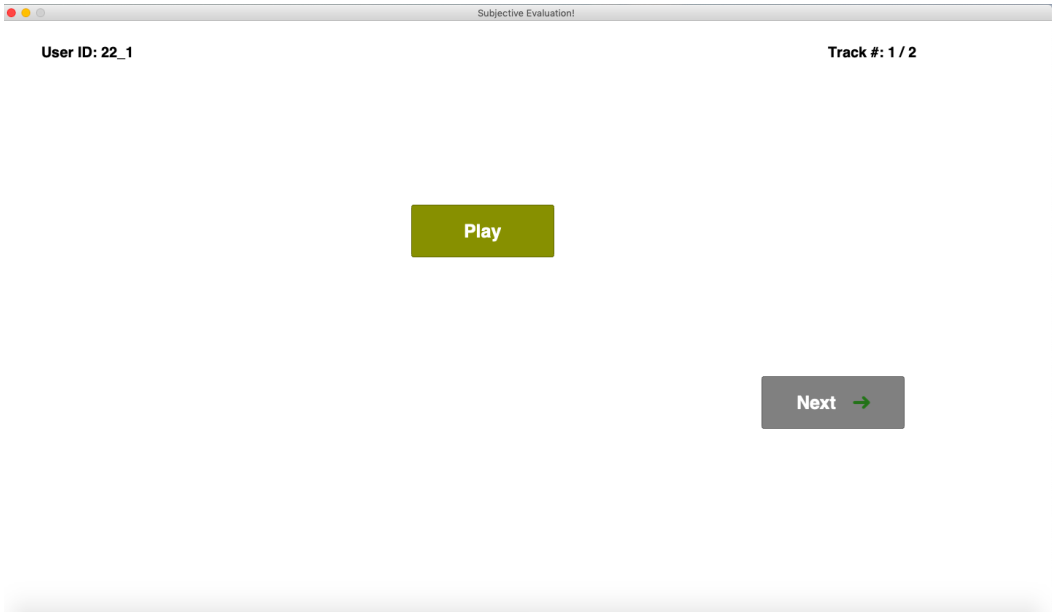
Fig. 6.1 The division of DEED into sets for the purpose of evaluating the data subjectively. (The number between brackets indicates the number of utterances.)

Evaluation task

In addition to the verbal instructions, on-screen instructions were presented as well on both approaches as follows: "1.Before you start the evaluation, you will be presented with 2 audio recordings of a person speaking in a neutral state so you can get a sense of the speaker's speaking style. 2.When the evaluation starts, you will be presented with audio recordings of a person speaking with different emotions. 3.After you listen to each recording, you will be asked to choose which emotion you felt was expressed. The instructions screen is shown in Figure 6.2.a. As mentioned in the instructions, to help participants get used to the speaker style of speaking, participants were trained using two recordings from that speaker speaking in the neutral emotion before the evaluation of each set began. These two neutral recordings were not among the evaluation set. The training screen is shown in Figure 6.2.b. Each recording was played only once. Participants were asked to choose an emotion using a forced-choice response format. The options were: angry, happy, sad, fear, disgust, surprise, and neutral. The options were disabled until the whole recording was played to insure that participants listened to the whole recording before making a choice. It also helped in preventing participants from moving quickly through some stimuli or skipping some. The response evaluation screen before and after playing the recording is shown in Figures 6.3.a and 6.3.b, respectively.

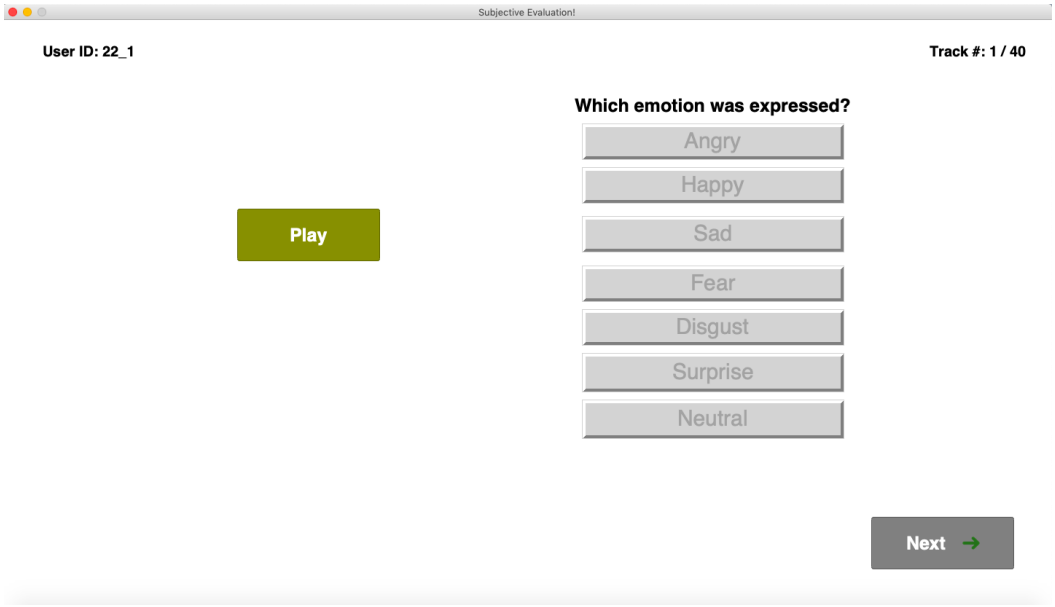


(a)

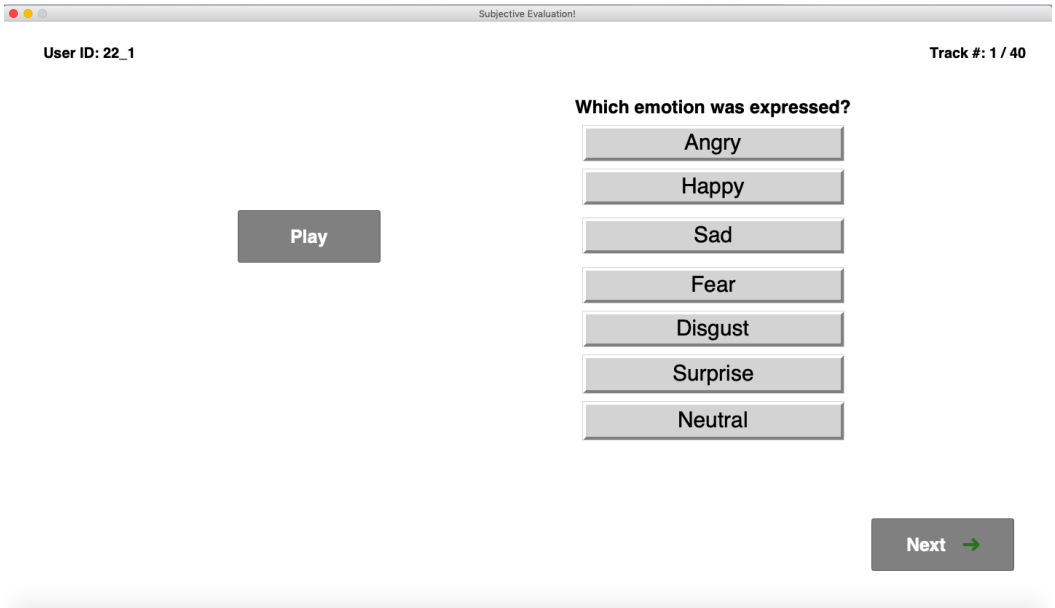


(b)

Fig. 6.2 Subjective evaluation - (a) instructions screen and (b) training screen.



(a)



(b)

Fig. 6.3 Subjective evaluation - evaluation screen (a) before playing the recording and (b) after playing the recording.

Speaker	Accuracy (SD)	Recall	Precision	F-score
DS01F	22.81 (0.31)	20.83	20.48	19.93
DS02F	56.25 (1.46)	54.11	54.32	52.90
DS04F	40.00 (0.62)	35.65	35.28	33.88
DS03M	43.23 (2.60)	40.65	41.29	39.51

Table 6.2 Average subjective evaluation performance (%) for 7 emotion classes on DEED-dysarthric speech.

Speaker	Accuracy (SD)	Recall	Precision	F-score
TS09F	64.25 (3.32)	61.71	63.02	60.78
TS13F	55.75 (5.04)	54.57	56.58	53.43
TS16F	56.25 (6.66)	54.29	55.62	53.13
TS17F	51.25 (5.18)	48.00	52.04	47.09
TS20F	51.25 (4.81)	45.86	47.42	44.68
TS06M	55.00 (3.16)	50.00	51.81	48.68
TS11M	58.00 (4.23)	54.43	56.98	52.89
TS18M	56.12 (3.37)	53.97	55.72	52.69

Table 6.3 Average subjective evaluation performance (%) for 7 emotion classes on subset of DEED-typical speech.

6.3 Results

All the DEED-dysarthric speech recordings, 320 utterances, were each evaluated 12 times. While the DEED-typical speech recordings included in the evaluation, 640 utterances, were each evaluated 5 times. The average performance in the evaluation on speakers with dysarthria and typical speakers for the 7 emotions by all the participants are given in Table 6.2 and Table 6.3, respectively.

The performance on typical speech is generally better than on dysarthric speech, as would be expected, on all speakers except on speaker DS02F, where the result is comparable to the results on some speakers with typical speech or even better than some of them. Nevertheless, the results on dysarthric speech are all above chance level (14%), which indicates the ability of listeners to perceive emotions communicated by speakers with dysarthria. Table 6.4 illustrates the average recall performance per emotion on all speakers with dysarthria. High accuracy was achieved for 'anger'. A good accuracy was also achieved for 'surprise', 'sad', and 'neutral'. While 'happy', 'disgust', and 'fear' are less accurate in being perceived.

As discussed in Chapter 4, DEED and SAVEE, a British English emotional database (Jackson and Haq, 2011) share a lot of similarities such as the language used for the recordings

Emotion	DS01F	DS02F	DS04F	DS03M
Anger	22.50	93.33	70.00	70.00
Surprise	16.67	55.00	35.00	36.67
Disgust	7.50	29.17	20.83	25.83
Fear	14.17	35.83	11.67	21.67
Happy	14.17	41.67	10.83	27.50
Sad	34.17	52.50	30.83	41.67
Neutral	36.67	71.25	70.42	61.25

Table 6.4 Average subjective evaluation recall (%) results per emotion on all speakers with dysarthria.

Database	Number of speakers	Average accuracy (SD)	Minimum accuracy	Maximum accuracy
SAVEE	4	66.45 (9.17)	53.20	73.70
DEED-Typical	8	55.97 (4.15)	51.00	64.25

Table 6.5 Comparison of the subjective evaluation performance (%) on SAVEE and a subset of DEED-typical speech.

and the stimuli set. The main differences, apart from DEED being a parallel database of dysarthric and typical speech, are: the number and gender of speakers as SAVEE has 4 male speakers only, speakers in DEED are not actors while the speakers in SAVEE are actors, and the number of utterances in DEED is 80 per speaker which is a subset of the 120 utterances recorded per speaker in SAVEE. Table 6.5 presents the performance of the subjective evaluation over all four actors on SAVEE where each utterance were evaluated ten times and the performance of the subjective evaluation over the randomly chosen eight speakers from DEED where each utterance were evaluated five times. Given the above stated differences, it is hard to directly compare the performance. However, it tells about the quality level of DEED.

In addition to general performance metrics, confusion matrices are very important as they help in giving more insight to the recognition performance. They help in highlighting which emotions appear easier and harder to recognise and which emotions are more easily confused. Figure 6.4 presents the averaged confusion matrices on all speakers with dysarthria, where the rows present the actual emotions and the columns present the recognised emotions. From Figure 6.4, it is observed that, i) for all speakers, except for speaker DS01F, 'anger' is never

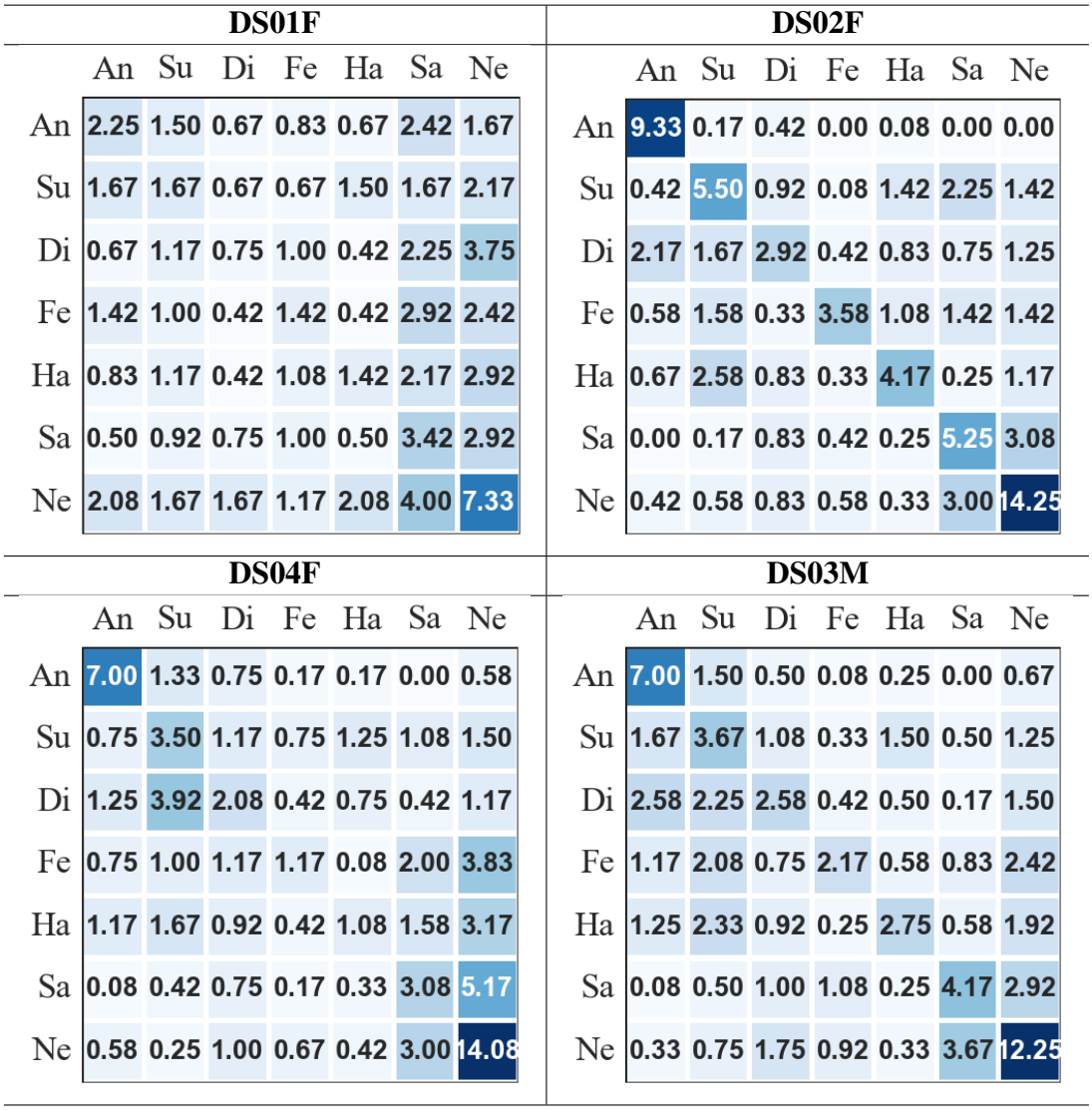


Fig. 6.4 Average confusion matrices of the subjective evaluation for each speaker with dysarthria. (rows= actual emotions and columns= recognised emotions, An= angry, Su= surprise, Di= disgust, Fe= fear, Ha= happy, Sa= sad, and Ne= neutral).

confused with 'sad' and 'sad' is rarely confused with 'anger', ii) for all speakers,'sad' and 'neutral' are most frequently confused with each other, and iii) for all speakers, except speaker DS01F, 'anger' appears to be the easiest emotion to recognise while 'happy', 'disgust', and 'fear' appear to be the most difficult emotions to recognise, for all speakers. For speaker DS01F, it is observed that participants perceived most of the emotions as either 'sad' or 'neutral'.

The averaged confusion metrics for typical speakers are presented in Figure 6.5, where the rows present the actual emotion and the columns present the recognised emotions. It is observed that for all speakers, i) 'anger' is rarely confused with 'sad' and 'sad' is never confused with 'anger', ii)'sad' and 'neutral' are are most frequently confused with each other, and iii) for all speakers, 'anger' appears to be the easiest emotion to recognise while 'happy', 'disgust', and 'fear' appear to be the most difficult emotions to recognise, and iv) 'happy' and 'surprise' are are most frequently confused for each other.

6.4 Discussion

Although, the overall recognition performance on typical speech was generally better than on dysarthric speech, the performance on the latter was all above chance level (14%), even for speaker DS01F, who has severe dysarthria and low speech intelligibility. This was also the case when looking at the recall of each emotion for speakers with dysarthria presented in Table 6.4, except for 'disgust', 'fear', 'happy' for speaker DS01F and 'fear' and 'happy' for speaker DS04F, where the performance were at or below chance level. The highest recognition performance on dysarthric speech was for speaker DS02F, who has mild dysarthria.

Based on the confusion matrices on both types of speech, most of the patterns of confusion were similar. On both types of speech, 'anger' was found to be the easiest emotion to recognise, while 'happy', 'disgust', and 'fear' appear to be among the most difficult emotions to recognise. This was also observed from the recall per emotion presented in Table 6.4. The results of recognising 'anger' does not completely align with the findings by Pell et al. (2006) who demonstrated the difficulty of English-speaking Canadians with dysarthria caused by PD to express emotions in their speech especially 'anger', 'happy', and 'disgust' were they were mostly perceived as 'neutral'. It also does not tally with the findings from the survey conducted in Chapter 3 where 'anger' was chosen by almost half of the respondents as the most difficult emotion to communicate from their perspective. It is important, however, to highlight the fact that these evaluation results were obtained from audio data only and the survey asked about communication generally. It could be that some emotions have higher *visual* component that make conveying them more easily or more difficult (depending on the

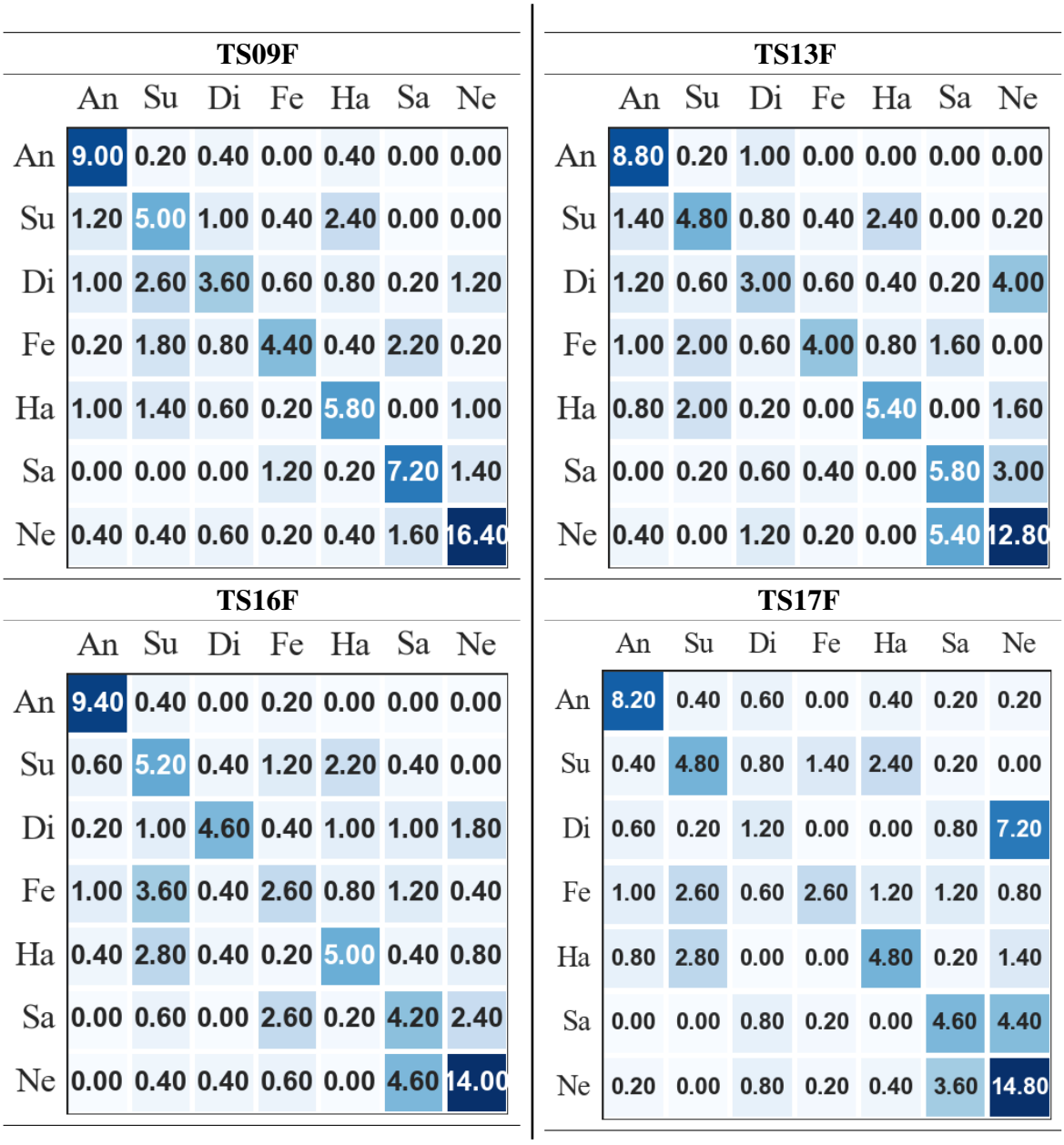


Fig. 6.5 Average confusion matrices of the subjective evaluation for typical speakers. (rows= actual emotions and columns= recognised emotions, An= angry, Su= surprise, Di= disgust, Fe= fear, Ha= happy, Sa= sad, and Ne= neutral).

TS20F

	An	Su	Di	Fe	Ha	Sa	Ne
An	7.60	1.20	0.40	0.00	0.60	0.20	0.00
Su	0.80	2.40	0.60	1.00	4.80	0.20	0.20
Di	0.20	1.20	1.80	0.40	1.20	1.00	4.20
Fe	0.40	1.80	1.00	2.40	0.60	1.20	2.60
Ha	1.80	2.00	0.80	0.20	3.80	0.20	1.20
Sa	0.00	0.20	0.20	0.20	0.00	5.20	4.20
Ne	0.20	0.60	0.40	0.00	0.40	0.60	17.80

TS06M

	An	Su	Di	Fe	Ha	Sa	Ne
An	7.80	1.00	0.80	0.00	0.40	0.00	0.00
Su	1.60	4.80	0.60	0.40	1.60	0.80	0.20
Di	1.60	1.00	3.00	1.40	0.60	1.60	0.60
Fe	0.40	2.60	1.00	4.00	0.00	0.60	1.40
Ha	0.00	5.40	0.00	1.60	2.20	0.40	0.40
Sa	0.00	0.20	0.60	1.20	0.00	4.20	3.80
Ne	0.40	0.20	0.60	0.00	0.00	0.80	18.00

TS11M

	An	Su	Di	Fe	Ha	Sa	Ne
An	8.40	1.00	0.40	0.00	0.00	0.00	0.20
Su	0.60	6.60	1.20	0.60	0.40	0.60	0.00
Di	2.00	2.00	4.40	0.20	0.60	0.40	0.40
Fe	2.20	2.20	1.00	1.40	0.40	1.20	1.60
Ha	1.20	3.20	0.20	0.80	3.80	0.20	0.60
Sa	0.00	0.80	0.40	2.60	0.00	5.20	1.00
Ne	0.00	0.20	0.80	0.20	0.00	2.20	16.60

TS18M

	An	Su	Di	Fe	Ha	Sa	Ne
An	8.40	0.80	0.80	0.00	0.00	0.00	0.00
Su	0.20	5.80	0.80	1.20	1.20	0.00	0.80
Di	1.00	2.20	3.80	1.00	0.40	1.00	0.60
Fe	0.80	3.20	1.20	2.80	0.40	0.80	0.80
Ha	0.00	2.20	0.00	0.40	3.40	0.60	3.40
Sa	0.00	0.00	0.20	1.20	0.00	6.40	2.20
Ne	0.20	1.20	1.20	0.60	0.80	1.60	14.40

Fig. 6.5 : continued

disability) in a face-to-face communication situation. It was also found from the confusion matrices that 'anger' and 'sad' are not considered confusing pairs while 'sad' and 'neutral' are considered confusing pairs. This aligns with the findings in the literature from subjective evaluation of emotional data on typical speech (Dai et al., 2009; Livingstone and Russo, 2018). In general, the results of this evaluation show that listeners were able to perceive different emotions expressed by people with dysarthria and typical speakers which aligns to the findings by Martens et al. (2011) who demonstrated the ability of speakers with dysarthria caused by PD to communicate emotions similar to typical speakers.

6.5 Conclusion

The initial plan was to recruit normal hearing participants from three different groups: participants who are familiar with dysarthric speech and familiar with a speaker/speakers with dysarthria in DEED, participants who are familiar with dysarthric speech but are not familiar with any of the speakers with dysarthria in DEED, and participants who are not familiar with dysarthric speech at all. The aim was to compare the effect of familiarity with the speech and speakers on the recognition of emotions. However, due to the situation of COVID-19 and the lockdown imposed, it was difficult to recruit enough participants from each group and analyse their results separately. A detailed analysis will be done in the future including participants from each group and applying statistics measures to check the significance, if any, of the participants familiarity level on the recognition performance.

Nevertheless, the conducted evaluation indicates that speakers with dysarthria in DEED were able to communicate different emotions. The overall recognition performance shows that participants in this study were able to recognise emotions spoken by speakers with dysarthria even for speaker DS01F, who has severe dysarthria and highly unintelligible speech. These results demonstrate this database will be a useful resources for understanding emotion communication by people with dysarthria. They also validate the use of the database in the acoustic and modelling studies presented in the thesis. These encouraging results together with the acoustic analysis results presented in the previous chapter motivate the development of automatic dysarthric speech emotion recognition model. This will be the focus of the following chapters.

Chapter 7

Towards the Automatic Recognition of Emotion in Dysarthric Speech

Part of the content of this chapter has been published in INTERSPEECH 2020 (Alhinti et al., 2020b).

7.1 Introduction

Acoustic analysis of dysarthric emotional speech, presented in Chapter 5, has shown the ability of some people with dysarthria to make some systematic changes in their speech when communicating emotions. This chapter will investigate the feasibility of automatically recognising emotions from dysarthric speech.

Given the importance of speech and emotions in effective communication and its importance in human-computer interaction (HCI), speech emotion recognition (SER) has evolved to enrich the use and benefit of the existing speech recognition systems. There are many increasing applications of SER in many fields including education (Bahreini et al., 2016; Jithendran et al., 2020), call centers (Gupta and Rajput, 2007; Pappas et al., 2015; Vidrascu and Devillers, 2007; Yoon and Park, 2007), mobile services (Hossain et al., 2016; Yoon et al., 2007), gaming (Jones and Sutherland, 2008, 2005), human-robot interaction (Chen et al., 2020b; Huahu et al., 2010) healthcare (Hossain, 2016; Hossain and Muhammad, 2017; Low et al., 2010; Van Lancker et al., 1989), and assistive technology (Garay et al., 2006). In addition to the applications of emotion recognition in AAC as have been presented in Section 2.4.3. Despite its importance, it is a very challenging task. The process of recognising emotions from speech is not a straightforward process for machines due to many factors including inter-speaker variability and the unavailability of an identified optimum set of

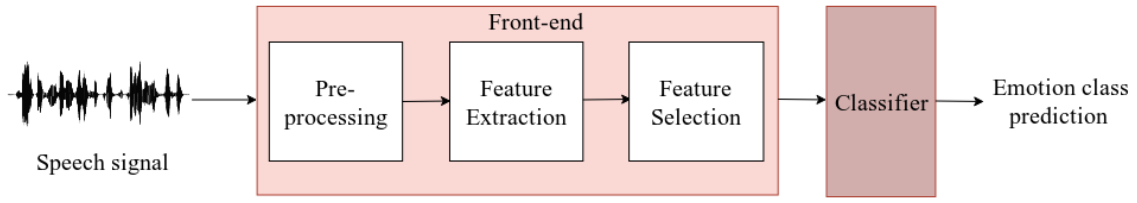


Fig. 7.1 Speech emotion recognition system general components.

acoustic features have not been identified. Due to the subjective nature of the perceptual task, listeners sometimes misinterpret an emotion when conveyed verbally (Parkinson, 1996).

SER model can be viewed as a composition of: a front-end and a classifier. Figure 7.1 depicts the general components of a SER model. The front-end consist of the feature extractor, this is also sometimes known as the parametrization process, and is responsible for obtaining different features from the speech signal that represent the data in a compact way by keeping the relevant information and discarding the irrelevant and misleading information which is a very challenging filtering process. The primary stages in the front-end subsystem are (Bishop, 2006; Jurafsky and Martin, 2000; Rabiner, 1993):

Signal pre-processing: There are many pre-processing techniques used prior to feature extraction including: framing, which is the process of dividing the speech signal into fixed length frames or segments to achieve stationarity, and windowing, which is the process of multiplying a window function such as Hamming window to the frames to minimise the spectral leakage in Fast Fourier Transform (FFT). Normalization and noise reduction are also other potential pre-processing techniques.

Feature extraction: Selecting a set of potential features that characterise the emotional aspects of the speech signal is one of the most important and challenging tasks. The chosen features contribute heavily to the performance of a SER systems. The literature includes investigations of many different set of features used for the task of SER. However, there is no agreement yet on the best set of features. There are two main types of features that can be extracted from a signal: local and global features (Rao et al., 2013). Local features, also called short-term features are used to capture the temporal information (dynamics) from the signal where features are extracted from the segmental level of an utterance. They are important as emotion related features are not uniformly present in all segments of an utterance (Rao et al., 2013). Anger for example is dominantly perceivable from the beginning of an utterance, while surprise for example is dominantly perceivable from the last part of an utterance. Global features on the other hand, also known as functionals,long-term,

or suprasegmental features, are used to represent the gross statistics of the extracted features from an utterance including minimum, maximum, range, mean, and standard deviation values.

Local and global features used for the purpose of SER can be categorised under these four categories: prosodic features, spectral features, voice quality features, and Teager Energy Operator (TEO) based features (Akçay and Oğuz, 2020). Prosodic features, sometimes called paralinguistic features, are the set of features that humans can perceive. F0, duration, and energy, are the most widely used prosodic features. Spectral features are frequency based features obtained by using Fourier transform on the time domain signal. Spectral features are primarily determined by the shape of the vocal tract when the sound was produced. Mel Frequency Cepstral Coefficients (MFCCs), Linear Prediction Cepstral Coefficients(LPCCs), and formant frequencies are examples of spectral features. Voice quality can be defined as the auditory colouring characteristics of someone's voice (Keller, 2004; Kent and Ball, 2000). Jitter, shimmer, and harmonics to noise ratio (HNR) are examples of voice quality features commonly used in SER tasks. TEO based features are mostly used to detect stress and anger in speech. TEO shows the instantaneous changes in the amplitude and frequency of the signal which enables the capture of energy fluctuations (Kaiser, 1990; Sundaram et al., 2003). Usually, a combination of features from theses categories are used to obtain better SER results.

Feature selection and dimension reduction: The curse of dimensionality is one of the main problems in machine learning due to the many features extracted and the lack of a certain set of features that best model emotions (Bishop, 2006). Therefore, feature selection and dimensionality reduction are considered to be one of the feasible solutions to this problem. In addition they help in reducing training time and overfitting, which can affect the performance of the classifier. Feature selection is the process of selecting a subset of all the features that contribute most in the prediction problem and eliminate redundant and irrelevant features without changing the features. While feature reduction transforms high dimensional features into lower a dimension. There are a number of feature selection and dimensionality reduction techniques available. The Least Absolute Shrinkage and Selection Operator (LASO) (Tibshirani, 1996) and the Principle Component Analysis (PCA) (Jolliffe, 2011) are among the most popular techniques used, where the former is a feature selection and the later is a dimentilaity reduction technique.

The output of the front-end is then fed as the input to the classifier, which is usually based on a machine learning approach that is responsible for classifying the emotion expressed in

the input. Classifiers get trained on a set of data and use what they learned to classify new samples.

The rest of this chapter is organised as follows: Section 7.2 reviews a set of popular speech emotion recognition techniques proposed in the literature. Section 7.3 presents the first attempt to develop an automatic dysarthric speech emotion recognition system. The baseline results on the collected database, DEED, using different classification approaches are discussed.

7.2 Speech emotion recognition techniques

There are many machine learning algorithms including classical classifiers and deep learning algorithms that have been used for the purpose of classifying emotions from speech. The literature includes many studies investigating the performance of different classifiers on different databases, using different features sets. Table 7.1 reviews some of these studies including the used databases, features, and classifiers along with the obtained results.

7.2.1 Classical classification algorithms

Classification algorithms take the training data X as input and the labels Y as output and return a mapping function between X and Y $f_x : X \rightarrow Y$ that represents the relation between them. The mapping function is then used to classify unseen instances, that is instances from the test set. There are many classification algorithms including Support Vector Machine (SVM), Gaussian Mixture Model (GMM), Logistic Regression (LR), Decision Trees (DT), Adaptive Boosting (AdaBoost), K-Nearest Neighbor (KNN), Hidden Markov Model (HMM), and ensemble methods (Bishop, 2006).

Support Vector Machine (SVM)

The SVM algorithm is widely used in speech emotion classification. It is a discriminative supervised learning model. It uses labeled training data to generate an optimal hyperplane that finds the maximum margin that separates data points that belong to different classes. SVMs are widely used for tasks involving pathological speech as they are particularly well-suited to sparse data domains (Fauvel et al., 2006, 2008; Kodrasi et al., 2020a,b). Considering a binary classification problem, the SVM algorithm defines decision boundaries that separate data from both classes. Then it finds the points that are the closest to the decision boundary. These points are known as support vectors. The perpendicular distance between the line and the support vectors, known as the margin, is computed. The hyperplane with the maximum

margin is selected as the optimal hyperplane. This hyperplane is then used to classify test data Bishop (2006). In the case of having non linearly separable data, kernel functions are used by the SVM to map the data into a higher dimensional space where the hyperplane is more easily defined. There are a number of different types of kernels used with SVMs where they differ in their way of mapping the data. Some of the common ones are linear, polynomial, and RBF kernels (Bishop, 2006). In addition, soft margin is also used by SVMs to address the non linearly separable data, where the SVM is modified to allow some points to be miscalssified with a penalty. This could be seen as relaxing the hard margin constraint. The value of the penalty increases with the distance from the decision boundary. This is implemented using slack variables one for each training data point (Bishop, 2006; Murphy, 2012). A regularisation parameter, C, is used to balance the trade-off between maximising the margin and minimising the errors in the training set. The bigger the value of C, the more penalty given for misclassification leading into defining narrower margin. Gamma is another parameter that is used for non linear hyperplanes, which controls the spread of the kernel. The higher the gamma value is, the narrower the curve gets (Bishop, 2006; Murphy, 2012).

Although SVM is fundamentally a binary classifier, different methods have been proposed to upgrade it to a multiclass classifier. One-versus-the-rest and one-versus-one are some of the commonly used approaches. In the one-versus-the-rest approach, a single classifier is trained per class where the samples of that class are treated as positive samples and all other samples of other classes as negative. A sample is then classified into the class with the maximum score among all classifiers. While in the one-versus-one approach, a classifier is trained for every possible binary pair of classes using only samples from those classes. A sample is then classified into the class with the highest number of votes (Bishop, 2006; Murphy, 2012).

Gaussian Mixture Models (GMMs)

GMMs are common and powerful generative classifiers that are used in many fields and it is the most widely used mixture model. It is a widely used algorithm in speech emotion recognition tasks (Jermsittiparsert et al., 2020; Palo et al., 2020; Patel et al., 2017). GMM is a probabilistic model with the assumption that there is a finite number of Gaussian distributions in which all the data points were generated from. GMMs use the training data to learn the mixture model which is subsequently used to classify the test data. A d-dimensional multivariate gaussian is defined by a mean vector and a dxd covariance matrix. In the case of having latent variables (missing values), the values of the mean and the covariance are typically determined using the Expectation-Maximization (EM) algorithm (Bishop, 2006; Murphy, 2012). EM is an iterative optimisation algorithm composed of two main steps, the

Study	Details	Results
Koduru et al. (2020)	Study focus: Removing noise using filters in the pre-processing stage.	An accuracy of 70%, 65%, and 85% for SVM, LDA, and decision, respectively
	Databases: RAVDESS.	
	Features: Pitch, ZCR, energy, MFCC, MFCC global features, and DWT.	
	Emotions: Anger, happiness, sadness and neutral.	
	Classifiers: SVM, LDA, and decision tree.	
Hao et al. (2019)	Study focus: Experimenting the proposed optimisation (SMO) algorithm, nonlinear SVM based on sequential minimal optimisation algorithm, on speaker-dependent, speaker-independent and cross-corpus.	An accuracy of 84.57%, 85.15%, and 78.88% for SI in CASIS, Berlin, and cross database, respectivey.
	Databases: CASIA and Berlin.	
	Features: A set of Low level Descriptors (LLD) related to energy, pitch, formants and cepstral features.	
	Emotions: Anger, happiness, sadness, fear, and neutral.	
	Classifiers: Non linear SVM based on SMO.	
Sahu (2019)	Study Focus: Investigating the contribution of handcrafted features of different modalities (audio features and text feature) using different traditional classifiers and deep learning models and comparing them to end-to-end deep learning models.	The best performance was achieved when using ensembles of multiple models. Using only audio features, an accuracy of 56.6% was achieved with an ensemble of RF, XGB, and MLP. Using text features only and a combination of text and audio features, an accuracy of 64.9% and 70.1% was achieved with an ensemble of random forest, XGB, MLP, MNB, and logistic regression, respectively.
	Databases: IEMOCAP.	
	Features: Audio features: Pitch, harmonics, RMS energy, pause, and mean and standard deviation of the amplitude. Text features: Term frequency and inverse document frequency.	
	Emotions: Anger, happiness, sadness, fear, surprise, and neutral.	

Study	Details	Results
	Classifiers: Traditional classifiers: SVM, logistic regression, random forest, gradient boosting (XGB), naive-bayes, and different ensemble models. Deep learning models: Multi-layer perceptron (MLP) neural network and LSTM.	
Iqbal and Barua (2019)	Study Focus: Developing a SER system capable of recognizing emotions from real-time speech.	For RAVDESS and SAVEE databases, SVM classifer resulted in the best performance accuracy while in the real-time classification, gradient boosting achieved the highest accuracy.
	Databases: SAVEE, RAVDESS, and live recorded emotional data.	
	Features: ZCR, 13 MFCCs, 12 chroma vectors and chroma deviation, energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral rolloff.	
	Emotions: Anger, happiness, sadness and neutral.	
	Classifiers: SVM, KNN, and gradient boosting.	
Mao et al. (2019)	Study Focus: Investigating the performance of three HMM based systems for the task of speech emotion recognition.	On CASIA, 53.81% accuracy for SI and 91.32% for SD using SGMM-HMM and DNN-HMM, respectively. On Berlin, 87.62% weighted accuracy for SI using SGMM-HMM. On IEMOCAP, 62.23% unweighted accuracy and 62.28% weighted accuracy for SI using SGMM-HMM and DNN-HMM, respectively.
	Databases: CASIA, Berlin, and IEMOCAP.	
	Features: Pitch, MFCCs, and voicing probability.	
	Emotions: CASIA: Anger, happiness, sadness, fear, surprise, and neutral. Berlin (sbset): Anger, boredom, joy, sadness and neutral. IEMOCAP: Anger, happiness, sadness, and neutral.	
	Classifiers: GMM-HMMs, SGMM-HMMs, and DNN-HMMs.	
Xie et al. (2019b)	Study Focus: Investigating ways to effectively use LSTM output at all times rather than a single last moment to improve the performance of SER.	When applying the proposed attention mechanism in the feature and time dimension, an improved UAR of approximately 2.7%, 3%, and 0.8% on eNTERFACE, GEMEP, and CASIA, corpus, respectively, in comparison to using the
	Databases: CASIA, eNTERFACE, and GEMEP.	
	Features: Using only the frame-level features in the ComParE openS	

Study	Details	Results
	-MILE features. Instead of HNR, glottal noise energy and glottal harmonic energy are extracted separately.	same approach but without modifying the forget-gate. Using the proposed approach, an UAR of 89.6%, 57%, and 92.8% on eNTERFACE, GEMEP, and CASIA, respectively.
	Emotions: CASIA: Anger, happiness, sadness, fear, surprise, and neutral. eNTERFACE: Anger, happiness, sadness, disgust, fear, and surprise. GEMEP (subset): Hot anger, amusement, sadness, anxiety, joy, despair, panic fear, interest, relief, irritation, pride, pleasure.	
	Classifiers: LSTM with a modified forget-gate based on attention mechanism followed by fully connected layers.	
Dai et al. (2019)	Study Focus: Investigating the effectiveness of using center loss in the SER model to learn effective features from different length spectrograms.	A weighted accuracy of 65.40% and an unweighted accuracy of 66.86% when using mel-spectrogram as input. A weighted accuracy of 62.96% and an unweighted accuracy of 65.13% when using STFT spectrogram as input.
	Databases: IEMOCAP.	
	Features: log Short Time Fourier Transform (STFT) spectrogram or log Mel-spectrogram as input.	
	Emotions: Anger, excitement and happiness were merged, sadness, and neutral.	
	Classifiers: 2D CNN-Bi-RNN.	
Atmaja (2019)	Study Focus: Investigating different RNN-based models for dimensional SER.	eGeMAPS significantly outperformed the other feature set. LSTM outperformed DNN and GRU. A concordance coefficient correlation (CCC) score of 0.43, 0.11, and 0.36 for arousal, valence, and dominance by LSTM using eGeMaps feature set.
	Databases: IEMOCAP.	
	Features: Set 1: ZCR, energy, entropy of energy, Spectral Centroid, Spectral Spread, Spectral Entropy, Spectral Flux, Spectral Rollof, 13 MFCCs, F0, and Harmonics. Set2: eGeMAPS feature set.	
	Emotions: Dimensional classification using arousal, valence, and dominance dimensions.	

Study	Details	Results
	Classifiers: DNN, GRU, and LSTM.	
Latif et al. (2019)	Study Focus: Investigating opportunities to improve SER performance from raw speech by using CNN modeling capabilities.	Using data augmentation and leave-one-speaker-out, an UAR of 60.23% on IEMOCAP and 52.43% on MSP-IMPROV. The results outperformed some studies in the literature that used raw speech as input. The obtained results were comparable to other results obtained from handcrafted features using other classifiers such as SVM with eGeMAPS.
	Databases: IEMOCAP and MSP-IMPROV.	
	Features: raw speech waveform.	
	Emotions: IEMOCAP: Anger, excitement and happiness were merged, sadness, and neutral. MSP-IMPROV: Anger, happiness, sadness, and neutral.	
	Classifiers: 2D CNN-LSTM followed by fully connected layers.	
Ke et al. (2018)	Study Focus: Investigating the effect of using PCA with different dimensional gradients as a feature reduction technique on the classification results.	An accuracy of 76.67% and 75% for PCA+SVM with polynomial kernel and PCA+ANN, respectively.
	Databases: CASIA.	
	Features: ZCR, F0, RMS of energy, sounding probability, 12 MFCCs, and the first three formant frequency and bandwidth.	
	Emotions: Anger, happiness, sadness, fear, surprise, and neutral.	
	Classifiers: SVM with different kernels and ANN.	
Jacob (2017)	Study Focus: Investigating the effect of using emotional content and valence of emotions when modeling binary decision trees and logistic regression.	The highest classification accuracy was achieved for classifying surprise and happy emotions. An accuracy of 93.63% and 73% using decision tree with 7 features, and logistic regression with 8 features, respectively.
	Databases: Malayalam emotional speech database.	
	Features: first four formants and their bandwidths.	
	Emotions: Several binary classifications: emotional versus neutral, positive versus negative, and happy versus surprise.	
	Classifiers: Decision tree and logistic regression.	

Study	Details	Results
Wang et al. (2017)	Study Focus: Investigating the effect of applying two decision fusion methods to combine the predictions of multiple classifiers on the performance of a multiclass speech emotion classification system.	A macro average precision of 94.9, 94.2%, 93.9%, 93.8% for fusion method based on majority vote, SVM, adaboost, and random forest, respectively.
	Databases: Chinese EESDB.	
	Features: Acoustic features were extracted using wavelet analysis,then methods of information gain and sequential forward selection were used to reduce the features dimensions .	
	Emotions: Anger, happiness, sadness, disgust, and neutral.	
	Classifiers: SVM, adaboost, Random forest, proposed fusion methods of all 3 classifiers based on majority vote and weighted vote.	
Badshah et al. (2017)	Study Focus: Investigating the effectiveness of using CNN to extract prominant distinguishing features to recognise emotions from spectrograms.	An accuracy of 84.3%.
	Databases: Berlin.	
	Features: Spectrograms.	
	Emotions: Anger, happiness, sadness, boredom, disgust, fear, and neutral.	
	Classifiers: 2D CNN followed by fully connected layers.	
Erdal et al. (2016)	Study Focus: Investigating the effect of different neural networks archetictures in terms of hidden layers and hidden units for the task of SER.	Using leave-one-speaker-out, an accuracy of 65.5 %, 67.2%, and 59.5% using SVM, DNN using 4 hidden layers, RNN using 4 hidden layers.
	Databases: Berlin.	
	Features: MFCCs.	
	Emotions: Anger, happiness, sadness,boredom, disgust, fear, and neutral.	

Study	Details	Results
	Classifiers: SVM with RBF kernel, DNN and RNN.	
Loweimi et al. (2015)	Study Focus: Introducing an interface layer between the front-end and the classifier. Each utterance is represented using a fixed length super-vector generated from stacking the average posterior probabilities of the components of GMMs trained for all classes for each frame.	An accuracy of 87.6%.
	Databases: Berlin.	
	Features: MFCC, log-energy, delta and delta-delta.	
	Emotions: Anger, happiness, sadness, boredom, disgust, fear, and neutral.	
	Classifiers: SVM classifier with RBF kernel.	
Deng et al. (2014)	Study Focus: Employing domain adaptation method.	An AUR of 62.74% for SUSAS database and 64.18% for ABC database.
	Databases: FAU AEC, ABC, and SUSAS.	
	Features: INTERSPEECH 2009 Emotion Challenge feature set.	
	Emotions: two-class classification task: positive and negative.	
	Classifiers: Linear SVM.	
Sarker et al. (2014)	Study Focus: Applying the majority voting technique over four machine learning algorithms and selecting most prominent features using FCBF and FS selection algorithms.	SVM classifier performs the best while KNN performs the worst. An average accuracy of 84.19% for the majority voting technique.
	Databases: Berlin and EMA.	
	Features: Energy, 12 MFCCs, ACF, voicing probability, F0, and their deltas.	
	Emotions: Berlin (subset) and EMA: Anger, happiness, sadness, and neutral	
	Classifiers: NN, SVM, KNN, and decision tree.	

Study	Details	Results
Seehapoch and Wongth-anavasu (2013)	Study Focus: Experimenting different combinations of features.	The highest accuracy result was achieved using MFCC, F0, and energy. An accuracy of 89.80%, 93.57% and 98.00%, in Berlin, Japan and Thai, respectively.
	Databases: Berlin, Japan and Thai.	
	Features: F0, energy, MFCC, ZCR, and Linear Predictive Coding (LPC).	
	Emotions: Berlin and Japan databases :Anger happiness, sadness, boredom, disgust, fear, and neutral. Thai database: anger, fear, happiness, sadness, disgust, surprise and neutral.	
	Classifiers: Linear SVM.	
Lee et al. (2011)	Study Focus: Proposing hierarchical multiclass emotion classifier framework using combination of binary classifiers where the framework starts with classifying the easiest emotions.	Using leave one speaker out cross validation, an UAR of 48.37% and 58.46% for AIBO and IEMOCAP databases, respectively.
	Databases: AIBO and IEMOCAP.	
	Features: RMS energy, HNR, ZCR, pitch, and 12 MFCCs and their deltas.	
	Emotions: AIBO (subset): anger, emphatic, positive, rest, and neutral. IEMOCAP (subset): anger, happiness, sadness, and neutral.	
	Classifiers: Decision tree.	
Bitouk et al. (2010)	Study Focus: Analysing the performance of class-level spectral features, traditional spectral features and utterance level prosodic features.	An accuracy, when classifying 6 emotions, of 81.3% for Berlin using combined utterance level prosodic features and class level spectral features with rank search subset evaluation feature selection, and an accuracy of 46.1% for LCD using their introduced class level spectral features with group-wise feature selection.
	Databases: Berlin and Emotional Prosody Speech and Transcripts corpus (LDC).	
	Features: Different levels of spectral features.	
	Emotions: Berlin (subset) and LCD (subset): Anger, happiness, sad-	

Study	Details	Results
	ness, disgust, fear, and neutral. LCD full set: hot and cold anger, happiness, sadness, disgust,anxiety (fear), panic, despair, interest, elation, , pride , shame, contempt, boredom and neutral.	
	Classifiers: SVM classifer with RBF kernel.	
Busso et al. (2009)	Study Focus: Finding the F0 contour aspects of emotionally salient and investigating the use of neutral models to discrminate emotional speech.	An average accuracy of 77.3% for emotional vs neutral classification task using sentence level features
	Databases:EPSAT, EMA, GES, SES, and WSJ1.	
	Features: Different features of the F0 contour.	
	Emotions: WSJ1: neutral. EMA: Anger, happiness, sadness, and neutral. SES: Anger, happiness, sadness, surprise and neutral. GES: Anger, happiness, sadness, fear, disgust, boredom, and neutral. EPSAT: Happiness, sadness, boredom, disgust, anxious, panic, hot anger, cold anger, despair, elaition, interest, pride, shame, contempt, and neutral.	
	Classifiers: GMM with LDC.	
Dai et al. (2008)	Study Focus: Investigating the effectivness of using landmarks in addition to basic acoustic and prosodic features in the performance of SER model.	Over 90% accuracy when classifying hot anger and neutral, and over 80% accuracy when classifying happiness and sadness. The accuracy decreases to 62% and 49% when classifying 4 and 6 emotions, respectively.
	Databases: EPSAT.	
	Features: Landmarks: Glottis, sonorant, and burst. Syllable: Syllable rate, syllable number, landmarks per syllable, and syllable duration. Acoustic and prosdic features: pitch features, timing features, and energy features.	
	Emotions: Subset: Hot anger, happiness, sadness, panic, interest, and neutral.	

Study	Details	Results
	Classifiers: ANN.	
Ververidis and Kotropoulos (2005)	Study Focus: Proposing a model that reduces the computational burden of cross validation in sequential floating forward selection algorithm.	A probability of correct classification of 48.5% 56%, and 50.9% for GMM with single Gaussian density on the both genders, males, and females, respectively.
	Databases: 1300 utterances from the DES.	
	Features: Statistics of pitch, formants, and energy contours in addition to duration of the rising and falling slopes of the energy and pitch contours.	
	Emotions: Anger, happiness, sadness, surprise, and neutral.	
	Classifiers: GMM with different numbers of Gaussians densities and Bayes classifier.	

Table 7.1 List of studies in SER. (SD = speaker-dependent, SI = speaker-independent).

expectation (E-step) and the maximisation (M-step). After initialising the parameters, the means, covariances, and mixing coeffecients, the EM algorithm is applied. In the E-step, the missing values are estimated using the available data. In the M-step, the parameters are optimised given the estimated data from the E-step. After each iteration of the EM algorithm, the log likelihood is computed to measure the fit of the GMM. These steps are repeated until convergence is reached. The goal of the EM algorithm is to provide the maximum log likelihood parameter estimates of the modal. GMMs usually need large amount of data to perform well (Murphy, 2012; Rogers and Girolami, 2016).

Logistic Regression (LR)

LR is a predictive algorithm used for classification tasks that is based on probabilities. The probability of an event to occur is predicted by fitting the data to a logit function. The sigmoid function is responsible for mapping predicted real values to probability values between 0 and 1 (Murphy, 2012). The classification decision is made by setting a threshold. LR parameters are estimated using maximum likelihood estimation (MLE) approach. To compute the MLE, an optimisation algorithm is used. There are number of different optimisation algorithms including gradient descent and Newton's algorithm (Murphy, 2012). In terms of regularisation, L1 regularisation and L2 regularisation strategies are commonly used. Regularisation is very important in LR to avoid overfitting. LR is highly interpretable, does not need high computational resources, and easy to implement. Its performance highly depend on the data presentation. In the case of a multiclass classification problem, multinomial logistic regression is used where the sigmoid function is replaced by a softmax function. The softmax function returns the probability of each class, where the probability is in the range of 0 and 1 and the sum of these probabilities is equal to 1. The final output is the class with the highest probability (Murphy, 2012).

Decision Trees (DTs)

DT is one of the simplest discriminative classifiers that is very easy to understand and interpret. It has a flowchart (tree)-like structure where it recursively partitions a data set into smaller subdivisions using a number of rules that are applied at the node level (Bishop, 2006). It is a non-parametric method that does not require any assumption about the space distribution. It is composed of i) nodes, which test the value of an attribute, ii) branches that connects the nodes with each other, which correspond to the values of the attributes, and iii) leaf nodes, which represent the class labels. Finding the optimal structure of the tree is known to be NP-complete. Therefore, greedy optimisation algorithm is used to compute a

locally optimal MLE. When using a greedy algorithm a common approach used is to have a fully grown tree and then prune the tree back. Pruning helps in reducing overfitting. DT has the ability to select features automatically and reduce complexity (Hu et al., 2009). However, the instability behaviour of the classifier, where small variations are present in the data can lead to generating a different tree, is one of the problems of DT (Bishop, 2006).

Adaptive Boosting (AdaBoost)

Adaboost is a powerful sequential ensemble method (Bishop, 2006). It is one of the most widely used boosting algorithms that is applied in many fields. It combines the output of a number of weak classifiers to form one strong classifier. Decision tress are examples of weak learners in AdaBoost. In each round, the algorithm focus on previously misclassified instances by assigning more weight to hard to classify instances while less weight is assigned to the easy to handle instances. The final prediction is computed as the sum of the weighted predictions of the weak classifiers (Bishop, 2006). Having few parameters, ease of implementation, high performance, and less susceptiblity to overfitting with low noise data are the main advantages of using AdaBoost. However, it does not perform very well in noisy data (Bishop, 2006).

K-Nearest Neighbour (KNN)

KNN is a discriminative supervised learning algorithm and one of the simplest and most efficient classification techniques (Bishop, 2006; Murphy, 2012). It is a non parametric and instance based learning algorithm that performs the classification based on a distance function. Euclidean distance is a widely uses distance metric used (Bishop, 2006). For a given data instance, KNN computes the distance between this data instance and all the training data instances. For the k-nearest neighbour, a majority voting is applied. The test data instance is then assigned to the class that forms the majority. KNN is useful in different applications including SER tasks (Guo and Yu, 2019; Kapoor and Thakur, 2020; Qianqian et al., 2020). A major drawback of the KNN algorithm is its high computational cost especially when having a large dataset and that it doesn't perform well with high dimensional data (Murphy, 2012).

Hidden Markov Models (HMMs)

HMMs have been extensively used in the speech recognition and signal processing domains. They have also been successfully used in SER tasks (Mao et al., 2019). They are a graphical model of finite number of states that represents a probability distribution over a sequence of observations that are generated from hidden states (Murphy, 2012). HMM is a generative model that is based on a strong statistical foundation. It is one of the considerably effective

methods for capturing temporal dynamic characteristics. However, HMM has large number of unstructured parameters (Murphy, 2012).

Neural networks or Artificial Neural Networks (ANNs)

ANNs consist of many artificial neurons that are interconnected under a specific network architecture (Bishop, 2006). These neurons work together to generate an output from a given input. Neural networks can learn using supervised or unsupervised learning processes and with the presence of noise (Saravanan and Sasithra, 2014). The neural network structure have a minimum of three layers: an input layer, at least one intermediate hidden layer, and an output layer. Each layer contains a number of nodes. While the number of nodes in the input and output layers depends on the data representation and number of classes, respectively, the hidden layers can contain any number of required nodes (Bishop, 2006). Initially random weights are used on the connection of each layer to the next one. There are a number of different methods and algorithms developed to train the neural network. One of the common algorithms that has been popular in the SER field is the backpropagation (Revathi and Sasikaladevi, 2020; Shi and Song, 2010; Yang and Shi, 2019). Backpropagation is a supervised learning algorithm that uses a feed forward network to compute the output. The error is computed at the output level, and then the weights are adjust backward through the network to reduce the error. (Bishop, 2006; Dedgaonkar et al., 2012).

Ensemble methods

They are machine learning technique that is based on combining the output from multiple models into one predictive model (Murphy, 2012). The aim is to improve the overall performance, decrease bias, or decrease variance. Voting is typically used to combine the outputs from each ensemble classifier. Hierarchical classifier is one of the ensemble classifiers architectures where the input is fed to one model and its output is fed to the next model in a hierarchical manner (Vasuki and Aravindan, 2020). Another architecture is feeding the input data to all models and the final decision is obtained by comparing the results of all models (Prasomphan and Doungwichain, 2018). The performance of the ensembles are usually better than individual classifiers, however, they usually require high computational resources (Murphy, 2012).

7.2.2 Deep learning algorithms

Deep learning algorithms are machine learning algorithms that are mostly based on ANN. They are called "deep" as their structure can have hundreds of layers. They gain a lot of

attention in many fields including SER due to its high performance when dealing with large data in comparison to classical machine learning algorithms. Their ability to extract high level features from raw input, deal with unstructured data, and deliver high performance are some of the main strengths of using deep learning algorithms. However, they require a lot of computational resources, and large amounts of data to perform well because of the many more parameters to estimate. Having a black box nature are some of the main downsides of deep learning. There are many different deep learning algorithms. Convolutional Neural Networks (CNNs) and different variations of Recurrent Neural Networks (RNNs) are among the most widely used algorithms in the field of SER (Goodfellow et al., 2016; Murphy, 2012).

Convolutional Neural Networks (CNNs)

CNNs are deep neural networks that are mostly applied to data with a grid-like topology such as images which are two-dimensional grid and time-series data which are one-dimensional grid. As the name implies, the network employs convolution which is a special kind of linear operation. It is basically the process of combining two functions to get a third function. CNNs proved their success in capturing spatial and temporal dependencies from an input. A CNN is composed of input layer, output layer, and several other layers including convolution layer, pooling layer, and fully connected layer (Goodfellow et al., 2016). The convolution layer is responsible for producing the feature map, while the pooling layer is responsible for reducing the feature size by keeping only the dominant features (Goodfellow et al., 2016).

Recurrent Neural Networks (RNNs)

RNNs are successfully used to process sequential data such as time series data. They have two main advantages over other neural networks. First, it has an internal memory which is used to store information about what has been calculated previously and use that information to predict the future. The output from an RNN unit is fed to the next unit in addition to it being looped back to itself. Thus an RNN unit has two inputs, one from the present and the other is from the recent past. Second, RNN accepts inputs of arbitrary length (Goodfellow et al., 2016). However, RNNs suffer from the short memory problem which results in forgetting what was seen in longer sequences (Goodfellow et al., 2016). To overcome this issue, different architectures of RNNs were developed such as the Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) where new gates were introduced in their architectures. A common architecture of LSTM, for example, is composed of memory cell and input gate, output gate, and forget gate. These gates control the flow of information inside the network and learn what data is important to keep and what to forget. Thus, LSTM

and GRUs are better than RNNs in processing long sequences (Goodfellow et al., 2016).

There is no general agreement on what classification algorithm nor set of features that work the best for SER tasks. It is a challenging and complicated problems due to many factors including the effect of language, culture, and gender, lack of enough databases of sufficient sizes, missing databases in many languages, and database recording settings and nature differences, in addition to the ambiguity of emotions and inter-speaker variability. So, what works best on one database might not suit other databases.

A comparison among the different techniques is not easy. As can be seen from Table 7.1, researchers used different features, different databases, and different parameters values in their studies. One of the major determinants of the type of classification algorithm is the size of the database. For example, neural networks need large databases to insure better performance. Also, as mentioned above, the lack of agreement on the set of features that best discriminate emotions from speech makes it difficult to compare classification algorithms.

The findings from the above studies were all based on experiments made on typical speech. Therefore, it is unknown what set of features and classifiers will work best for dysarthric speech. The next section will present the first dysarthric speech emotion recognition system.

7.3 Automatic dysarthric speech emotion recognition

The aim of this set of experiments is to explore the feasibility of automatically recognising emotions from dysarthric speech. It sets the baseline results for SER on the dysarthric speech part of DEED. Setting a baseline based on general techniques, previously used for typical speech, is important to be able to compare it later on to techniques turned specifically for dysarthric speech data. These baseline experiments also give an insight into the level of difficulty of the classification problem, and the performance of different classifiers using the same feature set on the dysarthric speech data.

As has been discussed in Chapter 2 and 4, people use different terms to describe their emotions in everyday life. Therefore, two main approaches in describing the emotional space have emerged. The first approach uses a finite set of emotions to represent the emotional space. The second approach uses dimensions to represent possible states in the emotional space. Arousal and valence are examples of these dimensions. The former approach is known as the categorical (discrete) emotion models and the later is known as the dimensional emotion models (Bojanic et al., 2013). For these baseline experiments, both classification approaches have been adopted.

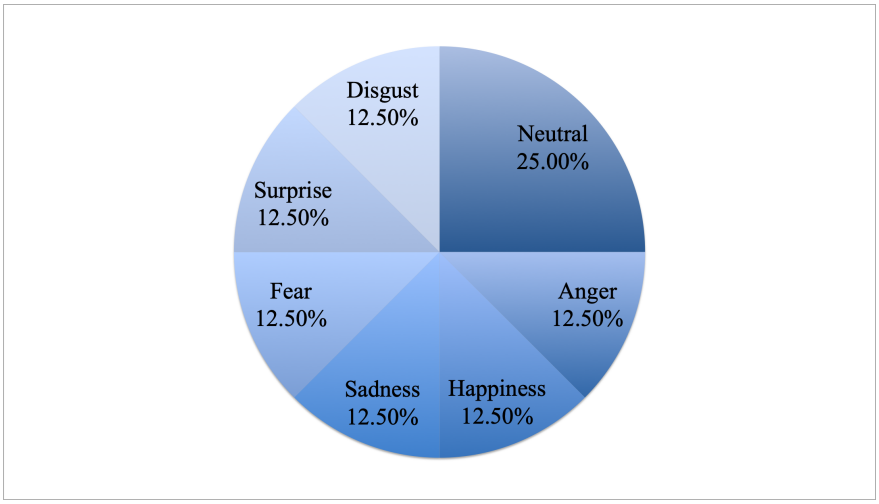


Fig. 7.2 The distribution of the emotion classes in DEED.

This section outlines the baseline experiments and presents all the details in terms of the database, the used feature set and classifiers. Finally, the results are presented and discussed.

7.3.1 Data

These experiments were carried out on the dysarthric speech part of DEED. All the details about the database in terms of the speakers, emotions, and recording settings can be found in Chapter 4. Figure 7.2 shows the distribution of the emotion classes in DEED. As can be seen, all classes has the same number of samples except for neutral where it has double the number of samples.

7.3.2 Feature extraction

Due to the gap in the exact relation between physical acoustic features and perceived features, there is no agreement yet on the set of features that best describes emotions in speech. As has been discussed in Section 7.1, the level of comparability between results reported in the literature is low. Apart from the different classifiers, evaluation strategies, and databases used, the diversity in the sets of features is high. Even when two studies use the same features, one or more of the following is usually found: the underlying parameters used in the extraction are different, the exact parameters are not reported, the selection and implementation of global features (functionals) are not the same, and/or the strategies used in features reduction are different. The choice of the set of features contributes heavily to the performance of the SER model and it is one of its main challenges (Atmaja, 2019; Wang et al., 2020). Since these experiments aim to develop a baseline model for the dysarthric SER,

25 Low Level Descriptors (LLD)	
Frequency related features	Log F0, jitter, formant 1, 2, and 3 frequency, and formant 1, 2, and 3 bandwidth
Energy related features	Shimmer, intensity, and HNR
Spectral related features	MFCC 1-4, spectral flux, alpha ratio, Hammarberg index, spectral slope 0–500 Hz, spectral slope 500–1500 Hz, formant 1, 2, and 3 relative energy, harmonic difference H1–H2, harmonic difference H1–A3
6 Temporal features	
Rate of loudness peaks, mean length of voiced and unvoiced regions, standard deviation of voiced and unvoiced regions, number of continuous voiced regions per second	

Table 7.2 eGeMAPS features (LLD and temporal features)

the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) standard feature set was chosen as the feature set (Eyben et al., 2016). These features have been suggested to comprise the majority of the features that are related to emotions. The eGeMAPS has been widely used as a benchmark for emotion recognition studies (Cummins et al., 2017; Neumann and Vu, 2017; Ringeval et al., 2015, 2016; Tian et al., 2016; Trigeorgis et al., 2016). It contains spectral, prosodic, cepstral, and voice quality information such as F0, jitter, shimmer, harmonic differences, and MFCC for a total of 88 features. Table 7.2 presents the low level descriptors (LLD) and the temporal features of eGeMAPS. A number of functionals, also called descriptive statistics, has been applied on these features such as the arithmetic mean and coefficient of variation. The details of these functionals can be found in (Eyben et al., 2016). Also, using a standard feature set helps in making the results reproducible. The features were automatically extracted using the openSMILE toolkit (Eyben et al., 2013) using the default parameters (Eyben et al., 2016). All features were standardised so that they have zero mean and unit variance.

Feature reduction

Classifier performance can increase when including additional features to the input up to a certain point. The performance, however, can decrease when further features are included especially if these features do not contribute much to class separability, mainly in the situations where there is lack of data (Murphy, 2012). As discussed earlier in this chapter, the curse of dimensionality is one of the main problems in machine learning and dimensionality reduction plays an important role in the classification performance. There are a number of dimensionality reduction techniques available. One of the widely used is the PCA

(Jolliffe, 2011; Murphy, 2012; Rogers and Girolami, 2016). PCA is a data preprocessing and dimensionality reduction technique. It is based on a mathematical algorithm that simplifies the complexity and reduces the dimensionality of the input vector, which is the features representing the data, while keeping most of the variation in it. Using PCA can improve the classifiers performance especially when applied on high dimensional feature vector (Howley et al., 2005).

To see the effect of using PCA on our data, a subset of the experiments were replicated using PCA with the top 30 components and compared to the results obtained with the full set of features (88 features).

7.3.3 Classification

Two approaches of emotion classification were performed: discrete and dimensional. In the discrete approach, 7-class and 4-class classification problem were reported. The 7-class classification problem included the full set of emotions while the 4-class classification problem included 'angry', 'happy', 'sad', and 'neutral' emotions only. In the dimensional approach, emotions were mapped to three classes along the valence axis, namely positive, neutral, and negative. The mapping of emotions to the three classes was adapted from Bojanic et al. (2013) where they mapped 5 emotions (anger, happiness, sadness, fear, and neutral) onto three classes. However, in DEED there are more emotions (surprise, and disgust). Therefore, the completed mapping process was achieved using the positions of these two emotions on the activation-evaluation space which resulted in mapping the disgust to the negative class and the surprise to the positive class. Table 7.3 represents the final mapping of emotions. The dimensional classification approach was also performed using the full set of emotions and using only four classes of emotions, namely 'angry', happy', 'sad', and 'neutral'.

For classification, a speaker-dependent approach was used where the model is trained and tested using the target speaker's speech characteristics. The results of each speaker is reported separately as have been done in other previous studies using different data sets (Espana-Bonet and Fonollosa, 2016; Jackson and Haq, 2011; Joy and Umesh, 2018). This helps in setting a clear baseline for each speaker. In this experimental study, the performance of the following classifiers were tested: One Versus Rest (OVR) SVM with RBF kernel, Logistic Regression (LR), Decision Tree (DT), and Adaboost. The selection of these traditional classifiers over deep learning models was based on the size of the data; as traditional classifiers usually require lesser data to work well while deep learning models are known to work better in the presence of sufficient training data. Figure 7.3 illustrates the structure of the experimental setup. For SVM, the regularization parameter (C) and the gamma coefficient of the kernel

were set to 5 and 0.01, respectively. For LR, the penalty and solver parameters were set to L2 and 'newton-cg', respectively. For Adaboost, the maximum number of estimators was set to 1200. The rest parameters were set to their default values. All classifiers were trained using Python Scikit-learn package (Pedregosa et al., 2011).

7.3.4 Performance evaluation

For evaluation, a five-fold cross-validation technique was used. Cross-validation is a common validation process used to evaluate machine learning models and increase the reliability of the results on the case of limited data (Murphy, 2012). The data in this approach is divided into five groups. For each group, the data of that group is held out as a test set where all the remaining groups are used as training sets. The splits in each fold were stratified to preserve the samples' distribution in each emotion. The resultant confusion matrix was formed by adding up the confusion matrices from all five folds. The overall performance of the classifier is determined by the average performance for all test sets. For each classifier, four performance metrics were calculated which are accuracy, unweighted average recall (UAR), unweighted average precision (UAP), and unweighted average F-score (UAF). These performance metrics were calculated using the following equations:

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + TrueNegative + FalseNegative} \tag{7.1}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{7.2}$$

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{7.3}$$

$$F - score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{7.4}$$

A True Positive is a result of correctly classifying a positive class instance. Similarly, a True Negative is a result of correctly classifying a negative class instance. Whereas a False Positive is a result of incorrectly classifying a negative class instance as positive. Similarly, False Negative is a result of incorrectly classifying a positive class instance as negative. Therefore, accuracy measures the correct predicted instances over all predicted instances. Recall measures the proportion of the actual positive instances that has been classified correctly by the classifier as positives. Precision measures the proportion of the positive classified instances that are really positives. F-score is an important measure to look at when a balance between recall and precision is needed as it takes into account both the False

Classes	Positive	Neutral	Negative
Emotions	Happiness, surprise	Neutral	Anger, fear, sadness, disgust

Table 7.3 Mapping of emotions onto 3 classes along the valence axis.

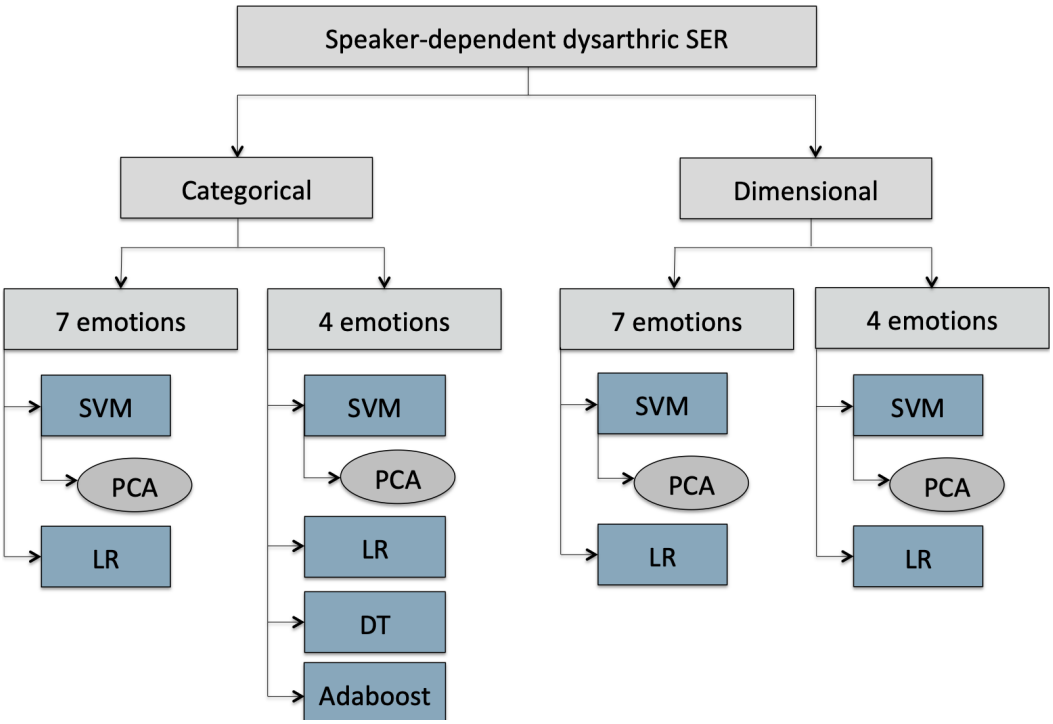


Fig. 7.3 Speaker-dependent experimental setup.

Positive and False Negative. More details about these measure and how they are calculated can be found in (Murphy, 2012).

7.3.5 Results and discussion

The experiments results for the categorical classification approach using 7 emotions and 4 emotions are presented in Table 7.4 and Figure 7.4, respectively. Similarly, the results for the dimensional classification approach using 7 and 4 emotions are presented in Table 7.5.

For all classification approaches using the full set of emotions and the reduced set, the performance of all classifiers are above chance performance, even for speaker DS01F who has severe dysarthria and low speech intelligibility. In fact, the results of classifying 7 emotions generally outperforms the human performance reported in Chapter 6 Section 6.3 Table 6.2 on all speakers except on speaker DS04F. The performance of the classifiers improved for all speakers when the number of emotions were reduced as would be expected. In all of the

experiments, highest classification performance is achieved on speaker DS02F. An accuracy of 67.50% using SVM and 90% using LR is achieved for the categorical classification approach using 7 and 4 emotions, respectively. This could be due to this speaker having mild dysarthria. The performance of the different classifiers are comparable. In fact, by looking into the 95% confidence interval of the classifiers accuracy in Figure 7.4, it could be inferred that the differences between the classifiers accuracy, for each speaker, are most generally not statistically significant.

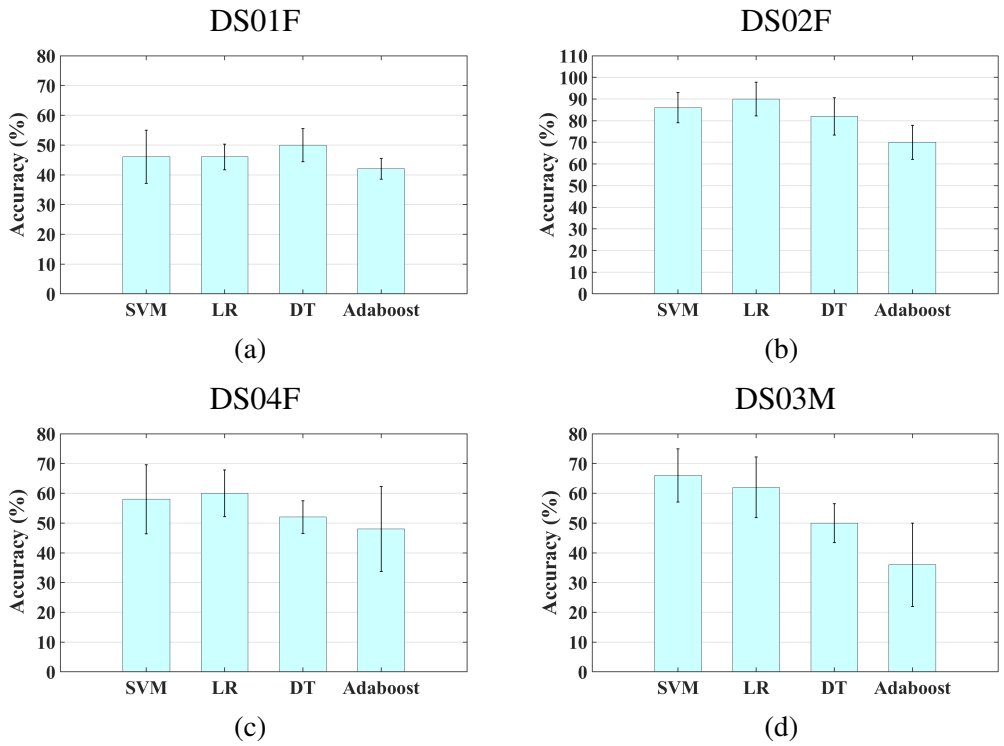
In addition to general performance metrics, confusion matrices helps in giving more information about the classification performance. For example, they illustrate what emotions are considered hard and which are easy to classify in the chosen feature space. Also, they can give an insight into what emotions are considered confusing, i.e., mostly get classified as another emotion. Figure 7.5 presents the confusion matrices for the categorical classification approach using 7 and 4 emotions for all speakers. For each speaker, the confusion matrix of the best classifier is presented where the rows present the actual emotions and the columns present the classified emotions. From Figure 7.5, it is observed that for all speakers, when classifying the full set of emotions and the reduced set, i) 'anger' is never confused with 'sad' and 'sad' is rarely confused with 'anger', ii) 'sad' is mostly confused with 'neutral', iii) for all speakers, except speaker DS01F, 'anger' and 'neutral' are mostly considered as non confusing pairs, and iv) 'anger' appears to be the easiest emotion to classify while 'happy' appears to be the most difficult emotion to classify. Similar results were observed on typical speech in (Dai et al., 2008; Yacoub, 2003). It is also observed that 'happy' and surprise' are mostly confused with each other. The biggest improvement is achieved for classifying 'happy' for all speakers when reducing the classification problem to 4 emotions.

Figure 7.6 presents the confusion matrices for the dimensional classification approach using 7 and 4 emotions for all speakers. Classifying emotions in the positive class, which includes ('happy' and 'surprise') and 'happy' when using 7 emotions and 4 emotions, respectively, appears to be the most difficult. This is consistent to the findings from the categorical classification. The biggest improvement is achieved for classifying 'neutral' for all speakers when reducing the classification problem to 4 emotions.

A comparison in terms of the classifier accuracy when using the full feature set (88 features) and the reduced set (30 features) obtained from using PCA is illustrated in Figure 7.7. As can be seen, generally, using PCA did not help in improving the SVM performance on these data. The reason could be that the dimensionality of the full set of features is not that high and most of the features have its contribution in distinguishing the different classes of emotions (Bishop, 2006). For more details on the classifier performance including recall, precision, and f-score when using PCA, the reader is referred to Appendix C.

Speaker	SVM				LR			
	Accuracy	**UAR**	**UAP**	**UAF**	**Accuracy**	**UAR**	**UAP**	**UAF**
DS01F	**28.75**	**27.14**	27.84	27.16	27.50	25.00	26.82	25.32
DS02F	**67.50**	**64.29**	62.55	62.57	63.75	60.00	59.77	59.65
DS04F	28.75	23.57	19.59	21.37	**37.50**	**32.86**	31.22	31.97
DS03M	**47.50**	**43.57**	44.07	42.83	43.57	41.43	40.77	40.58

Table 7.4 Speaker-dependent categorical classification results using 7 emotions.



Speaker	SVM			LR			DT			Adaboost		
	UAR	**UAP**	**UAF**	**UAR**	**UAP**	**UAF**	**UAR**	**UAP**	**UAF**	**UAR**	**UAP**	**UAF**
DS01F	42.50	42.13	42.12	41.25	42.64	41.44	**48.75**	**48.17**	**48.42**	28.75	19.64	21.57
DS02F	85.00	86.43	85.64	**90.00**	**91.88**	**90.60**	82.50	82.44	82.25	65.00	70.45	64.67
DS04F	53.75	49.91	51.31	**58.75**	**55.92**	**56.92**	46.25	43.71	44.75	32.50	36.78	33.90
DS03M	**62.50**	**64.42**	**62.60**	58.75	59.13	58.83	53.75	53.44	52.50	51.25	46.88	47.79

Fig. 7.4 Speaker-dependent categorical classification results using 4 emotions with error bars show the 95% confidence interval.

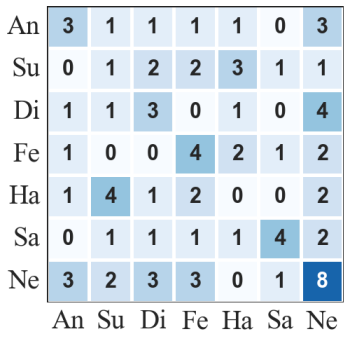
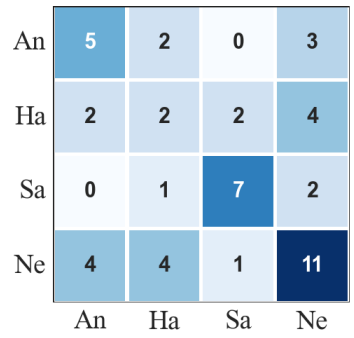
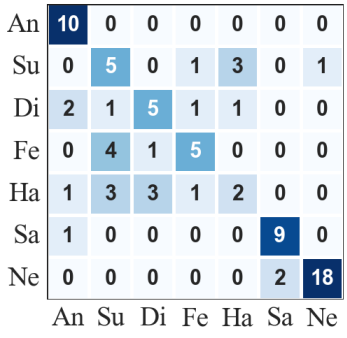
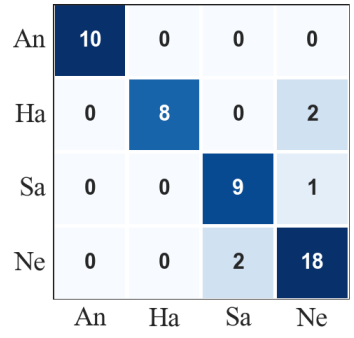
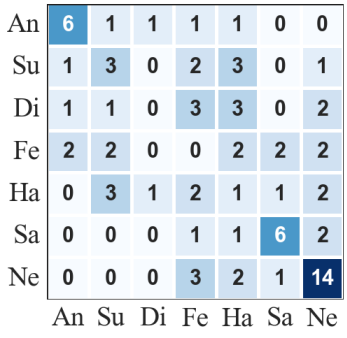
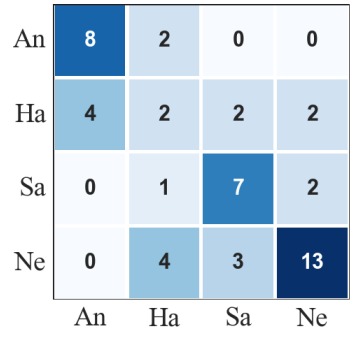
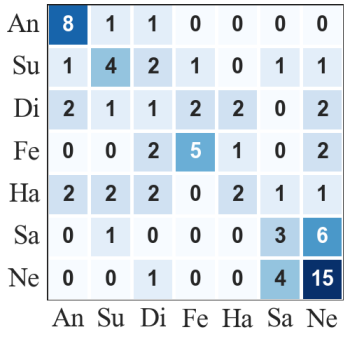
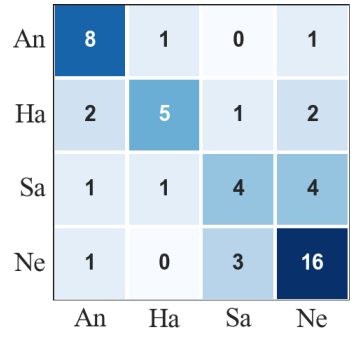
Speaker	7 emotions classification	4 emotions classification
DS01F	 SVM	 DT
DS02F	 SVM	 LR
DS04F	 LR	 LR
DS03M	 SVM	 SVM

Fig. 7.5 Confusion matrices of the categorical classification using 7 and 4 emotions. (rows= actual emotions and columns= classified emotions, An= angry, Su= surprise, Di= disgust, Fe= fear, Ha= happy, Sa= sad, and Ne= neutral).

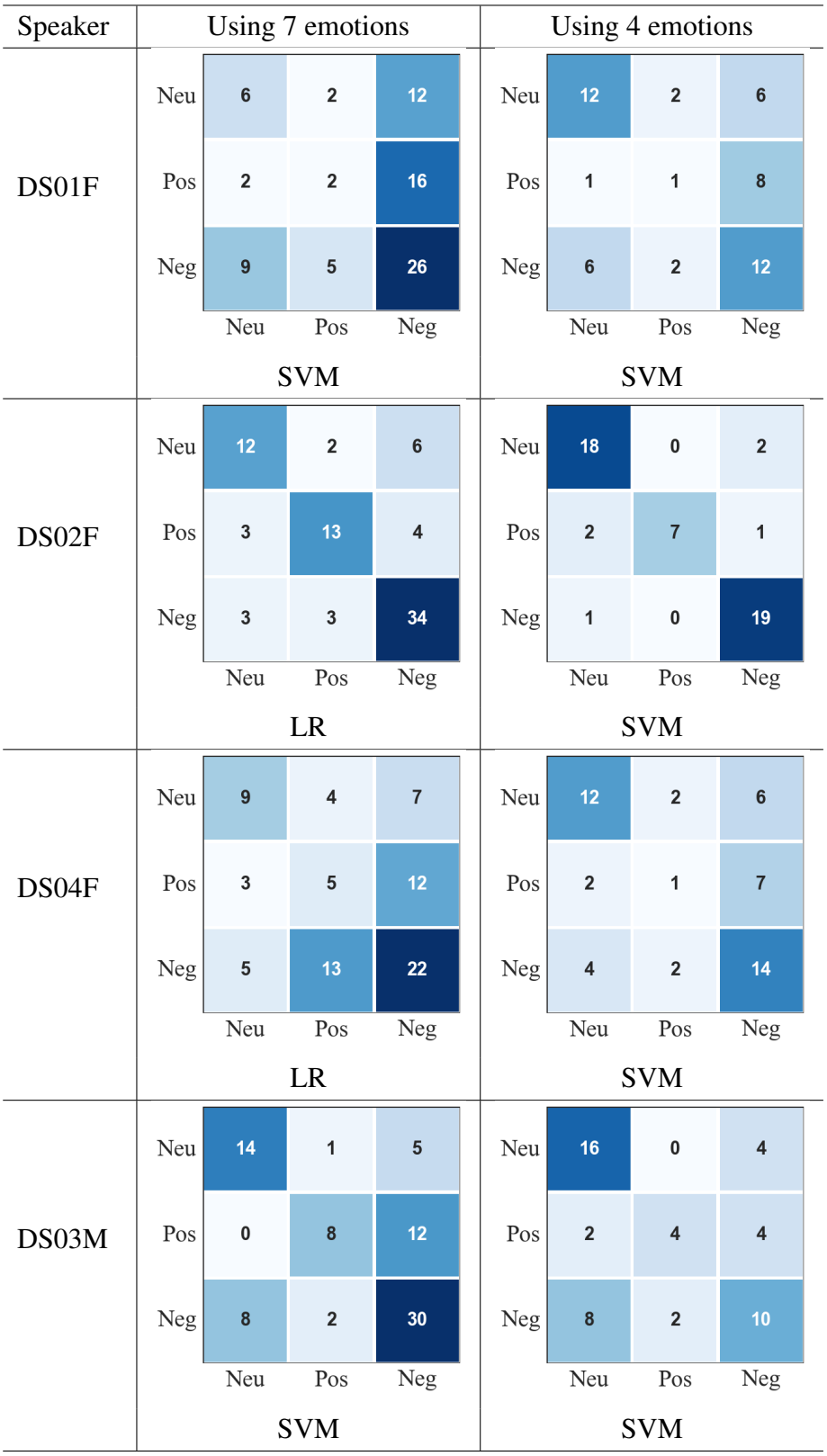
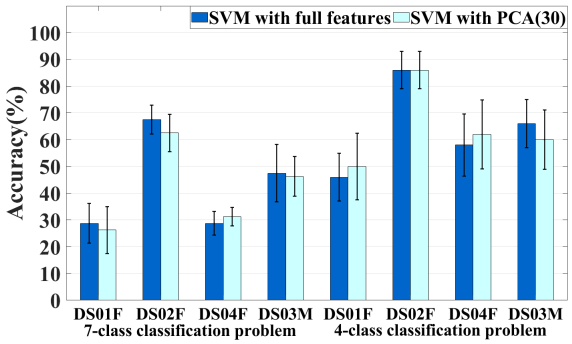


Fig. 7.6 Confusion matrices of the dimensional classification using 7 and 4 emotions. (rows= actual emotions and columns= classified emotions, Ne= neutral, Pos= positive, Neg= negative).

Speaker	Emotions	SVM				LR			
		Accuracy	**UAR**	**UAP**	**UAF**	**Accuracy**	**UAR**	**UAP**	**UAF**
DS01F	7 emotions	**42.50**	35.00	35.22	33.85	38.75	35.83	35.70	35.72
	4 emotions	**50.00**	43.33	43.10	42.35	48.00	43.33	43.00	42.73
DS02F	7 emotions	72.50	69.17	72.72	70.19	**73.75**	**70.00**	72.05	70.84
	4 emotions	**88.00**	**85.00**	90.69	86.88	80.00	76.67	82.34	87.09
DS04F	7 emotions	45.00	39.17	39.81	38.80	45.00	**41.67**	43.11	42.26
	4 emotions	**54.00**	**46.67**	46.17	45.36	46.00	43.33	44.03	43.33
DS03M	7 emotions	**65.00**	61.67	66.73	62.42	62.50	61.67	61.06	61.27
	4 emotions	**60.00**	**56.67**	61.25	57.40	54.00	51.67	51.35	51.48

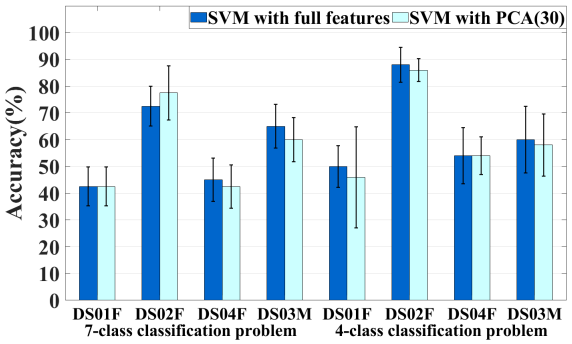
Table 7.5 Speaker-dependent dimensional classification results using 7 and 4 emotions.

Speaker-dependent accuracy using the categorical approach



(a)

Speaker-dependent accuracy using the dimensional approach



(b)

Fig. 7.7 Classification accuracy results when using the full features set and the reduced feature set with 95% confidence interval.

Developing a SER on typical speech is not the focus of this research, however, it would be useful to have a baseline results on the typical speech part of DEED. Setting a baseline is important to be able to compare it to the results obtained from the dysarthric SER. Also, it will give an insight into the level of difficulty of the classification problem. Therefore, using the same settings in terms of the feature set and classification tasks, a SER on the typical speech part of the DEED, was developed. The performance of two classification approaches, speaker-dependent and speaker-independent using two classifiers, OVR-SVM and LR,is presented in details in Appendix D. Using the categorical classification approach, the average accuracy of classifying 7 emotions is 50.45% and 47.44% using the speaker-dependent approach and speaker-independent approach, respectively. These classification performance are comparable to what is achieved on another emotional typical database, SAVEE (see Appendix D). These results show that this is a difficult task even when staying within the typical domain.

7.4 Conclusion

In this chapter, a review of the popular speech emotion recognition techniques used in the literature has been given. After that, a dysarthric SER model has been implemented and the results on four speakers with dysarthria have been presented and discussed. Given the nature of the dysarthric speech and its phonological and prosody dimensions limitations, the experiments in this chapter were conducted to investigate i) the feasibility of automatically recognizing emotions from dysarthric speech, and ii) what emotions in the dysarthric speech are found to be close to each other (confusing) in the chosen feature space. It was demonstrated that dysarthric speech emotion recognition could be possible. In fact, the results of recognising 7 and 4 emotions using categorical and dimensional classification approaches were very encouraging including the results of speaker DS01F who has a severe dysarthria and highly unintelligible speech. All results were above chance performance which confirms the initial findings in Chapter 6 that people with dysarthria may have control to perform systematic changes in their speech to communicate emotions. It was observed from all of the experiments that the performance of the classifiers increased when the number of emotions decreased. This is expected as the classification problem gets simpler in terms of the number of classes that the data will be classified into. Thus, generally speaking, the performance using 4 emotions is better than using 7 emotions and the performance using the dimensional classification approach is better than the categorical approach.

Most deep learning techniques requires large amount of data to be able to perform well. However, collecting dysarthric emotional speech is more challenging than typical speech data.

Therefore, given the encouraging results obtained from the speaker-dependent dysarthric SER models presented in this chapter, it would be interesting to investigate whether dysarthric SER model would benefit from being trained on typical speech data. In other words, investigating whether people with dysarthria share some similarities with typical speakers while expressing emotions and to what extent models trained on typical speech data can accurately classify emotions in dysarthric speech. This investigation will be the focus of the next chapter.

Since this research is more of an exploratory kind with the objective of setting suitable baseline results of common techniques, investigates what can be achieved, and looks into ways to improve the classifiers performance with the limited data in hand, it is going to be beneficial to carry on the experiments on a focused approach rather than running the experiments on multiple approaches using different sets of emotions. Therefore, the focus of the rest of this thesis is going to be on classifying 4 emotions: 'angry', 'happy', 'sad', and 'neutral' using the categorical approach. The selection of this set is guided by two main reasons. First, given that this is the first investigation using these data, starting with a smaller non-overlapping set can provide the base for a more focused initial exploration of the problem. Second, based on the findings of the conducted survey in Chapter 3, 'anger', 'happiness', and 'sadness' were chosen by people with dysarthria as the most important emotions in terms of being able to communicate them successfully (Alhinti et al., 2020a). 'Neutral' is included as a baseline condition.

Chapter 8

Automatic Dysarthric Speech Emotion Recognition Using Models Trained on Typical Speech

Part of the content of this chapter has been published in INTERSPEECH 2020 (Alhinti et al., 2020b).

8.1 Introduction

Employing state of the art SER methods directly to the domain of dysarthric SER is challenging due to several reasons. Mainly, these methodologies require large amounts of data to perform well. However, collecting large dysarthric emotional speech data is more challenging than emotional typical speech. Also, the high inter-speaker variability found in this group of speakers in comparison to typical speakers poses another challenge (Christensen et al., 2012, 2014; Darley et al., 1969a; Hawley et al., 2007; Ma et al., 2010; Wilson, 2000; Xiong et al., 2020; Yue et al., 2020b).

Based on i) the acoustic analysis performed in Chapter 5 where it was found that the changes made by speakers with dysarthria to some of the acoustic features when communicating emotions appear similar to those of typical speakers, despite speakers with dysarthria having a more limited phonetic and prosodic control, and ii) the performance of the targeted dysarthric SER models presented in Chapter 7, where the recognition results were very encouraging for all speakers, a tendency to explore the possibility of recognising dysarthric emotional speech using models trained on typical speech arise and forms the motivation of this chapter. The use of typical speech data to boost the performance on dysarthric speech data has been applied successfully to the domain of dysarthric Automatic Speech

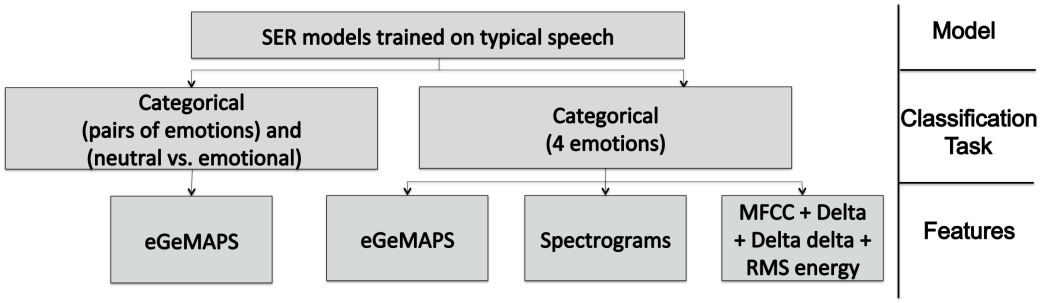


Fig. 8.1 Speaker-independent experimental setup.

Recognition (ASR) (Xiong et al., 2019). This motivates the investigation of whether people with dysarthria share some similarities with typical speakers while expressing emotions and to what extent SER models trained on typical speech can accurately classify emotions in dysarthric speech. Therefore, this chapter will present several speaker-independent SER models trained on typical speech data that either classify four emotions or pairs of emotions. The effect of training a model using a mixed emotional speech (typical and dysarthric) will also be investigated. Figure 8.1 presents the experiments setup. As can be seen, several feature sets will be investigated. Same to the experiments in Chapter 7, all the models in this chapter will be tested on the dysarthric speech part of DEED. All the details about the database in terms of the speakers, emotions, and recording settings can be found in Chapter 4. The rest of this chapter is structured as follows: Section 8.2 illustrates the training and test data. Section 8.3 presents dysarthric SER models that aim to distinguish pairs of emotions and discriminate emotional speech from neutral speech. Section 8.4 demonstrates the different dysarthric SER models that aim to classify 4 emotional states, namely 'angry', 'happy', 'sad', and 'neutral. Finally, a discussion and a conclusion is presented in Section 8.5.

8.2 Training and test data

It is widely believed that females express emotions differently than males do in many cultures (Brody and Hall, 2008). In some cultures, females are known to be more emotionally intense, more expressive, and more skilled in the employment of nonverbal cues in some emotions than males do (Briton and Hall, 1995; Hess et al., 2000; Plant et al., 2000; Robinson and Johnson, 1997; Timmers et al., 2003; Wood et al., 1989). Several studies, as well, showed the correlation of various acoustic measures, such as F0, with gender (Biemans, 2000; Chen et al., 2020a; Izadi et al., 2012; Mendoza et al., 1996; Teixeira and Fernandes, 2014). Therefore,

this was taken into consideration while choosing the training and test datasets (Brody and Hall, 2008). Thus, from DEED, a gender-based training data was chosen as follows: all female typical speakers were set as training data when the test data was set to a female speaker with dysarthria. The same method was used for male speakers.

8.3 Classifying two emotional states

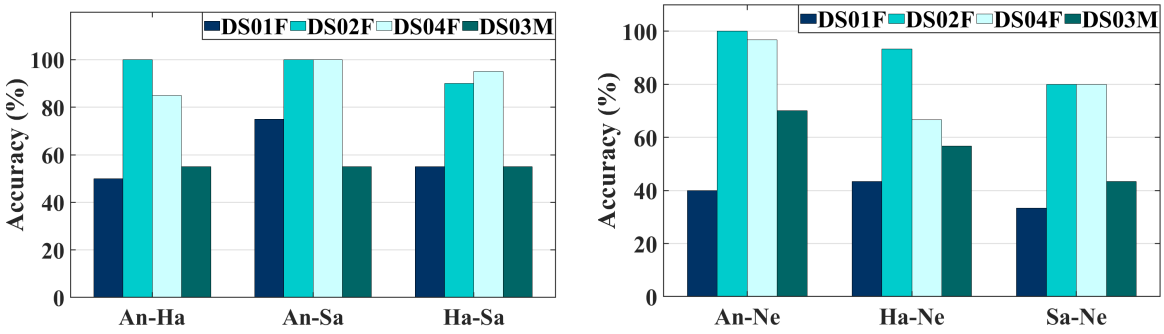
The aim of these experiments is to test the i) feasibility of classifying pairs of emotions and ii) how well dysarthric emotional speech can be distinguished from dysarthric neutral speech using speaker-independent models trained on typical speech data. This will enable an analysis of which emotions in the dysarthric speech are found to be close to each other (confusing) in the chosen feature space.

Feature extraction and classifiers

Similar to the feature set used in the speaker-dependent dysarthric SER model presented in Section 7.3 in the previous chapter, the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) standard feature set was chosen as the feature set (Eyben et al., 2016). It contains spectral, prosodic, cepstral, and voice quality information such as F0, jitter, shimmer, harmonic differences, and MFCC for a total of 88 features. For more details about this feature set, the reader is referred to Section 7.3.2. In terms of the classifier, a OVR-SVM with RBF kernel was used. SVMs are particularly well-suited to sparse data domains (Fauvel et al., 2006, 2008). The regularization parameter (C) and the gamma coefficient of the kernel were set to 5 and 0.01, respectively. These values were set based on grid search. The classifier was trained using Python Scikit-learn package (Pedregosa et al., 2011).

Performance evaluation and results

The results of classifying the following pairs: 'anger/happy', 'anger/sad', and 'happy/sad' for all speakers with dysarthria are presented in Figure 8.3 and 8.2.a. High classification accuracies are obtained for the two female speakers DS02F and DS04F with an accuracy of 100% achieved for distinguishing some of the pairs. This is not the case for speakers DS01F and DS03M where most of the results are at chance level or a little bit above chance except when recognising 'anger/sad' for speaker DS01F, where higher accuracy results are achieved. The most likely explanation is that speakers DS01F and DS03M are more severely dysarthric speakers than the the other two speakers. Figure 8.4 and 8.2.b present the results of distinguishing the three main emotions from neutral: 'anger/neutral',



(a) Results of classifying two emotional states

(b) Results of discriminating emotional speech from neutral speech

Fig. 8.2 The result of classifying pairs of emotions. (An= angry, Ha= happy, Sa= sad, Ne= neutral).

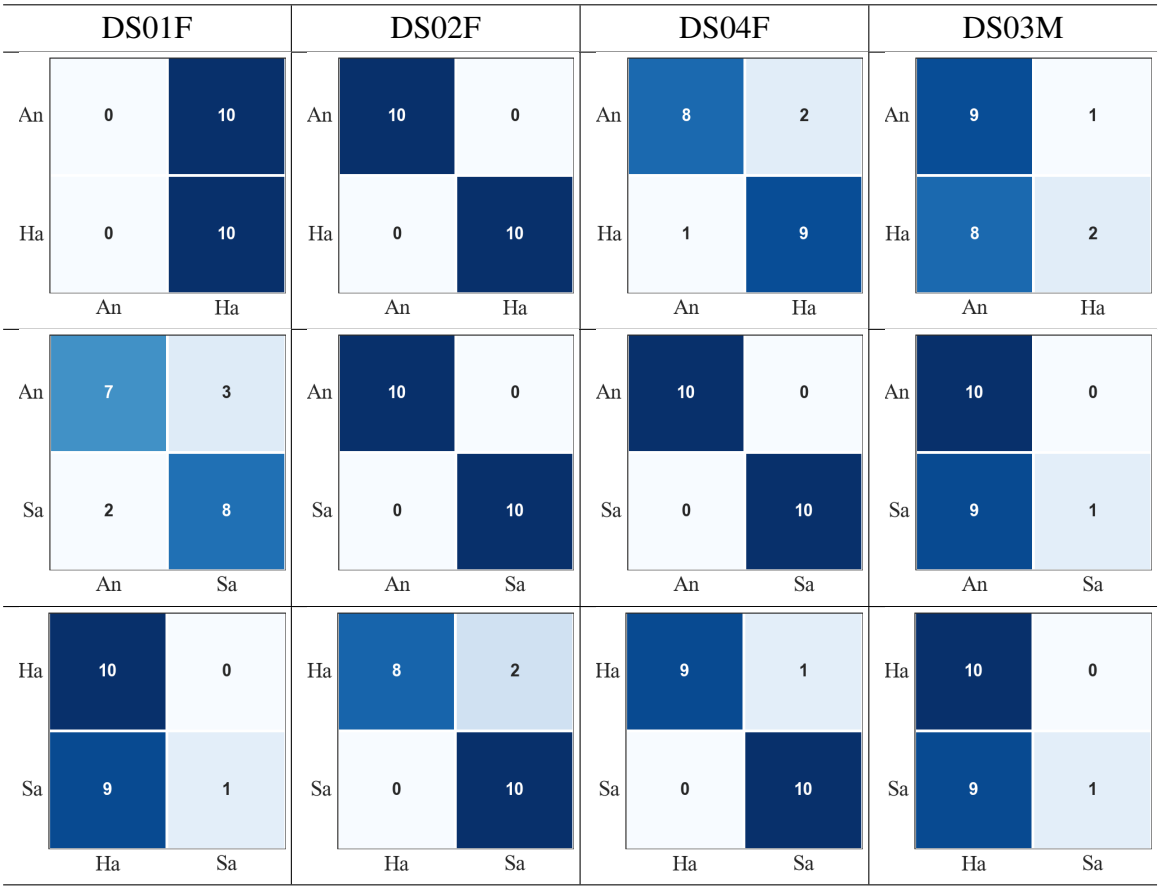


Fig. 8.3 Results of recognising pairs of emotions when trained using speaker-independent gender-dependent emotional typical speech. (rows = actual emotions and columns = classified emotions, An= angry, Ha= happy, Sa= sad).

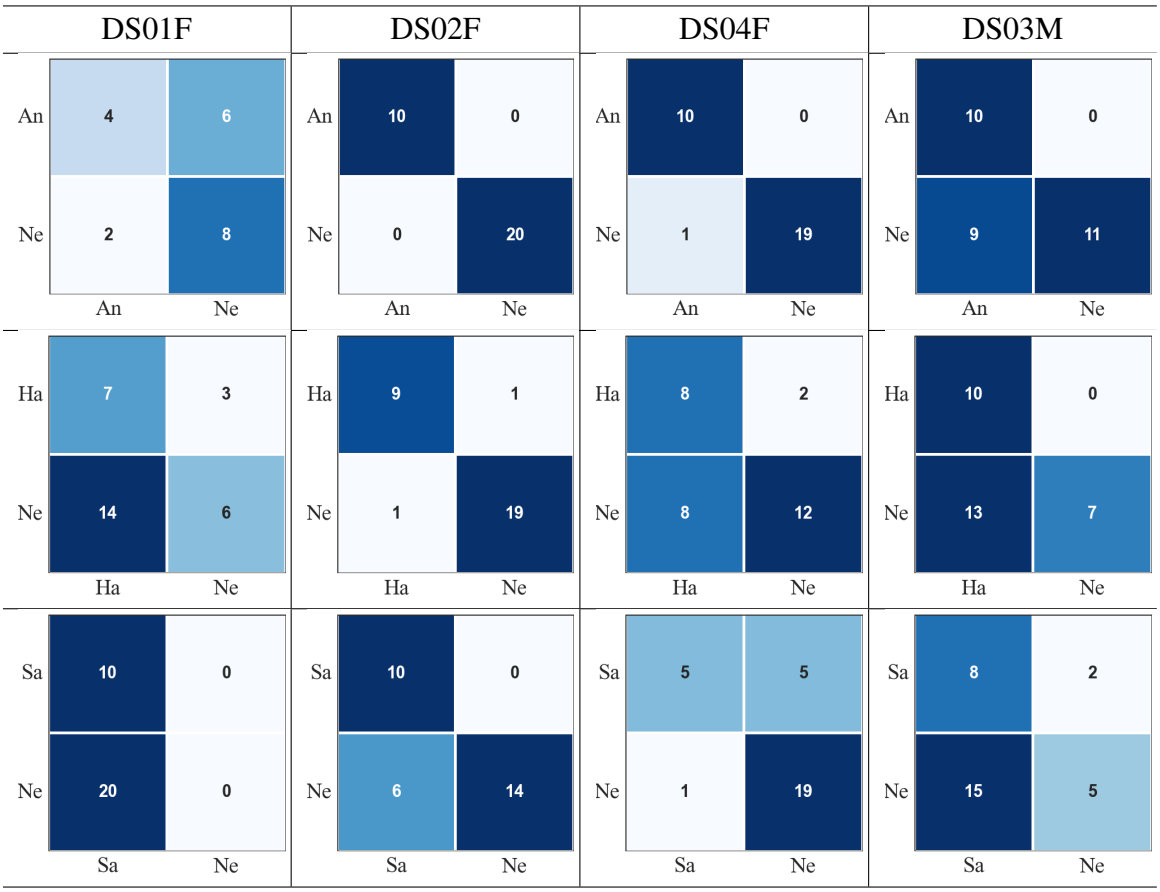


Fig. 8.4 Results of discriminating emotional speech from neutral speech when trained using gender-dependent emotional typical speech. (rows = actual emotions and columns = classified emotions, An= angry, Ha= happy, Sa= sad, Ne= neutral).

'happy/neutral', and 'sad/neutral'. The classifications accuracy of speaker DS01F are below chance level performance for all three pairs of emotions. For the other speakers, the results of discriminating emotional from non emotional speech are mostly very good. On average, it is found that the easiest pair to discriminate is 'anger/neutral'. Similar results were reported on typical speech in (Dai et al., 2008; Yacoub, 2003).

8.4 Classifying four emotional states

The aim of these experiments is to test the feasibility of classifying four emotional states using different speaker-independent models trained on typical speech data. As can be seen from Figure 8.1, three main experimental setups using different feature sets and classifiers have been used to test the feasibility of this task. A discussion of each one of them in terms of

the features and the classifiers used and the results obtained is presented below as categorised in Figure 8.1.

8.4.1 Experimental setup 1: eGeMAPS feature set

Feature extraction and classifiers

As in the previously reported experiments, the eGeMAPS feature set will be used. Also, in terms of the classifiers, the performance of the same four classifiers used in the speaker-dependent dysarthric SER model, presented in the previous chapter, were tested, SVM with RBF kernel, Logistic Regression (LR), Decision Tree (DT), and Adaboost. Choosing the same feature set and same classifiers allows a direct comparison between the two models; speaker-dependent models trained on the speech characteristics of each speaker with dysarthria (presented in the previous chapter) and speaker-independent models trained on typical speech (presented in this chapter). For SVM, the regularization parameter (C) and the gamma coefficient of the kernel were set to 5 and 0.01, respectively. For LR, the penalty and solver parameters were set to l2 and 'newton-cg', respectively. For Adaboost, the maximum number of estimators was set to 1200. Thes hyperparameter values were set based on grid search. All other hyperparameters were set to their default values. All classifiers were trained using Python Scikit-learn package (Pedregosa et al., 2011).

Performance evaluation and results

For each classifier, four performance metrics were calculated which are accuracy, Unweighted Average Recall (UAR), Unweighted Average Precision (UAP), and Unweighted Average F-score (UAF). Table 8.1 presents the classification results of all the speaker-independent models. A comparison between the performance of the speaker-dependent models obtained from the previous chapter,(from Figure 7.4), and the speaker-independent models trained on typical speech is depicted in Figure 8.5. For each speaker, the confusion matrix of the best classifier, which is not the same for each speaker, is presented in Figure 8.6 where the rows present the actual emotions and the columns present the classified emotions.

As can be seen from Table 8.1 and Figure 8.5, the performance of all classifiers for all speakers are above chance performance including, DS01F, who has a severe level of dysarthria and highly unintelligible speech. The highest recognition accuracy of 88% is achieved when training the model on typical speech data for speaker DS02F using Adaboost classifier. It is expected that the targeted speaker-dependent models give better performance given that they were trained using the target speaker's voice characteristics. However, from Figure 8.5 it is observed that for some speakers, the speaker-independent model outperforms

Speaker	**SVM**				**LR**			
	Accuracy	**UAR**	**UAP**	**UAF**	**Accuracy**	**UAR**	**UAP**	**UAF**
DS01F	32.00	28.75	21.21	23.21	42.00	33.75	27.08	28.46
DS02F	78.00	83.75	83.30	80.48	76.00	81.25	80.10	87.90
DS04F	**62.00**	**66.25**	63.76	62.78	54.00	61.25	62.50	53.67
DS03M	34.00	36.25	41.42	30.15	36.00	41.25	40.08	32.40
Speaker	**DT**				**Adaboost**			
	Accuracy	**UAR**	**UAP**	**UAF**	**Accuracy**	**UAR**	**UAP**	**UAF**
DS01F	**44.00**	**38.75**	31.11	34.5	26.00	30.00	55.32	17.86
DS02F	82.00	86.25	82.31	83.36	**88.00**	**87.5**	89.16	87.98
DS04F	40.00	43.75	44.01	40.29	56.00	55.00	66.59	50.74
DS03M	34.00	38.75	38.49	31.37	**40.00**	**40.00**	45.77	35.21

Table 8.1 Speaker-independent gender-dependent classification results of 4 classes of emotions using eGeMAPS feature set.

or has a very close performance to the speaker-dependent one such as the DT and Adaboost for speaker DS02F, SVM and Adaboost for speaker DS04F, and Adaboost for speaker DS03M (see Figure 7.4 in Chapter 7 for more details on the speaker-dependent results). From Figure 8.6, it is observed that for all speakers 'anger' is never confused with 'sad'. This aligns with the findings from the speaker-dependent models reported in the previous chapter in Section 7.3.5. Comparing the best models from the speaker-dependent and speaker-independent approaches, a big improvement is achieved for classifying 'happy' for speaker DS04F after training the model on typical speech. For speaker DS03M, the performance of classifying 'happy' is poor and it is highly confused with 'anger' when the model is trained on typical speech in comparison to the performance of the best speaker-dependent model. For all speakers, except speaker DS01F, 'anger' is never confused with 'neutral'. Similar to the findings from the speaker-dependent models, 'anger' appears to be the easiest emotion to classify. Similar results were observed on typical speech in (Dai et al., 2008; Yacoub, 2003).

The relatively good results when testing the dysarthric speakers on the speaker-independent gender-dependent models are encouraging. This indicates that a good level of typical-like emotion specific information is being successfully expressed. This might be predicted given the results from the statistical analysis of some acoustic features presented in Chapter 5, where no significant difference is found in some features for the condition (speech type) factor. However, of course not entirely but also this might be predicted from the relatively good results obtained from the subjective evaluation of the dysarthric speech in DEED presented

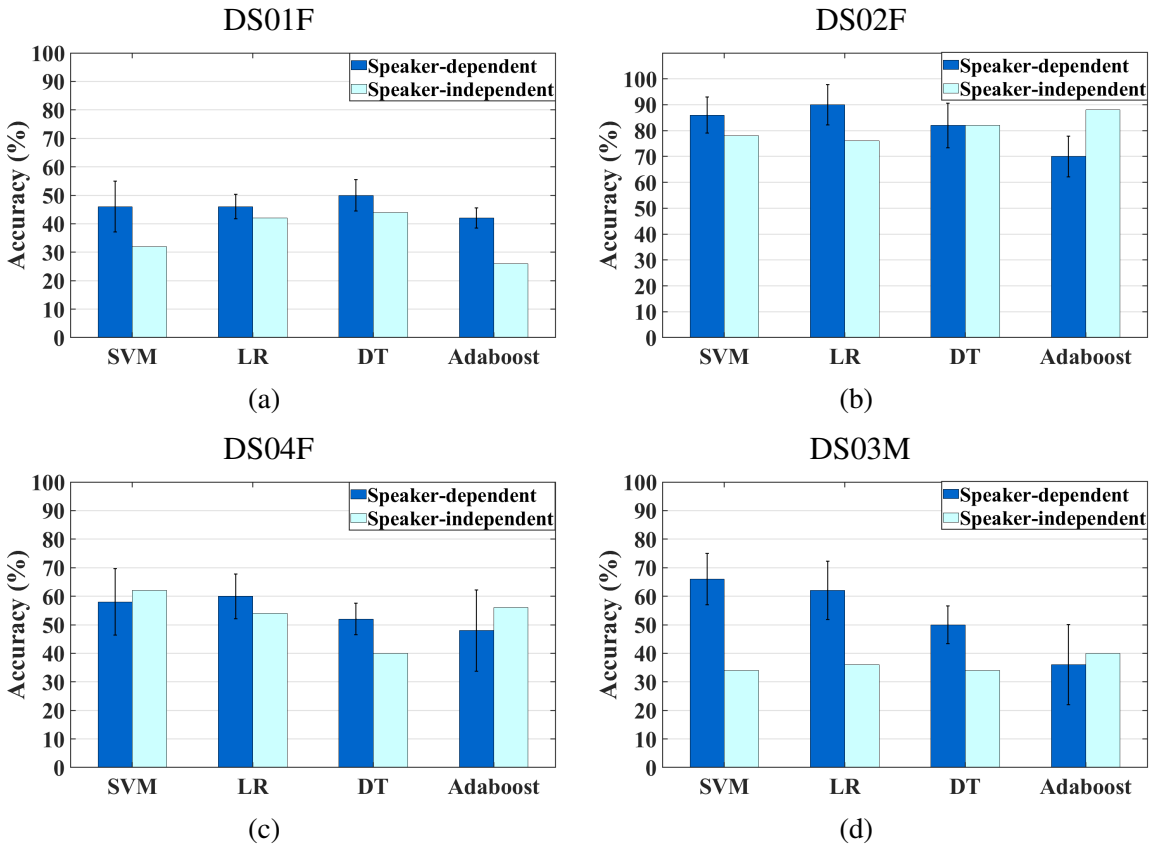


Fig. 8.5 Comparison between the classification accuracy results for speaker-dependent models with 95% confidence interval for the 5 folds and speaker-independent gender-dependent models using eGeMAPS feature set.

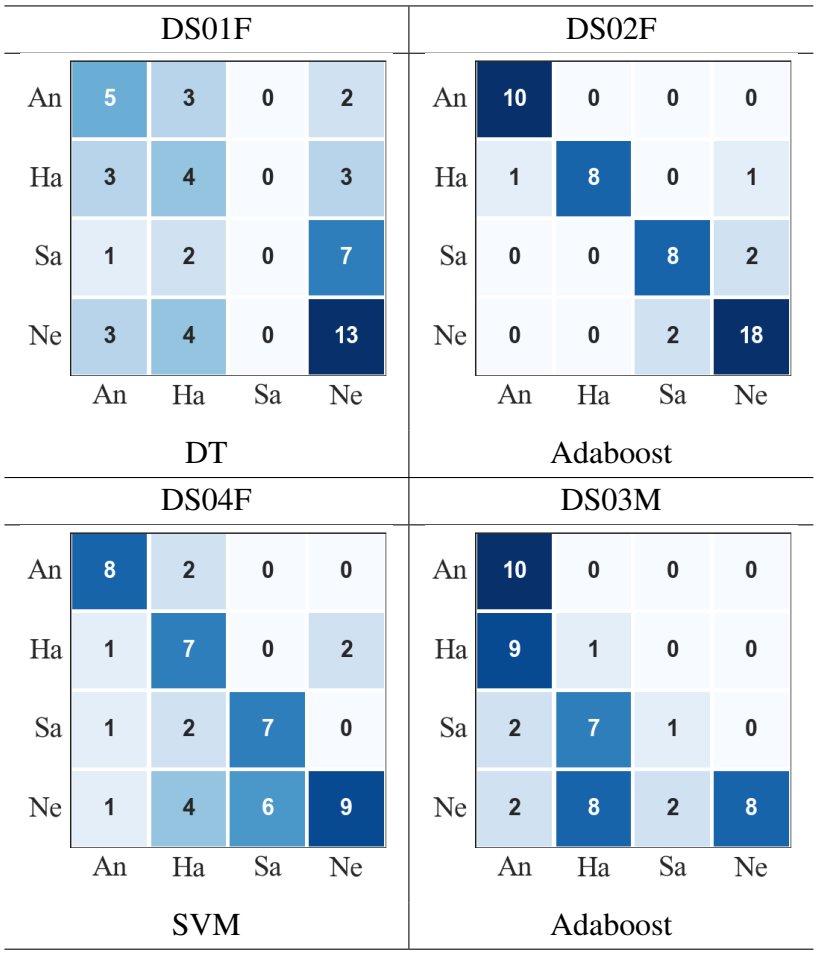


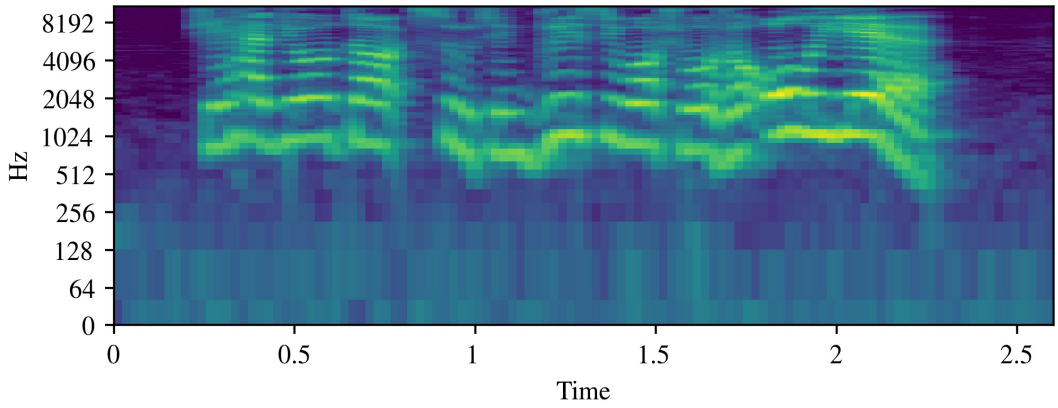
Fig. 8.6 Confusion matrices of the speaker-independent classification using eGeMAPS feature set. (rows= actual emotions and columns= classified emotions, An= angry, Ha= happy, Sa= sad, Ne= neutral).

in Chapter 6. In comparison, testing the typical speakers on the speaker-independent models using leave-one-speaker-out approach presented in Appendix D Table D.3, where the model is trained on all typical speakers data except one speaker who was held as a test set, achieved an average accuracy of 59.81%, showing that this is a difficult task even when staying within the typical domain.

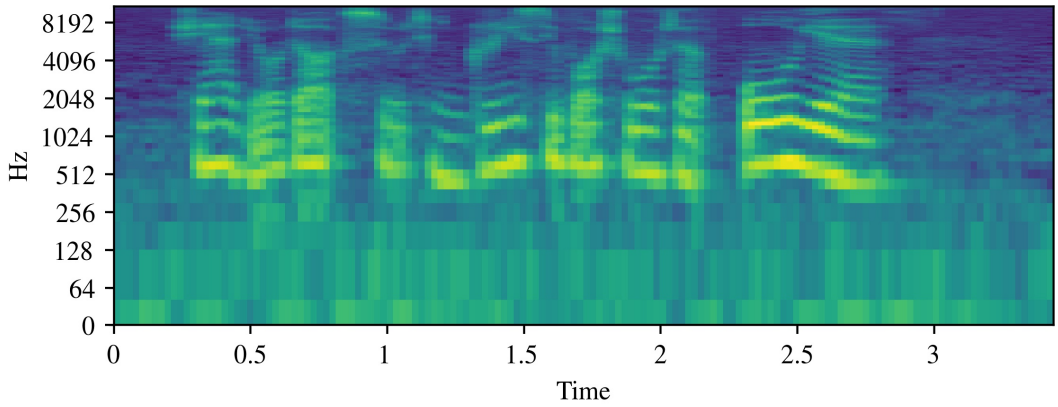
8.4.2 Experimental setup 2: Spectrograms

Feature extraction and classifiers

As have been discussed previously in Section 7.3.2, finding direct and clear relations between emotions and specific acoustic features is a very challenging task. Therefore, it is



(a) Angry



(b) Sad

Fig. 8.7 Mel-spectrograms of the same utterance by the same speaker, DS02F, spoken in (a) angry and (b) sad.

common to use speech spectrograms which encode acoustic features such as F0 of the whole utterance over using individual low level or high level parameters. Spectrograms are visual representation of the variation of intensity across frequency over time. In other words, it shows the signal intensity at different frequencies over time. They are a two-dimensional representation with a third dimension indicated by colors. The x axis represents the time while the y axis represents the frequency. The third dimension represents the amplitude of a specific frequency at a specific time. The intensity of the color indicates the amplitude of a frequency, where dark colors correspond to low amplitudes and bright colors correspond to high amplitudes. Spectrograms are computed by applying Fast Fourier transform (FFT) to each window of the divided speech waveform, where these windows are usually overlapped (Allen and Rabiner, 1977; Smith III, 2011).

Although spectrograms can be more susceptible to being affected by channel distortions, they proved their suitability for acoustic content representation in many speech analysis

tasks including speaker identification (Liu et al., 2018; Xie et al., 2019a), speech recognition (Agrawal and Ganapathy, 2017; Zhang et al., 2017), sound event classification (Dennis et al., 2011), and SER (Badshah et al., 2017; Fayek et al., 2017; Hajarolasvadi and Demirel, 2019; Han et al., 2020; Lech et al., 2018; Lim et al., 2016; Liu et al., 2020; Ma et al., 2018; Satt et al., 2017; Stola et al., 2018; Wang et al., 2020). Figure 8.7 presents two log scale Mel-spectrograms samples from DEED of the same utternace, ('She had your dark suit in greasy wash water all year'), spoken by the same speaker, DS02F, who has dysarthria, once in angry (Figure 8.7.a) and the other in sad (Figure 8.7.b). As can be seen in the angry spectrogram high and clear energy is found in high frequencies compared to the sad spectrogram.

Therefore, spectrograms were chosen as the feature set were discriminative features are learnt automatically and directly from spectrograms. In this setup, the log scale Mel-spectrogram was computed using Librosa package on python (McFee et al., 2015). The Mel scale is widely used in speech recognition and emotion recognition tasks. A sequence of overlapping Hanning windows were applied to each speech signal with window size of 128ms and window shift of 32ms. All generated narrowband spectrograms were resized to 64 x 64 using the flow_from_dataframe method from the Keras ImageDataGenerator class. The generated spectrograms were fed into two classifiers, OVR-SVM and two dimensional CNN. Since the generated spectrograms have three-dimensional shape and SVMs only accept one-dimensional input, spectrograms were reshaped accordingly. For SVM, an RBF kernel was chosen. The regularization parameter (C) and the gamma coefficient of the kernel were set to 5 and 0.01, respectively. These values were set based on grid search.

For the CNN, Figure 8.8 shows the proposed model architecture. The model consisted of two convolutional layers, one fully connected layer, and a softmax layer. The input of the network was (64 x 64 x 3) for 64 x 64 spectrograms images in RGB. Features were extracted from the initial convolutional layers using convolution operations. The first and second layers had a 32 and 64 (3 x 3) kernels, respectively. Rectified Linear Units (ReLU) was chosen as the activation function. ReLU has been proven to work well in neural networks. The second layer was followed by max-pooling layer of size 2 x 2 and a dropout layer with probability of retention p = 0.25. A flatten layer was added to connect the convolutional layers with the fully connected (dense) layer. The fully connected layer had a 128 nodes with ReLU as the activation function. Another dropout layer with probability of retention p = 0.5 was added. The last layer was the output layer which had 4 nodes, one node for each possible class (outcome). Softmax was chosen as the activation function which produces a vector that presents the probability of potential outcomes where the sum of these probabilities equals to one. The final prediction is based on the outcome with the highest probability. The model was implemented using Keras supported by Tensorflow backend (Chollet et al., 2015). For

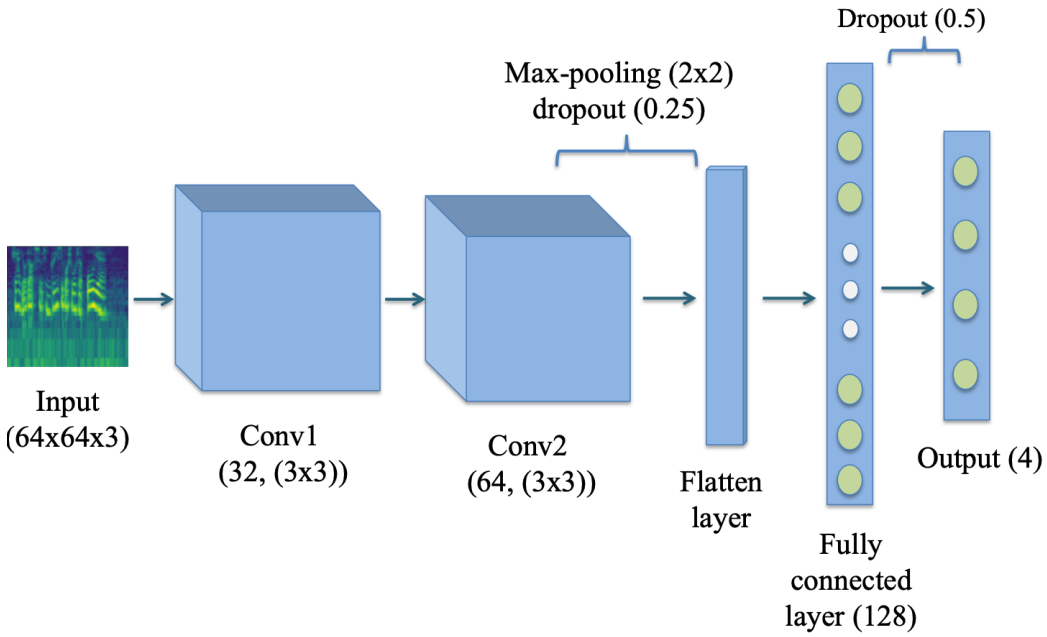


Fig. 8.8 2D CNN model architecture. (Conv = 2D-convolutional layer).

compilation, 'Adam' was chosen as the optimisation algorithm with a learning rate of 0.001 and 'categorical crossentropy' was chosen as the loss function. These hyperparameter values were set based on grid search. All other hyperparameters were set to their default values.

Performance evaluation and results

For each classifier, the same four performance metrics were calculated which are accuracy, UAR, UAP, and UAF. CNN models were trained using 11 epochs and 13 epochs for female and male speakers models, respectively as the network started to overfit after that. The batch size was 32. For model validation, Leave-One-Speaker-Out approach was used and the average performance of all runs is reported. Table 8.2 presents the classification results of all the speaker-independent models that were trained on gender-dependent typical speakers data using SVM and CNN. Figure 8.9 presents the confusion matrices of both models per speaker. For the CNN models, the average confusion matrix of all runs were computed and presented.

As can be seen from Table 8.2, the performance of all classifiers for all speakers are above chance performance except for speaker DS01F when using the SVM classifier. In terms of the recall, a better performance is achieved using the CNN for the female speakers with dysarthria DS01F, DS02F, and DS04F. In particular, using the CNN for speaker DS01F

Speaker	SVM				CNN			
	Accuracy	UAR	UAP	UAF	Accuracy	UAR	UAP	UAF
DS01F	20.00	25.00	5.00	8.33	**41.67**	**30.11**	44.99	27.14
DS02F	76.00	70.00	90.63	72.44	74.00	**71.39**	76.25	72.59
DS04F	66.00	63.75	75.26	62.37	**66.67**	**66.11**	67.67	65.17
DS03M	**66.00**	**60.00**	70.22	55.88	54.17	53.42	55.40	49.54

Table 8.2 Speaker-independent gender-dependent classification results of 4 classes of emotions using spectrograms.

highly improved the performance from being random to above random. The highest recall achieved is 71.39% for speaker DS02F using CNN.

By looking into the confusion matrices presented in Figure 8.9, it is observed that for all speakers, using either classifiers, 'anger' is never confused with 'sad', except for speaker DS01F and 'sad' is never confused with 'anger', except for speaker DS03M . This aligns with the findings from the speaker-dependent models reported in the previous chapter in Section 7.3.5. By comparing the performance of the two classifiers, SVM and CNN, an improvement is achieved for classifying 'sad' for all speakers when using CNN. For all speakers, except speaker DS01F, 'anger' is never confused with 'neutral' which is also observed in the previous experiment presented in Section 8.4.1 when using eGeMAPS feature set. Similar to the findings from the speaker-dependent models and speaker-independent models using the eGeMAPS feature set, 'anger' appears to be the easiest emotion to classify. Similar results were reported on typical speech in (Dai et al., 2008; Yacoub, 2003).

8.4.3 Experimental setup 3: MFCCs

Feature extraction and classifiers

Although MFCCs are the most used and probably the best known feature representation in speech recognition and speaker identification tasks, they have been found to be effective in speech emotion classification tasks as well and have become more or less a standard (Alghifari et al., 2018; Kathiresan and Dellwo, 2019; Kishore and Satish, 2013; Koo et al., 2020; Palo et al., 2018; Prasetya et al., 2019; Selvaraj et al., 2016; Wang et al., 2020).

Since MFCCs are static features; meaning they do not incorporate temporal dynamics of the signal which is important for emotion recognition, MFCC's first and second derivatives (deltas and delta-deltas) are usually computed to overcome this issue. After testing a number of different analysis settings, using Hanning window, a frame size of 128 ms and frame shift

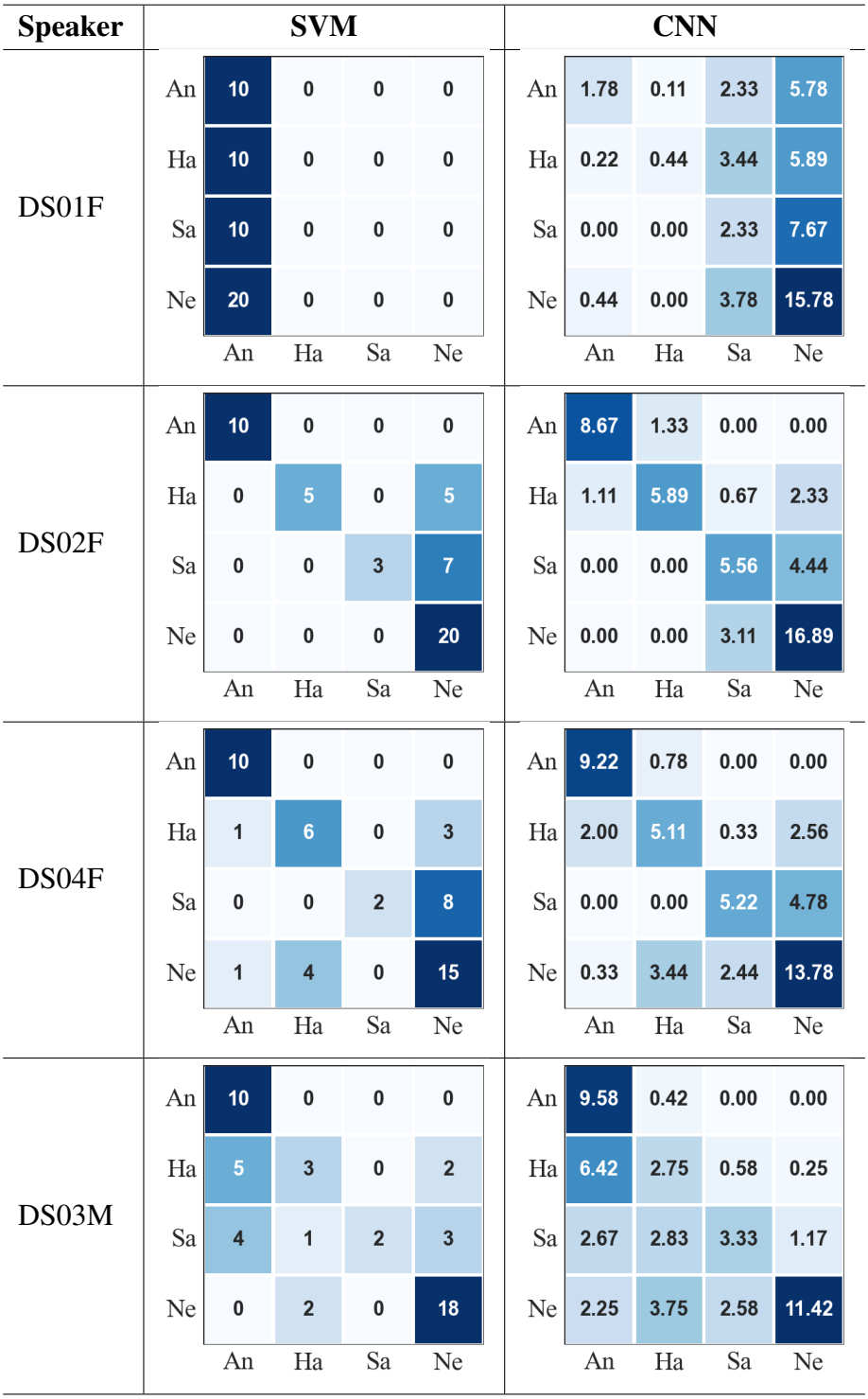


Fig. 8.9 Confusion matrices of the Speaker-independent gender-dependent classification results of 4 classes of emotions using spectrograms. (rows= actual emotions and columns= classified emotions, An= angry, Ha= happy, Sa= sad, and Ne= neutral).

of 32 ms were chosen for this experiment. From each frame, 40 MFCCs with their deltas and delta-deltas were extracted in addition to the RMS energy. The RMS energy was added given its high importance when communicating different emotional states as was learned from the acoustic analysis in Chapter 5. Features were extracted using Librosa package on python McFee et al. (2015). To have a fixed length feature representation from the variable length frame level speech features, the following five statistical functionals were applied to each feature: minimum, maximum, mean, standard deviation, and range, and their values were stacked to form the final feature vector. The features were standardised by removing the mean and scaling to unit variance using the statistics computed from the training dataset, which were then used for standardising the features in the testing dataset.

In terms of the classifiers, the performance of the same four classifiers used in experiment setup 1 were chosen: SVM with RBF kernel, LR, DT, and Adaboost. For SVM, the regularization parameter (C) and the gamma coefficient of the kernel were set to 5 and 0.01, respectively. For LR, the penalty and solver parameters were set to l2 and 'newton-cg', respectively. For Adaboost, the maximum number of estimators was set to 1200. These hyperparameter values were set based on grid search. All other hyperparameters were set to their default values. All classifiers were trained using the Python Scikit-learn package (Pedregosa et al., 2011). In addition, the performance of 1D CNN and LSTM were tested.

For the CNN, Figure 8.10 shows the standard configured model architecture. The model consisted of two 1D convolutional layers with 64 parallel feature maps with kernal size set to 3 and ReLU as the activation function, followed by a dropout layer with a probability of retention p = 0.5 and a pooling layer of size 2 . The learned features are then flattened to one vector and passed to a fully connected layer with 100 nodes and ReLU activation function. The last layer, softmax layer, was the output layer which had 4 nodes, one node for each possible class (outcome). Softmax was chosen as the activation function which produces a vector that presents the probability of potential outcomes where the sum of these probabilities equals to one. The final prediction is based on the outcome with the highest probability. The model was implemented using Keras supported by Tensorflow backend (Chollet et al., 2015). For compilation, 'RMSprop' was chosen as the optimisation algorithm with a learning rate of 0.00001 and 'categorical_crossentropy' was chosen as the loss function. All other parameters were set to their default values.

Since traditional machine learning algorithms and some deep learning networks such as CNN can only accept fixed length representation of the input data, statistical functions were applied to the frame level features. On the other hand, LSTM models can accept variable length representation of the input data. Therefore, for this network, the frame-level features were fed directly as input without applying statistical functions preserving the temporal

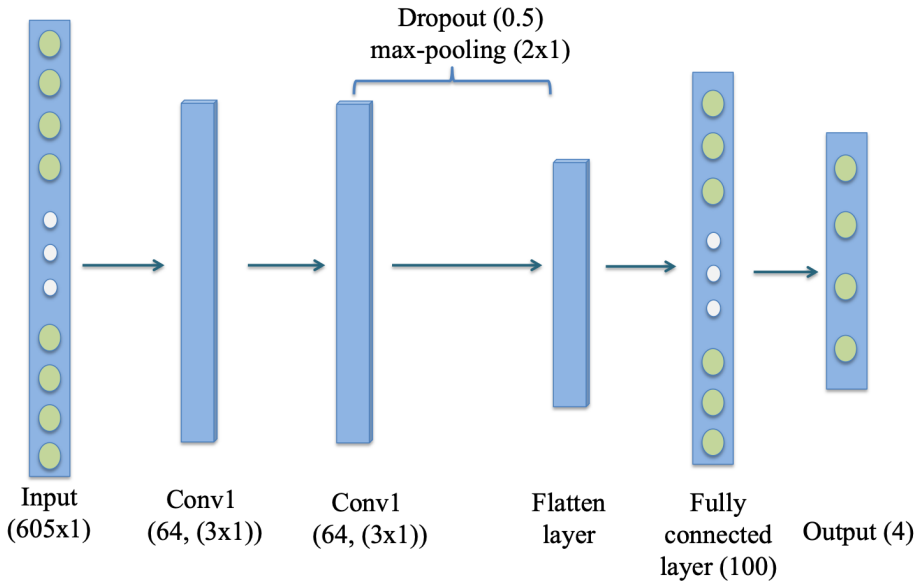


Fig. 8.10 1D CNN model architecture. (Conv = 1D-convolutional layer).

information in the original speech. The dimensionality of the features changes depending on the speech length. Figure 8.11 presents the used LSTM model structure. As can be seen, the model consisted of two LSTM layers with 128 LSTM units in the hidden layers. The input had the dimension of [32, *timestep*, 121], where 32 is the batch size, *timestep* is the number of frames, and 121 is the number of extracted features. The return_sequences argument was set to "True" to ensure that the full sequence of outputs is returned at every time step. The next layer is a dropout layer with a probability of retention p = 0.5. After that, two fully connected layers were added were the first one had 100 nodes and ReLU activation function and the second one, the output layer, had 4 nodes, one node for each possible outcome, and softmax activation function. The model was implemented using Keras supported by Tensorflow backend (Chollet et al., 2015). For compilation, 'Adam' was chosen as the optimisation algorithm with a learning rate of 0.001 and 'categorical_crossentropy' was chosen as the loss function. All other parameters were set to their default values.

Performance evaluation and results

For each classifier, the same four performance metrics were calculated which are accuracy, UAR, UAP, and UAF. CNN and LSTM models were trained using 11 epochs and 13 epochs for female and male speakers models, respectively with a batch size of 32. For model validation, Leave-One-Speaker-Out approach was used and the average performance of all runs is reported. Table 8.3 presents the classification results of all the speaker-independent

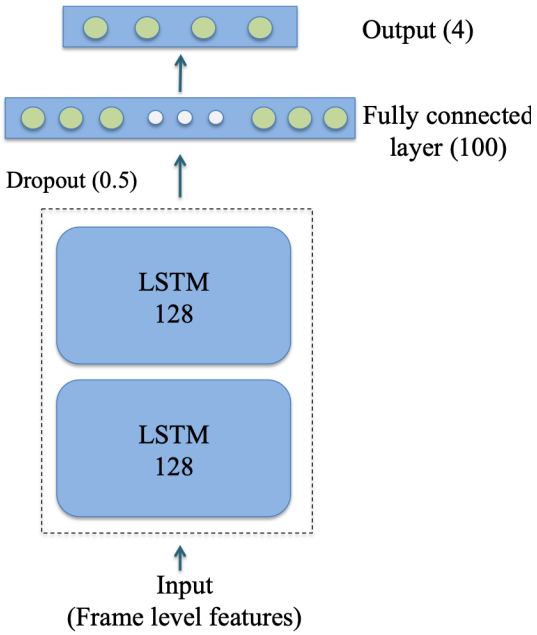


Fig. 8.11 LSTM model architecture.

models that were trained on gender-dependent typical speakers data using all classifiers. Figure 8.12 presents the confusion matrices of the best model in terms of the recall for each speaker. For CNN and LSTM models, the average confusion matrix of all runs was computed and presented.

As can be seen from Table 8.3, the performance of most classifiers per speaker are above chance performance. The highest recall achieved is 83.75% for speaker DS02F using Adaboost classifier. It is observed that in this experiment, deep learning models are outperformed by one of the traditional classifiers, except for speaker DS03M where the best performance is achieved when using LSTM. This could be explained by the size of the training data, as traditional classifiers usually work better than deep learning models in the case of scarcity of training data while deep learning models are known to work better in the presence of sufficient training data. By comparing the two deep learning models, 1D CNN and LSTM, it is observed that generally the LSTM performance is better. This could be explained by the LSTM ability to learn long-term dependency information which is critical for emotion recognition.

By looking into the confusion matrices presented in Figure 8.12, it is observed that for all speakers, 'anger' is never confused with 'sad', except for speaker DS01F. 'Happy' is highly confused with 'anger'. It is also observed that 'anger' is never confused with 'neutral'. Similar to the findings from the speaker-dependent models presented in Section 7.3.5, 'anger'

Speaker	SVM				LR			
	Accuracy	**UAR**	**UAP**	**UAF**	**Accuracy**	**UAR**	**UAP**	**UAF**
DS01F	26.00	**32.50**	20.98	24.73	24.00	30.00	20.48	22.08
DS02F	62.00	67.50	72.38	63.74	60.00	70.00	77.47	61.81
DS04F	46.00	50.00	47.83	44.76	46.00	**53.75**	47.88	46.24
DS03M	24.00	28.75	39.04	20.03	24.00	30.00	16.69	20.53
Speaker	**DT**				**Adaboost**			
	Accuracy	**UAR**	**UAP**	**UAF**	**Accuracy**	**UAR**	**UAP**	**UAF**
DS01F	30.00	22.50	16.11	18.77	28.00	31.25	27.80	21.88
DS02F	40.00	42.50	36.99	38.29	82.00	**83.75**	83.84	83.12
DS04F	32.00	36.25	35.91	29.55	38.00	42.50	63.92	37.71
DS03M	22.00	22.50	31.68	17.53	32.00	38.75	43.74	27.00
Speaker	**1D CNN**				**LSTM**			
	Accuracy	**UAR**	**UAP**	**UAF**	**Accuracy**	**UAR**	**UAP**	**UAF**
DS01F	19.78	23.75	12.11	11.88	39.11	29.72	25.06	24.41
DS02F	53.56	50.42	40.16	42.77	67.56	62.36	62.84	58.35
DS04F	47.11	45.14	34.02	37.78	45.33	43.75	40.43	37.76
DS03M	35.17	34.69	32.25	24.60	50.50	**49.00**	45.92	42.74

Table 8.3 Speaker-independent gender-dependent classification results of 4 classes of emotions using MFCCs features.

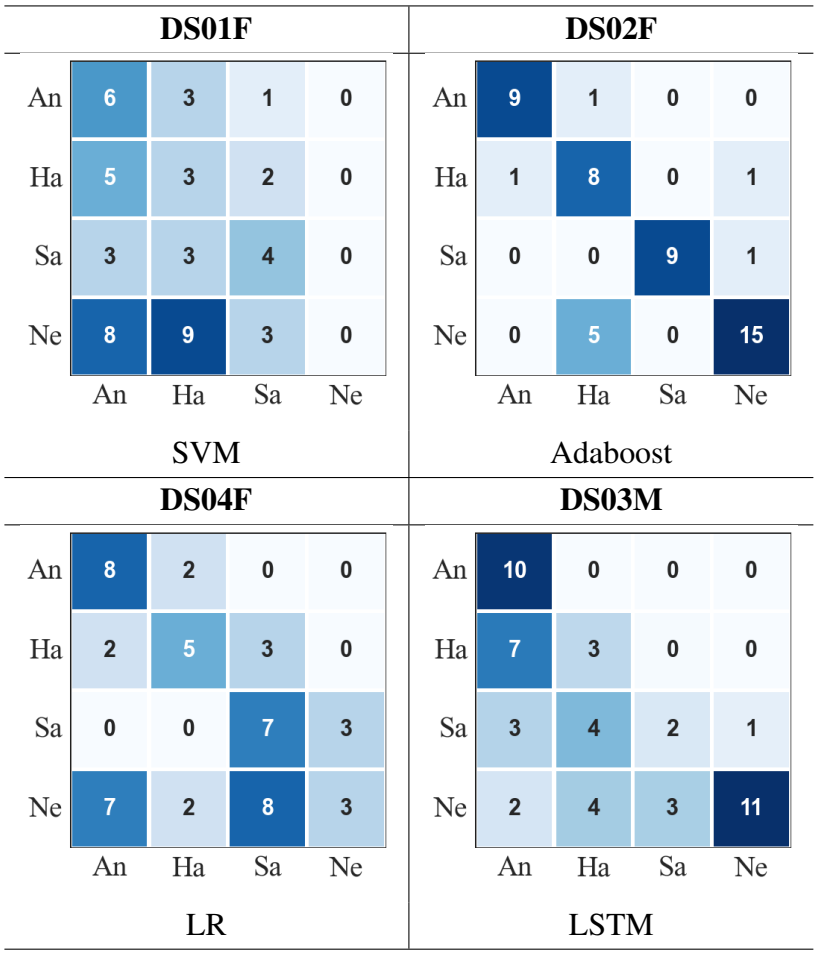


Fig. 8.12 Confusion matrices of the speaker-independent gender-dependent classification using MFCC feature set. (rows= actual emotions and columns= classified emotions, An= angry, Ha= happy, Sa= sad, Ne= neutral).

appears to be the easiest emotion to classify. Similar results were observed on typical speech in (Dai et al., 2008; Yacoub, 2003).

8.5 Discussion and Conclusion

Given the nature of the dysarthric speech and its phonological and prosody differences with typical speech, the experiments in this work were conducted to investigate i) the feasibility of automatically recognising emotions from dysarthric speech using models trained on typical speech, ii) whether there are similarities between emotional typical speech and emotional dysarthric speech, and iii) what emotions in the dysarthric speech are found to be close to each other (confusing) in the chosen feature space. It was demonstrated that using models

trained on emotional typical speech, dysarthric speech emotion recognition could be possible. In fact, the results of recognising four emotional states were above chance performance for all speakers including, DS01F, who has a severe level of dysarthria and low speech intelligibility. The highest recognition accuracy of 88% was achieved when training the model on typical speech data for speaker DS02F using eGeMAPS feature set and Adaboost classifier. For speakers Ds01F and DS02F the highest classification performance was achieved when using eGeMAPS feature set while for speakers DS04F and DS03M, the highest classification performance was achieved when using Spectrograms as feature set.

From recognising pairs of emotions, it was generally found that 'anger/sad' and 'anger/ neutral' are the easiest pairs to recognise. This could be justified by the distant positions of these pairs of emotions in the arousal-valence space. High accuracy results were achieved for most of the speakers, with an accuracy of 100% achieved for some of them.

It is observed that the performance of the different classifiers used vary among speakers and features. The performance of the deep learning techniques applied in the last experiment were mostly outperformed by the performance of one of the traditional classifiers. This is expected due to the size of the training data as traditional classifiers usually require lesser data to work well. As deep learning models are known to work better in the presence of sufficient training data, it would be interesting to apply some adaptation techniques to increase the size of the training data and compare the performance of these models.

The performance of the models trained on typical speech vary among speakers with dysarthria. The difference is still observed even within the group of speakers with PD. Speakers DS02F and DS04F seem to share strong similarities with typical speech in their way of expressing emotions although this may not mean that these similarities are exactly the same for those two speakers. The fact that it is possible to classify emotions from dysarthric speech using models of typical speech is encouraging from a technological perspective, because it means there is less need to collect speaker specific data in order to be able to recognise a majority of people with dysarthria. Collecting sufficient amounts of emotional typical speech to train such a speaker-independent model, is generally easier than collecting speaker-dependent data for a particular target dysarthric speaker. In addition, working with these larger typical data sets enables the use of more sophisticated deep learning techniques that may improve the model's performance.

Using the same features as in experiment setup 1, the performance of two classifiers, SVM and LR have been investigated when using two other different training sets. The first training set includes the data of all the typical speakers, male and female in DEED. The second training set includes mix data of typical speakers and speakers with dysarthria. It has been found that the performance among speakers vary. For example, for speakers DS01F and

DS04F, training the model on mixed data (typical and dysarthric) degraded the classification performance, while for speaker DS03M training the model with gender-independent mixed data improved the classification performance in comparison to training the model with only gender-dependent typical speech. The detailed performance of these models can be found in Appendix E. Using mixed data of typical and dysarthric emotional speech in the training process might have a potential in improving the classification results. However, more investigations needs to be carried out using different sets of the training data and investigating other features and classification techniques to adequately assess its effect.

Collecting dysarthric emotional data is a very challenging task. However, although a limited number of speakers with dysarthria are included in this study, promising results have been found. The recognition results may be improved by investigating the performance of other feature sets and other classifiers, training the model with only typical speakers close in age to the speaker with dysarthria, and apply augmentation techniques to increase the size of the data.

Chapter 9

Conclusion

Emotions play a critical and important role as a relation regulatory factor in the continuous complexity of social life. A person's emotional responses sway other people's reactions and future encounters (Parkinson, 1996).

Dysarthria, one of the most common speech disorders, may not only result in producing less intelligible speech, but it may make it hard to convey emotions in the speech in a way that can be understood clearly and easily by others. This may in time increase the potential of them being socially withdrawn (Hartelius et al., 2008; Walshe and Miller, 2011).

Recently, there has been a concerted effort on developing and improving dysarthric automatic speech recognition (ASR) which concentrates on recognising the *verbal* part of the dysarthric speech. A number of studies have also investigated the prosodic control ability of speakers with dysarthria in different tasks such as signalling questions-statement contrast. A little has been done in *perceptually* assessing the ability of people with dysarthria to convey emotions through their speech. However, *objectively* assessing their ability to convey emotions in their speech through suprasegmental and prosodic features remains unexplored. This thesis has investigated this for people with dysarthria caused by cerebral palsy and Parkinson's disease (PD) motivated by a long term goal to improve the potential for voice-input communication aids. The overarching aim is to make these communication aids more sensitive to specific cues in the vocalization signal produced by the speaker with dysarthria and add expressiveness to the produced synthesised voice. Section 9.1 summarises the research reported in this thesis along with its main findings. In alignment of the findings of this research, directions for future work is suggested in Section 9.2. Finally, a SWOT (strengths, weaknesses, opportunities, and threats) analysis is conducted in Section 9.3 that identifies the major strengths and weaknesses of this research, in addition to the opportunities and threats.

9.1 Summary of thesis

This thesis has investigated how people with dysarthria convey emotions in their speech and the ability to automatically recognise these emotions. These investigations have been underpinned by a novel database collected for this purpose. Since this is a relatively new research area, a good understanding of a number of related points needed to be established beforehand.

As little is known about the ability of people with dysarthria caused by cerebral palsy to communicate emotions, a preliminary study in the form of a survey was designed and distributed to explore the difficulty, importance, and methodology used to communicate emotions by speakers with dysarthria from the speakers' point of view. The survey helped in achieving a better understanding of the problem and defining the scope of the research (see Chapter 3). The survey showed that people with dysarthria find more difficulties when communicating emotions with unfamiliar people in comparison to familiar people. Thus, having a VIVOCA that could assist with the communication of emotions and other nonverbal information could be beneficial.

Establishing a proper database of emotional dysarthric speech has been an essential part for this research for several reasons. Mainly, research on automatic typical speech emotion recognition is usually done using emotional databases. Despite the existence of a few databases of dysarthric speech, the fact that these are not emotional databases make them not suitable to be used in analysing emotions in dysarthric speech, nor developing automatic emotion recognition models. As it has been an interest for this research to see whether the acoustic signalling of emotions of speakers with dysarthria differ to that used by typical speakers, it was essential to record both types of speech, typical and dysarthric. Chapter 4 presented all the details related to the collected parallel multimodal emotional database of dysarthric speech and typical speech (DEED). The database will be made publicly available for research purposes in the near future.

The ability of people with dysarthria to make some systematic changes in their speech when communicating emotions was first explored by performing an acoustic analysis of some potential features and analyse them using statistical models (see Chapter 5). The results indicated that some people with dysarthria, even with severe dysarthria, are able to control some aspects of the suprasegmental and prosodic features of their speech when communicating emotions. Despite speakers with dysarthria having more limited articulatory and prosodic control, it was observed that the changes to the analysed features made by speakers with dysarthria appeared to be similar to those of typical speakers when communicating different emotions.

When databases of typical emotional speech are collected, the perceptual recognition on the collected database is usually reported. The ability of listeners to perceptual recognise the expressed emotion provides confidence that the speech samples accurately convey the emotion. Therefore, a subjective evaluation on the dysarthric speech part of DEED and a subset of the typical speech part was performed (see Chapter 6). Knowing the human performance on determining emotions in dysarthric speech not only helped in determining the task difficulty level for humans but also in setting a benchmark for automatic emotion recognition models. The human performance on the dysarthric speech part of DEED was promising, even for speaker DS01F, who has severe dysarthria and highly unintelligible speech. This indicated that the speakers with dysarthria in this study have the ability to express some emotions through their speech in a way others can recognise.

Next, Chapter 7 presented the first attempt to examine the feasibility of automatically recognising emotions in dysarthric speech using the collected database, DEED. Two classification approaches were examined: discrete and dimensional. Using speaker-dependent models, the performance of different classifiers using the same feature set, eGeMAPS, when classifying the full set of emotions (7-classes) and the reduced set of emotions (4-classes) were reported and discussed. Emotion classification by itself is a very challenging task even when staying within the typical speech domain. This was furtherly confirmed by the performance of the classifiers on the typical speech part of DEED. The performance of the classifiers, however, on the dysarthric speech part of DEED were very encouraging for all speakers, with better results obtained when classifying the reduced set of emotions. In fact, A high accuracy of 67.50% using SVM and 90% using LR is achieved on one of the speakers for the categorical classification approach using 7 and 4 emotions, respectively. The baseline results reported in this chapter were twofold. Firstly, allowing future comparisons to improved classification techniques. Secondly, giving an insight into the level of difficulty of the classification problem.

Given the successful application of using typical speech data to boost the performance on dysarhtric speech data in the domain of dysarthric automatic speech recognition (ASR), Chapter 8 examined the applicability of applying this to the domain of dysarthric speech emotion recognition (SER). Several speaker-independent dysarthric SER models trained on typical speech data and on mixed data (typical and dysarthric) to classify emotions in dysarthic emotional speech data were developed. The results of classifying two emotional states and four emotional states were reported. For the classification of four emotional states, a number of different traditional classifiers and deep learning models where trained and their results were discussed. The relatively good results achieved for recognising four emotional states were very promising for all speakers. A recognition accuracy of 88% was achieved

when training the model on typical speech data for one of the speakers with dysarthria. This shows the potential of recognising the emotions conveyed in dysarthric speech using models trained on emotional typical speech. Thus, the existing databases of emotional speech can be used, probably for the majority of people with dysarthria. This indicates the potential of this being implemented without the need for large new databases.

9.2 Future work

The above section summarises the work conducted in this thesis along with its main findings. As this research explored a totally new area, it could be seen as the starting point to a more deep focused investigation of the research problem. Several future directions for this work are suggested below.

- Increasing the size of the recorded database, DEED, by recruiting and recording more speakers with dysarthria and more aged-matched typical speakers. Having more data will allow the uses of state of the art SER methods which will enhance the emotion classification performance.
- Comparing the effect of familiarity with the dysarthric speech and speakers on the classification of emotions, which was planned to be done within this work. However, due to the situation of the COVID-19 and the lockdown imposed this was not possible. This would imply augmenting the evaluation done in Chapter 6 by recruiting more participants from three different groups: participants who are familiar with the dysarthric speech and familiar with a speaker/speakers with dysarthria in DEED, participants who are familiar with the dysarthric speech but are not familiar with any of the speakers with dysarthria in DEED, and participants who are not familiar with the dysarthric speech at all.
- Applying different data augmentation techniques on DEED and exploring the classification performance of different deep learning models.
- Developing dysarthric ASR on DEED. As there are large number of sentences, which are not usually captured in the existing databases of dysarthric speech.
- Developing a VIVOCA system similar to the one presented in Figure 1.1 in Chapter 1. This will require adding a number of subsystems together including dysarthric ASR, dysarhric SER, and emotional speech synthesis. As a VIVOCA product now exists, which is licensed by UoS (Therapy Box), adding the emotion aspect and incorporating emotional synthesis now has a high potential.

- Exploring the visual part of DEED. Investigating the ability of people with dysarthria perceptually and objectively to communicate emotions through facial expressions and gesture. If they were able, then investigate the effect of combining audio and video emotional cues to the classification performance.
- Investigating approaches to enhance the performance of communication aids by looking into ways to improve the recognition of the context of the communication situation. This includes recognising who is present, what is the topic being discussed, and the place where the communication taking place in.
- Publishing the database and collaborating with other people in the field. Multiple options are currently considered for releasing the database, including the possibility of releasing it with an emotion recognition "challenge" at INTERSPEECH.

9.3 SWOT analysis

SWOT analysis helps in emphasising the strengths and opportunities and reducing potential weaknesses and threats. Below is a point by point list of the strengths, weaknesses, opportunities, and threats.

9.3.1 Strengths

- **S1, Novel database:** the establishment of a parallel multimodal emotional database of dysarthric speech and typical speech, which is a first of its kind, is considered one of the main strengths of this research. As it didn't only allow the investigations of this research to be made but it will allow and encourage other people in the field to collaborate and investigate more related aspects once the database is made publicly available.
- **S2, Ability of people with dysarthria to communicate emotions in their speech:** the results from the different investigations made in this research indicated the ability of some people with dysarthria caused by cerebral palsy and PD to communicate emotions in their speech. The consistent and reliable changes made by speakers with dysarthria when conveying different emotions were found to be enough for people to accurately perceive these emotions.
- **S3, Ability of automatically classifying emotions in dysarthric speech:** a significant strength of this research is that it showed that in addition to the ability of people

to perceive emotion in dysarthric speech, these emotions can be picked up using automatic processing.

- **S4, Feasibility of automatically classifying emotions in dysarthric speech using models trained on typical speech:** the fact that it is possible to classify emotions from dysarthric speech using models of typical speech is another main strength of this research. The encouraging results of these models indicated the existence of some similarities between emotional typical speech and emotional dysarthric speech. From a technological perspective this means there is less need to collect speaker specific data in order to be able to recognise a majority of people with dysarthria's emotions.

9.3.2 Weaknesses

- **W1, Limited number of speakers:** the generalisability of the findings of this research is difficult to be made due to the limited number of speakers with dysarthria included in this study. However, it showed that people with dysarthria may have enough control to communicate intentions, gain attention, and convey emotions.

- **W2, Age-range mismatch between speakers with dysarthria and typical speakers:** Although there are few typical speakers who falls in the same age-range of speakers with dysarthria in DEED, the age range of typical speakers is a lot wider. This is considered one of the weaknesses of this research as normal aging can affect some acoustic characteristics of speakers, which can affect the accuracy of direct comparisons between speakers with dysarthria and typical speakers and the performance of the automatic classifiers that are trained on DEED-typical speech part. However, a direct comparison of some acoustic features was made to compare speakers with dysarthria to only closely age-matched typical speakers to asses the effect of age range in DEED.

9.3.3 Opportunities

As this is a first of its kind research, a number of opportunities arise. The reader is referred to the previous section, Section9.2 for a list of some of the main opportunities and future directions of this research.

9.3.4 Threats

- **T1, Difficulty to increase the size of the database:** recording more speakers with dysarthria is essential to generalise the results and allow the use of state of the art SER methods. However, recruiting more people with dysarthria to record their emotional speech is very challenging and it is not fast and easy process. Thus, when publishing the data, all the recording technical details will be published as well to allow other people in the field to collaborate in the recording process.
- **T2, Data sharing commercial violations:** the data will be published under a non-commercial provision license, freely available for research purposes. Having the data online could lead to a non authorised commercial use of the data which violates the data sharing license.

In conclusion, the work presented in this thesis showed that people with different etiology and severity of dysarthria can communicate different emotions through their speech. A significant and important finding is that these emotions can be picked up using automatic processing. This could open the door to this being implemented in a VIVOCA system. The recognition accuracy achieved on some speakers' data were very high and close to what is achieved on typical speech. The relatively good results when testing the dysarthric speakers on models trained on typical speech were very encouraging. This indicates that a good level of typical-like emotion specific information is being successfully expressed. Also, the results obtained from the different experiments presented in this thesis demonstrate this database will be a useful resources in the field of dysarthric emotion classification research field.

References

Robert P Abelson and Vello Sermat. Multidimensional scaling of facial expressions. *Journal of Experimental Psychology*, 63(6):546, 1962.

Purvi Agrawal and Sriram Ganapathy. Speech representation learning using unsupervised data-driven modulation filtering for robust ASR. In *INTERSPEECH*, pages 2446–2450, 2017.

Mehmet Berkehan Akçay and Kaya Oğuz. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116, 2020.

Muhammad Fahreza Alghifari, Teddy Surya Gunawan, and Mira Kartiwi. Speech emotion recognition using deep feedforward neural network. *Indonesian Journal of Electrical Engineering and Computer Science*, 10(2):554–561, 2018.

Lubna Alhinti, Heidi Christensen, and Stuart Cunningham. An exploratory survey questionnaire to understand what emotions are important and difficult to communicate for people with dysarthria and their methodology of communicating. *International Journal of Psychological and Behavioral Sciences*, 14(7):187–191, 2020a.

Lubna Alhinti, Stuart Cunningham, and Heidi Christensen. Recognising emotions in dysarthric speech using typical speech data. *INTERSPEECH*, pages 4821–4825, 2020b.

Lubna Alhinti, Heidi Christensen, and Stuart Cunningham. Acoustic differences in emotional speech of people with dysarthria. *Speech Communication*, 126:44–60, 2021.

Jont B Allen and Lawrence R Rabiner. A unified approach to short-time fourier analysis and synthesis. *Proceedings of the IEEE*, 65(11):1558–1564, 1977.

Kai Alter, Erhard Rank, Sonja A Kotz, Erdmut Pfeifer, Mireille Besson, Angela D Friederici, and Johannes Matiasek. On the relations of semantic and acoustic properties of emotions. In *14th International Congress of Phonetic Sciences*, pages 2121–2125. University of California, 1999.

Noam Amir, Samuel Ron, and Nathaniel Laor. Analysis of an emotional speech corpus in hebrew based on objective criteria. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

Beth M Ansel and Raymond D Kent. Acoustic-phonetic contrasts and intelligibility in the dysarthria associated with mixed cerebral palsy. *Journal of Speech, Language, and Hearing Research*, 35(2):296–308, 1992.

Bagus Tris Atmaja. Rnn-based dimensional speech emotion recognition. *ASJ Autum Meeting. Acoustical Society of Japan*, pages 743–744, 2019.

James R Averill. Are there basis emotions? in the eye of the beholder. In Paul Ekman and Richard J Davidson, editors, *The Nature of Emotion Fundamental Questions*, pages 7–14. Oxford University Press, Inc., 1994.

Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 International Conference on platform Technology and Service (PlatCon)*, pages 1–5. IEEE, 2017.

Kiavash Bahreini, Rob Nadolski, and Wim Westera. Towards real-time speech emotion recognition for affective e-learning. *Education and Information Technologies*, 21(5): 1367–1386, 2016.

Rainer Banse and Klaus R Scherer. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3):614, 1996.

David R Beukelman, Pat Mirenda, et al. *Augmentative and alternative communication*. Paul H. Brookes Baltimore, MD, 3 edition, 2005.

David R Beukelman, Susan Fager, Laura Ball, and Aimee Dietz. AAC for adults with acquired neurological conditions: A review. *Augmentative and Alternative Communication*, 23(3): 230–242, 2007.

Monique Biemans. *Gender variation in voice quality*. Netherlands Graduate School of Linguistics, 2000.

Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

Dmitri Bitouk, Ragini Verma, and Ani Nenkova. Class-level spectral features for emotion recognition. *Speech Communication*, 52(7-8):613–625, 2010.

Paul Boersma and David Weenink. Praat: doing phonetics by computer [computer program]. version 6.1. 07. *URL http://www. praat. org/. Retrieved November*, 26, 2019.

Michael Boiger and Batja Mesquita. The construction of emotion in interactions, relationships, and cultures. *Emotion Review*, 4(3):221–229, 2012.

Milana Bojanic, Milan Gnjatovic, Milan Secujski, and Vlado Delic. Application of dimensional emotion model in automatic emotional speech recognition. In *2013 IEEE 11th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 353–356, 2013.

Stephanie A Borrie, Megan J McAuliffe, Julie M Liss, Cecilia Kirk, Gregory A O'Beirne, and Tim Anderson. Familiarisation conditions and the mechanisms that underlie improved recognition of dysarthric speech. *Language and Cognitive Processes*, 27(7-8):1039–1055, 2012.

Stephanie A Borrie, Kaitlin L Lansford, and Tyson S Barrett. Generalized adaptation to dysarthric speech. *Journal of Speech, Language, and Hearing Research*, 60(11):3110–3117, 2017.

Michael P Boyle. Personal perceptions and perceived public opinion about stuttering in the united states: Implications for anti-stigma campaigns. *American Journal of Speech-Language Pathology*, 26(3):921–938, 2017.

Caterina Breitenstein, Diana Van Lancker, and Irene Daum. The contribution of speech rate and pitch variation to the perception of vocal emotions in a german and an american sample. *Cognition & Emotion*, 15(1):57–79, 2001.

Nancy J Briton and Judith A Hall. Beliefs about female and male nonverbal communication. *Sex Roles*, 32(1-2):79–90, 1995.

Leslie R Brody and Judith A Hall. Gender and emotion in context. In Michael Lewis, Jeannette M Haviland-Jones, and Lisa Feldman Barrett, editors, *Handbook of Emotions*, pages 395–408. Guilford press, 3 edition, 2008.

Kate Bunton and Gary Weismer. The relationship between perception and acoustics for a high-low vowel contrast produced by speakers with dysarthria. *Journal of Speech, Language, and Hearing Research*, 2001.

Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*, 2005.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4): 335, 2008.

Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):582–596, 2009.

Henry H Calero. *The power of nonverbal communication: How you act is more important than what you say*. Silver Lake Publishing, 2005.

Gerald J Canter. Speech characteristics of patients with parkinson's disease: I. intensity, pitch, and duration. *Journal of Speech and Hearing Disorders*, 28(3):221–229, 1963.

Fuling Chen, Roberto Togneri, Murray Maybery, and Diana Tan. An objective voice gender scoring system and identification of the salient acoustic measures. *INTERSPEECH*, pages 1848–1852, 2020a.

Luefeng Chen, Wanjuan Su, Yu Feng, Min Wu, Jinhua She, and Kaoru Hirota. Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. *Information Sciences*, 509:150–163, 2020b.

François Chollet et al. Keras. github; 2015, 2015. URL https://github.com/fchollet/keras.

Heidi Christensen, Stuart Cunningham, Charles Fox, Phil Green, and Thomas Hain. A comparative study of adaptive, automatic recognition of disordered speech. In *INTERSPEECH*, 2012.

Heidi Christensen, MB Aniol, Peter Bell, Phil D Green, Thomas Hain, Simon King, and Pawel Swietojanski. Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech. In *INTERSPEECH*, pages 3642–3645, 2013.

Heidi Christensen, I Casanueva, S Cunningham, P Green, and Thomas Hain. Automatic selection of speakers for improved acoustic modelling: Recognition of disordered speech with sparse data. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 254–259. IEEE, 2014.

Soo-Jin Chung. Expression and perception of emotion extracted from the spontaneous speech in korean and in english. *Dissertation from ILPGA, Sorbonne Nouvelle University*, 2000.

Gerald L Clore. Why emotions are felt? In Paul Ekman and Richard J Davidson, editors, *The Nature of Emotion Fundamental Questions*, pages 103–111. Oxford University Press,Inc., 1994a.

Gerald L Clore. Why emotions are never unconscious. In Paul Ekman and Richard J Davidson, editors, *The Nature of Emotion Fundamental Questions*, pages 285–290. Oxford University Press,Inc., 1994b.

Kathryn P Connaghan and Rupal Patel. The impact of contrastive stress on vowel acoustics and intelligibility in dysarthria. *Journal of Speech, Language, and Hearing Research*, 60 (1):38–50, 2017.

Kathryn P Connaghan, Chelsea Wertheim, Jacqueline S Laures-Gore, Scott Russell, and Rupal Patel. An exploratory study of student, speech–language pathologist and emergency worker impressions of speakers with dysarthria. *International Journal of Speech-Language Pathology*, pages 1–10, 2020.

Leda Cosmides. Invariances in the acoustic expression of emotion during speech. *Journal of Experimental Psychology: Human Perception and Performance*, 9(6):864–881, 1983.

Roddy Cowie and Ellen Douglas-Cowie. Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In *Fourth International Conference on Spoken Language Processing*, 1996.

Nicholas Cummins, Bogdan Vlasenko, Hesam Sagha, and Björn Schuller. Enhancing speech-based depression detection through gender dependent vowel-level formant features. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 209–214. Springer, 2017.

Dongyang Dai, Zhiyong Wu, Runnan Li, Xixin Wu, Jia Jia, and Helen Meng. Learning discriminative features from spectrograms using center loss for speech emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7405–7409. IEEE, 2019.

Keshi Dai, Harriet J Fell, and Joel MacAuslan. Recognizing emotion in speech using neural networks. *Telehealth and Assistive Technologies*, 31:38, 2008.

Keshi Dai, Harriet Fell, and Joel MacAuslan. Comparing emotions using acoustics and human perceptual dimensions. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*, pages 3341–3346. ACM, 2009.

Frederic L Darley, Arnold E Aronson, and Joe R Brown. Clusters of deviant speech dimensions in the dysarthrias. *Journal of Speech and Hearing Research*, 12(3):462–496, 1969a.

Frederic L Darley, Arnold E Aronson, and Joe R Brown. Differential diagnostic patterns of dysarthria. *Journal of Speech, Language, and Hearing Research*, 12(2):246–269, 1969b.

Frederic L Darley, Arnold Elvin Aronson, and Joe Robert Brown. *Motor speech disorders*. Saunders, 1975.

Charles Darwin. *The expression of the emotions in man and animals*. John Murray, 1872.

Poorna Banerjee Dasgupta. Detection and analysis of human emotions through voice and speech pattern processing. *International Journal of Computer Trends and Technology*, 2017.

Assel Davletcharova, Sherin Sugathan, Bibia Abraham, and Alex Pappachen James. Detection and analysis of emotion from speech signals. *Procedia Computer Science*, pages 91–96, 2015.

Marc S De Bodt, Maria E Hernández-Diaz Huici, and Paul H Van De Heyning. Intelligibility as a linear combination of dimensions in dysarthric speech. *Journal of Communication Disorders*, 35(3):283–292, 2002.

Nivja H De Jong and Ton Wempe. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2):385–390, 2009.

Miet De Letter, Patrick Santens, Marc De Bodt, Paul Boon, and John Van Borsel. Levodopa-induced alterations in speech rate in advanced parkinson's disease. *Acta Neurologica Belgica*, 106(1):19, 2006.

Suruchi G Dedgaonkar, Anjali A Chandavale, and Ashok M Sapkal. Survey of methods for character recognition. *International Journal of Engineering and Innovative Technology (IJEIT)*, 1(5):180–189, 2012.

Jun Deng, Zixing Zhang, Florian Eyben, and Björn Schuller. Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, 21(9):1068–1072, 2014.

J Dennis, HD Tran, and Haizhou Li. Spectrogram image feature for sound event classification in mismatched conditions. *IEEE Signal Processing Letters*, 2(18):130–133, 2011.

Sylvia Dickson, Rosaline S Barbour, Marian Brady, Alexander M Clark, and Gillian Paton. Patients' experiences of disruptions associated with post-stroke dysarthria. *International Journal of Language & Communication Disorders*, 43(2):135–153, 2008.

Vipula Dissanayake, Haimo Zhang, Mark Billinghurst, and Suranga Nanayakkara. Speech emotion recognition 'in the wild' using an autoencoder. *INTERSPEECH*, pages 526–530, 2020.

Ellen Douglas-Cowie, Roddy Cowie, and Marc Schröder. A new emotion database: considerations, sources and scope. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

Ellen Douglas-Cowie, Nick Campbell, Roddy Cowie, and Peter Roach. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1-2):33–60, 2003.

Joseph R Duffy. *Motor speech disorders-E-Book: substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2013.

Paul Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska Symposium on Motivation*. University of Nebraska Press, 1971.

Paul Ekman. *The face of man: Expressions of universal emotions in a New Guinea village*. Garland STPM Press New York, 1980.

Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992a.

Paul Ekman. Facial expressions of emotion: New findings, new questions, 1992b.

Paul Ekman. Are there basic emotions? *Psychological Review*, 99:550–553, 1992c.

Paul Ekman. *Emotions revealed: recognizing faces and feelings to improve communication and emotional life.* Times Books/Henry Holt and Co, 2003.

Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124, 1971.

Paul Ekman, E Richard Sorenson, and Wallace V Friesen. Pan-cultural elements in facial displays of emotion. *Science*, 164(3875):86–88, 1969.

Paul Ekman, Wallace V Friesen, Maureen O'sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 53(4):712, 1987.

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

Rana El Kaliouby and Peter Robinson. The emotional hearing aid: an assistive tool for children with asperger syndrome. *Universal Access in the Information Society*, 4(2): 121–134, 2005.

Inger S Engberg, Anya Varnich Hansen, Ove Andersen, and Paul Dalsgaard. Design, recording and verification of a danish emotional speech database. In *Fifth European Conference on Speech Communication and Technology*, 1997.

Mehmet Erdal, Markus Kächele, and Friedhelm Schwenker. Emotion recognition in speech with deep learning architectures. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 298–311. Springer, 2016.

Cristina Espana-Bonet and José AR Fonollosa. Automatic speech recognition with deep neural networks for impaired speech. In *International Conference on Advances in Speech and Language Technologies for Iberian Languages*, pages 97–107. Springer, 2016.

Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 835–838. ACM, 2013.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016.

Grant Fairbanks and Wilbert Pronovost. An experimental study of the pitch characteristics of the voice during the expression of emotion. *Communications Monographs*, 6(1):87–104, 1939.

Mathieu Fauvel, Jocelyn Chanussot, and Jon Atli Benediktsson. Evaluation of kernels for multiclass classification of hyperspectral remote sensing data. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages II–II. IEEE, 2006.

Mathieu Fauvel, Jón Atli Benediktsson, Jocelyn Chanussot, and Johannes R Sveinsson. Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles. *IEEE Transactions on Geoscience and Remote Sensing*, 46(11):3804–3814, 2008.

Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92:60–68, 2017.

Raul Fernandez and Rosalind W Picard. Modeling drivers' speech under stress. *Speech Communication*, 40(1-2):145–159, 2003.

Davis Foulger. Models of the communication process. *Evolutionary Media*, 2004.

Andrew S Fox, Regina C Lapate, Alexander J Shackman, and Richard J Davidson. *The nature of emotion: fundamental questions*. Oxford University Press, 2018.

Daniel Joseph France, Richard G Shiavi, Stephen Silverman, Marilyn Silverman, and M Wilkes. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering*, 47(7):829–837, 2000.

Valerie Freeman. Attitudes toward deafness affect impressions of young adults with cochlear implants. *The Journal of Deaf Studies and Deaf Education*, 23(4):360–368, 2018.

Melanie Fried-Oken, David R Beukelman, and Karen Hux. Current and future aac research considerations for adults with acquired cognitive and communication impairments. *Assistive Technology*, 24(1):56–66, 2012.

Nestor Garay, Idoia Cearreta, Juan Miguel López, and Inmaculada Fajardo. Assistive technology and affective mediation. *Human Technology: an Interdisciplinary Journal on Humans in ICT Environments*, 2006.

Jane Mertz Garcia and Michael P Cannito. Influence of verbal and nonverbal contexts on the sentence intelligibility of a speaker with dysarthria. *Journal of Speech, Language, and Hearing Research*, 39(4):750–760, 1996a.

Jane Mertz Garcia and Paul A Dagenais. Dysarthric sentence intelligibility: Contribution of iconic gestures and message predictiveness. *Journal of Speech, Language, and Hearing Research*, 41(6):1282–1293, 1998.

JM Garcia and MP Cannito. Top down influences on the intelligibility of a dysarthric speaker: Addition of natural gestures and situational context. *Disorders of Motor Speech*, pages 67–87, 1996b.

Mengzhe Geng, Xurong Xie, Shansong Liu, Jianwei Yu, Shoukang Hu, Xunying Liu, and Helen Meng. Investigation of data augmentation techniques for disordered speech recognition. *INTERSPEECH*, pages 696–700, 2020.

A Ghio, D Robert, C Grigoli, M Mas, C Delooze, C Mercier, and F Viallet. F0 characteristics in parkinsonian speech: contrast between the effect of hypodopaminergy due to parkinson's disease and that of the therapeutic delivery of l-dopa. *Revue de laryngologie-otologie-rhinologie*, 135(2):63–70, 2014.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016.

Steven L Gordon. Social structural effects on emotions. *Research Agendas in the Sociology of Emotions*, pages 145–179, 1990.

Michael Grimm, Kristian Kroschel, Emily Mower, and Shrikanth Narayanan. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49(10-11): 787–800, 2007.

Wentao Gu, Ping Fan, and Weiguo Liu. Acoustic analysis of mandarin speech in parkinson's disease with the effects of levodopa. In *International Seminar on Speech Production*, pages 211–224. Springer, 2017.

Dandan Guo, Hongzhi Yu, Axu Hu, and Yanbing Ding. Statistical analysis of acoustic characteristics of tibetan lhasa dialect speech emotion. In *SHS Web of Conferences*, volume 25, page 01017. EDP Sciences, 2016.

Jianting Guo and Haiyan Yu. Using speech emotion recognition to preclude campus bullying. In *International Conference on Machine Learning and Intelligent Communications*, pages 728–734. Springer, 2019.

Purnima Gupta and Nitendra Rajput. Two-stream emotion recognition for call center monitoring. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.

Noushin Hajarolasvadi and Hasan Demirel. 3d cnn-based speech emotion recognition using k-means clustering and spectrograms. *Entropy*, 21(5):479, 2019.

Vicki L Hammen and Kathryn M Yorkston. Speech and pause characteristics following speech rate reduction in hypokinetic dysarthria. *Journal of Communication Disorders*, 29 (6):429–445, 1996.

Wenjing Han, Tao Jiang, Yan Li, Björn Schuller, and Huabin Ruan. Ordinal learning for emotion recognition in customer service calls. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6494–6498. IEEE, 2020.

Meng Hao, Yan Tianhao, and Yuan Fei. The svm based on smo optimization for speech emotion recognition. In *2019 Chinese Control Conference (CCC)*, pages 7884–7888. IEEE, 2019.

Shlomo Hareli, Osnat Moran-Amir, Shlomo David, and Ursula Hess. Emotions as signals of normative conduct. *Cognition & Emotion*, 27(8):1395–1404, 2013.

Jinni Harrigan, Robert Rosenthal, Klaus R Scherer, and Klaus Scherer. *New handbook of methods in nonverbal behavior research*. Oxford University Press, 2008.

Lena Hartelius, Marie Elmberg, Rebecca Holm, Ann-Sofie Lövberg, and Stiliani Nikolaidis. Living with dysarthria: evaluation of a self-report questionnaire. *Folia Phoniatrica et Logopaedica*, 60(1):11–19, 2008.

Mark S Hawley, Pam Enderby, Phil Green, Stuart Cunningham, and Rebecca Palmer. Development of a voice-input voice-output communication aid (vivoca) for people with severe dysarthria. In *International Conference on Computers for Handicapped Persons*, pages 882–885. Springer, 2006.

Mark S Hawley, Pam Enderby, Phil Green, Stuart Cunningham, Simon Brownsell, James Carmichael, Mark Parker, Athanassios Hatzis, Peter O'Neill, and Rebecca Palmer. A speech-controlled environmental control system for people with severe dysarthria. *Medical Engineering & Physics*, 29(5):586–593, 2007.

Mark S Hawley, Stuart P Cunningham, Phil D Green, Pam Enderby, Rebecca Palmer, Siddharth Sehgal, and Peter O'Neill. A voice-input voice-output communication aid for people with severe speech impairment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 21(1):23–31, 2013.

Ursula Hess, Sacha Senécal, Gilles Kirouac, Pedro Herrera, Pierre Philippot, and Robert E Kleck. Emotional expressivity in men and women: Stereotypes and self-perceptions. *Cognition & Emotion*, 14(5):609–642, 2000.

D Jeffery Higginbotham, Howard Shane, Susanne Russell, and Kevin Caves. Access to AAC: present, past, and future. *Augmentative and Alternative Communication*, 23(3):243–257, 2007.

M Shamim Hossain. Patient state recognition system for healthcare using speech and facial expressions. *Journal of Medical Systems*, 40(12):272, 2016.

M Shamim Hossain and Ghulam Muhammad. Emotion-aware connected healthcare big data towards 5g. *IEEE Internet of Things Journal*, 5(4):2399–2406, 2017.

M Shamim Hossain, Ghulam Muhammad, Mohammed F Alhamid, Biao Song, and Khaled Al-Mutib. Audio-visual emotion recognition using big data towards 5g. *Mobile Networks and Applications*, 21(5):753–763, 2016.

Nazia Hossain and Mahmuda Naznin. Sensing emotion from voice jitter. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, pages 359–360, 2018.

Tami J Howe. The icf contextual factors related to speech-language pathology. *International Journal of Speech-Language Pathology*, 10(1-2):27–37, 2008.

Tom Howley, Michael G Madden, Marie-Louise O'Connell, and Alan G Ryder. The effect of principal component analysis on machine learning accuracy with high dimensional spectral data. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 209–222. Springer, 2005.

Lihong Hu, Zifan Yu, and Yanfang Liu. An algorithm of decision-tree generating automatically based on classification. In *2009 First International Workshop on Education Technology and Computer Science*, volume 1, pages 823–827. IEEE, 2009.

Xu Huahu, Gao Jue, and Yuan Jian. Application of speech emotion recognition in intelligent household robot. In *2010 International Conference on Artificial Intelligence and Computational Intelligence*, volume 1, pages 537–541. IEEE, 2010.

Jian Huang, Jianhua Tao, Bin Liu, and Zheng Lian. Learning utterance-level representations with label smoothing for speech emotion recognition. *INTERSPEECH*, pages 4079–4083, 2020.

Kun-Yi Huang, Chung-Hsien Wu, Qian-Bei Hong, Ming-Hsiang Su, and Yi-Hsuan Chen. Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5866–5870. IEEE, 2019.

Katherine C Hustad, Tabitha Jones, and Suzanne Dailey. Implementing speech supplementation strategies. *Journal of Speech, Language, and Hearing Research*, 2003.

Judy Illes, EJ Metter, WR Hanson, and S Iritani. Language production in parkinson's disease: acoustic and linguistic considerations. *Brain and Language*, 33(1):146–160, 1988.

Aseef Iqbal and Kakon Barua. A real-time emotion recognition from speech using gradient boosting. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 1–5. IEEE, 2019.

Ignasi Iriondo, Roger Guaus, Angel Rodríguez, Patricia Lázaro, Norminanda Montoya, Josep M Blanco, Dolors Bernadas, Josep Manel Oliver, Daniel Tena, and Ludovico Longhi. Validation of an acoustical modelling of emotional expression in spanish using speech synthesis techniques. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

Farzad Izadi, Ramin Mohseni, Ahmad Daneshi, and Nazila Sandughdar. Determination of fundamental frequency and voice intensity in iranian men and women aged between 18 and 45 years. *Journal of Voice*, 26(3):336–340, 2012.

Carroll E. Izard. *The face of emotion*. Appleton-Century-Crofts, 1971.

Carroll E Izard. Innate and universal facial expressions: evidence from developmental and cross-cultural research. *American Psychological Association*, 1994.

Rhonda J. Holmes, Jennifer M. Oates, Debbie J. Phyland, and Andrew J. Hughes. Voice characteristics in the progression of parkinson's disease. *International Journal of Language & Communication Disorders*, 35(3):407–418, 2000.

Rachael E Jack, Oliver GB Garrod, and Philippe G Schyns. Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biology*, 24(2): 187–192, 2014.

Philip Jackson and Sanaul Haq. Surrey audio-visual expressed emotion (SAVEE) database, Apr 2011. URL www.ee.surrey.ac.uk/Personal/P.Jackson/SAVEE/.

Agnes Jacob. Modelling speech emotion recognition using logistic regression and decision trees. *International Journal of Speech Technology*, 20(4):897–905, 2017.

Abhishek Jaywant and MARC D PELL. Listener impressions of speakers with parkinson's disease. *Journal of the International Neuropsychological Society: JINS*, 16(1):49, 2010.

Kittisak Jermsittiparsert, Abdurrahman Abdurrahman, Parinya Siriattakul, Ludmila A Sundeeva, Wahidah Hashim, Robbi Rahim, and Andino Maseleno. Pattern recognition and features selection for speech emotion recognition model using deep learning. *International Journal of Speech Technology*, pages 1–8, 2020.

A Jithendran, P Pranav Karthik, S Santhosh, and J Naren. Emotion recognition on e-learning community to improve the learning outcomes using machine learning concepts: A pilot study. In *Smart Systems and IoT: Innovations in Computing*, pages 521–530. Springer, 2020.

Philip N Johnson-Laird and Keith Oatley. Basic emotions, rationality, and folk theory. *Cognition & Emotion*, 6(3-4):201–223, 1992.

Tom Johnstone and Klaus R Scherer. Vocal communication of emotion. *Handbook of Emotions*, 2:220–235, 2000.

Ian Jolliffe. *Principal component analysis*. Springer, 2011.

Christian Jones and Jamie Sutherland. Acoustic emotion recognition for affective computer gaming. In *Affect and Emotion in Human-Computer Interaction*, pages 209–219. Springer, 2008.

Christian Martyn Jones and James Sutherland. Creating an emotionally reactive computer game responding to affective cues in speech. *HCI Proceedings*, pages 1–2, 2005.

Neethu Mariam Joy and S Umesh. Improving acoustic models in torgo dysarthric speech database. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(3): 637–645, 2018.

Daniel Jurafsky and James H Martin. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice-Hall, 2000.

James F Kaiser. On a simple algorithm to calculate the'energy'of a signal. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 381–384. IEEE, 1990.

L Kaiser. Communication of affects by single vowels. *Synthese*, 14(4):300–319, 1962.

Preeti Kapoor and Narina Thakur. Emotion recognition using Q-KNN: a faster KNN approach. In *International Conference on Innovative Computing and Communications*, pages 759–768. Springer, 2020.

Inger Karlsson, T Banziger, Jana Dankovicová, Tom Johnstone, Johan Lindberg, Håkan Melin, Francis Nolan, and K Scherer. Speaker verification with elicited speaking styles in the verivox project. *Speech Communication*, 31(2-3):121–129, 2000.

A. Karpf. *The human voice: the story of a remarkable talent*. Bloomsbury, 2007. ISBN 9780747585374.

Todd B Kashdan, Anjali Mishra, William E Breen, and Jeffrey J Froh. Gender differences in gratitude: Examining appraisals, narratives, the willingness to express emotions, and changes in psychological needs. *Journal of Personality*, 77(3):691–730, 2009.

Thayabaran Kathiresan and Volker Dellwo. Cepstral derivatives in MFCCs for emotion recognition. In *2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP)*, pages 56–60. IEEE, 2019.

Mary Katsikitis and I Pilowsky. A controlled quantitative study of facial expression in parkinson's disease and depression. *Journal of Nervous and Mental Disease*, 1991.

Xianxin Ke, Yujiao Zhu, Lei Wen, and Wenzhen Zhang. Speech emotion recognition based on svm and ann. *International Journal of Machine Learning and Computing*, 8(3):198–202, 2018.

Eric Keller. The analysis of voice quality in speech processing. In *International School on Neural Networks, Initiated by IIASS and EMFCSC*, pages 54–73. Springer, 2004.

Dacher Keltner and Jonathan Haidt. Social functions of emotions at four levels of analysis. *Cognition & Emotion*, 13(5):505–521, 1999.

Dacher Keltner and Ann M Kring. Emotion, social function, and psychopathology. *Review of General Psychology*, 2(3):320, 1998.

Daniel Kempler and Diana Van Lancker. Effect of speech task on intelligibility in dysarthria: A case study of parkinson's disease. *Brain and Language*, 80(3):449–464, 2002.

Raymond D Kent and Martin John Ball. *Voice quality measurement*. Singular, 2000.

Joseph Keshet. Automatic speech recognition: A primer for speech-language pathology researchers. *International Journal of Speech-Language Pathology*, 20(6):599–609, 2018.

Joann Keyton. *Communication and organizational culture: A key to understanding work experiences*. Sage, 2011.

Yelin Kim, Honglak Lee, and Emily Mower Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3687–3691. IEEE, 2013.

Charles E Kimble and Donald A Olszewski. Gaze and emotional expression: The effects of message positivity-negativity and emotional intensity. *Journal of Research in Personality*, 14(1):60–69, 1980.

KV Krishna Kishore and P Krishna Satish. Emotion recognition in speech using mfcc and wavelet features. In *2013 3rd IEEE International Advance Computing Conference (IACC)*, pages 842–847. IEEE, 2013.

Ina Kodrasi, Michaela Pernon, Marina Laganaro, and Hervé Bourlard. Automatic discrimination of apraxia of speech and dysarthria using a minimalistic set of handcrafted features. In *Proc. Annual Conference of the International Speech Communication Association, Shanghai, China*, 2020a.

Ina Kodrasi, Michaela Pernon, Marina Laganaro, and Hervé Bourlard. Automatic and perceptual discrimination between dysarthria, apraxia of speech, and neurotypical speech. *arXiv preprint arXiv:2011.07542*, 2020b.

Anusha Koduru, Hima Bindu Valiveti, and Anil Kumar Budati. Feature extraction algorithms to improve the speech emotion recognition rate. *International Journal of Speech Technology*, pages 1–11, 2020.

Hyejin Koo, Soyeong Jeong, Sungjae Yoon, and Wonjong Kim. Development of speech emotion recognition algorithm using MFCC and prosody. In *2020 International Conference on Electronics, Information, and Communication (ICEIC)*, pages 1–4. IEEE, 2020.

Guus de Krom. Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. *Journal of Speech, Language, and Hearing Research*, 38(4):794–811, 1995.

Mansi Kumbhakarn and Bageshree Sathe-Pathak. Analysis of emotional state of a person and its effect on speech features using praat software. In *2015 International Conference on Computing Communication Control and Automation*, pages 763–767. IEEE, 2015.

Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, and Julien Epps. Direct modelling of speech emotion from raw speech. *INTERSPEECH*, 2019.

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

Guylaine Le Dorze, Lisa Ouellet, and John Ryalls. Intonation and speech rate in dysarthric speech. *Journal of Communication Disorders*, 27(1):1–18, 1994.

Margaret Lech, Melissa Stolar, Robert Bolia, and Michael Skinner. Amplitude-frequency analysis of emotional speech using transfer learning and classification of spectrogram images. *Adv. Sci. Technol. Eng. Syst. J*, 3:363–371, 2018.

Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53(9-10):1162–1171, 2011.

Robert W Levenson. Human emotion: a functional view. *The Nature of Emotion: Fundamental Questions*, 1:123–126, 1994.

Robert W Levenson. The intrapersonal functions of emotion. *Cognition & Emotion*, 13(5): 481–504, 1999.

Robert W Levenson, Paul Ekman, Karl Heider, and Wallace V Friesen. Emotion and autonomic nervous system activity in the minangkabau of west sumatra. *Journal of Personality and Social Psychology*, 62(6):972, 1992.

Margaux Lhommet and Stacy C Marsella. *Expressing emotion through posture*, volume 273. The Oxford Handbook of Affective Computing, 2014.

Xi Li, Jidong Tao, Michael T Johnson, Joseph Soltis, Anne Savage, Kirsten M Leong, and John D Newman. Stress and emotion classification using jitter and shimmer features. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages IV–1081. IEEE, 2007.

Wootaek Lim, Daeyoung Jang, and Taejin Lee. Speech emotion recognition using convolutional and recurrent neural networks. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–4. IEEE, 2016.

Julie M Liss and Gary Weismer. Selected acoustic characteristics of contrastive stress production in control geriatric, apraxic, and ataxic dysarthric speakers. *Clinical Linguistics & Phonetics*, 8(1):45–66, 1994.

Jiateng Liu, Wenming Zheng, Yuan Zong, Cheng Lu, and Chuangao Tang. Cross-corpus speech emotion recognition based on deep domain-adaptive convolutional neural network. *IEICE TRANSACTIONS on Information and Systems*, 103(2):459–463, 2020.

Lei Liu, Meng Jian, and Wentao Gu. Prosodic characteristics of mandarin declarative and interrogative utterances in parkinson's disease. In *INTERSPEECH*, 2019.

Zheli Liu, Zhendong Wu, Tong Li, Jin Li, and Chao Shen. Gmm and cnn hybrid method for short utterance speaker recognition. *IEEE Transactions on Industrial informatics*, 14(7): 3244–3252, 2018.

Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

Lu-Shih Alex Low, Namunu C Maddage, Margaret Lech, Lisa B Sheeber, and Nicholas B Allen. Detection of clinical depression in adolescents' speech during family interactions. *IEEE Transactions on Biomedical Engineering*, 58(3):574–586, 2010.

Erfan Loweimi, Mortaza Doulaty, Jon Barker, and Thomas Hain. Long-term statistical feature extraction from speech signal and its application in emotion recognition. In *International Conference on Statistical Language and Speech Processing*, pages 173–184. Springer, 2015.

Marko Lugger and Bin Yang. The relevance of voice quality features in speaker independent emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages IV–17. IEEE, 2007.

Fred C Lunenburg. Communication: the process, barriers, and improving effectiveness. *Schooling*, 1(1):1–11, 2010.

Joan K-Y Ma, Tara L Whitehill, and Susanne Y-S So. Intonation contrast in cantonese speakers with hypokinetic dysarthria associated with parkinson's disease. *Journal of Speech, Language, and Hearing Research*, 2010.

Joan KY Ma and Rüdiger Hoffmann. Acoustic analysis of intonation in parkinson's disease. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

Xi Ma, Zhiyong Wu, Jia Jia, Mingxing Xu, Helen Meng, and Lianhong Cai. Emotion recognition from variable-length speech segments using deep learning on spectrograms. In *INTERSPEECH*, pages 3683–3687, 2018.

Brenda Major and Laurie T O'brien. The social psychology of stigma. *Annual Review of Psychology*, 56:393–421, 2005.

Shuiyang Mao, Dehua Tao, Guangyan Zhang, PC Ching, and Tan Lee. Revisiting hidden markov models for speech emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6715–6719. IEEE, 2019.

Heidi Martens, Gwen Van Nuffelen, Patrick Cras, Barbara Pickut, Miet De Letter, and Marc De Bodt. Assessment of prosodic communicative efficiency in parkinson's disease as judged by professional listeners. *Parkinson's Disease*, 2011, 2011.

Heidi Martens, Gwen Van Nuffelen, Marc De Bodt, Tomas Dekens, Lukas Latacz, and Werner Verhelst. Automated assessment and treatment of speech rate and intonation in dysarthria. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, pages 382–384. IEEE, 2013.

Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas. The enterface'05 audio-visual emotion database. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pages 8–8. IEEE, 2006.

David Matsumoto, Mark G Frank, and Hyi Sung Hwang. *Nonverbal communication: Science and applications: Science and applications*. Sage, 2013.

David Ricky Matsumoto. *Culture and psychology*. Brooks/Cole Pub, 1996.

Evalynn J Mazurski and Nigel W Bond. A new series of slides depicting facial expressions of affect: A comparison with the pictures of facial affect series. *Australian Journal of Psychology*, 45(1):41–47, 1993.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, volume 8, pages 18–25, 2015.

David Mcnaughton and Diane Nelson Bryen. Aac technologies to enhance participation and access to meaningful societal roles for adolescents and adults with developmental disabilities who require aac. *Augmentative and Alternative Communication*, 23(3):217–229, 2007.

Elvira Mendoza, Nieves Valencia, Juana Muñoz, and Humberto Trujillo. Differences in voice quality between men and women: use of the long-term average spectrum (ltas). *Journal of Voice*, 10(1):59–66, 1996.

Xavier Menendez-Pidal, James B Polikoff, Shirley M Peters, Jennie E Leonzio, and H Timothy Bunnell. The nemours database of dysarthric speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1962–1965. IEEE, 1996.

Kinfe Tadesse Mengistu and Frank Rudzicz. Adapting acoustic and lexical models to dysarthric speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4924–4927. IEEE, 2011.

Freeman Miller and Steven J Bachrach. *Cerebral palsy: A complete guide for caregiving*. JHU Press, 2017.

Kevin P Murphy. Machine learning: A probabilistic perspective. adaptive computation and machine learning, 2012.

Iain R Murray and John L Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2):1097–1108, 1993.

Kathleen F Nagle, Tanya L Eadie, and Kathryn M Yorkston. Everyday listeners' impressions of speech produced by individuals with adductor spasmodic dysphonia. *Journal of Communication Disorders*, 58:1–13, 2015.

Michael Neumann and Ngoc Thang Vu. Attentive convolutional neural network based speech emotion recognition: a study on the impact of input features, signal length, and acted speech. *INTERSPEECH*, 2017.

Michael Neumann and Ngoc Thang Vu. Improving speech emotion recognition with unsupervised representation learning on unlabeled speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7390–7394. IEEE, 2019.

Andrew Ortony and Terence J Turner. What's basic about basic emotions? *Psychological Review*, 97(3):315, 1990.

Hemanta Kumar Palo, Mahesh Chandra, and Mihir Narayan Mohanty. Recognition of human speech emotion using variants of mel-frequency cepstral coefficients. In *Advances in Systems, Control and Automation*, pages 491–498. Springer, 2018.

Hemanta Kumar Palo, Debasis Behera, and Bikash Chandra Rout. Comparison of classifiers for speech emotion recognition (ser) with discriminative spectral features. In *Advances in Intelligent Computing and Communication*, pages 78–85. Springer, 2020.

Dimitris Pappas, Ion Androutsopoulos, and Haris Papageorgiou. Anger detection in call center dialogues. In *2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 139–144. IEEE, 2015.

Brian Parkinson. Emotions are social. *British Journal of Psychology*, 87(4):663–683, 1996.

Pavitra Patel, Anand Chaudhari, Ruchita Kale, and M Pund. Emotion recognition from speech with gaussian mixture models & via boosted gmm. *International Journal of Research In Science & Engineering*, 3, 2017.

R Patel. Prosodic control in severe dysarthria: preserved ability to mark the question-statement contrast. *Journal of Speech, Language, and Hearing Research*, 45(5):858–870, 2002a.

Rupal Patel. Phonatory control in adults with cerebral palsy and severe dysarthria. *Augmentative and Alternative Communication*, 18(1):2–10, 2002b.

Rupal Patel. Acoustic characteristics of the question-statement contrast in severe dysarthria due to cerebral palsy. *Journal of Speech, Language, and Hearing Research*, 2003.

Rupal Patel. The acoustics of contrastive prosody in adults with cerebral palsy. *Journal of Medical Speech-Language Pathology*, 12(4):189–193, 2004.

Rupal Patel and Pamela Campellone. Acoustic and perceptual cues to contrastive stress in dysarthria. *Journal of Speech, Language, and Hearing Research*, 2009.

Rupal Patel and Carrie Watkins. Stress identification in speakers with dysarthria due to cerebral palsy: An initial report. *Journal of Medical Speech-Language Pathology*, 15(2): 149–160, 2007.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct): 2825–2830, 2011.

Marc D Pell, Henry S Cheang, and Carol L Leonard. The impact of parkinson's disease on vocal-prosodic communication from the perspective of listeners. *Brain and Language*, 97 (2):123–134, 2006.

Brian Pentland, Thomas K Pitcairn, John M Gray, and William Riddle Jr. The effects of reduced expression in parkinson's disease on impression formation by health professionals. *Clinical Rehabilitation*, 1(4):307–312, 1987.

Thomas K Pitcairn, Susan Clemie, John M Gray, and Brian Pentland. Impressions of parkinsonian patients from their recorded voices. *International Journal of Language & Communication Disorders*, 25(1):85–92, 1990a.

Thomas K Pitcairn, Susan Clemie, John M Gray, and Brian Pentland. Non-verbal cues in the self-presentation of parkinsonian patients. *British Journal of Clinical Psychology*, 29(2): 177–184, 1990b.

Christopher J Plack, Daphne Barker, and Deborah A Hall. Pitch coding and pitch processing in the human brain. *Hearing Research*, 307:53–64, 2014.

E Ashby Plant, Janet Shibley Hyde, Dacher Keltner, and Patricia G Devine. The gender stereotyping of emotions. *Psychology of Women Quarterly*, 24(1):81–92, 2000.

Larry J Platt, Gavin Andrews, Margrette Young, and Peter T Quinn. Dysarthria of adult cerebral palsy: I. intelligibility and articulatory impairment. *Journal of Speech, Language, and Hearing Research*, 23(1):28–40, 1980.

Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350, 2001.

Muhammad Ramadhan Prasetya, Agus Harjoko, Catur Supriyanto, et al. Speech emotion recognition of indonesian movie audio tracks based on MFCC and SVM. In *2019 International Conference on Contemporary Computing and Informatics (IC3I)*, pages 22–25. IEEE, 2019.

Sathit Prasomphan and Surinee Doungwichain. Detecting human emotions in a large size of database by using ensemble classification model. *Mobile Networks and Applications*, 23 (4):1097–1102, 2018.

Jesse Prinz. Which emotions are basic. *Emotion, Evolution, and Rationality*, 69:88, 2004.

Graham Pullin and Shannon Hennig. 17 ways to say yes: Toward nuanced tone of voice in aac and speech technology. *Augmentative and Alternative Communication*, 31(2):170–180, 2015.

LI Qianqian, Fuji Ren, Xiaoyan Shen, and Xin Kang. Speech emotion recognition based on data enhancement in time-frequency domain. In *International Symposium on Artificial Intelligence and Robotics*, volume 11574, page 115740R. International Society for Optics and Photonics, 2020.

Lawrence Rabiner. *Fundamentals of speech recognition*. PTR Prentice Hall, 1993.

Viviana Mendoza Ramos, Hector A Kairuz Hernandez-Diaz, Maria E Hernandez-Diaz Huici, Heidi Martens, Gwen Van Nuffelen, and Marc De Bodt. Acoustic features to characterize sentence accent production in dysarthric speech. *Biomedical Signal Processing and Control*, 57:101750, 2020.

K Sreenivasa Rao, Shashidhar G Koolagudi, and Ramu Reddy Vempada. Emotion recognition from speech using global and local prosodic features. *International Journal of Speech Technology*, 16(2):143–160, 2013.

A Revathi and N Sasikaladevi. Emotion recognition from speech using perceptual filter and neural network. In *Neural Networks for Natural Language Processing*, pages 78–91. IGI Global, 2020.

Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic. Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 3–8. ACM, 2015.

Fabien Ringeval, Erik Marchi, Charline Grossard, Jean Xavier, Mohamed Chetouani, David Cohen, and Björn Schuller. Automatic analysis of typical and atypical encoding of spontaneous emotion in the voice of children. In *INTERSPEECH*, pages 1210–1214, 2016.

Michael D Robinson and Joel T Johnson. Is it emotion or is it stress? gender stereotypes and the perception of subjective experience. *Sex Roles*, 36(3-4):235–258, 1997.

Simon Rogers and Mark Girolami. *A first course in machine learning*. CRC Press, 2016.

Kristin M Rosen, Raymond D Kent, Amy L Delaney, and Joseph R Duffy. Parametric quantitative acoustic analysis of conversation produced by speakers with dysarthria and healthy speakers. *Journal of Speech, Language, and Hearing Research*, 2006.

Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Language resources and evaluation*, 46(4):523–541, 2012.

James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161, 1980.

James A Russell and Merry Bullock. Multidimensional scaling of emotional facial expressions: similarity from preschoolers to adults. *Journal of Personality and Social Psychology*, 48(5):1290, 1985.

Jan Rusz, Roman Cmejla, Hana Ruzickova, and Evzen Ruzicka. Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinson's disease. *The journal of the Acoustical Society of America*, 129(1):350–367, 2011.

Gaurav Sahu. Multimodal speech emotion recognition and ambiguity resolution. *arXiv preprint arXiv:1904.06022*, 2019.

K Saravanan and S Sasithra. Review on classification based on artificial neural networks. *International Journal of Ambient Systems and Applications (IJASA)*, 2(4):11–18, 2014.

Md Kamruzzaman Sarker, Kazi Md Rokibul Alam, and Md Arifuzzaman. Emotion recognition from speech based on relevant feature and majority voting. In *2014 International Conference on Informatics, Electronics & Vision (ICIEV)*, pages 1–5. IEEE, 2014.

Aharon Satt, Shai Rozenberg, and Ron Hoory. Efficient emotion recognition from speech using deep learning on spectrograms. In *INTERSPEECH*, pages 1089–1093, 2017.

Disa A Sauter, Frank Eisner, Andrew J Calder, and Sophie K Scott. Perceptual cues in nonverbal vocal expressions of emotion. *Quarterly Journal of Experimental Psychology*, 63(11):2251–2272, 2010.

Klaus R Scherer. Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99(2):143, 1986.

Klaus R Scherer. Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1-2):227–256, 2003.

Klaus R Scherer and Grazia Ceschi. Lost luggage: a field study of emotion–antecedent appraisal. *Motivation and emotion*, 21(3):211–235, 1997.

Klaus R Scherer, Judy Koivumaki, and Robert Rosenthal. Minimal cues in the vocal communication of affect: Judging emotions from content-masked speech. *Journal of Psycholinguistic Research*, 1(3):269–285, 1972.

Klaus R Scherer, Harvey London, and Jared J Wolf. The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality*, 7(1):31–44, 1973.

Klaus R Scherer, Stanley Feldstein, Ronald N Bond, and Robert Rosenthal. Vocal cues to deception: A comparative channel approach. *Journal of Psycholinguistic Research*, 14(4): 409–425, 1985.

KR Scherer and G Ceschi. Studying affective communication in the airport: The case of lost baggage claims. *Personality and Social Psychology Bulletin*, 26(3):327–339, 2000.

U Scherer, H Helfrich, and Klaus R Scherer. Paralinguistic behaviour: internal push or external pull? In *Language*, pages 279–282. Elsevier, 1980.

Björn Schuller, Ronald Müller, Manfred Lang, and Gerhard Rigoll. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *Ninth European Conference on Speech Communication and Technology*, 2005.

Björn Schuller, Stefan Steidl, and Anton Batliner. The INTERSPEECH 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *INTERSPEECH*, 2013.

Thapanee Seehapoch and Sartra Wongthanavasu. Speech emotion recognition using support vector machines. In *2013 5th International Conference on Knowledge and Smart Technology (KST)*, pages 86–91. IEEE, 2013.

Maheshwari Selvaraj, R Bhuvana, and S Padmaja. Human speech emotion recognition. *International Journal of Engineering & Technology*, 8:311–323, 2016.

Claude E Shannon and Warren Weaver. *The mathematical theory of communication*. University of illinois Press, Urbana, 1949.

Ying Shi and Weihua Song. Speech emotion recognition based on data mining technology. In *2010 Sixth International Conference on Natural Computation*, volume 2, pages 615–619. IEEE, 2010.

Joel Shor, Dotan Emanuel, Oran Lang, Omry Tuval, Michael Brenner, Julie Cattiau, Fernando Vieira, Maeve McNally, Taylor Charbonneau, Melissa Nollstadt, et al. Personalizing asr for dysarthric and accented speech with limited data. *INTERSPEECH*, 2019.

Marcia C Smith, Melissa K Smith, and Heiner Ellgring. Spontaneous and posed facial expression in parkinson's disease. *Journal of the International Neuropsychological Society*, 2(5):383–391, 1996.

Juliu O Smith III. *Spectral audio signal processing*. W3K Publishing, 2011.

Jennifer L Spielman, Joan C Borod, and Lorraine O Ramig. The effects of intensive voice treatment on facial expressiveness in parkinson disease: preliminary data. *Cognitive and Behavioral Neurology*, 16(3):177–188, 2003.

Konstantin Stanislavsky, ER Hapgood, and J Gielgud. An actor prepares. new york, ny: theatre arts, 1936.

Piotr Staroniewicz and Wojciech Majewski. Polish emotional speech database–recording and preliminary validation. In *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions*, pages 42–49. Springer, 2009.

Sheila Steinberg. *Introduction to communication course book 1: The basics*, volume 1. Juta and Company Ltd, 1995.

Melissa Stola, Margaret Lech, Robert S Bolia, and Michael Skinner. Acoustic characteristics of emotional speech using spectrogram image classification. In *2018 12th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pages 1–5. IEEE, 2018.

Nithya Sundaram, Brett Y Smolenski, and R Yantorno. Instantaneous nonlinear teager energy operator for robust voiced–unvoiced speech classification. *Speech Processing*, 2003.

Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. Two-step acoustic model adaptation for dysarthric speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6104–6108. IEEE, 2020.

João Paulo Teixeira and Paula Odete Fernandes. Jitter, shimmer and HNR classification within gender, tones and vowels in healthy voices. *Procedia Technology*, 16:1228–1237, 2014.

Therapy Box. VocaTempo. URL https://therapy-box.co.uk/vocatempo.

Erin H Thompson and James A Hampton. The effect of relationship status on communicating emotions through touch. *Cognition & Emotion*, 25(2):295–306, 2011.

Leimin Tian, Johanna Moore, and Catherine Lai. Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 565–572. IEEE, 2016.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Monique Timmers, Agneta Fischer, and Antony Manstead. Ability versus vulnerability: Beliefs about men's and women's emotional behaviour. *Cognition and Emotion*, 17(1): 41–63, 2003.

Juhani Toivanen, Teija Waaramaa, Paavo Alku, Anne-Maria Laukkanen, Tapio Seppänen, Eero Väyrynen, and Matti Airas. Emotions in [a]: a perceptual and acoustic study. *Logopedics Phoniatrics Vocology*, 31(1):43–48, 2006.

Frank J Tolkmitt and Klaus R Scherer. Effect of experimentally induced stress on vocal parameters. *Journal of Experimental Psychology: Human Perception and Performance*, 12(3):302, 1986.

Silvan Tomkins. *Affect imagery consciousness: Volume I: The positive affects*. Springer Publishing Company, 1962.

Silvan Tomkins. *Affect imagery consciousness: Volume II: The negative affects*. Springer Publishing Company, 1963.

John Tooby and Leda Cosmides. The evolutionary psychology of the emotions and their relationship to internal regulatory variables. In *Handbook of Emotions, eds*. Citeseer, 2008.

Nim Tottenham, James W Tanaka, Andrew C Leon, Thomas McCarry, Marcella Nurse, Todd A Hare, David J Marcus, Alissa Westerlund, BJ Casey, and Charles Nelson. The nimstim set of facial expressions: judgments from untrained research participants. *Psychiatry Research*, 168(3):242–249, 2009.

MTD Training. *Effective communication skills*. Bookboon, 2012.

George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204. IEEE, 2016.

Jürgen Trouvain and Khiet Phuong Truong. Comparing non-verbal vocalisations in conversational speech corpora. In *Proceedings of the 4th International Workshop on Corpora for Research on Emotion Sentiment and Social Signals (ES3 2012)*. European Language Resources Association (ELRA), 2012.

Renée Van Bezooijen, Stanley A Otto, and Thomas A Heenan. Recognition of vocal expressions of emotion: A three-nation study to identify universal characteristics. *Journal of Cross-Cultural Psychology*, 14(4):387–406, 1983.

Egon L Van Den Broek, Viliam Lisỳ, Joris H Janssen, Joyce HDM Westerink, Marleen H Schut, and Kees Tuinenbreijer. Affective man-machine interface: unveiling human emotions through biosignals. In *International Joint Conference on Biomedical Engineering Systems and Technologies*, pages 21–47. Springer, 2009.

Diana Van Lancker, Cathleen Cornelius, and Jody Kreiman. Recognition of emotional-prosodic meanings in speech by autistic, schizophrenic, and normal children. *Developmental Neuropsychology*, 5(2-3):207–226, 1989.

P Vasuki and Chandrabose Aravindan. Hierarchical classifier design for speech emotion recognition in the mixed-cultural environment. *Journal of Experimental & Theoretical Artificial Intelligence*, pages 1–16, 2020.

Rudolph F Verderber. *Communicate!* Wadsworth Publishing Company, 1990.

Dimitrios Ververidis and Constantine Kotropoulos. A review of emotional speech databases. In *Proc. Panhellenic Conference on Informatics (PCI)*, pages 560–574, 2003.

Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech classification using gaussian mixture models and the sequential floating forward selection algorithm. In *2005 IEEE International Conference on Multimedia and Expo*, pages 1500–1503. IEEE, 2005.

Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181, 2006.

Laurence Vidrascu and Laurence Devillers. Five emotion classes detection in real-world call center data: the use of various types of paralinguistic features. In *Proc. Inter. Workshop on Paralinguistic Speech Between Models and Data, ParaLing*, 2007.

Margaret Walshe and Nick Miller. Living with acquired dysarthria: the speaker's perspective. *Disability and Rehabilitation*, 33(3):195–203, 2011.

Jianyou Wang, Michael Xue, Ryan Culhane, Enmao Diao, Jie Ding, and Vahid Tarokh. Speech emotion recognition with dual-sequence LSTM architecture. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6474–6478. IEEE, 2020.

Kunxia Wang, Zongcheng Chu, Kai Wang, Tongqing Yu, and Li Liu. Speech emotion recognition using multiple classifiers. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data*, pages 84–93. Springer, 2017.

Sandra P Whiteside. Simulated emotions: an acoustic study of voice and perturbation measures. In *Fifth International Conference on Spoken Language Processing*, 1998.

SP Whiteside. Acoustic characteristics of vocal emotions simulated by actors. *Perceptual and motor skills*, 89(3_suppl):1195–1208, 1999.

John Wilson, Bronagh Blaney. Acoustic variability in dysarthria and computer speech recognition. *Clinical Linguistics & Phonetics*, 14(4):307–327, 2000.

Wendy Wood, Nancy Rhodes, and Melanie Whelan. Sex differences in positive well-being: A consideration of emotional style and marital status. *Psychological Bulletin*, 106(2):249, 1989.

Weidi Xie, Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Utterance-level aggregation for speaker recognition in the wild. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5791–5795. IEEE, 2019a.

Yue Xie, Ruiyu Liang, Zhenlin Liang, Chengwei Huang, Cairong Zou, and Björn Schuller. Speech emotion classification using attention-based lstm. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(11):1675–1685, 2019b.

Feifei Xiong, Jon Barker, and Heidi Christensen. Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition. In *Speech Communication; 13th ITG-Symposium*, pages 1–5. VDE, 2018.

Feifei Xiong, Jon Barker, and Heidi Christensen. Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5836–5840. IEEE, 2019.

Feifei Xiong, Jon Barker, Zhengjun Yue, and Heidi Christensen. Source domain data selection for improved transfer learning targeting dysarthric speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7424–7428. IEEE, 2020.

Sherif Yacoub. Recognition of emotions in interactive voice response systems. In *EuroSpeech 2003, 8th European Conference on Speech Communication Technologies. Geneva, Switzerland, Sept 1-4,*, 2003.

Ningning Yang and Fuqian Shi. Speech emotion recognition based on back propagation neural network. In *Proceedings of the 3rd International Conference on Information Technology and Intelligent Transportation Systems (ITITS 2018) Xi'an, China, September 15-16, 2018*, volume 314, page 216. IOS Press, 2019.

Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Zhigang Deng, Sungbok Lee, Shrikanth Narayanan, and Carlos Busso. An acoustic study of emotions expressed in speech. In *Eighth International Conference on Spoken Language Processing*, 2004.

Emre Yilmaz, Mario Ganzeboom, Lilian Beijer, Catia Cucchiarini, and Helmer Strik. A dutch dysarthric speech database for individualized speech therapy research. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 792–795, 2016.

Won-Joong Yoon and Kyu-Sik Park. A study of emotion recognition and its applications. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 455–462. Springer, 2007.

Won-Joong Yoon, Youn-Ho Cho, and Kyu-Sik Park. A study of speech emotion recognition and its application to mobile services. In *International Conference on Ubiquitous Intelligence and Computing*, pages 758–766. Springer, 2007.

KM Yorkston, C Bombardier, and VL Hammen. Dysarthria from the viewpoint of individuals with dysarthria. *Motor speech disorders: Advances in assessment and treatment*, pages 19–36, 1994.

Zhengjun Yue, Heidi Christensen, and Jon Barker. Autoencoder bottleneck features with multi-task optimisation for improved continuous dysarthric speech recognition. In *INTERSPEECH*. International Speech Communication Association (ISCA), 2020a.

Zhengjun Yue, Feifei Xiong, Heidi Christensen, and Jon Barker. Exploring appropriate acoustic and language modelling choices for continuous dysarthric speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6094–6098. IEEE, 2020b.

Emmanuelle Zech and Bernard Rimé. Is talking about an emotional experience helpful? effects on emotional recovery and perceived benefits. *Clinical Psychology & Psychotherapy*, 12(4):270–287, 2005.

Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2008.

Ling Zhang, Ying Wei, Shuangwei Wang, Di Pan, Shili Liang, and Tingfa Xu. Specific two words lexical semantic recognition based on the wavelet transform of narrowband spectrogram. In *2017 First International Conference on Electronics Instrumentation & Information Systems (EIIS)*, pages 1–6. IEEE, 2017.

Huan Zhou, Kai Liu, and PRC Shenzhen. Speech emotion recognition with discriminative feature learning. *INTERSPEECH*, pages 4094–4097, 2020.

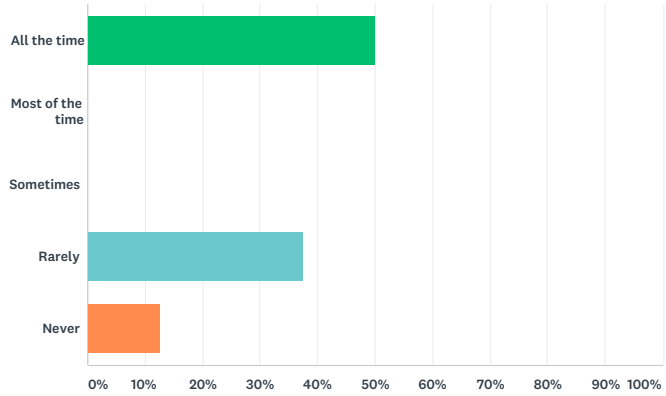
Zoom Video Communications Inc. Security guide. zoom video communications inc., 2016. URL https://d24cgw3uvb9a9h.cloudfront.net/static/81625/doc/Zoom-Security-White-Paper.pdf.

Appendix A

Survey - Questions and Results

Q1 How often do you use a communication aid with familiar people?

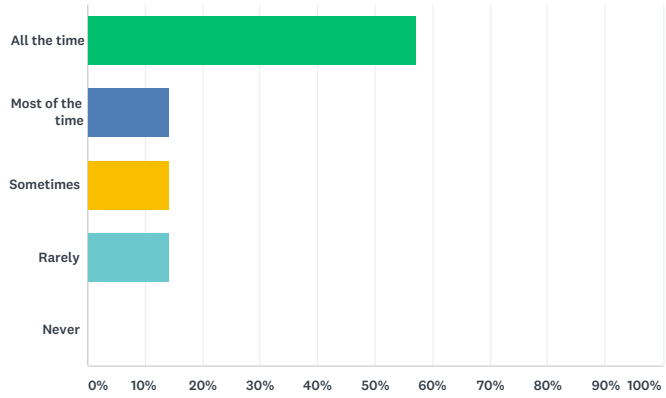
Answered: 8 Skipped: 0



ANSWER CHOICES	RESPONSES	
All the time	50.00%	4
Most of the time	0.00%	0
Sometimes	0.00%	0
Rarely	37.50%	3
Never	12.50%	1
TOTAL		8

Q2 How often do you use a communication aid with unfamiliar people?

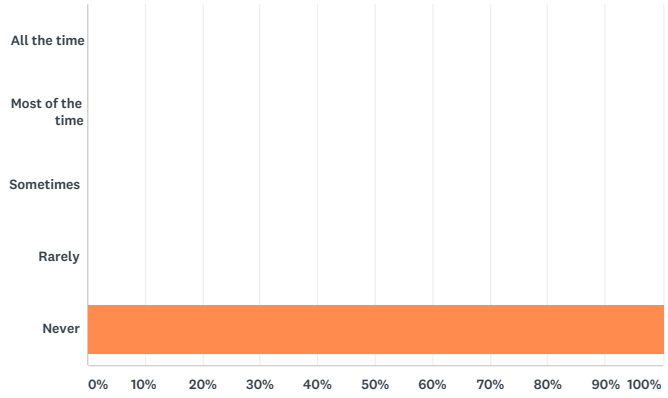
Answered: 7 Skipped: 1



ANSWER CHOICES	RESPONSES	
All the time	57.14%	4
Most of the time	14.29%	1
Sometimes	14.29%	1
Rarely	14.29%	1
Never	0.00%	0
TOTAL		7

Q3 How often do you use a communication aid with unfamiliar people?

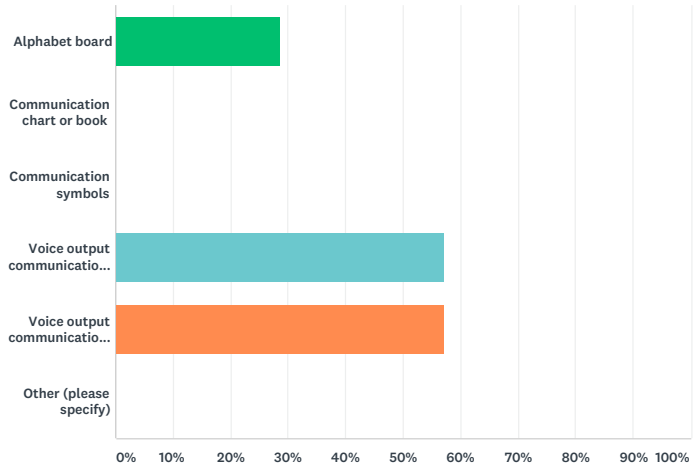
Answered: 1 Skipped: 7



ANSWER CHOICES	RESPONSES	
All the time	0.00%	0
Most of the time	0.00%	0
Sometimes	0.00%	0
Rarely	0.00%	0
Never	100.00%	1
TOTAL		1

Q4 What are the communication aids that you use? (please, choose all applicable).

Answered: 7 Skipped: 1



ANSWER CHOICES	RESPONSES	
Alphabet board	28.57%	2
Communication chart or book	0.00%	0
Communication symbols	0.00%	0
Voice output communication aid	57.14%	4
Voice output communication aid app on tablet	57.14%	4
Other (please specify)	0.00%	0
Total Respondents: 7		

#	OTHER (PLEASE SPECIFY)	DATE
	There are no responses.	

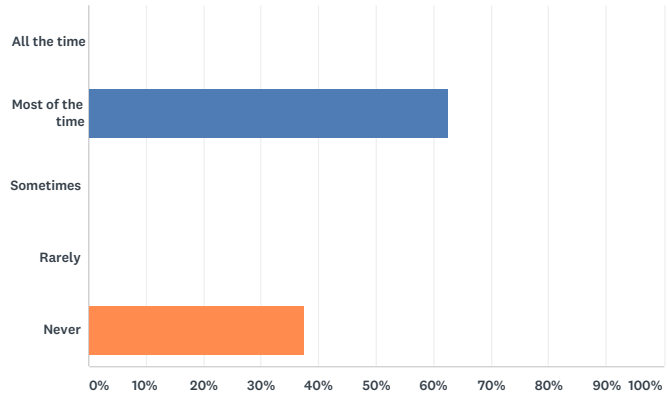
Q5 Could you please tell us why you don’t use communication aids?

Answered: 1 Skipped: 7

#	RESPONSES	DATE
1	speech is quicker	6/15/2018 9:23 PM

Q6 Without using communication aids, familiar people normally understand your speech …….

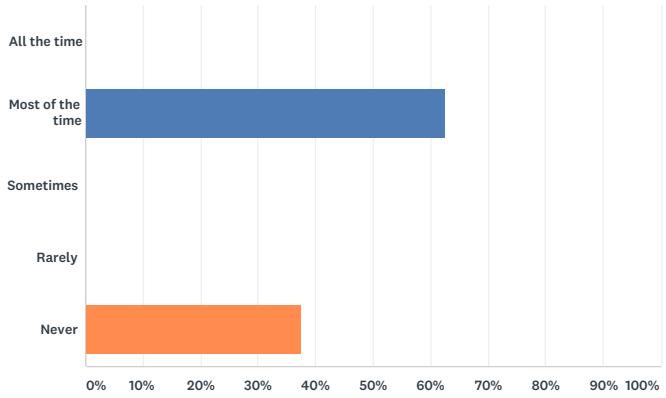
Answered: 8 Skipped: 0



ANSWER CHOICES	RESPONSES	
All the time	0.00%	0
Most of the time	62.50%	5
Sometimes	0.00%	0
Rarely	0.00%	0
Never	37.50%	3
TOTAL		8

Q6 Without using communication aids, familiar people normally understand your speech ……..

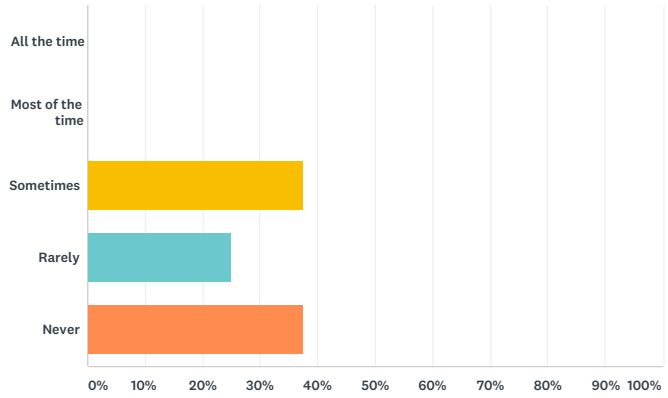
Answered: 8 Skipped: 0



ANSWER CHOICES	RESPONSES	
All the time	0.00%	0
Most of the time	62.50%	5
Sometimes	0.00%	0
Rarely	0.00%	0
Never	37.50%	3
TOTAL		8

Q7 Without using communication aids, unfamiliar people normally understand your speech ……..

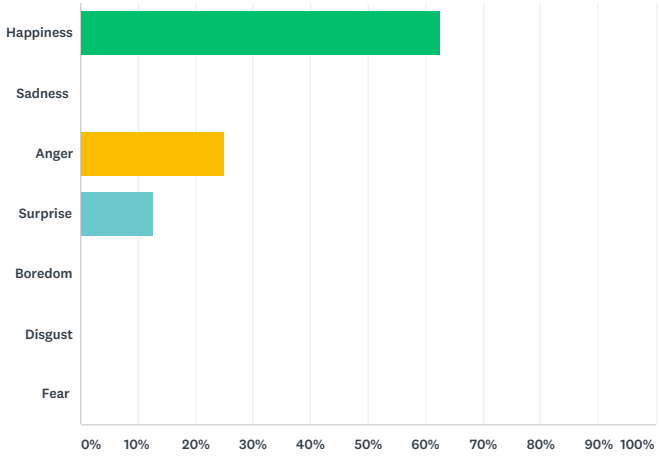
Answered: 8 Skipped: 0



ANSWER CHOICES	RESPONSES	
All the time	0.00%	0
Most of the time	0.00%	0
Sometimes	37.50%	3
Rarely	25.00%	2
Never	37.50%	3
TOTAL		8

Q8 From the following emotions, what emotion do you feel is overall the most important to communicate in your everyday life?

Answered: 8 Skipped: 0



ANSWER CHOICES	RESPONSES	
Happiness	62.50%	5
Sadness	0.00%	0
Anger	25.00%	2
Surprise	12.50%	1
Boredom	0.00%	0
Disgust	0.00%	0
Fear	0.00%	0
TOTAL		8

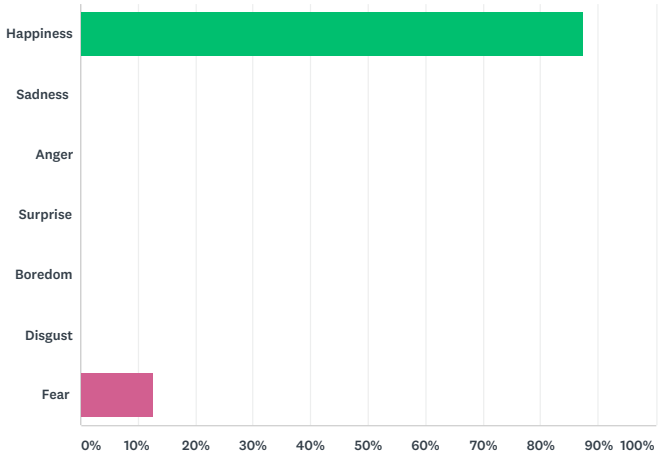
Q9 Why do you think the emotion you selected in the previous question is important?

Answered: 8 Skipped: 0

#	RESPONSES	DATE
1	It is very important that the afflicted person's request for help and support is delivered in the same manner as how they would expect to receive this help and support.	6/25/2018 9:52 PM
2	all emotions are important	6/22/2018 4:17 PM
3	rude	6/21/2018 3:39 PM
4	Te physiological interaction between dysarthria and anger	6/15/2018 9:27 PM
5	People need to know that I am happy with them so they want to come back and be with me.	6/11/2018 8:35 PM
6	I want people to think that I'm a positive person.	5/25/2018 4:54 PM
7	because it's I	5/22/2018 12:06 PM

Q10 What emotion do you feel is the most useful to try to communicate in your social life?

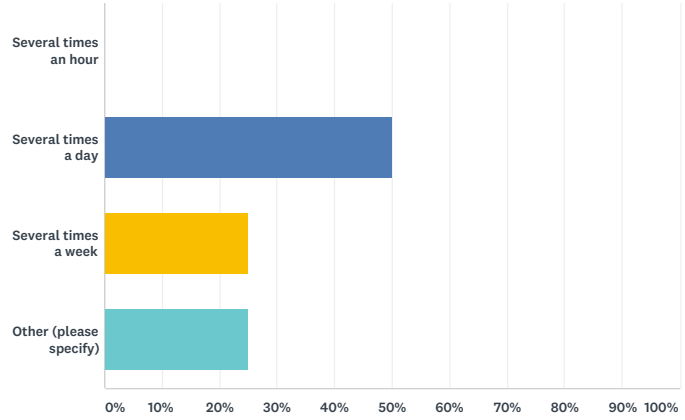
Answered: 8 Skipped: 0



ANSWER CHOICES	RESPONSES	
Happiness	87.50%	7
Sadness	0.00%	0
Anger	0.00%	0
Surprise	0.00%	0
Boredom	0.00%	0
Disgust	0.00%	0
Fear	12.50%	1
TOTAL		8

Q11 How often do you try to communicate the emotion you selected in the previous question?

Answered: 8 Skipped: 0

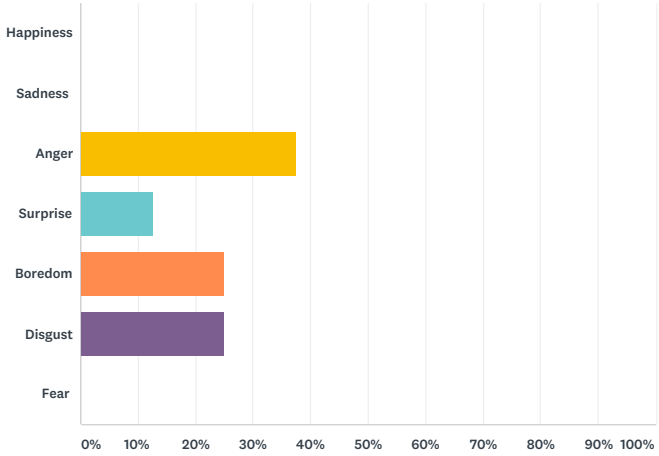


ANSWER CHOICES	RESPONSES	
Several times an hour	0.00%	0
Several times a day	50.00%	4
Several times a week	25.00%	2
Other (please specify)	25.00%	2
TOTAL		8

#	OTHER (PLEASE SPECIFY)	DATE
1	Every time I communicate with someone I always try and do this with a smile, I also make a point of offering my help.	6/25/2018 9:55 PM
2	When provoked	6/15/2018 9:29 PM

Q12 What emotion do you feel is the most difficult for you to communicate to familiar people?

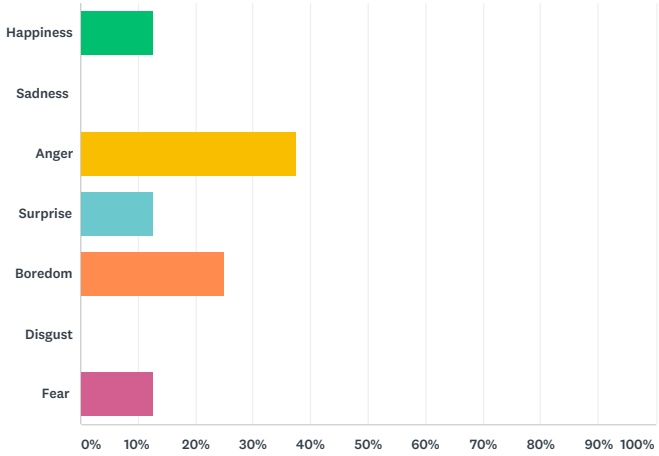
Answered: 8 Skipped: 0



ANSWER CHOICES	RESPONSES	
Happiness	0.00%	0
Sadness	0.00%	0
Anger	37.50%	3
Surprise	12.50%	1
Boredom	25.00%	2
Disgust	25.00%	2
Fear	0.00%	0
TOTAL		8

Q13 What emotion do you feel is the most difficult for you to communicate to unfamiliar people?

Answered: 8 Skipped: 0



ANSWER CHOICES	RESPONSES	
Happiness	12.50%	1
Sadness	0.00%	0
Anger	37.50%	3
Surprise	12.50%	1
Boredom	25.00%	2
Disgust	0.00%	0
Fear	12.50%	1
TOTAL		8

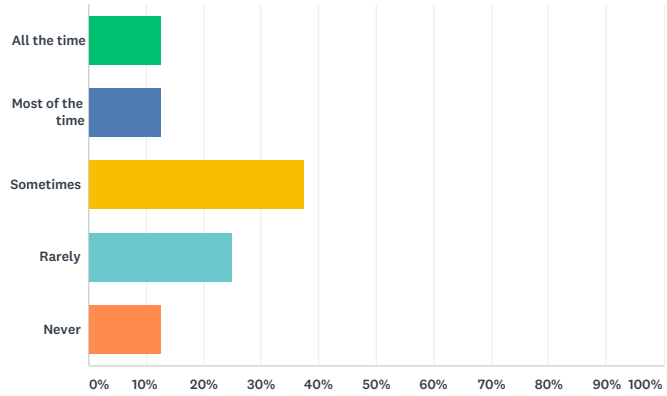
Q14 Can you please give a situation where you tried to communicate the emotion you selected in the previous question and it was difficult for you to convey or for others to understand?

Answered: 8 Skipped: 0

#	RESPONSES	DATE
1	I would never show disgust to someone who I'm familiar with, people who I am unfamiliar with will never see me angry. Disgust often comes with anger-related behaviour.	6/25/2018 10:00 PM
2	this survey is confusing	6/22/2018 4:19 PM
3	was	6/21/2018 3:40 PM
4	A P.A. was cooking my tea when a picture of a snake appeared on my Facebook, my P.A. did not realise I needed them because I frightened.	6/15/2018 11:47 PM
5	The LightWriter has no tone so even if i'm not angry with someone, the LightWriter makes me sounds like i am	6/11/2018 8:38 PM
6	It is an emotion that it hard for people to detect in me. It doesn't happen often.	5/25/2018 4:59 PM
7	When I was in residential respite .	5/22/2018 12:14 PM
8	how I am feeling about my situation at that time	5/21/2018 2:00 PM

Q15 How often do you feel it is difficult to convey an emotion to familiar people?

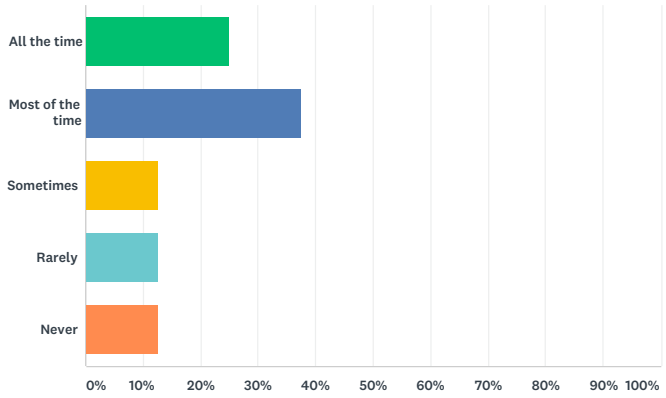
Answered: 8 Skipped: 0



ANSWER CHOICES	RESPONSES	
All the time	12.50%	1
Most of the time	12.50%	1
Sometimes	37.50%	3
Rarely	25.00%	2
Never	12.50%	1
TOTAL		8

Q16 How often do you feel it is difficult to convey an emotion to unfamiliar people?

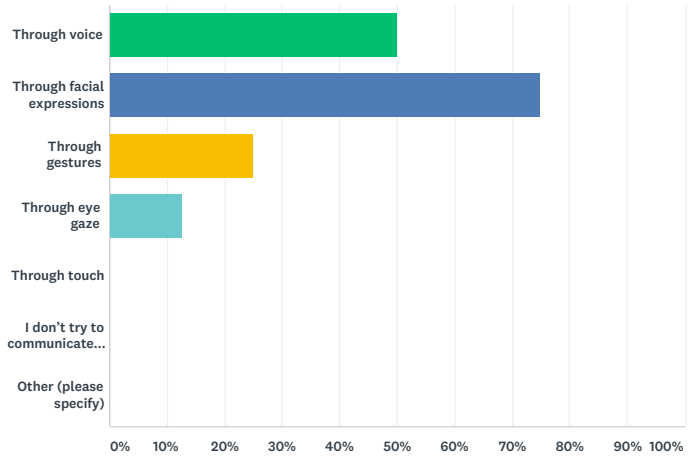
Answered: 8 Skipped: 0



ANSWER CHOICES	RESPONSES	
All the time	25.00%	2
Most of the time	37.50%	3
Sometimes	12.50%	1
Rarely	12.50%	1
Never	12.50%	1
TOTAL		8

Q17 How do you communicate your emotions to familiar people? (please, choose all applicable)

Answered: 8 Skipped: 0



ANSWER CHOICES	RESPONSES	
Through voice	50.00%	4
Through facial expressions	75.00%	6
Through gestures	25.00%	2
Through eye gaze	12.50%	1
Through touch	0.00%	0
I don't try to communicate emotions.	0.00%	0
Other (please specify)	0.00%	0
Total Respondents: 8		

#	OTHER (PLEASE SPECIFY)	DATE
	There are no responses.	

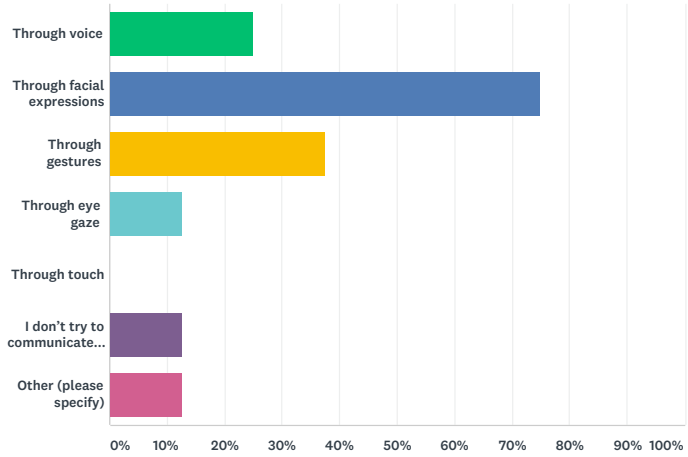
Q18 Please let us know how you use the chosen channel/s from the previous question to communicate emotions (for example: raise the voice tone to communicate anger)

Answered: 8 Skipped: 0

#	RESPONSES	DATE
1	I am mute, I can grunt but this could be associated with disgust. My facial expressions speak much more than words and this is commonly the case for people without any communication barriers.	6/25/2018 10:05 PM
2	lets chat about it	6/22/2018 4:20 PM
3	h	6/21/2018 3:41 PM
4	raise voice	6/15/2018 11:50 PM
5	Use of tone and body language	6/11/2018 8:39 PM
6	Voice tone. Increased body gestures.	5/25/2018 5:01 PM
7	smiling or frowning to show happy and angry. shouting if I need attention or am frustrated.	5/22/2018 12:21 PM
8	smile or do not try to use my communication aid	5/21/2018 2:02 PM

Q19 How do you communicate your emotions to unfamiliar people? (please, choose all applicable)

Answered: 8 Skipped: 0



ANSWER CHOICES	RESPONSES	
Through voice	25.00%	2
Through facial expressions	75.00%	6
Through gestures	37.50%	3
Through eye gaze	12.50%	1
Through touch	0.00%	0
I don't try to communicate emotions	12.50%	1
Other (please specify)	12.50%	1
Total Respondents: 8		

#	OTHER (PLEASE SPECIFY)	DATE
1	Communication aid	6/11/2018 8:39 PM

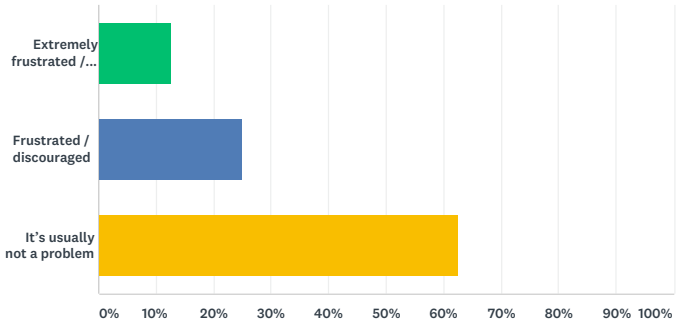
Q20 Please let us know how you use the chosen channel/s from the previous question to communicate emotions (for example: raise the voice tone to communicate anger)

Answered: 8 Skipped: 0

#	RESPONSES	DATE
1	My answer would be the same.	6/25/2018 10:06 PM
2	lets chat about it	6/22/2018 4:21 PM
3	tf	6/21/2018 3:42 PM
4	All three anger	6/15/2018 11:52 PM
5	The LightWriter is difficult to express emotion as it has no tone	6/11/2018 8:40 PM
6	Because of my work as a minister I wouldn't always show my true emotions.	5/25/2018 5:04 PM
7	facial expression, but it is harder	5/22/2018 12:22 PM
8	smile or try to use my communication aid	5/21/2018 2:02 PM

Q21 If the people you are communicating with have difficulty understanding your emotion, how does that make you feel?

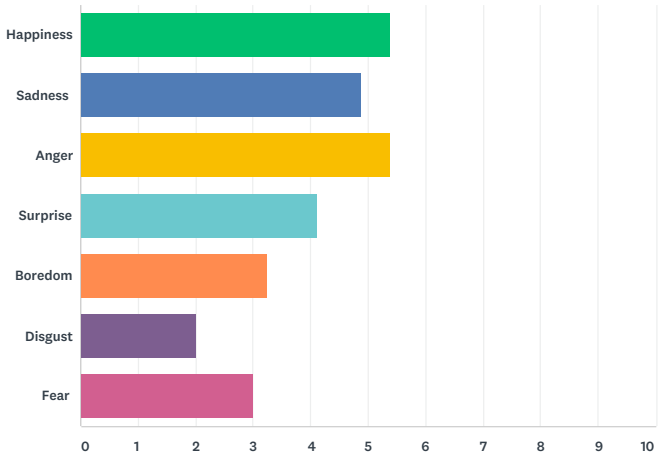
Answered: 8 Skipped: 0



ANSWER CHOICES	RESPONSES	
Extremely frustrated / extremely discouraged	12.50%	1
Frustrated / discouraged	25.00%	2
It's usually not a problem	62.50%	5
TOTAL		8

Q22 For 1 being the most important and 7 being the least important, please number the following emotions according to their importance to you in terms of being able to communicate them successfully.

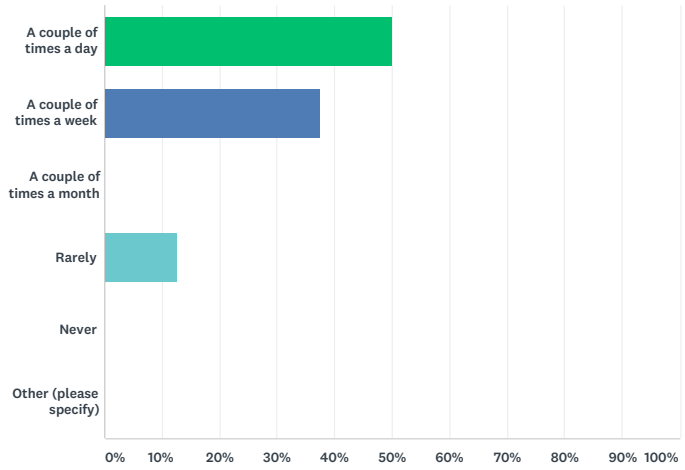
Answered: 8 Skipped: 0



	1	2	3	4	5	6	7	TOTAL	SCORE
Happiness	50.00% 4	12.50% 1	12.50% 1	0.00% 0	12.50% 1	0.00% 0	12.50% 1	8	5.38
Sadness	25.00% 2	25.00% 2	12.50% 1	12.50% 1	0.00% 0	25.00% 2	0.00% 0	8	4.88
Anger	25.00% 2	37.50% 3	12.50% 1	12.50% 1	0.00% 0	12.50% 1	0.00% 0	8	5.38
Surprise	0.00% 0	0.00% 0	37.50% 3	37.50% 3	25.00% 2	0.00% 0	0.00% 0	8	4.13
Boredom	0.00% 0	0.00% 0	12.50% 1	25.00% 2	50.00% 4	0.00% 0	12.50% 1	8	3.25
Disgust	0.00% 0	0.00% 0	12.50% 1	0.00% 0	0.00% 0	50.00% 4	37.50% 3	8	2.00
Fear	0.00% 0	25.00% 2	0.00% 0	12.50% 1	12.50% 1	12.50% 1	37.50% 3	8	3.00

Q23 How often do you find yourself interacting in social situations such as in the church, restaurants, sports, party, etc.

Answered: 8 Skipped: 0

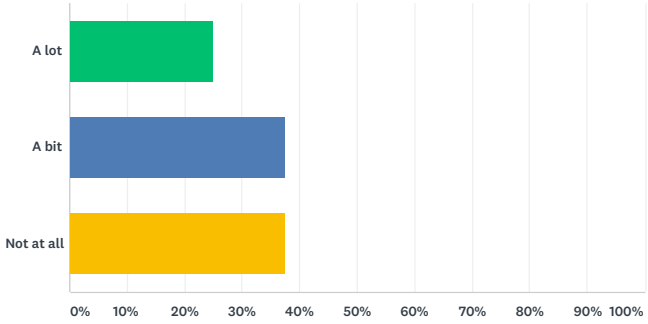


ANSWER CHOICES	RESPONSES	
A couple of times a day	50.00%	4
A couple of times a week	37.50%	3
A couple of times a month	0.00%	0
Rarely	12.50%	1
Never	0.00%	0
Other (please specify)	0.00%	0
TOTAL		8

#	OTHER (PLEASE SPECIFY)	DATE
	There are no responses.	

Q24 How much do you rely on others when you cannot communicate an emotion?

Answered: 8 Skipped: 0



ANSWER CHOICES	RESPONSES	
A lot	25.00%	2
A bit	37.50%	3
Not at all	37.50%	3
TOTAL		8

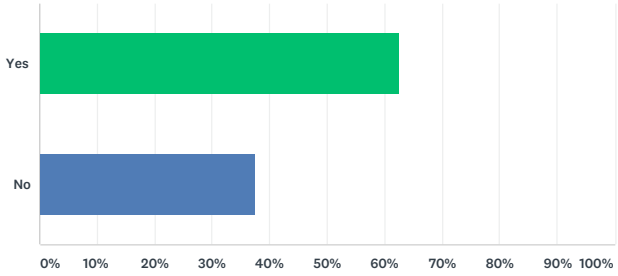
Q25 What is your age?

Answered: 8 Skipped: 0

#	RESPONSES	DATE
1	47	6/25/2018 10:12 PM
2	44	6/22/2018 4:23 PM
3	65	6/21/2018 3:43 PM
4	51	6/16/2018 12:04 AM
5	24	6/11/2018 8:42 PM
6	23	5/25/2018 5:08 PM
7	17	5/22/2018 12:32 PM
8	43	5/21/2018 2:04 PM

Q26 Do you rely on somebody to help you with your communication needs in addition to your communication aid, if you are using one?

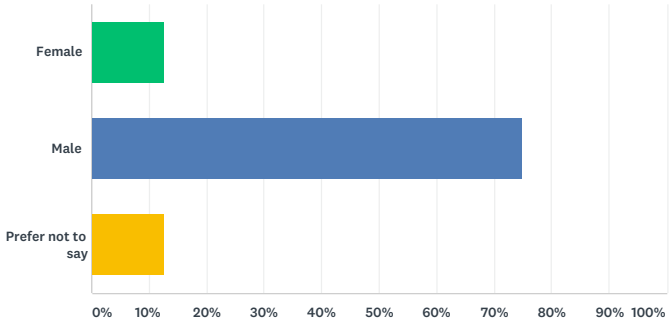
Answered: 8 Skipped: 0



ANSWER CHOICES	RESPONSES	
Yes	62.50%	5
No	37.50%	3
TOTAL		8

Q27 What is your gender?

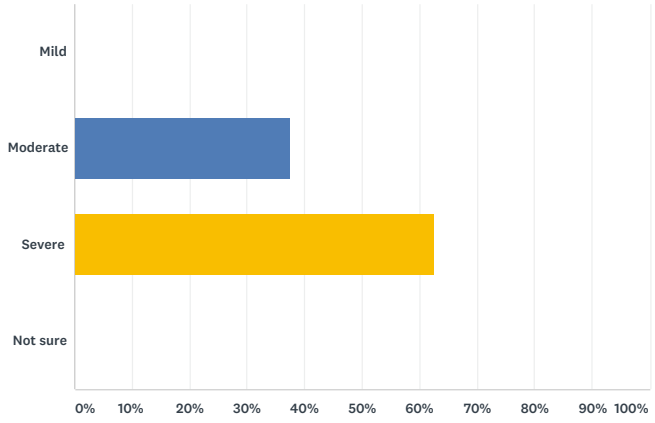
Answered: 8 Skipped: 0



ANSWER CHOICES	RESPONSES	
Female	12.50%	1
Male	75.00%	6
Prefer not to say	12.50%	1
TOTAL		8

Q28 How would you classify your level of speech impairments?

Answered: 8 Skipped: 0



ANSWER CHOICES	RESPONSES	
Mild	0.00%	0
Moderate	37.50%	3
Severe	62.50%	5
Not sure	0.00%	0
TOTAL		8

Appendix B

DEED Sentences

List of Dysarthric Expressed Emotion Database(DEED) sentences for Anger, Disgust, Fear, Happiness, Sadness, Surprise and Neutral emotions

Color Code	
	Common sentences
	Emotion specific
	Generic sentences

DEE sentence number	Emotion	DEED Sentence
1	A	She had your dark suit in greasy wash water all year.
2	A	Don't ask me to carry an oily rag like that.
3	A	Will you tell me why?
4	A	Who authorized the unlimited expense account?
5	A	Destroy every file related to my audits.
6	A	Cory and Trish played tag with beach balls for hours.
7	A	He will allow a rare lie.
8	A	Withdraw all phony accusations at once.
9	A	Right now may not be the best time for business mergers.
10	A	A few years later the dome fell in.
11	D	She had your dark suit in greasy wash water all year.
12	D	Don't ask me to carry an oily rag like that.
13	D	Will you tell me why?

14	D	Please take this dirty table cloth to the cleaners for me.
15	D	The small boy put the worm on the hook.
16	D	Basketball can be an entertaining sport.
17	D	How good is your endurance?
18	D	Barb burned paper and leaves in a big bonfire.
19	D	If the farm is rented, the rent must be paid.
20	D	Laboratory astrophysics.
21	F	She had your dark suit in greasy wash water all year.
22	F	Don't ask me to carry an oily rag like that.
23	F	Will you tell me why?
24	F	Call an ambulance for medical assistance.
25	F	Tornadoes often destroy acres of farm land.
26	F	Straw hats are out of fashion this year.
27	F	That diagram makes sense only after much study.
28	F	Special task forces rescue hostages from kidnappers.
29	F	Will Robin wear a yellow lily?
30	F	The pulsing glow of a cigarette.
31	H	She had your dark suit in greasy wash water all year.
32	H	Don't ask me to carry an oily rag like that.
33	H	Will you tell me why?
34	H	Those musicians harmonize marvelously.
35	H	The eastern coast is a place for pure pleasure and excitement.
36	H	Project development was proceeding too slowly.
37	H	The oasis was a mirage.

38	H	Are your grades higher or lower than Nancy's?
39	H	Serve the coleslaw after I add the oil.
40	H	He would not carry a brief case.
41	S	She had your dark suit in greasy wash water all year.
42	S	Don't ask me to carry an oily rag like that.
43	S	Will you tell me why?
44	S	The prospect of cutting back spending is an unpleasant one for any governor.
45	S	The diagnosis was discouraging; however, he was not overly worried.
46	S	Before Thursday's exam, review every formula.
47	S	They enjoy it when I audition.
48	S	John cleans shellfish for a living.
49	S	He stole a dime from a beggar.
50	S	American newspaper reviewers like to call his plays nihilistic.
51	Su	She had your dark suit in greasy wash water all year.
52	Su	Don't ask me to carry an oily rag like that.
53	Su	Will you tell me why?
54	Su	The carpet cleaners shampooed our oriental rug.
55	Su	His shoulder felt as if it were broken.
56	Su	The viewpoint overlooked the ocean.
57	Su	I'd ride the subway, but I haven't enough change.
58	Su	The clumsy customer spilled some expensive perfume.
59	Su	Grandmother outgrew her upbringing in petticoats.
60	Su	Salvation reconsidered.
61	N	The best way to learn is to solve extra problems.

62	N	Calcium makes bones and teeth strong.
63	N	Greg buys fresh milk each weekday morning.
64	N	He always seemed to have money in his pocket.
65	N	No return address whatsoever.
66	N	She had your dark suit in greasy wash water all year.
67	N	Don't ask me to carry an oily rag like that.
68	N	Will you tell me why?
69	N	Who authorized the unlimited expense account?
70	N	Destroy every file related to my audits.
71	N	Please take this dirty table cloth to the cleaners for me.
72	N	The small boy put the worm on the hook.
73	N	Call an ambulance for medical assistance.
74	N	Tornadoes often destroy acres of farm land.
75	N	The carpet cleaners shampooed our oriental rug.
76	N	His shoulder felt as if it were broken.
77	N	The prospect of cutting back spending is an unpleasant one for any governor.
78	N	The diagnosis was discouraging; however, he was not overly worried.
79	N	Those musicians harmonize marvelously.
80	N	The eastern coast is a place for pure pleasure and excitement.

Appendix C

Speaker-dependent Dysarthric SER Using PCA

Speaker	Emotions	SVM with PCA			
		Accuracy	UAR	UAP	UAF
DS01F	7 emotions	26.25	22.86	24.07	22.99
	4 emotions	50.00	46.25	53.45	46.33
DS02F	7 emotions	62.50	59.29	56.41	56.44
	4 emotions	86.00	85.00	87.68	86.17
DS04F	7 emotions	31.25	27.14	20.55	23.35
	4 emotions	62.00	58.75	53.72	55.38
DS03M	7 emotions	46.25	41.43	40.77	39.64
	4 emotions	60.00	57.50	58.24	57.07

Table C.1 The effect of using PCA on the speaker-dependent categorical classification approach using 7 and 4 emotions.

Speaker	Emotions	SVM with PCA			
		Accuracy	UAR	UAP	UAF
DS01F	7 emotions	42.50	35.83	35.36	34.90
	4 emotions	46.00	40.00	40.00	38.69
DS02F	7 emotions	77.50	75.00	77.19	75.58
	4 emotions	86.00	81.67	89.72	83.79
DS04F	7 emotions	42.50	38.33	38.66	38.34
	4 emotions	54.00	46.67	45.62	45.84
DS03M	7 emotions	60.00	56.67	59.25	56.58
	4 emotions	58.00	53.33	58.33	53.89

Table C.2 The effect of using PCA on the speaker-dependent dimensional classification approach using 7 and 4 emotions.

Appendix D

SER on the Typical Speech part of DEED

Using the same settings in terms of the feature set, eGeMAPS, and classification tasks, categorical and dimensional, used in the dysarthric SER presented in Chapter 7 Section 7.3, a SER on the typical speech part of the DEED, was also developed. The performance of two classification approaches, speaker-dependent and speaker-independent using two classifiers, OVR-SVM and LR, was evaluated. In the speaker-dependent approach, the model is trained and tested using the target speaker's voice characteristics where the average result for all speakers is reported. In the speaker-independent approach, leave-one-speaker-out approach where the model is trained on all typical speakers data except one speaker who was held as a test set. In this experimental study, the performance of the following two classifiers were tested: OVR-SVM with RBF kernel and logistic regession (LR). For SVM, the regularization parameter (C) and the gamma coefficient of the kernel were set to 5 and 0.01, respectively. For logistic regression, the penalty and solver parameters were set to l2 and 'newton-cg', respectively. The rest parameters were set to their default values. All classifiers were trained using Scikit-learn package (Pedregosa et al., 2011). For evaluation, 5-fold cross-validation technique was used for the speaker-dependent approach. The overall performance of the classifiers on speaker-dependent and speaker-independent approaches are determined by the average performance for all test sets.

Emotions	**SVM**				**LR**			
	Accuracy	**UAR**	**UAP**	**UAF**	**Accuracy**	**UAR**	**UAP**	**UAF**
Using 7 emotions	50.45	46.50	46.75	45.78	**55.61**	**50.11**	51.63	50.16
Using 4 emotions	67.71	64.17	65.89	63.77	**67.71**	**64.35**	66.83	64.82

Table D.1 Speaker-dependent classification results on DEED-typical using the categorical approach.

Emotions	SVM				LR			
	Accuracy	**UAR**	**UAP**	**UAF**	**Accuracy**	**UAR**	**UAP**	**UAF**
Using 7 emotions	65.10	62.89	65.21	63.47	**65.28**	**64.52**	65.00	64.53
Using 4 emotions	**68.00**	**64.12**	67.15	63.56	64.00	61.03	62.44	61.16

Table D.2 Speaker-dependent classification results on DEED-typical using the dimensional approach.

Database	Emotions	SVM				LR			
		Accuracy	**UAR**	**UAP**	**UAF**	**Accuracy**	**UAR**	**UAP**	**UAF**
DEED	Using 7 emotions	**47.44**	**45.07**	45.77	44.90	45.95	43.57	43.90	43.60
	Using 4 emotions	59.81	56.37	57.30	56.43	57.71	55.06	55.38	55.19
SAVEE	Using 7 emotions	38.75	35.12	39.81	34.82	38.12	36.19	39.01	36.84
	Using 4 emotions	**60.00**	**56.46**	59.42	56.25	59.00	56.46	60.50	56.94

Table D.3 Speaker-independent classification results on DEED-typical and SAVEE using the categorical approach.

Tables D.1 and D.2 present the classification performance using the speaker-dependent approach for the categorical classification and dimensional classification, respectively. Tables D.3 and D.4 present the classification performance using the speaker-independent approach for the categorical classification and dimensional classification, respectively.

In order to see where this database stands in comparison to previously published emotional databases on typical speech, the same approach was tested on SAVEE database, a British English emotional database (Jackson and Haq, 2011) and the results are reported in Tables D.3 and D.4. As discussed in Chapter 4, DEED and SAVEE database share a lot of similarities such as the language used for the recordings and the stimuli set. The main differences, apart from DEED being a parallel database of dysarthric and typical speech, are: DEED has a total of 21 male and female speakers while SAVEE has 4 male speakers only, speakers in

Database	Emotions	SVM				LR			
		Accuracy	**UAR**	**UAP**	**UAF**	**Accuracy**	**UAR**	**UAP**	**UAF**
DEED	Using 7 emotions	**63.39**	**59.92**	62.38	60.81	60.54	58.29	59.18	58.62
	Using 4 emotions	61.81	58.33	60.59	58.98	61.90	60.08	60.16	60.10
SAVEE	Using 7 emotions	56.04	51.53	57.08	52.28	55.00	54.86	55.45	54.25
	Using 4 emotions	69.33	65.00	73.38	66.83	**70.67**	**68.61**	70.45	69.05

Table D.4 Speaker-independent classification results on DEED-typical and SAVEE using the dimensional approach.

DEED are not actors while the speakers in SAVEE are actors, and the number of utterances in DEED is 80 per speaker which is a subset of the 120 utterances recorded per speaker in SAVEE. As can bee seen, the performance is considered to be comparable. The results also show that SER is a challenging task even when staying within the typical domain.

Appendix E

Speaker-independent Dysarthric SER

Train data	Test data	SVM				LR			
		Accuracy	**UAR**	**UAP**	**UAF**	**Accuracy**	**UAR**	**UAP**	**UAF**
All typical speakers	DS01F	34.00	30.00	30.11	24.89	22.00	27.50	18.80	21.30
	DS02F	72.00	81.25	78.98	73.89	58.00	71.25	68.69	55.88
	DS04F	52.00	60.00	64.68	52.74	50.00	60.00	63.54	49.60
	DS03M	36.00	41.25	46.19	32.75	42.00	51.25	54.91	39.26
Gender-based typical speakers + DS02F + DS04F	DS01F	20.00	23.75	17.61	10.61	26.00	26.25	45.91	22.92
Gender-based typical speaker + DS04F	DS02F	80.00	85.00	83.36	82.22	80.00	83.75	82.53	82.11
Gender-based typical speaker + DS01F + DS04F		76.00	81.25	81.19	78.35	78.00	80.00	80.92	79.35
Gender-based typical speaker + DS02F	DS04F	58.00	58.75	61.12	57.77	54.00	53.75	61.19	54.80
Gender-based typical speaker + DS01F + DS02F		62.00	60.00	61.60	59.00	54.00	56.25	56.13	54.75
All typical speakers + DS02F + DS04F	DS03M	42.00	45.00	49.20	38.94	40.00	48.75	52.87	36.77
All typical speakers + DS01F + DS02F + DS04F		38.00	40.00	44.64	34.49	42.00	51.25	54.88	38.60

Table E.1 Categorical classification results of 4 classes of emotions using eGeMAPS feature set.