Virginia Commonwealth University
VCU Scholars Compass

Theses and Dissertations | Graduate School

2021

USING DEEP LEARNING-BASED FRAMEWORK FOR CHILD SPEECH EMOTION RECOGNITION

Gerald N. Onwujekwe
Virginia Commonwealth University

Follow this and additional works at: https://scholarscompass.vcu.edu/etd

Part of the Business Commons, Communication Commons, Counseling Commons, Engineering Commons, and the Social Work Commons

© The Author

Downloaded from
https://scholarscompass.vcu.edu/etd/6736

This Dissertation is brought to you for free and open access by the Graduate School at VCU Scholars Compass. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of VCU Scholars Compass. For more information, please contact libcompass@vcu.edu.

www.manaraa.com

USING DEEP LEARNING-BASED FRAMEWORK FOR CHILD SPEECH EMOTION RECOGNITION

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

By

Gerald Onwujekwe

Director: Dr. Victoria Yoon
Professor, Information Systems

Department of Information Systems
Virginia Commonwealth University
Richmond, VA

June 2021

©Gerald .N. Onwujekwe 2021 All Rights Reserved

www.manaraa.com

TABLE OF CONTENTS

Front Page

Table of Contents i

Abstract iii

Dissertation Committee iv

Chapter 1 – Introduction 1

1.0 Problem Setting 1

1.1 Research Motivation 4

1.2 Research Issues 8

1.3 Research Questions 10

1.4 Research Contribution and Significance 12

1.5 Organization of the Dissertation 13

Chapter 2 – Background 14

2.1 The Limbic System 14

2.2 Human Speech Production System 16

2.3 The Process of Speech Production 18

2.4 Properties of Emotions 18

Chapter 3 – Literature Review 21

3.1 Theories of Emotion 21

3.1.1 Early Theories 22

3.1.2 Physiological Theories 25

3.1.3 Cognitive Theories 27

3.2 Techniques for Speech Emotion Recognition 29

3.3 Speech Emotion Recognition with Age Detection 37

Chapter 4 – Deep Learning-Based Child Speech Emotion Recognition Framework 40

4.1 Research Objectives 40

4.2 Research Method – Design Science Research in Information Systems 44

www.manaraa.com

4.3 Design Science Research Guidelines 47

4.4 Proposed Design Artifact 50

4.4.1 Ideation Component 51

4.4.2 Codetermination Component 56

4.4.3 Affectation Component 60

4.4.4 Proposed Working Memory Recurrent Network (WMRN) 61

4.5 Instantiation 66

Chapter 5 – Evaluation **71**

5.1 Experiment 1: Identification of the Best Features for Speaker Age Group Recognition 71

5.2 Experiment 2: Emotion Recognition 80

5.3 Experiment 3: Spectrogram Versus Audio Features 93

5.4 Experiment 4: Experiment with the Proposed Framework 97

5.5 Experiment 5: Performance Analysis for the Affectation Component 100

Chapter 6 – Conclusion **106**

6.1 Summary of Research and Research Findings 106

6.2 Answers to research questions 107

6.3 Implications of the Study 109

6.4 Limitations of the Study 113

6.5 Future Work 113

References 115

www.manaraa.com

Abstract

Biological languages of the body through which human emotion can be detected abound including heart rate, facial expressions, movement of the eyelids and dilation of the eyes, body postures, skin conductance, and even the speech we make. Speech emotion recognition research started some three decades ago, and the popular *Interspeech Emotion Challenge* has helped to propagate this research area. However, most speech recognition research is focused on adults and there is very little research on child speech. This dissertation is a description of the development and evaluation of a child speech emotion recognition framework. The higher-level components of the framework are designed to sort and separate speech based on the speaker's age, ensuring that focus is only on speeches made by children. The framework uses Baddeley's Theory of Working Memory to model a Working Memory Recurrent Network that can process and recognize emotions from speech. Baddeley's Theory of Working Memory offers one of the best explanations on how the human brain holds and manipulates temporary information which is very crucial in the development of neural networks that learns effectively.

Experiments were designed and performed to provide answers to the research questions, evaluate the proposed framework, and benchmark the performance of the framework with other methods. Satisfactory results were obtained from the experiments and in many cases, our framework was able to outperform other popular approaches. This study has implications for various applications of child speech emotion recognition such as child abuse detection and child learning robots.

www.manaraa.com

Dissertation Committee

Dr. Victoria Yoon – Chair

Dr. Kweku-Muata Osei-Bryson – Member

Dr. Paul Brooks – Member

Dr. Yeongin Kim – Member

www.manaraa.com

CHAPTER 1

INTRODUCTION

1.0 Problem Setting

Humans show affection and feelings using several ways, such as gesticulation, movement of the body, changes in the tone of voice, and speech. Research on speech recognition and speech emotion recognition has gained considerable ground in recent times especially with the advent of machine learning and deep learning. It has become possible for machines to use artificial intelligence techniques to recognize speech made by humans, translate these speeches from one language to another, identify the person who made the speech, estimate the age of the speaker and even detect the underlying emotions in the speech. Recognizing speech made by humans, performing transcription, and conversion from one written language to another is generally called speech recognition. It is defined as an interdisciplinary subfield of computational linguistics and computer science that develops methodologies and technologies that enable the recognition and translation of spoken language into text by computers ("Speech Recognition," 2021). It is also referred to as automatic speech recognition (ASR) or computer speech recognition. A specific niche of speech recognition that deals with translating speech to lexical texts is called speech to text (STT). Speech recognition has found several applications that are impacting the way we live on a daily basis, ranging from the virtual assistants that have permeated our mobile devices, the web, and our homes to the transcription solutions for meetings and conferences, podcasts, and TV programming. Most mobile platforms have virtual assistants on internet-enabled mobile devices with the capacity to connect to the internet, large databases, and digital repositories to deliver quick information and updates and help people perform simple tasks. Virtual assistants are also deployed on websites to assist in answering questions from visitors and generally enhance their experience. Virtual assistants are also deployed as smart speakers that deliver on-demand content to users via speech commands.

Identifying the person who made the speech is called speaker recognition. (Furui, 2008) defined speaker recognition as the process of automatically recognizing who is speaking by using the speaker-specific information included in speech waves to verify identities being claimed by people accessing systems; that is, it enables access control of various services by voice. Speaker

www.manaraa.com

recognition uses qualities and characteristics in the voice of the speaker as a key distinguishing feature for recognition. Other terms that are often associated with speaker recognition are speaker identification and speaker verification. These terms are often confusing to the uninitiated and we will attempt to clarify them by giving a definition. Speaker identification is a process of matching the speech characteristics from an unknown source with the database of speech characteristics within a system in order to determine the legitimacy of the unknown speaker. The unknown speaker is identified when there is a fit within some certain thresholds permitted by the system, between the utterance of the unknown speaker and a known speech model within the system. Speaker verification on the other hand is the process of confirming or rejecting a specific identity claimed by an unknown speaker. Speaker recognition and its sub-niches have various applications which include access control and biometric authentication, perimeter defense, telephone or voice banking, and voice calling. Speaker recognition solutions can also be adapted in forensics where the speaker recognition system is used to determine if a suspect actually made a speech (also called trace) associated with a crime. This type of system is called forensic speaker recognition (Drygajlo, 2012).

Estimating the age of the speaker from their speech features is referred to as speaker age estimation or speaker age recognition. Applications of speaker age recognition technology include adapting call queue music, age-specific advertisement, and target marketing of specific age groups. The studies of age recognition have been often combined with gender recognition from speech. While gender recognition is a categorical classification with two classes and thus relatively easier to solve, age recognition is continuous and more difficult to solve; hence, age recognition research usually creates groups for specific ranges of age(Chaudhari & Kagalkar, 2015). The typical range used in the grouping is 5 or 10 years. However, since research in this area has continued to mature, systems that could recognize specific age numbers with a good degree of accuracy may not be too far on the horizon.

The ways that humans interact with computing technologies have continued to advance with the advancements in human-computer interactions. The ways that we interact with these technologies in recent years have become even more conversational with the advent of natural language processing and speech recognition systems, such as those available in virtual assistants. However, is it ever possible that these systems will someday understand our dialogical emotions on the go as

المنارة للاستشارات
www.manaraa.com

a human company would do? Detecting the underlying emotions in speech is called speech emotion recognition. Speech emotion recognition (SER) can be defined as the use of computing technologies to analyze human speech and decode the embedded emotion irrespective of lexical and semantic context. Even though this idea has existed for a long time, the first pioneering, seminal work in this research domain is considered to be the study of (Dellaert et al., 1996). The authors use statistical pattern recognition and maximum likelihood Bayes classifier as the key components and to address the challenge of paucity of data, they applied a majority voting technique to obtain emotional classification. They highlight some of the challenges that speech emotion recognition still faces to date, including the need to have a held-out dataset and to train and validate on different datasets. They also noted that emotional features can be speaker-dependent. In other words, different speakers may have different emotional features for a given emotion say anger. There is no model for anger that will fit every speaker because there are nuances to the ways people express anger in their speech. There is also the challenge of appropriately labeling the dataset. Putting emotion labels on a speech could be very subjective as different people could perceive emotions differently. What may sound to one labeler as a neutral emotion, for example, may sound to another labeler as a different emotion. It is very common for labelers of emotional speech to never come to an agreement about the emotional label of a given speech. Some researchers, such as (Al-Talabani et al., 2015), have even argued that speech emotion recognition should not be approached as a classification problem. They claimed that they are able to demonstrate the presence of different spectrums of emotions in a single speech file that is supposed to have only one class. In the FAU AIBO emotional speech corpus, for example, the labelers have to reach a compromise to accept a given emotional label when three out of the five labellers agree (Steidl, 2009). In some cases, the duration of the speech is too short for a human listener to make sense of the emotional content. In the FAU AIBO corpus, for example, the average speech length is 1.7 seconds. Another challenge with speech emotion recognition that has been highlighted in previous research is the limitation of learning a person's emotion from speech. Only emotions that are expressed can be analyzed. In other words, SER system can only attempt to analyze emotional expressions, not emotional feelings. Some emotions are felt which may not be expressed and at other times, people could suppress the real emotion that they are feeling without expressing it. Such internalized emotional feelings remain hidden and cannot be detected with

www.manaraa.com

speech emotion recognition systems.

Speech emotion recognition research has focused mainly on adult speech, and there is little research that has attempted to do SER for children's speech. There are several reasons for the lack of SER for children. Most adult SER use speech that is performed by actors who put on emotions based on a script. Asking children to act emotions from a script would be very challenging to do. Secondly, certain extreme emotions, such as fear or anger, may still be under development in young children and these emotions may not be well expressed in their speech or may be hard to detect even when expressed. Child speech emotion recognition can be applied in child learning robots, child abuse detection, and many other applications.

1.1 Research Motivation

One of the motivations for this research is the increasing importance of human-computer interaction (HCI). HCI is growing so rapidly even though it had its humble beginnings like many fields. According to John Carrol who is one of the founders of the human-computer interaction field, the original focus of the field is on generic computer user behaviors, but it has advanced to include social and affective computing, cognitive science, human factor engineering, and artificial intelligence. Artificial intelligence, which started to gain recognition in the 1950s after the works of the British scientist Alan Turing who published a ground-breaking paper on the potentials of creating a computer that can think and learn, has fast-tracked the rate of development of the HCI field. Practitioners and researchers in human-computer interaction believe that it is a veritable tool towards achieving the utopian goal that the interaction between a computing device and the user should closely match a human-to-human dialogic conversation. While natural language processing has helped machines to understand human language, speech emotion recognition is crucial for machines to understand human emotions – both of which are key components necessary to achieve the idealistic similitude of machine-to-human conversation matching human-to-human conversation.

We have highlighted the challenge of labeling emotional speech data where perceived emotions can sometimes be subjective and labelers struggle to agree on specific labels. This is a problem that many other speech emotion recognition research has reported (Lotfian, 2018) that can cause the labeling process to be lengthy and time-consuming. However, there has not been much that

www.manaraa.com

is said about how this problem can be solved. We think that patterns and groups can be discovered from extensively testing the labeled data and once those patterns are established, they can be used to effectively label the remaining data by using techniques from unsupervised methods such as clustering and density estimation. This technique is very useful when there are massive amounts of data that have been collected but do not have emotional labels, such as data collected from TV programming and podcasts. This technique is also useful when there is a high degree of variability between the classes and patterns. An example would be a situation where the interest is to label data as negative, positive, and neutral emotions. In this scenario, unlabelled examples that show strong clustering in the input space should be mapped to similar features and descriptors in the labeled instances.

While speech emotion recognition research has been ongoing for nearly three decades, most of the efforts have been concentrated on speech emotion recognition for adults while speech emotion recognition for children has been largely neglected. We have highlighted that one of the reasons for this is the difficulty in obtaining emotional speech data for children. This in turn could be as a result of the many layers of protection and regulation when children are involved in research or collecting research data. In the US for example, the basic federal regulation for the protection of humans as research subjects applies to both adults and children but further regulations such as the subpart D of the HHS (Department of Health and Humans Services) regulations are specifically crafted to protect children as research subjects (Hicks, 2019). Furthermore, institutional policies and IRB reviewings tend to be more rigorous when children are involved as research subjects. A future bottleneck to children getting involved in research is that parental consent is usually required. Research, in general, requires every subject to give informed consent and willingness to participate. Children by definition are not able to give informed consent (Hicks, 2019). Parents must provide this informed consent and the child is expected to assent to that consent. Unfortunately, in a majority of cases where the parent cannot identify any direct benefits of the research to their wards, consent is not given. These layers of procedure and bottlenecks that a research must undergo for children to participate in collecting research data will have contributed to the paucity of children's emotional speech data and the lack of child speech emotion recognition research in general.

المنارة للاستشارات
www.manaraa.com

Most of the speech dataset available for adult's speech emotion recognition are acted speech where emotions are fabricated rather than stimulated from natural causes. Due to the abundance of these acted speeches from adult speakers, research that uses them is not lacking. However, asking children to simulate emotions this way is rather a daunting task. Secondly, children of certain ages may not be emotionally mature enough to produce emotional expressions that are suitable for speech emotion recognition research. Henri Wallon's Theory of Early Child Development puts emotional development in the second stage of a child's development after the sensorimotor stage (Veer, 1996). Younger children may not be capable of expressing emotions as strongly and distinctly as an adult could. Despite these challenges with children's emotional data, we believe that more can be done to collect emotional corpus from children that are readily available for research. FAU AIBO children emotional corpus engaged a learning robot that was being manipulated by a human behind the scene, to elicit emotional interactions between the robot and the children. Computer games programmed to provoke certain emotional responses at different stages of the game is also another tool that can be used. Little research has used datasets that features age groups that could be considered as young adults such as (Chen et al., 2012) where age groups were split into ages less than 20 as young adults, 21-60 as adults and above 60 as seniors, however, the dataset is proprietary and not publicly available, and the age group categorized as young adults are not strictly children. Palo et al., 2017 is one of the few studies that has given adequate attention to child speech emotion recognition.

The fact that there are multiple ways that child speech emotion recognition systems could be used in real-life applications is another key motivation. These applications include child abuse detection, child learning robots, child counseling services, educational games, and estimating the level of engagement of role-playing games. Traditionally, the child abuse detection research domain has been spearheaded by efforts from researchers in the medical field especially pediatricians, family physicians, emergency department workers, social workers, and psychologists with school teachers as research participants. However, with the advent of artificial intelligence with its ability to recognize age, gender, and emotions using speech, facial expressions, and other physiological features, there is a burgeoning potential for child speech emotion recognition systems to be applied in detecting cases of child abuse. Robots have continued to play an increasing role in children's learning and education. (Pachidis et al., 2019) report on the growing market for

www.manaraa.com

educational robots and the different roles they play in pedagogy. Learning robots have been associated with the development of cognitive skills for mathematical thinking and logical reasoning. With these increased roles, integrating child speech emotion recognition in robotics learning systems that could understand the emotions of the child and serve up appropriate content would increase levels of engagement and learning for the child. This construct is also valid for educational games where contents could be adapted to the emotions to increase participation and involvement. Generally, the goal of child counselors is to provide emotional and goal support for the child while working through challenges that affect the child's thoughts, feelings, and behaviors. Child speech emotion recognition systems can be integrated into the tools that child counselors use to better support abnormal children needing their services. Child speech emotion recognition systems could also be used in building smart speakers. It has been reported that children enjoy using smart speakers (Lau et al., 2018). They use these devices to ask simple and direct questions and also for jokes and games. The questions arise, should we make smart speakers that are specifically designed for children just as already been done for tablet devices, or should smart speakers be equipped with speech recognition systems to recognize when a child is interacting with the device? This is necessary to serve content and give responses that are appropriate and tailored for children. Furthermore, smart speakers as they are made currently may not able to recognize when a user is frustrated or happy for example. If they do, it is not known from daily usage experience that they do something about it. Although Amazon and Google have been talking about integrating speech emotion recognition into their voice assistants, there has not been any significant changes in daily usage experience. It may well be that the technology is still in the works (Dormehl, 2021). For example, smart speakers should tell when someone is frustrated due to repeated speech commands that were wrongly translated and should offer soothing options. It should be able to hear you cough while giving a speech command and after responding to the command, should throw in, "By the way, would you like to order cough drops with one-hour delivery?" (Scott, 2021).

المنارة للاستشارات
www.manaraa.com

1.2 Research Issues

Developing a solution that detects child speech emotion using artificial intelligence methods requires the analysis of input data. Such input data could come directly from the child, the caregiver, or a third party, such as a child counselor, social worker, or pediatrician. The data could be in the form of textual data or in the form of speech data where the underlying emotional expressions in the speech could be analyzed (Palo et al., 2017). Emotions can be detected from written words, facial expressions, and speech. In this study, we focus on the branch of artificial intelligence research that recognizes emotion from speech, called speech emotion recognition.

One of the challenges of SER is finding the best speech features for accurately recognizing emotions (Al-Talabani, 2015). This is mostly because speech signals carry other information that is unrelated to emotion. Secondly, several thousand-dimensional features can be extracted from a given speech signal. For example, (Sarma et al., 2019) have 3683-dimensional features, (Yu et al., 2017) used 1845 features, and (Sahu et al., 2017) extracted 1582 features for emotion recognition. In this domain, there is an ongoing debate about which features work best for emotion recognition. To further complicate this debate, adult and children's speech characteristics are markedly different. In other words, the speech features that work best in identifying emotions for an adult speaker do not necessarily work well for a child speaker. Identifying and isolating speech features that are most discriminative for a child speaker becomes an additional challenge. The high number of speech features that are often used in SER research presents another issue. The large speech features result in a high dimensional vector, which leads to a large number of input variables in the training data, requiring a dimensionality reduction technique to lower the dimensional representation of the features. Fatima & Erzin (2017) did not total up the speech features that were extracted in their experiment but informed that 50 dimensions were retained after applying principal component analysis (PCA) and that more than 90% of the data were preserved. PCA, Linear discriminant analysis (LDA), and Auto-encoders were used in (Sahu et al., 2017) to reduce the 1582-dimensional features extracted from speech signals. The subsampled features generated from these three dimensionality reduction techniques were fed into a support vector machine (SVM), and the results were compared with the values obtained from using the full features. Using unweighted average recall (UAR) as the evaluation metrics, the full features yielded 57.88%, and the next best performing dimension reduction technique, which

www.manaraa.com

is auto-encoder yielded 53.92%, nearly 8% performance degradation.

PCA and LDA yielded even worse performances of 43.12% and 48.67%, respectively. This result leads to one inevitable conclusion that using dimensionality reduction impairs the performance of the SER model and should be avoided when possible. This is a well-known issue in the research community and there have been attempts to find effective methods that reduce feature dimensions but do not severely impact the performance of the model. Wu et al., (2017) proposed a lobe-dependent convolutional neural network while (Sahu et al., 2017) espoused the use of code vectors. Rather than finding an efficient dimension reduction technique, we explore alternative methods for performing SER. Another issue is that most speech emotion recognition datasets are acted speech that was created in a controlled, unnatural environment, performed by professional actors who can simulate the target emotions. The environment is described as unnatural because they do not express realistic emotions that are encountered in real-life situations and spontaneous conversations. Attempts have been made to collect more natural emotions by using elicitation techniques and by recording calls in call centers; however, SER dataset remains mostly acted speech and they could be problematic in translating models into real-life applications (Lotfian, 2018). To address this problem, data collected in a natural environment is used in training and testing the child speech emotion recognition model. These emotions in the data used in this research are naturally elicited from children who were interacting with a pet robot.

www.manaraa.com

1.3 Research Questions

In this section, we cover the fundamental questions that this research seeks to answer. Seeking answers to research questions requires that the researcher proceed by undertaking a descriptive study or experimental study (Raphael et al., 2011).

The overarching research question of this study is:

How can an IT artifact be developed that could identify children's emotions in their speech?

In addressing the overarching research questions, the following five specific research questions arise:

RQ 1: What speech features provide a good fit for identifying a child speaker from a mix of speech signals containing adult and child speakers?

RQ 2: What speech features provide a good fit for recognizing the emotion of a child speaker

RQ 3: What are the advantages of using spectrogram over audio features in child speech emotion recognition models?

RQ 4: Can an effective child speech emotion recognition framework be developed using deep learning informed by theories of working memory?

RQ 5: How does the proposed framework perform compared to other approaches?

The first research question, RQ1, is very fundamental as it helps to isolate a child speaker from an adult speaker. This is a very complex process as several other characteristics are embedded in a speech signal, making speaker age recognition a very challenging problem (Qawaqneh et al., 2017). Furthermore, datasets for speech emotion recognition mainly feature adult speakers. There are very limited datasets available that feature child speakers for researchers in this field to experiment with. Several speech features have been used for age recognition, which will be covered in the literature review section; however, it is not very definitive which features or combination of features are best for identifying that a speaker is a child rather than an adult.

www.manaraa.com

Other studies such as (Chaudhari & Kagalkar, 2015; Markitantov & Verkholyak, 2019) performed speaker age recognition. It is expected that the results for the experiments conducted to answer this question and other research questions be independent of cultural context. In our experiments, we used a non-English dataset as one of our datasets to further show that the artifact can generalize to non-English speaking cultures.

Research question 2, RQ 2, guides us in the investigation of speech features that yield good performance for child speech emotion recognition. Audio data contains a series of information that could be used for many research purposes. For example, audio data forms the input for speech recognition, speaker recognition, speech emotion recognition, music genre classification, and other research domains. Often, features that are relevant for the research objective are extracted from the dataset and used in conducting experiments. This question allows us to investigate the speech features that are most relevant for child speech emotion recognition from all the potential speech features that could be abstracted from the speech data. Palo et al., (2017) used clustering technique to distinguish features that vary with age. He concludes that his findings might help the computer game and multimedia industries in targeting certain age groups, a focus that is distinct from our objectives in this work. Traditional stream of SER research use features extracted from audio samples to perform emotion recognition. However, spectrograms, which is a visual image representing the energy content or 'loudness' of speech signals as it varies with time over a range of frequencies, have also been used (Ma et al., 2018; Satt et al., 2017). Research question RQ 3 allows us to further investigate the potentials of spectrograms for child speech emotion recognition.

The RQ 4 research question leads to the investigation of how deep learning techniques could be used for child speech emotion recognition without the need for feature extraction or the use of spectrograms. Researchers sometimes refer to this idea as end-to-end solution because the speech data is used directly to train the network model and the need for hand-crafted features with its high dimensional vector that is difficult to train is eliminated. Here, we will develop a novel deep learning architecture that will be trained directly using child speech data for emotion recognition.

In RQ 5 research question, we evaluate the performance of the novel deep learning-based

www.manaraa.com

framework developed for child speech emotion recognition. We compare the performance of this framework with other approaches such as using speech features and other deep learning methods.

1.4 Research Contributions and Significance

Speech emotion recognition research has made some steady progress in the last decade owing to the advent and accessibility of machine learning, deep learning techniques, the assistant high computational power, and massive amounts of data. However, not all of the research efforts translate into applications that could help solve real problems. In this work, we develop a Working Memory Recurrent Network (WMRN), a type of neural network that is based on Baddeley's Theory of Working Memory. This neural network is novel to this study and has not been used in any other study. Speech emotion recognition models are sometimes unable to generalize well when deployed in real-life applications even though they performed well in the lab because they were trained and tested with acted speech. These are speech performed by professional actors where emotions are simulated rather than stimulated from natural causes. This study uses natural speech data to train the deep learning model. The challenge with acted speech is not just with the actors alone but also with the methodology where actors are given scripted sentences and asked to repeat these sentences each time with a given emotion. This goes against many of the foremost theories of emotion such as the cognitive appraisal theory (more in chapter three) where a stimulus must first happen, followed by a cognitive appraisal of the stimulus before physiological reactions and emotions take place. By incorporating real data where emotions are not contrived but naturally happened due to stimulus in the environment, we will be able to overcome this problem and produce a system that is better primed for the real world.

Baddeley's model of working memory has been used extensively in clinical psychology to explain the working memory of the human brain. To the best of our knowledge, this theoretical model has not been used in any other Information Systems (IS) research. This work brings this model into the domain of IS and blazes the trail for other possibilities and potentials of Baddeley's model of working memory in IS research. No other research in IS to be best of our knowledge has this theoretical underpinning. Other studies have done speech emotion recognition as a stand-alone study, while others have done speaker recognition with age classification as the unitary focus.

www.manaraa.com

1.5 Organization of the Dissertation

The rest of this dissertation is organized as follows: Chapter two presents a background discussion on human speech production and the human brain for emotion creation. Chapter three dives into a comprehensive review of the extant literature on speech emotion recognition. It includes discussions on some definitional concepts that form the foundation of speech emotion recognition and builds from thereon to the recent techniques used in SER studies. Theories of emotion that frame the IT artifact are also discussed. Chapter four introduces the deep-learning based child speech emotion recognition framework and offers a detailed discussion of the techniques and methods employed to develop the artifact. The fifth chapter presents the evaluation of the artifact and chapter six concludes the dissertation. The research questions raised guide the evaluation of the artifact.

www.manaraa.com

CHAPTER 2

BACKGROUND

The background section covers knowledge areas that form the foundation of concepts and ideas presented in other chapters. The background concepts that are discussed are the limbic system, human speech production system, the process of speech production, and properties of emotions. The limbic system is the comprehensive name of the portions of the brain that is involved in making and feeling of emotions. The human speech production system highlights the organs and other components of the human anatomy that is involved in making speech and the four procedures that involve these organs. The process of speech production focuses on the models that are available in the literature that explains how speech is produced and the properties of emotions explain how emotions can be represented in continuous 2-dimensions and 3-dimensions.

2.1 The Limbic System

The limbic system is a part of the brain that was first named by Paul Maclean in his 1952 studies. Though the idea of a limbic lobe was originally introduced by the French scientist Paul Broca in 1878 to describe the cortical part of the brain that includes the Para-hippocampal gyrus, James Papez, the father of papez-circuit associated this part of the brain with emotional roles. As such, the limbic system is also called the emotional brain. Many scientists believe that the limbic system is one of the earliest parts of the brain to develop in animals. It is considered the oldest part of the brain and is also known as the paleomammalian brain. The limbic system is responsible for emotion, learning, memory, and other functionalities that are necessary for everyday living and basic survival such as flight or fight responses, stress reactivity, and caregiving. Most studies agree that the limbic system is made of four major components – the hippocampus, the amygdala, the hypothalamus, and the thalamus. These components and other minor components of the limbic system are all interconnected by neural pathways.

www.manaraa.com

The *hippocampus* is a well-researched area of the brain, and it comes in a pair, one on each side of the brain hemisphere. It is curved quite like the curvature of the earlobe and it has also been described to resemble a sea horse. The hippocampus is responsible for understanding our spatial orientation with the environment, helping us understand where we are in the 3-dimensional space with respect to other objects in our environment. The hippocampus is also involved in forming episodic (memory of events) and declarative memory (memory of facts), where both types of memories are later stored away in the cerebral cortex of the brain. The hippocampus is also associated with a procedure called neuroplasticity, an ability of the brain to re-organize itself

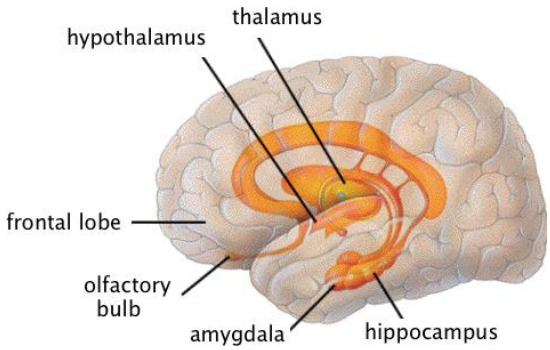


Figure 2: The limbic system (credit: webspace.ship.edu)

and change, form new neural circuits and pathways, rewire itself, and even develop new neurons (Bergland, 2017).

The *amygdala* like the hippocampus is also a pair, one on each side of the hemisphere. The amygdala has been called the emotional center of the brain and is associated with rapid emotional changes (LeDoux, 1992). It is responsible for processing fear conditions and other negative emotions such as anxiety and anger, and also some positive pleasure emotions (Baxter & Murray, 2002). The amygdala also plays a role in the linking of the emotional intensity of events to memory, where events that tend to be emotionally intense, whether negative or positive emotion last longer in the brain cells where they are stored.

The *thalamus* is like a hub through which neural and sensory information passes on their way to the forebrain or cerebral cortex. The thalamus plays a major role in emotional expression mainly due to its many connections to other parts of the limbic system.

www.manaraa.com

The *hypothalamus* works like the thermostat that maintains a certain temperature in the refrigerator which turns on or off as it monitors the fluctuation of the temperature. The hypothalamus is associated with the regulation of anger, pain, pleasure, sexual satisfaction, and hunger. It is one of the busiest parts of the brain as it is connected to several other sensors all over the body. It sends messages to the rest of the body using the autonomic nervous system and is connected to the pituitary gland for the regulation of the release of several hormones into the blood.

2.2 Human Speech Production System

Human speech happens as a result of the complex interaction of certain organs of the body that include the lungs, diaphragm, glottis, pharynx, nose, tongue, and lips. The movement and dilation of these organs allow results in speech production which starts from the lungs and ends on the lips. Four procedures are necessary for speech production – respiration, phonation, resonation, and articulation. *Respiration* involves the movement of air through the nostrils in and out of the lungs as a result of the expansion and contraction of the lungs. The role of respiration in speech production is to produce air and as such, some studies such as Abhang et al., 2016, do not include respiration as part of the speech production procedures. Zveglic, (2014) defines *phonation* as the production of sound from the oscillation of the vocal folds (or 'vocal cords') and resonance of the vocal tract, both of which are essential for normal speech and singing function. It is the process that converts air pressure from the lungs into sounds. Phonation occurs at the larynx also known as the voice box and it is comprised of muscles and cartilages including the epiglottis. The thyroid cartilage, which is one of the cartilages of the larynx, forms Adam's apple that is visible on some people's neck region. The larynx is central to the production of speech. Some studies have suggested that phonation occurs as a continuum that starts from the voiceless to complete glottal closing (Gordon & Ladefoged, 2001; Ladefoged, 1971). During phonation, the vocal fold vibrates and slices the continuous glottal airstream into tiny puffs of air. The rate at which these puffs move into the supraglottal region ranges between 98 cycles per second (hertz) to 262 hertz and this range is called the fundamental frequency (Resonation, 1998).

www.manaraa.com

Different phonations are produced depending on the size of the aperture between the two vocal folds attached to the arytenoid cartilages. They include voiceless, breathy, modal (normal), creaky, and falsetto (Gordon & Ladefoged, 2001; Pietrowicz et al., 2017).

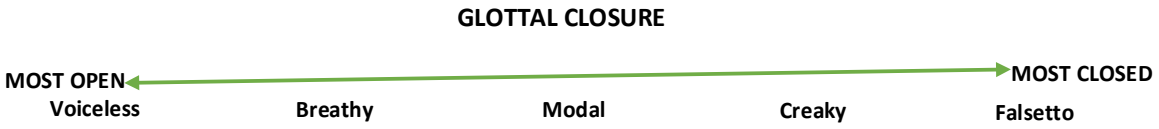


Figure 3: Phonation Continuum

Phonation continuum plays a huge role in speech emotion recognition. Breathy phonation is associated with the emotional expression of intimacy and politeness (Ito, 2002) while creaky phonation is indicative of high emotional states such as happiness or anger. As the closing and opening of the vocal folds resonate the air coming from the glottis, the resonating air in turn causes the components around the larynx to resonate as well. This results in a phenomenon called *resonance***.** All the components that resonate as a result of the resonation of the air column are called resonators. Resonators in the human body include the face, the nasals, the mouth, the skull, the throat, and the chest. The size and shape of the oral cavity, vocal tract, and nasal cavity modifies the sound, resulting in different resonant frequencies. *Articulation* is the movement of anatomical structures to shape the air passing into the vocal tract to produce distinguishable speech sounds. The anatomical parts that are involved in articulation are called articulators. Sometimes, they are divided into movable articulators and immovable articulators. Immovable articulators are those that cannot be moved by the muscles of the body such as the teeth and the hard palate. Moveable articulators include the tongue, the lips, the jaw, and the velum.

www.manaraa.com

2.3 The Process of Speech Production

Different models of speech production are found in the literature. Researchers generally agree on the overall architecture of speech production but tend to disagree on the number of stages involved. Fromkin's five-stage model (Fromkin, 1971) believes that there are five stages of speech production. The model was designed from the results of her extensive experiment in speech errors. Garreth's model (Garrett, 1980) is made up of three stages of semantic level, sentence level, and motor level. In the semantic level, the intended preverbal message is created in the brain, the sentence level is the stage where the words that make up the sentences are selected and the order and grammatical arrangement of the words are performed. The motor level is the stage where the nervous system actuates the organs involved in speech production to produce audible speech signals. The Bock and Levelt model (Bock & Levelt, 1994) is one of the modern frameworks for understanding speech production and it is very similar to Garreth's model. It is divided into four stages of message level, functional level, position level, and phonological encoding level. The message level matches with Garreth's model semantic level as the level where the thoughts to be conveyed is formed. At the functional level, words are selected to be used to convey the thoughts. This process is also referred to as lemmatization. The phonological encoding ensures that sound units are formed and arranged in the proper order and sent to the articulators to produce audible verbal messages.

Speech production is also divided into conceptualization, formation, and articulation states (Levelt, 1993). Conceptualization is the beginning of the speech production process where ideas that we intend to convey as speech first comes to the brain in the form of thought. At this stage, the message is called a preverbal message. As the thoughts of what to speak are formed in the brain, neural information is sent through the nervous system to the organs involved in speech production, which then prepares and assumes the rightful shape and position. The linguistic form of the preverbal ideas is formed in the formation state. The articulators as described previously then articulates the ideas.

2.4 Properties of Emotions

The two-dimensional representation of emotion: The circumplex model of emotion (Russell, 1980) was developed to be used to represent emotion discreetly. Before the development of the

www.manaraa.com

circumplex model, emotion types, and affective states have been viewed as separate from each other independently. The model suggests that affective states are interrelated and represents them as a set of dimensions that vary in a related way. Although the circumplex model was developed in English-speaking western climes, later studies such as (Russell et al., 1989) where the model was used in Greek, Chinese and Polish cultures have demonstrated that the model transcends cultural bounds. The circumplex model shows that affective states are as a result of two neurophysiological systems – the valence system and the arousal system. This idea is in contradistinction to the major theories of emotion that hold that each emotion is distinct and are evoked by distinctive physiological changes and neural patterns. The arousal is represented by a vertical line with the north end of the line indicating high arousal (high activation) and the south end of the line low arousal (low activation). The valence is represented by a horizontal line where the west end of the line is negative valence and the east end is positive valence. The (0,0) axis of the two lines represents neutral valence and neutral arousal.

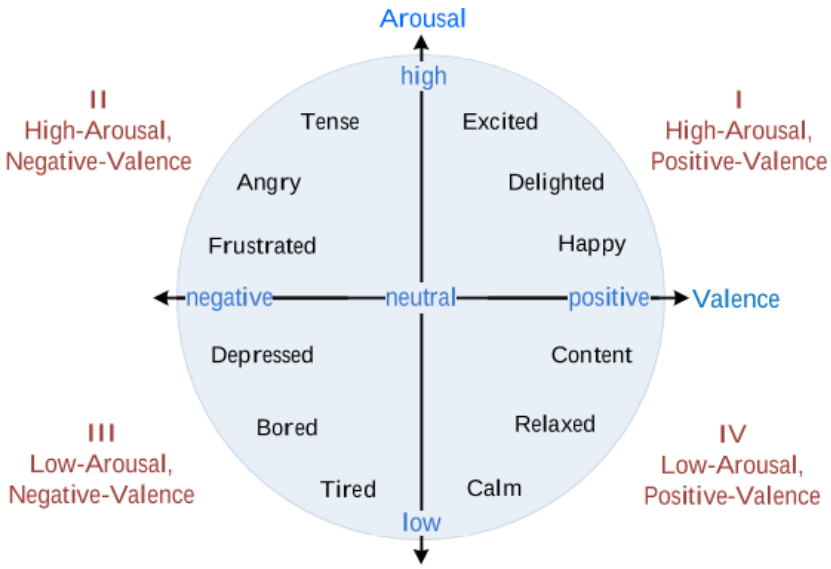


Figure 4: Circumplex Model of Emotion. Credit: (Liu et al., 2018)

Valence is a measure of how positive or negative the emotion is. Some affective states can be extremely positive or extremely negative and others could be mildly positive or mildly negative. Arousal is the intensity of the emotion, ranging from very calm, to exciting and agitating (Kensinger & Schacter, 2007). Emotions with very high valence and arousal tend to be logged into long-term memory than those closer to the neutral line (Kensinger & Corkin, 2003). This concept

www.manaraa.com

also applies to speech emotion recognition where it is easier for deep learning models to detect emotions with high arousals such as ecstasy or rage.

Three-dimensional representation of emotion: PAD

PAD Model: Mehrabian & Russell, (1974) developed the PAD three-dimensional model for representing continuous emotion. PAD is an acronym for Pleasure, Arousal, and Dominance. Pleasure is a measure of how enjoyable or pleasant the emotion is. It differentiates positive and negative emotions. For example, joy is pleasant while anger is not pleasant. The Pleasure-axis has negative and positive values, positive for pleasant emotions, and negative for unpleasant emotion, while zero value is a neutral emotion. Arousal is a measure of excitation or lack thereof of the emotion. It is also a measure of how active or passive the emotion is. For example, ecstasy and joy are both positive emotions but ecstasy has a higher activation or arousal than joy. Dominance is a measure of how domineering or submissive the emotion is. Fear and anger for example are both negative emotions but fear is submissive while anger is domineering, hence anger has more dominance than fear. Huang et al., (2017) used this PAD model to study how online games spread through Word of Mouth. They found that the experience of a gamer will lead to Word of Mouth that induces pleasure, arousal, or dominance.

المنارة للاستشارات
www.manaraa.com

CHAPTER 3

LITERATURE REVIEW

Chapter three examines the literature that has been published on the concepts that are pivotal to this dissertation. They include the theories of emotion, techniques for SER, and speech emotion recognition with age. Under the theories of emotion, we will examine some of the theories that have been put forward by scientists to explain the concept, physiology, and psychology of human emotions. We will cover the early theories, the physiological theories, and the cognitive theories. The physiological and the cognitive theories specifically Plutchik's theory and Lazarus cognitive mediational theory have implications for speech emotion recognition. Next, we review the literature on the techniques that are used for SER and highlight some strengths and limitations. We review studies that added age detection on speech emotion recognition.

3.1 Theories of Emotion

We start the literature review by taking a deep dive into the theories of emotion. The reason we do this is that understanding the theories of emotion deepens our understanding of the different components and features that we use to recognize emotion in speech. It allows us to make a distinction between the features that are grounded in theory, enabling us to view them with the proper lenses. Many theories have been espoused to explain human emotion, all of which we cannot cover, however, we will cover some of the most important and the most influential theories especially as it relates to speech emotion. These theories fall into major categories, starting from those that were proposed many years ago called the Early Theories. Other categories include the Physiological Theories and Cognitive Theories.

www.manaraa.com

3.1.1 Early Theories

James-Lange Theory: This theory was developed by two early psychologists William James and Carl Lange who were working independently at the time but were able to arrive at a similar result that is today known as the James-Lange theory. It is also one of the best known of the many theories of emotion. James-Lange theory was very predominant in the 1900s and generated such controversies that have led to the development of other theories by other researchers who couldn't agree with the James-Lange theory. It suggests that emotions occur as a result of the body's physiological responses to external stimuli. When we receive an external stimulus, the stimulus is going to trigger some physiological changes in the body, and the changes as they happen, make up our emotional experience. According to the James-Lange theory, the physiological changes that happen that are associated with external stimulation are what lead to the emotion that we experience.

One of the criticisms of the James-Lange theory is that physiological changes alone are not sufficient to cause emotions. Experiments were carried out to show that some emotions happen without any accompanying physiological changes (Marañon, 1924). Other critics have also noted that physiological changes that were induced artificially do not always result in emotion and humans with sensory and muscular losses were still able to experience emotion (Söderkvist et al., 2018).

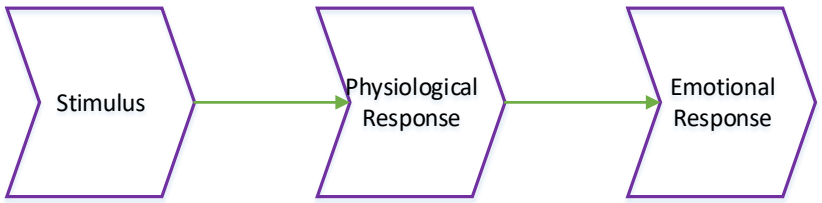


Figure 5: James-Lange Theory

www.manaraa.com

Cannon-Bard Theory: Walter Cannon was one of the chief opposers of the theory put forward by William James and it is only natural that he would develop an alternative theory (Cannon, 1927). Some of the major highlights of Cannon's criticisms of James-Lange theory include;

- Laboratory creation of physiological changes does not always lead to an emotion.
- When the organs involved in these physiological changes are separated by medical surgery from the central nervous system, emotions still occur.
- The emotions at times seem to happen faster than the physiological changes itself and hence the physiological changes cannot be the emotion.

He performed experiments that demonstrated that the brain is the chief organ that is involved in physiological responses and feelings. He was of the position that physiological responses cannot be depended upon to know the type of emotion because different emotions may be formed by one type of physiological response. For example, a heart that is racing and thumping could be out of fear or out of anger. In this theory, Walter Cannon and his student Philip Bard states that both the physiological changes and the emotion happen simultaneously in response to an external stimulus. The feeling of emotion involves two essential parts of the nervous system which are the autonomic nervous system and the cortex, the autonomic nervous system being responsible for the arousal, and the cortex, the emotion production. An external excitement stimulates neuroreceptors that relay impulses to the thalamus of the brain. The thalamus releases neural impulses to the cortex that causes the conscious experience of emotions and also sends neural impulses to the hypothalamus that causes physiological changes. We experience the emotion and the physiological changes simultaneously. Because of the central role that the thalamus plays in this theory of emotion, it is also called the thalamic theory of emotion (Tee, 2020).

www.manaraa.com

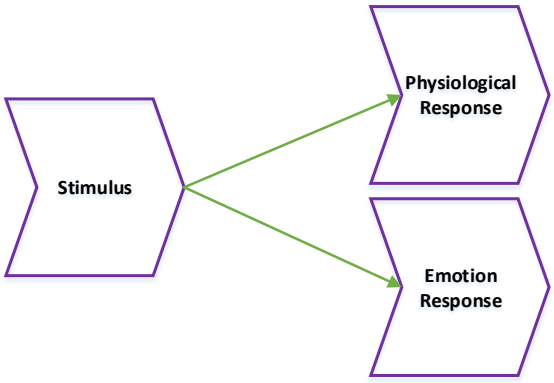


Figure 6: Cannon-Bard Theory

Papez's Theory: James Papez's theory of emotion was the next to challenge the status quo that was set by the dominant preceding theory of Cannon-Bard. This theory, like that of Cannon-Bard, also emphasized certain aspects of the brain as the chief controlling factor in experiencing emotion. In his work (Papez, 1937), he described the Papez circuit as a neural pathway in the brain that is involved in experiencing emotion. He believed that emotional expression (behavior) and emotional experience (feeling) can be dissociated from each other and that both are controlled by different aspects of the brain that are interconnected. Emotional expression is controlled by the hypothalamus and emotional experience by the cortex. After his work on the Papez circuit, several researchers have experimented on this area of the brain, exploring its structure and functions, leading to its new name as the limbic system (Chow et al., 2018). The major contribution of Papez's theory is its differentiation of emotional feeling and emotional expression. It is the expressed emotion that comes out through speech and other body languages that can be detected using machine learning and deep learning. Unexpressed emotions cannot be detected.

www.manaraa.com

3.1.2 Physiological Theories

Lindsley's Theory: According to this theory, emotions are expressed in three distinct channels; cortical channels, visceral channels, and somatomotor channels. The whole concept of emotion is a mechanism of arousal and activation which involves the reticular, thalamic, and cortical systems. Lindsley agrees with Papez that the limbic system controls emotional expressions and emotional behaviors. Lindsley's research heavily involved the use of electroencephalogram (EEG) techniques hence, his thesis is leaned towards neural arousal in general that goes beyond the concept of emotion. He believes that the arousal of different parts of the brain results in different emotional arousal. For example, worry and anxiety show emotional arousal at the cortical level while sweating and tears show encephalic and brain stem arousal (Lindsley, 1970).

Maclean's Theory: His theory lays special emphasis on the importance of the amygdala and the hippocampus on the emotional experience (feeling), though he agrees with Papez's theory that all the elements of the limbic system are involved in emotion. The agreement of Maclean's theory and Papez's theory has led to an amalgamation that is sometimes referred to as the Maclean-Papez limbic theory (Bhattacharyya, 2017). The limbic system is located right above the brainstem and under the brain cortex. Even though the limbic system is a really small part of the brain in terms of size, it is responsible for very important roles that make life meaningful and enjoyable. A concept of the brain called the triune brain that divides the brain into three parts calls the limbic system the emotional brain (Pogliano, 2017). Though Maclean's earlier studies underscore the limbic system as a whole as being responsible for emotions, his latter studies began to differentiate the responsibilities of the different components of the limbic system. He noted that a disorder of the hippocampus could cause paranoia, mood swings, and distortion of perception. Even though most of MacLean's thesis emphasizes the physiological constructs of emotion, that is the functions, mechanisms, and activities of the parts of the brain that are involved in emotion, he does touch on the experiential and the subjective part of emotion, arguing that the study of the subjective side of emotion is necessary to bring about a holistic understanding of the phenomenon.

www.manaraa.com

Plutchik's Theory: This theory is psycho-evolutionary in nature and is theorized to include not just humans but other forms of life such as animals (Plutchik, 1980). it provides very well-known emotional classification approaches that have influenced the field of speech emotion recognition. He argues that there are eight (8) basic emotions that are primordial in biological life and they are anger, fear, sadness, disgust, surprise, anticipation, trust, and joy. These are the primary emotions and all other forms of emotions are a derivative, mixture, combinations, and compounds of these eight. He further theorized that emotions can be conceptualized in pairs of opposites, that emotions vary in degree of similarity, and that emotions also vary in intensity and levels of arousal. His theory helped to discretize and digitize emotional representation that has allowed machines to understand emotions using artificial intelligence techniques. He defines emotions as, 'an infrared, complex sequence of reactions, including cognitive evaluation, subjective change, and autonomic and neural arousal impulses to action, the resulting behavior affecting the precipitating stimulus'. His theory represents emotions as having three dimensions of intensity, similarity, and polarity. For example, ecstasy and joy have the same similarity and polarity but are only different in intensity. So also, is anger and rage. Ecstasy and grief are opposites while happiness and joy are similar. Emotional reactions according to Plutchik serve to communicate intent and motivation between members of the same biological group. He further developed the wheel of emotions that helps to give people a clearer understanding of the emotions they are feeling. The wheel of emotion (Abbasi & Beltiukov, 2019) is made up of the eight primary emotions set in such a way that any emotion is set opposite of the reverse or the contrast of that emotion. For example, joy is set opposite to sadness and fear opposite to anger. These primary emotions are represented with various colors, while a combination of two primary emotions has no color representation. The wheel of emotions also uses shades of colors to represent the intensity of the emotion. Lighter shades represent lower intensity and darker shades represent high intensity. For example, serenity which is a lower intensity of joy has a lighter shade for the color of joy, and ecstasy which is a higher intensity of joy has a darker shade.

www.manaraa.com

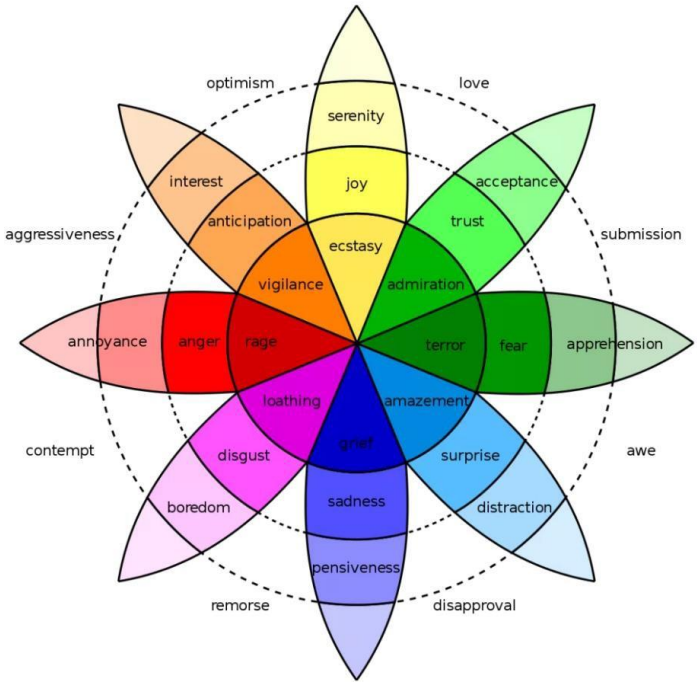


Figure 7: Plutchik's Wheel of Emotions (credit: wikiversity.org)

Plutchik's theory of emotions also covered the concept of dyads which are emotions that presents a combination of two other emotions. The four groups of dyads are the primary dyad, the secondary dyad, the tertiary dyad, and the opposite emotions. The primary dyads are made of emotions that are one petal separated from each other, for example, Optimism = Joy + Anticipation. Secondary dyads are two petals apart from each other, for example, Hope = Trust + Anticipation. Tertiary dyads are separated by three petals, for example, Delight = Joy + Surprise. Finally, opposite emotions are four petals apart, for example, Trust ≠ Disgust. The idea that emotions can be derived brings the subject into the domain of discreet and mathematical concepts where derivatives are tenable. Derivatives are applied in speech features that are used for speech emotion recognition.

3.1.3 Cognitive Theories

These are theories of emotion that involve cognition in the process of emotion. These theories hold that the emotion we experience is determined by our cognitive appraisal of the external stimulus. They argue that there are intentional elements to emotions.

www.manaraa.com

Schachter Theory: Schachter holds the view that emotions are determined by physiological arousal and cognitive factors (Reisenzein, 1983). The essence of the theory is that an external stimulus leads to the arousal of our nervous system. Our cognitions now give interpretation and direction to this physiological arousal based on the situation that brought about the physiological arousal and the way we perceive things and our past experiences. Schachter performed an experiment where he injected his subject with epinephrine – adrenaline that stimulated physiological changes in the subject such as an increase in the rate of heartbeat and muscular blood flow. This experiment did not lead the subjects to experience any emotion, leading to the position that physiological changes do not cause emotional experience but could support it.

Even though Schachter's theory recognizes the role of physiological arousal in emotion, unlike the James-Lange theory, they argue that physiological arousal alone cannot be responsible for emotional responses as the same physiological arousal could lead to a variety of emotions. This theory focuses on the interaction of physiological arousal and cognitive evaluation and appraisal of the arousal; hence it is often referred to as the two-factor theory of emotion (Dror, 2017). In other words, when we feel arousal as a result of an external stimulus, we must cognitively identify and label this arousal in order to feel the emotion.

The works of Schachter have led to several other studies in this direction but it has also attracted some notable criticism. A simple anecdotal criticism is that there are instances when emotions are experienced before we think about them. This suggests that not all emotional experiences can be explained using Schachter's theory of emotion. Leventhal, (1974) sees Schachter's works as an effort to combine cognition and emotion, where cognition is allowed to interpret emotional stimulus, detect arousal, and also label the emotion.

Lazarus Cognitive Mediational Theory: Richard Lazarus' works focused on the interaction of stress, cognition, and emotion. Just like the Schachter theory, there is a major focus on the role of appraisal in this theory where appraisal is the ability of the mind to consciously assess and interpret situations. However, his theory also touched on the interplay of culture and emotions where he argues that culture can affect emotion;

- In the way and manner that we receive emotional stimulus.

www.manaraa.com

- By directly altering emotional expression.
- By determining social relationships and judgments.
- By highly ritualized behavior (Strongman, 2003).

To use an example to illustrate this theory, when a gunshot is fired, this external stimulus could result in increased heart rate, shock, adrenaline surge, muscular tension, and other physiological changes. For this individual, the emotional reaction would be fright and panic and he would move away from the direction of the gunshot. For another person with a different culture and personal experiences, the gunshot stimulus in their experience means that someone is in trouble and requires help. Physiologically, he will have a heightened sense of alertness and readiness and instead of moving away from the sound of the gunshot, would move towards it. Instead of feeling frightened and panicky, he feels sympathy and concern for someone who could be in trouble. In Lazarus' theory, thoughtful evaluation and cognitive appraisal precede the emotion and the physiological changes. The stimulus evokes cognition and this cognition is moderated by cultural and experiential factors. The cognition evaluates this stimulus and the result of the evaluation determines our physiological changes and the emotions we feel. As a result of this conscious mental evaluation of the stimulus, this theory and other related theories are called cognitive appraisal theory (Moors et al., 2013).

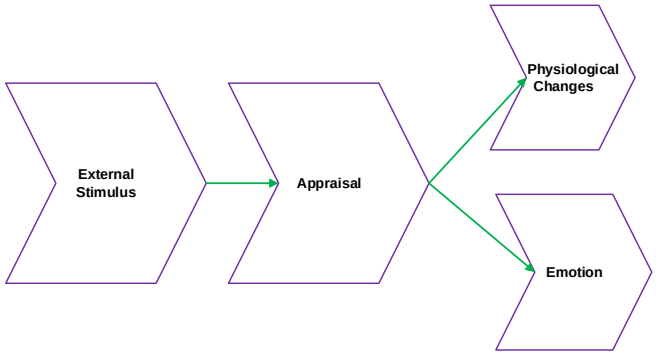


Figure 8: Lazarus cognitive mediational theory

3.2 Techniques for Speech Emotion Recognition

Research in the area of SER has proliferated in recent times. Despite this, SER problems are far from being solved. Learning emotion from images such as facial emotion recognition is less intricate than learning emotion from speech.

www.manaraa.com

Many facial features and pointers could be used to develop a recognition model, however, in the case of speech, the audio signal contains other information and it is very challenging to isolate the best features that could help to accurately recognize emotions. This is one of the greatest challenges of SER studies. Studies that incorporate other modes of learning human emotion such as body language, text scripts (lexical features), facial images with speech data in other to recognize emotions are identified in the literature as multimodal studies (Caridakis et al., 2007; Chae et al., 2018; Fatima & Erzin, 2017; C. W. Lee et al., 2018; Sebastian & Pierucci, 2019).

In designing an SER study, a decision will have to be made on the length of speech that will constitute a unit of input data that will be fed into the recognition model. Based on this, there are segment level, utterance level, sentence level, and dyadic SER studies. In studies (Gideon et al., 2019; J. Han et al., 2018; K. Han et al., 2014; Mao et al., 2019; Zhang et al., 2018) that used segment level analysis, input speech signals are split up into smaller segments with a fixed frame and then shifting the split in time order. The typical split used in most studies is a 25-millisecond frame window with a 10-millisecond shift (hop size) (Khorram et al., 2017; Latif et al., 2019; Le et al., 2017; J.-L. Li & Lee, 2018; Neumann & Vu, 2017; Parry et al., 2019). This is also called the fixed frame-size and rate technique (FFSR). Though this method has led to very successful SER models, it has been criticized for causing overlaps for periodic parts of speech such as vowels, leading to insertion errors, and overall degradation of performance (B. Lee & Cho, 2016). The alternative solution to the challenges of using FFSR is using the variable frame rate (VFR) (Huang et al., 2018; B. Lee & Cho, 2016; Ma et al., 2018; Tan & Lindberg, 2010). The VFR can analyze speech frames at a shorter interval such as 2-3miliseconds which in effect could improve the performance of the recognition system but introduces computational burdens due to repetitive calculations. The key challenge with segment level analysis is that the emotional state in all segments may not be true for that of the entire utterance, in other words, all segments within a given utterance may not have the emotional ground truth label that is assigned to that utterance (J. Lee & Tashev, 2015). As a result, some studies use some type of aggregation or voting method (Chatziagapi et al., 2019; J. Lee & Tashev, 2015; Mao et al., 2019) to scale up the segment level classification into utterance level emotion recognition. Furthermore, detailed annotation of speech emotion datasets is a very

www.manaraa.com

laborious and ambiguous work and most datasets only have utterance level labels and not segment level labels.

Other studies omit splitting the speech input data into segments and prefer to analyze them at the level of utterance (Cho et al., 2018; Fayek et al., 2017; K. Han et al., 2014; Mao et al., 2019; Sebastian & Pierucci, 2019; B. Wang et al., 2019). For the purposes of speech emotion analysis, an utterance is a sequence of words (Mary, 2019; Rao & Koolagudi, 2013a). Utterances may be created from long continuous speech using word boundaries or syllable boundaries. However, most speech emotion datasets have audio data already split and annotated at the utterance level. Utterance level analysis extracts high-level features at the utterance level and does not account for temporal frame-by-frame feature variations that are present in the data. Three approaches are used for utterance level analysis in literature. One approach (Sebastian & Pierucci, 2019) performs feature extraction using the utterance level audio data and feeds this extracted feature vector into a machine learning classifier to obtain categorical emotion classification. Another approach (Z.-Q. Wang & Tashev, 2017) feeds the raw utterance level audio data into a deep learning system for SER. A third approach is to perform segment level analysis and aggregate the result to perform utterance level emotion classification. Using the first approach, Sebastian & Pierucci, (2019) utilized the openSMILE toolkit to extract low-level features (LLDs) such as pitch, Mel-spectra, loudness, and Mel frequency cepstral coefficients (MFCC). The high-level statistical derivatives of the LLDs such as min, max, mean and standard deviation were also calculated and these features put together, form the input to the convolutional neural network classifier. Using the second approach, Z.-Q. Wang & Tashev, (2017) fed all the frames in a given utterance into the deep learning network, and a pooling operation is used to downsample the dimension and obtain an utterance level representation. Several types of pooling operations could be used such as min-pooling, max-pooling, or mean-pooling though the authors report that mean-pooling yielded the best result. The mean-pooled vector representation is fed into the softmax layer of the deep learning network or a separate kernel extreme learning machine (ELM) to obtain categorical emotion classification.

Sentence level (Jeon et al., 2011; Rao & Koolagudi, 2013a) SER study is not as prominent as the segment and utterance level variants. Sentence level SER attempts to extract the speech emotion

www.manaraa.com

using the sentence as the basic unit of input. It is a technique that appears to be fading away and is rarely used in recent studies. Dyadic SER (Fatima & Erzin, 2017; Hazarika et al., 2018; J.-L. Li & Lee, 2018; B. Wang et al., 2019; J. Zhao et al., 2019) or dialog level SER is the method of analyzing the emotion that is present when two individuals or groups are engaged in a dialogic conversation. This is a basic one-on-one, face-to-face human verbal interaction that provides a basic means of exchanging ideas, ideals, information, and fostering mutual understanding. There is mutual interdependency that emerges almost automatically at the inception of dyadic interaction. This mutual interdependency is evident in the dialogists' internal state such as their emotions and perceptions and also their expressions such as speech, facial expressions, and body language. While speech recognition from one speaker has useful applications in patient care, automated call assistants, and expert systems, other types of applications require the understanding of the transactional dynamics of emotion in human-to-human communication. Dyadic studies are tasked with understanding the emotion of each speaker in the dialogue and also tracking the emotion contagion phenomenon (Barsade, 2002) where one speaker can influence the emotion of the other and bring it into resonance with his own emotion and vice versa, influencing and affecting each other's emotional states and behaviors. Dyadic studies sometimes incorporate other modes of emotion recognition such as body motion (Fatima & Erzin, 2017) and facial expression (Caridakis et al., 2007) to achieve a more robust recognition model.

SER research could also be grouped in the way they handle the speech data which is typically distributed as audio files in standard SER datasets. Some studies utilize the raw audio, others convert the raw audio into spectrograms while others handcraft speech emotion features from the raw audio. Using raw audio data as input to the SER model is also called an end-to-end SER. Studies that use this method (K. Han et al., 2014; Latif et al., 2019; Sarma et al., 2019, 2018; Tang et al., 2018; Tarantino et al., 2019; Trigeorgis et al., 2016) split the raw speech data into smaller sizes and fed them as input to a deep learning model, usually a convolutional neural network (CNN) (Tang et al., 2018), a recurrent neural network or a combination of both (Latif et al., 2019; Sarma et al., 2019; Trigeorgis et al., 2016). Latif et al., (2019) built a two-stage recognition model that utilizes a CNN at the first stage where the convolution layers are used as

www.manaraa.com

a filterbank to extract spectral features from the raw speech. The optimal number of parallel convolutional layers was determined empirically as the number is very significant in the overall performance of the solution. Max pooling was shown to perform better than average pooling and l_2 pooling and was adopted in the network. The second stage is the emotion recognizer that uses a combination of a deep neural network, CNN, and Long short-term memory (LSTM), a variation of the recurrent neural network to classify the emotions.

Spectrogram is a pictorial view of sound which includes speech, music, and other sounds presented in a frequency-time and/or amplitude-time graph. They can visually reveal noise interferences present in audio data and allow for these interferences to be isolated and removed. Spectrograms retain the emotional features of speech (Yenigalla et al., 2018) and some studies (Guo et al., 2018; Y. Li et al., 2019; Luo et al., 2018; Satt et al., 2017; Tang et al., 2018; Yenigalla et al., 2018; Zhang et al., 2018; Zheng et al., 2015) prefer to use the spectrogram rather than raw audio or speech features. The standard method for generating spectrograms from speech signals is by applying Short Term Fourier Transform (STFT) to the audio frames (Yenigalla et al., 2018; Zhang et al., 2018). The obtained frequency domain signal is windowed using the Hanning window to improve frequency resolution and reduce spectral leakages.

Emotional features can be manually extracted from the spectrogram and fed into the SER systems such as in (Guo et al., 2018) while at other times such as in (Y. Li et al., 2019; Yenigalla et al., 2018; Zhang et al., 2018), the spectrograms are fed directly into the system. The size of each spectrogram segment has been found to significantly affect the performance of the recognition model especially in the scenario where the spectrograms are directly fed into the system (Luo et al., 2018). The network can learn different contextual emotional features based on the spectrogram size. Typically, global features are learned from larger spectrograms while local features from smaller-sized spectrograms.

Manually crafting emotional characteristics from speech signals is called feature extraction. This is the most dominant method of SER until recently when deep learning methods are beginning

www.manaraa.com

to show superior performance. Some studies perform feature extraction and feed the extracted feature vector into a deep learning network for categorical emotion classification. Features that can be extracted and used to detect emotion are grouped into (i) excitation source features (ii) spectral or cepstral features, and (iii) prosodic features (Gamage et al., 2017; Rao & Koolagudi, 2013b).

Excitation Source Features: During the mechanism of speech production, the subglottal airstream coming out of the lungs vibrates the vocal fold, and the vocal fold cuts through this continuous airstream, producing a periodic excitation of the vocal tract. The articulators in the supraglottal region eventually help to produce speech sounds. Researchers have shown that the vocal tract excitation at the glottal region possesses emotional characteristics and such characteristics can be used for SER studies (Choudhury et al., 2018; Gangamohan et al., 2014; Kadiri et al., 2020; Koolagudi et al., 2012; Prasanna & Govind, 2010; Pravena & Govind, 2017; Rao & Koolagudi, 2013b). Kadiri et al., (2020) used an aggregate of features to develop an SER that is based on excitation source features only. The method that they used to extract the features include zero frequency filtering, linear prediction residual technique, and short-term Fourier transform, and the extracted features include instantaneous fundamental frequency, strength of excitation, energy of excitation, and ratio analysis of high and low-frequency spectral energies. Zero frequency filtering process involves passing speech signals through two zero frequency resonators used to dampen all other speech frequencies except zero frequencies. For the output signal that passed through the two resonators, the average pitch period is computed and local mean subtracted to achieve the zero-frequency filtered signal (Murty et al., 2009). A global template was built based on these features and their statistical derivations as obtained from three emotional states (anger, happiness, and sadness) and neutral state. To test new speech data for emotion, the system will extract and normalize features from the speech signal and match it to the global template. The biggest drawback of this template matching solution is that the neutral state of a speaker (or new speech signal for testing) must be obtained and registered in the system. This neutral state will be compared with the neutral state in the global template and then computation is performed to adjust for the dynamic range of the new speaker. The system is then ready to recognize other emotional state of the new speaker. Choudhury et al., (2018) on the other hand, did not base their SER system on excitation features only but added spectral features.

www.manaraa.com

They combined the zero-frequency filtered (ZFF) signal along with the linear prediction (LP) residual signal as their excitation source signal. From the ZFF signal, they extracted epoch locations, strength of epoch, slope of strength of epoch, and instantaneous frequency. These features and the cepstral feature are combined into a feature vector that is fed into a random forest (RF) and sequential minimal optimization (SMO) classifiers for emotion recognition.

Spectral or Cepstral Features: Most speech signals are captured in the time domain. However, they can be converted to frequency domain using a variant of Fourier transform called short-term Fourier transform. These signals are typically chopped into a small segment of 20-30miliseconds each. The features that are obtained from these small segmented frequency domain representations of speech signals are called spectral features. These spectral features can form the basis of emotion recognition in some SER studies (Choudhury et al., 2018; Daneshfar et al., 2020; Demircan & Kahramanli, 2018; Pohjalainen et al., 2016; Rao & Koolagudi, 2013b, 2013b; Sun et al., 2015; Yadav et al., 2018; Yogesh et al., 2017). Important spectral features have been proposed over the years and some of the prominent ones used in SER studies include energy (X. Li & Akagi, 2018), pitch (Verma & Mukhopadhyay, 2016), formant frequency, Mel-frequency cepstral coefficients (MFCC) (Kerkeni et al., 2019; X. Li & Akagi, 2018; Syed et al., 2018), modulation spectral features (MSF) (Kerkeni et al., 2019; X. Li & Akagi, 2018), linear prediction cepstral coefficients (LPCC) (Yeh et al., 2011), perceptual linear prediction coefficients (PLPC) (Syed et al., 2018), discrete wavelet transform (DWT) (Campo et al., 2016; Lalitha et al., 2015). The MFCC is one of the most prominent spectral features used in SER studies because it reduces the impact of noise in the speech signal better when compared to other methods (Sun et al., 2015). Daneshfar et al., (2020) developed an SER system using the spectral features in combination with prosodic features and fed the dimensionality-reduced feature vector into a gaussian elliptical basis function (GEBF) neural network for classification. Specifically, MFCC, PLPC, and perceptual minimum variance distortionless response (PMVDR) coefficients were extracted using different techniques, resulting in a multi-dimensional feature vector. The multidimensionality was reduced using the point mass function weighted quantum-based

www.manaraa.com

particle swarm optimization (pQPSO) method. The GEBF neural network takes the dimensionality-reduced feature vector as input and provides categorical emotional classes. The main challenge with this technique is dimension reduction where some relevant data containing emotional information will be lost. Furthermore, the authors report that the system suffers from high computation cost and high memory requirements due to an iterative step involved in the dimensionality reduction stage.

Prosodic Features: Prosodic features are features that are obtained over a longer speech signal such as sentences, phrases, words, and syllables. The focus of prosodic features is the overall quality and characteristics of the speech signal and in the literature, sometimes they are called long-term features. Words and syllable boundaries are identified in a speech signal using vowel onset points (VOP) (Vuppala et al., 2012). Prosodic features that have been used for SER include stress, loudness, intonation, rhythm, pitch, amplitude, duration, zero-crossing rate (ZCR), energy, and their statistical derivatives (Daneshfar et al., 2020; Koolagudi et al., 2011; X. Li & Akagi, 2018; Palo & Mohanty, 2020; Y. Wang et al., 2008; Zhou et al., 2009; Zvarevashe & Olugbara, 2020). Statistical derivatives are measures such as mean, min, max, variance, and standard deviation. Other prosodic features such as shimmer, gitter, and fundamental frequency are useful for age detection of the speaker in SER studies (Qawaqneh et al., 2017; Trigeorgis et al., 2016). Zvarevashe & Olugbara, (2020) used *jaudio* tool to extract the prosodic features of energy, fundamental frequency, and zero-crossing rate from the RAVDESS and SAVEE audio database for emotion recognition. Based on the results obtained from previous studies, they proposed a hybrid acoustic features (HAF) that consist of a combination of prosodic and spectral features that have yielded good results in published studies. The HAF features were benchmarked with MFCC features. Bagging ensemble learning was used to build a support vector machine (SVM), multilayer perceptron (MLP), and Random Forest (RF) classifiers for emotion classification. The downside of this methodology is that ensemble learning requires the concatenation of two or more machine learning techniques which result in high computational cost and training times, a drawback that was acknowledged by the authors. Deep learning architectures in general tend to overcome this challenge.

www.manaraa.com

3.3 Speech Emotion Recognition with Age Detection:

We examine SER studies that also performed age detection (Goel et al., 2018; Qawaqneh et al., 2017; Verma & Mukhopadhyay, 2016; Z.-Q. Wang & Tashev, 2017). Qawaqneh et al., (2017) built an SER solution with age detection using a deep neural network, BNF extractor, and I-vector classifier. They fed MFCC features into the bottleneck deep neural network (DNN) to extract what they called transformed features and also used the DNN for dimensionality reduction, utilizing the non-linear activation functions of the DNN for this purpose. Phoneme labels are the basis of building the transformed features. With the phoneme labels, specific characteristics of speakers are preserved in the speech signal. The transformed feature was fed into I-vector and DNN for classification where the solution was able to classify 65% of children's speech correctly. Verma & Mukhopadhyay, (2016) extracted pitch, intensity, and MFCC and their statistical derivatives as features that were fed to a two-stage support vector machine (SVM). The first stage uses the features to classify the two groups of speakers in the dataset and in the second stage, SVMs classify the speaker's emotions. Z.-Q. Wang & Tashev, (2017) used a voice activity detector (VAD) to remove unvoiced audio frames from the speech dataset and they fed the resultant utterance level frames to a DNN with mean pooling to obtain utterance-level representations that were fed to an extreme learning machine (ELM) for utterance-level classification. A proprietary dataset was used for this research and due to the limited number of samples, ELM was used to replace the softmax layer in the DNN to obtain a more robust recognition of the classes. Kaya et al., (2017) compared the performance of humans and machines in detecting emotions from children's speech. Pediatric students were selected to listen to emotional speech from the EmoChildRu database for children aged 3-7 years and determine if the emotion is comfortable, uncomfortable, or neutral. Duration, energy, and spectral features were extracted from the speech signal using the INTERSPEECH 2010 Computational Paralinguistics Challenge (ComParE) as baseline set. Linear and radial basis function (RBF) was utilized for dimension reduction and normalization of the feature set. Kernel ELM was used as the classification. Human evaluators in Chen et al., (2012) relied on arousal intensities in the speech signal to perform age classification into young (<20), adult (21-60), and senior (>60) categories. Fundamental frequency was used as a key differentiator in the age as children tend to have a higher fundamental frequency than adults.

www.manaraa.com

Other features such as ZCR, MFCC, and spectral centroid were also extracted and fed to an SVM classifier trained on a proprietary dataset. The major drawback of this study is the use of a small, self-collected speech dataset that consists of 336 audio clips. Even though the study also used a more standardized dataset, the berlin emotion speech dataset, it was only used for gender recognition as all the speakers were adults. Shaqra et al., (2019) used hierarchical modeling to develop an SER system with age and gender recognition. Three separate MLP classifiers were used for emotion recognition and age detection. The first level determines a dichotomous younger (<27) and older (>=27) categories. The two categories are fed into separate MLP classifiers for emotion detection. Feature sets obtained directly from the OpenSmile toolkit were used, and it was found that the eGeMAPs feature set was sufficient for gender detection while the large emotion feature set performed better for age detection. Feature sets are a group of features that have been found to work well in previous studies and they are now bundled together and given a name. Goel et al., (2018) fed raw speech to a deep neural network to obtain age and emotion classification. The front end of the deep neural network is made of convolutional layers for feature learning while the later layers are recurrent layers for temporal modeling. The dataset used to train the model is a proprietary dataset that was not disclosed. The paper in essence is a description of a commercial software. Palo et al., (2018) used fuzzy c- means and k-means clustering techniques to show how various speech features such as fundamental frequency, pitch, speech rate, energy, and others vary with age. Based on the results obtained, it is easy to see features that vary significantly with the age of the speaker and incorporate these features in an SER system with age recognition. Chaudhari & Kagalkar, (2015) used MFCC features only to train an SVM and Gaussian mixture model (GMM) classifiers to recognize age and emotion on the aGender dataset. PCA was used to reduce the dimensionality of the 'supervectors' created with the Gaussian mixture model (GMM) before it was fed into the model. While MFCC has been known to perform well in emotion recognition, it is not clear that MFCC features are proven to perform well for age detection from

www.manaraa.com

speech signals. Palo et al., (2017) seek to distinguish speaker's emotions using speech features that vary with age. The selected speech features are MFCC, log energy, and speech rate. K-means clustering was used to separate the children's speech data and adult speech data and log energy was found to be the most distinguishing feature for the children group and the adult group. This study is useful in that the focus is on finding the key features in a speech that would be used to distinguish a child speaker from an adult speaker, the main drawback is that dozens of features could be extracted from speech signal but only three of them were evaluated. H. Zhao et al., (2020) proposed a hierarchical learning architecture for SER with age and gender detection. Pretrained sub-models were created for age detection and gender detection, and the learning weights from the sub-models were transferred to the main model for emotion recognition. The normalized feature set was fed into the sub-models first before the main model. Several parameter transfer settings were used between the sub-models and the main model and a DNN was also benchmarked with the hierarchical model. Though the results were encouraging, selecting features that are known to give good results with age detection may work better than using a bundled feature set. Yücesoy, (2020) did not use extracted MFCC and delta-coefficients features directly as most other studies do but used a Gaussian mixture model (GMM) like Chaudhari & Kagalkar, (2015) to build a supervector for better modeling of the age and gender classes. Nuisance attribute projection (NAP) was used to clean out channel effects and the resultant vector was fed to an SVM for classification.

www.manaraa.com

CHAPTER 4

Deep Learning-Based Child Speech Emotion Recognition Framework

In this chapter, we outline and discuss the objectives of the research. Design science in information systems is reviewed, and the guidelines for design science research as espoused by (Hevner et al., 2004) is presented. The current research is then mapped to the guidelines for design science research. The design artifact, which consists of three components, is proposed. At the core of the design artifact is a deep learning framework that is based on a theory from clinical psychology that explains how the human brain learns.

4.1 Research Objectives

This research seeks to achieve the following objectives:

1. Identify the best features for speaker age group recognition.
2. Identify the best features for child speech emotion recognition.
3. Assess the advantages of using spectrograms over audio features for child speech emotion recognition.
4. Develop a deep learning-based, child speech emotion recognition framework from raw audio.
5. Evaluate the performance of the proposed framework against other existing approaches.

Identify the best features for speaker age group recognition.

A child speech emotion recognition system must first recognize speech made by children and separate them from speech made by adults. As a result, recognizing the speaker age group is a very crucial first step in this work because It allows us to filter out speech made by adults and focus only on speech that was made by children to further detect the underlying emotion in the speech. Speech is a natural human ability that involves a complex working together of several human organs. Speech signal is a speech file that could be presented in time domain and frequency domain (Rao & Koolagudi, 2013b). It contains numerous features that could be used for several research objectives one of which is speaker age group recognition. Speaker age group recognition is the ability to recognize the age group of a person based on his speech. Speaker age group recognition

www.manaraa.com

is a challenging task to perform using machine learning or deep learning techniques (Abhang et al., 2016). Machine learning and deep learning techniques at the current maturity can predict the speaker age group fairly accurately when the range between age groups is wide enough such as between 1-10, 10- 20, or 20-30 years; however, recognition accuracy deteriorates as the range gets narrower, similar to the way humans can only make a rough estimate age based on external features (Ma et al., 2017). At the current maturity level, it is difficult to achieve a pinpoint age determination from an artificial intelligence model using speech signals alone. A combination of other features such as facial features and other physiological features will lead to better age determination using artificial intelligence techniques (Chae et al., 2018). For our purposes in this research, we perform speaker age group recognition in other to isolate a child speaker from an adult speaker. This separation is an important first step to ensure that we are dealing with child speakers only. The United Nations Convention on the Rights of the Child is a treaty on human rights that defines the rights of children and that also defines a child as any person that is under the age of eighteen (18) (OHCHR, 1990). We will use this definition as a boundary landmark between the binary classification of child age group and adult age group. Several speech features have been used in the speaker age group recognition including low-level speech features such as time-domain energy, harmonic energy, Mel-frequency cepstral coefficient, perceptual linear predictive features, zero-crossing rate, etc., however, there is no semblance of unanimity between researchers in working in this domain as to the best features to use for speaker age group recognition. Markitantov & Verkholyak, (2019) believes Mel-spectrogram and Mel-frequency cepstral coefficient (MFCC) with their first and second-order derivative is best while (Grzybowska & Kacprzak, 2016) prioritizes LSP frequency, fundamental frequency, jitter, and shimmer as key features. Palo et al., (2018) used pitch and speech rate among others while (Palo et al., 2017) favors log of energy as a good distinguishing feature between a child speaker and an adult speaker. In addressing RQ1, we will perform a series of experiments involving speech features to identify the features that best recognize a child speaker.

www.manaraa.com

Identify the best features for child emotion recognition.

The essential task in this objective is to identify significant speech features that contain substantial emotional information in the speech signal. Although certain widely used speech features such as Mel-frequency cepstral coefficient (MFCC) could be generally used across both domains, some researchers suggest that there is some range of speech features that are best used for emotion recognition (Assunção et al., 2019; Navyasri et al., 2018). Some of the features that have been used for emotion recognition include spectral centroid, spectral pitch chroma, spectral skewness, instantaneous fundamental frequency, strength of excitation, linear predictor coefficient (LPC), pitch, and energy among others (Khalil et al., 2019; Navyasri et al., 2018; Swain et al., 2018). It has been reported that models that were trained for speech emotion recognition for adults were found to perform very poorly when they were used to recognize emotion from children's speech (Palo et al., 2017). This poses a further challenge in this research domain for efforts to be made to advance knowledge for children's speech emotion recognition. This is a growing niche within the overall field of speech recognition. Use cases and applications such as educational robots and decision support for child psychologists and counselors are still being identified. The variance between speech features for emotion recognition between adults and children as identified by Palo et al., 2017 adds a new dimension to our research challenge as we cannot use the well-researched speech features for adult emotion recognition in a plug-and-play manner for child speech emotion recognition. To seek the answers for RQ2, we will perform a series of experiments to determine which features are best suited for emotion recognition in children's speech.

Assess the advantages of using spectrograms over audio features

When building an artificial intelligence model for speech emotion recognition, there are mainly three data options that could be used as input to the model – features extracted from speech, the raw speech itself, and spectrograms. The first option is the most-researched option; however, there is relatively less research available for using spectrograms and raw speech as input to a child speech emotion recognition model. Spectrograms are a useful way of visualizing sound signals. They have been described as sound images (Sharan & Moir, 2015). The appeal of using spectrograms for child speech emotion recognition is its potential to overcome the multi-

www.manaraa.com

dimensional audio features that complicate the learning process (Kerkeni et al., 2019). Spectrograms are created from raw sound signals using constant Q transform (CQT) or short-term Fourier transform (STFT) (Mao et al., 2019; Syed et al., 2018; Tang et al., 2018). While the conventional audio signal is presented in the time domain as the variation of amplitude or loudness over time, spectrogram presents sounds in the frequency domain where we can observe the variation of frequency and energy over time. Frequency is typically represented on the vertical axis, time on the horizontal axis, and energy as the third dimension is represented with color variations. Some studies have demonstrated the effectiveness of using spectrograms as input to speech emotion recognition models (Satt et al., 2017; Sharan & Moir, 2015; Yenigalla et al., 2018); however, these research are focused on general speech emotion recognition (which mainly involve adult speakers) without any major focus on children's speech. Furthermore, most datasets available for speech emotion recognition contain adult speech and there are very few data sources for children's speech. In this objective, we will build a child speech emotion recognition model using spectrograms to determine the effectiveness of the technique for child emotion recognition over the use of audio features.

Develop a deep learning-based, end-to-end child speech emotion recognition framework

Building a deep learning-based framework for emotion recognition from raw audio is often called end-to-end emotion recognition (Li et al., 2019; Tang et al., 2018). This is because the architecture is built to learn directly from the raw audio files without any intermediate procedure of creating spectrograms or extracting features from the raw audio. End-to-end learning eliminates the need for feature extraction or spectrogram generation. End-to-end learning techniques use deep neural networks such as convolutional neural networks, deep belief networks (DBN), recurrent neural network (RNN) deep Boltzmann machines (DBM), and long short-term memory (LSTM). A combination of CNN + RNN and CNN + LSTM is also used for end-to-end emotion recognition. End-to-end emotion recognition where neural networks learn the representation that best suits the task at hand directly from raw audio data could lead to improved performance, at least theoretically. What is clear so far is that end-to-end learning generally requires less procedure and fewer lines of code for data processing; however, it takes more skill to develop a deep neural network framework used for end-to-end emotion recognition than developing standard machine

www.manaraa.com

learning models. In addressing this objective, we develop a framework using Baddeley's Theory of Working Memory that will be used to perform child speech emotion recognition.

Evaluate the performance of the proposed framework against other existing approaches

In this objective, we evaluate the performance of the proposed framework and conduct comparative performance analysis with the models built using standard deep learning architectures using CNN and LSTM, and their combinations. Standard metrics for quantifying the performance of models include Accuracy of Classification, Area Under Curve (AUC) of Receiver Operating Characteristics (ROC), Precision-Recall-F-measure and Confusion Matrix (Antipov et al., 2017; Mena, 2012; Verma & Mukhopadhyay, 2016). We will use classification accuracy, precision-recall-f-measure, loss curve, and confusion matrix to access the performance of the framework developed.

Research Question	**Maps to**
RQ 1	Objective 1
RQ 2	Objective 2
RQ 3	Objective 3
RQ 4	Objective 4
RQ 5	Objective 5

Table 1: Research Questions mapped to Research Objectives

4.2 Research Method - Design Science Research in Information Systems

Broadly speaking, the world of science generally consists of the natural sciences and the artificial sciences referred to by Simon Herbert as Sciences of the Artificial (Simon, 1969). While the natural sciences broadly consist of the physical sciences and biological sciences and deals with the nature, behavior, characteristics, and interaction of naturally existing phenomena, the artificial sciences deal with systematic knowledge of man-made artifacts as opposed to naturally occurring phenomena. The end product of the artificial sciences is to create an artifact. A new theory or a refinement of an existing theory could also be a byproduct. An artifact provides utility that solves a human problem and enhances human existence. The idea of artificial sciences as espoused by

www.manaraa.com

Simon is termed design science in information systems discipline. The research method adopted in this research is design science. The end product is to create a design artifact that provides utility in identifying the emotions of children from their speech. According to the design science methodology, design artifacts could be guided by laws of artificial sciences and they do not violate natural laws that govern the natural sciences. In reality, the knowledge of natural laws often aids the design of artifacts, and the kernel theory that design artifacts are based on could directly be a law of natural science or be moderated by it (March & Smith, 1995). The nature of design artifacts in the field of information systems has evolved. March & Smith, (1995) argue that design science research outputs can be classed into four categories of constructs, methods, models, and instantiation. The concept of constructs is slightly more difficult to grasp, and the definition is presented from a broad IS perspective. Schon, (1984) offers that constructs provide a language in which problems and solutions are defined and communicated, (March & Smith, 1995) adds that, "constructs constitute a conceptualization used to describe problems within the domain and to specify their solutions. They form the specialized language and shared knowledge of a discipline or sub-discipline." Hevner et al., (2004) provide more perspective, adding that, "constructs provide the vocabulary and symbols used to define problems and solutions. They enable the construction of models or representations of the problem domain." The constructs in our research method include raw audio, speech signals, audio features, spectrograms, dataset, convolutional neural network, deep learning, long short-term memory, etc. These constructs are used to define our problem and solution domain. A Model is a set of representations that shows a logical relationship and/or interaction between constructs. Models are built from constructs and they represent the problem or the solution to the problem within a specified domain. Method is comprised of a series of steps that work together to complete a task that contributes to the solution of the problem. Several methods could work together in a predefined manner to bring about a problem solution or a single method could be all that is required to solve a problem. The building blocks of methods are the constructs and languages that have been defined and are operational within the solution domain. Instantiation is the realization of an artifact in its environment. The utility that the instantiation provides attests to the effectiveness of the models and methods that make up the instantiation. While (Hevner et al., 2004) and (Peffers et al., 2007) retained constructs, methods, models, and instantiation as the products of design science research, (Peffers et al., 2012) adds

www.manaraa.com

algorithm and framework as design science artifact types. They make a distinction between methods and algorithms by arguing that while methods are, "actionable instructions more conceptual" in nature, algorithms are more of "formal logical instructions." Framework is simply defined as a "meta-model." Design science research methodology as espoused by (Hevner et al., 2004) is adopted in this dissertation.

There are several reasons why the design science methodology is the best choice for this research. The methodology has been described by (Hevner et al., 2004) as, 'fundamentally a problem-solving paradigm.' The alternative behavioral science paradigm also according to the same authors is involved with, 'developing and justifying theories (principles and laws) that explain and predict organization and human phenomena.' This research is not aimed at explicating any organizational or human phenomena but rather at creating an artifact for child speech emotion detection. It is, therefore, more suited to the design science paradigm than the alternative paradigm. Secondly, the design science artifacts, 'are represented in a structured form that may vary from software, formal logic, and rigorous mathematics to informal natural language descriptions.' The artifact produced in this work has a form that users can interact with and underneath the user-interactive frontend is the backend that is mostly logic, mathematics, and algorithms. The design science research processes include *build and evaluate.* Design methodology makes room for artifacts to go through series of iterations of building and evaluation before the final product is realized. From the outset, it was clear that the research questions are such that optimal solutions cannot be achieved in one go but requires modification and rerun. Our proposed artifact would undergo several iterations and modifications before the final design was attained.

Design science artifacts must be evaluated to demonstrate utility and effectiveness. According to (Peffers et al., 2012), evaluation methods include logical argument, expert evaluation, technical experiment, subject-based experiment, action research, prototype, case studies, and illustrative scenario. Logical argument is surface explications and arguments made in favor of the artifact's utility. It should be grounded in well-known and accepted theory, principles, laws, and guidelines. Expert evaluation involves the use of subject matter experts and authoritative personnel in the assessment of the artifact. Technical Experiment according to (Peffers et al., 2012) is, 'a performance evaluation of an algorithm implementation using real-world data, synthetic data, or no data, designed to evaluate the technical performance, rather than its performance in relation

www.manaraa.com

to the real world.' Subject-based Experiment involves the use of subjects, usually humans or animals, to assess if a hypothetical proposition is true. Action research involves the insertion of an artifact into a real-world environment to substantiate its effect. This is usually very time-consuming and often requires collaboration and agreement between the researcher and industry personnel. Prototype is a preliminary typification or instantiation to demonstrate the utility of the artifact. Illustrative scenario is the exemplification of the utility or suitability of an artifact based on specific real or synthetic scenarios. In this research, we use technical experiment as our evaluation method. It is the most suitable method as artificial intelligence solutions require the use of data to train models. The downside is that technical experiments are used to access technical performance rather than real-world performance. Beyond the scope of this dissertation, we hope to further advance the artifact and evaluate it with real-word evaluation methods such as subject-based experiment and case study.

4.3 Design Science Research Guidelines.

Hevner et al. (2004) suggest the following seven guidelines:

- **Guideline 1: Design as an Artifact**
- **Guideline 2: Problem Relevance**
- **Guideline 3: Design Evaluation**
- **Guideline 4: Research Contribution**
- **Guideline 5: Research Rigor**
- **Guideline 6: Design as a Search Process**
- **Guideline 7: Communication of Research**

We will discuss the dissertation with respect to these guidelines.

www.manaraa.com

Design as an Artifact:

Certain artifact types must be produced for a research to qualify as design science research. Those artifact types include constructs, methods, models, algorithms, frameworks, and instantiation (Peffers et al., 2012). The artifact is created to provide utility, value, or service and to solve an existing organizational or human problem. In this research, we develop a framework that supports emotion detection from children's speech. Furthermore, we propose a working memory recurrent network (WMRN) for child speech emotion recognition.

Problem Relevance

Problem has been defined as, "the differences between a goal state and the current state…" (Hevner et al., 2004). Goal state refers to the utopian *to-be* state that we aim to achieve and the current state is the present *as-is* state that is ridden with challenges and problems that we are aspiring to solve. The goal state for child speech emotion recognition is to create IT artifacts that could be used to solve real-world problems such as a decision support system for a child therapist and applications that could support challenged children. This is a real-world problem that is very relevant and that attracts the interests of governmental and non-governmental child welfare organizations such as UNICEF.

Design Evaluation

Peffers et al., (2012) espouse different methods for the evaluation of a design science artifact. We will adopt the technical experiment method for the evaluation of the framework. They define technical experiment as, "a performance evaluation of an algorithm implementation using real-world data, synthetic data, or no data, designed to evaluate the technical performance, rather than its performance in relation to the real world." Rigorous evaluation is the way to demonstrate the effectiveness and utility of the design artifact. In our evaluation, we will use both real-world data (unsolicited speech from everyday living) and synthetic (acted speech) data to evaluate the performance of the design artifact.

www.manaraa.com

Research Contributions

Design science research must make a contribution that is distinct and clearly defined. Hevner et al., (2004) propose that design science research should provide research contribution in the area of novelty, generality, and significance. Novelty speaks to something new and interesting. Exploiting findings in speech synthesis and clinical psychology, we develop a model, named "a working memory recurrent network (WMRN)" to recognize emotions from children's speech. Generality speaks to an artifact that transcends the bounds of the environment where it was created, an artifact that is useful in other circumstances. All components and constructs of the design artifact in this study such as child, speech, and emotion transcend the environment where the artifact is created. For example, the language of the speech construct that is used in the design artifact is not limited to the English language. In other words, the artifact is useful even when the child's speech is in other languages, making the artifact generalize well in other non- English-speaking climes.

Research Rigor

Research rigor is demonstrated in the logical formulation of the components of the framework and the working memory recurrent network, the effective use of the design science methodology, and the method used in the evaluation of the framework. Furthermore, this logical formulation is based on a kernel theory – Baddeley's theory of working memory that is widely accepted and used in clinical psychology. According to Hevner, success in rigor, "is predicated on the researcher's skilled selection of appropriate techniques to develop or construct a theory or artifact and the selection of appropriate means to justify the theory or evaluate the artifact". The selection of Baddeley's theory of working memory is justified by its ubiquitous use in the field of clinical psychology. Neural networks, in general, are originally conceptualized and inspired by the way the human brain and the nervous system work, the best theories to use to model neural networks are the theories that best explain the human nervous system and the brain. Furthermore, we selected an appropriate means to evaluate the framework by benchmarking it with commonly used neural network types such as LSTM.

www.manaraa.com

Design as a Search Process

As observed by (Hevner et al., 2004), achieving an optimal design of an artifact often requires iteration. In this study, we perform experimental iterations between several techniques that have been used for speech emotion recognition. In each of these techniques, there were sub-experiments, mostly a search process that was performed to achieve the optimal performance for a given technique. Furthermore, the framework had to build upon existing knowledge components that are domiciled in extant literature that can only be harnessed through a careful search process. In the search process, we first explored extant literature to find out what has been published on speech emotion recognition. While there were many papers found, most of them were focused on adult speech and not child speech. Developing a model for child speech emotion recognition requires building from the ground up as those models that work for adult speech emotion are known to not work well for child speech emotion recognition (Palo et al., 2017). These search processes allow us to eliminate areas that have been heavily researched and focus our work on less published and more novel areas.

Communication of the Research

Research communication has to include satisfactory information for both managerial and technical personnel. Technical persons must find enough information to attempt to replicate the research within their context and managerial people should also find the content of the research satisfying. In writing this dissertation, we make efforts to satisfy both viewpoints. In communicating this research, attempts will be made to recreate components of the research that will be publishable in a journal. This dissertation will also be cataloged in online research databases via the university library.

4.4 Proposed Design Artifact –Deep Learning-based Framework for Child Speech Emotion Recognition

The design artifact consists of three components – the Ideation Component, the Codetermination Component, and the Affectation Component, as shown below. These three components together accomplish the objectives of this study.

www.manaraa.com

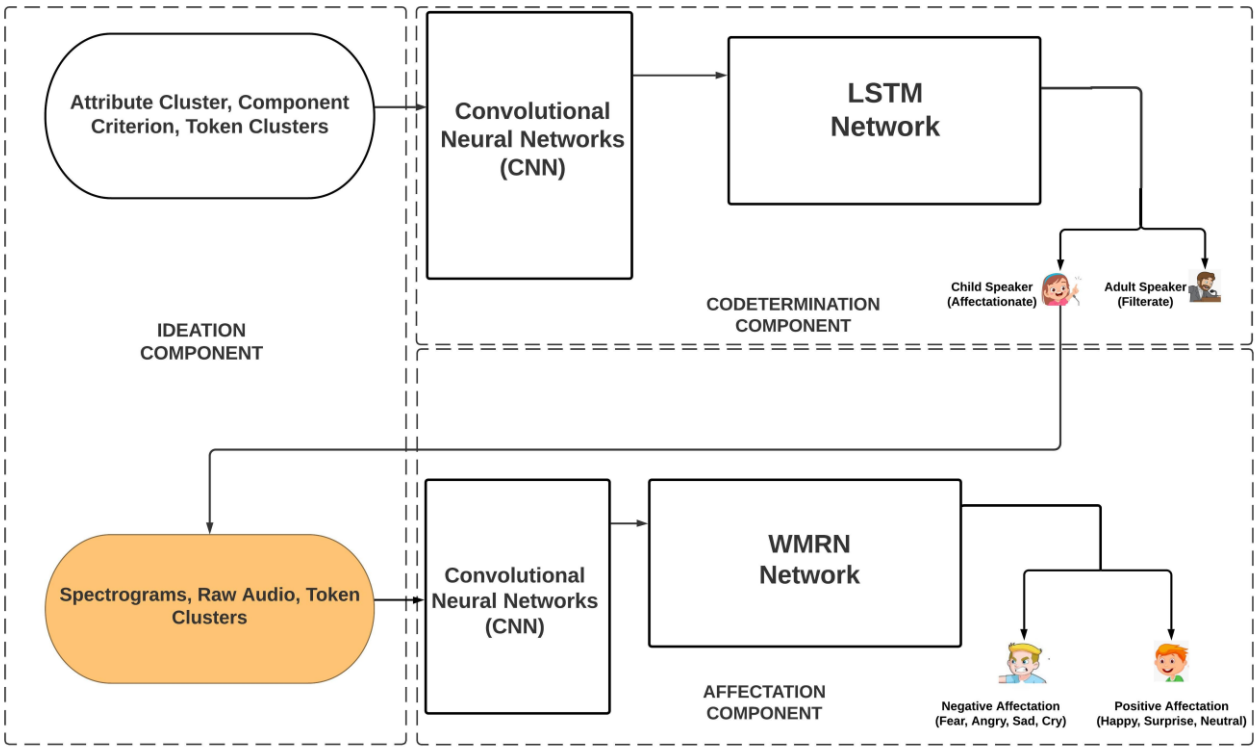


Figure 9: High-level architecture of the proposed framework for Child Speech Emotion Recognition

4.4.1 IDEATION COMPONENT

The ideation component guides a researcher who is looking into the solution space of the research problem and searching for the best features to be used in training the deep learning models to achieve the research objectives. The solution space involves looking at extant literature to find what is feasible and what has been done in the past. It also involves finding the best methods to engineer and control key experiment variables in other to achieve the desired results. In this research, the key experiment variable is the speech data. The way and manner that the speech data is used will affect the result. We frame the speech data as consisting of several attributes that are clustered within the speech audio file. We call this the attribute cluster. Each of these clustered attributes of the speech could be harnessed and used for a different objective. Such objectives include speech emotion recognition, speaker recognition, speaker identification, music classification, music genre recognition, audio fingerprinting, automatic music tagging, audio segmentation, speech processing, and synthesis, and many others.

www.manaraa.com

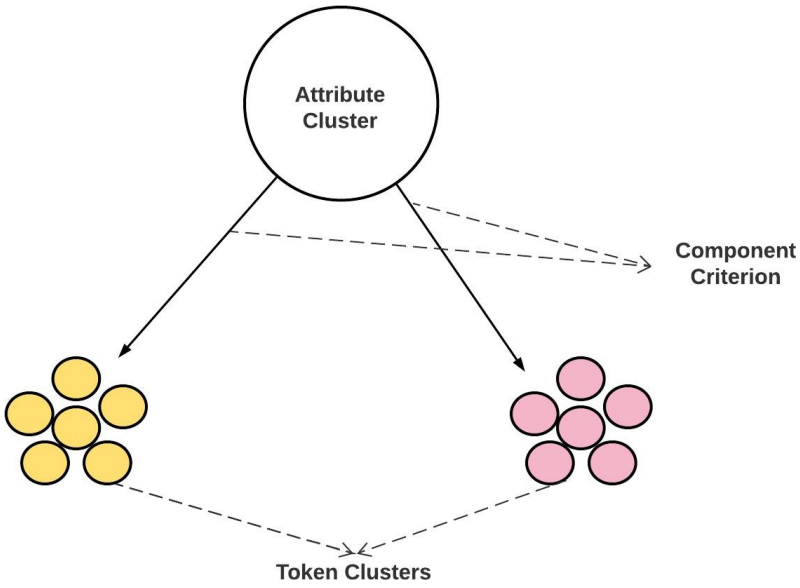


Figure 10: Ideation Component

Next, a decision is made on the attributes or features within the attribute cluster that will work best for the research objective. These attributes that have been selected to be used for the research objective are selected based on certain decision factors. The factors upon which the decision to use certain features or attributes are based on are the component criterion. Component criterion is the guidelines, knowledge, measures, and rules that have been used to identify specific attribute(s) or features from the attribute cluster that are effective for given research objectives. Components criterion can be obtained by experimental trials, theoretical knowledge, and related studies. In this study, we combine experimental trials and knowledge from related studies to find the component criterion for our research objectives. The token clusters are attributes that have been abstracted from the attribute clusters using the component criterion. The token clusters are essentially the features that will be used to train our deep learning models. The token clusters for age group classification tie to research question one (RQ 1) while token clusters for child speech emotion recognition tie to our research question two (RQ 2) where we identify the speech features that work best for child speech emotion recognition.

Mel Frequency Cepstral Coefficient (MFCC)

MFCC is a popular feature used in speech signal analysis (Qawaqneh et al., 2017). Time-domain

www.manaraa.com

speech signals can be analyzed using a short analysis window where each speech signal is split into smaller frames. These smaller speech frames are passed through a process known as fast Fourier transform in order to generate spectrums. In other words, we can represent speech signals as a series or sequence of spectrums. The fast Fourier transform is a form of discrete Fourier transform and to obtain the spectrums from the frames, we take the fast Fourier transform on frames to compute speech signal $S_t(k)$ for given ith number of frame and N and k length of fast Fourier transform, using the following equation;

$$Si(k)=\sum_{n=1}^{N} si(n)h(n)e^{-j2\pi kn/N} \qquad 1\leq k\leq K,$$

where $s_i(n)$ is the signal and $h(n)$ is the analysis window for $\forall n \in \{1,\ldots,N\}$,

The spectrum or spectral vector is a frequency-amplitude representation of the speech signal. Typically, the amplitude is mapped to a color scale where 255 would represent a black or other dark color and 0 would represent white or other light colors. Hence, higher amplitudes will result in a darker color and lower amplitude in a lighter color. These color scale representations would be used to represent the amplitude variation rather than an amplitude line graph. The sequence of the color representation for the duration of the speech signal is the spectrogram.

It is possible and often the case that for speech spectrums, the amplitude is represented in the log domain rather than linear values of amplitude. The log domain of the amplitude is measured in decibels (dB).

The peak of each wave or cycle carries the identity of the sound and it is called the dominant frequency or the formant. The curved line connecting the formants is the spectral envelope. The spectral envelope is used extensively for various speech recognition applications. Taking the inverse fast Fourier transform (IFFT) of the spectrum in the log domain yields the cepstrum. The IFFT equation is represented as follows;

$$si(n)h(n)= 1/N\sum_{k=1}^{N} Si(k)e^{j2\pi nk/N} \qquad 1\leq n\leq N$$

www.manaraa.com

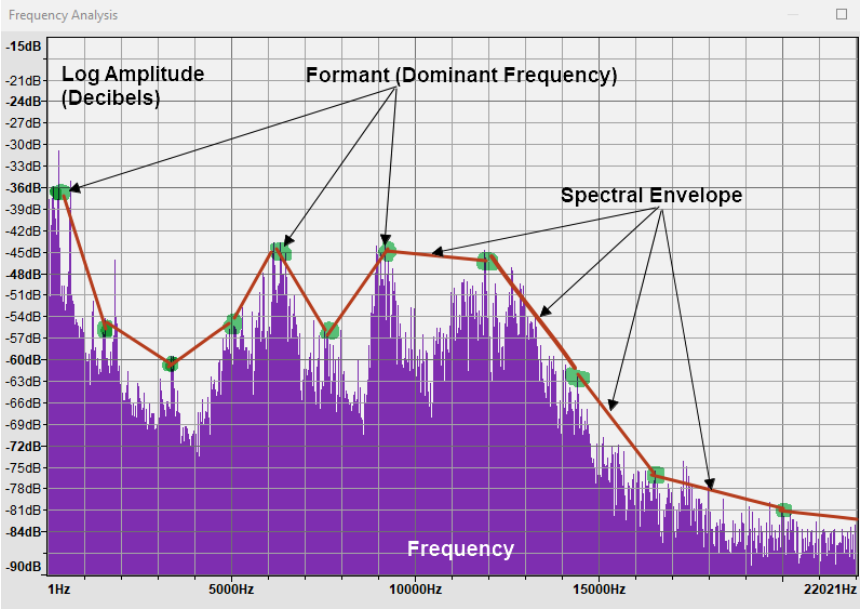


Figure 11: Speech Spectrum

The cepstrum consists of high-frequency components and low-frequency components. The low-frequency components represent the spectral envelope. The spectral envelope can be extracted using a low-frequency filter. Taking the IFFT yield a signal that goes back to the time domain of the original signal. However, this time domain is often called quefrency to denote an inversion of frequency (Oppenheim & Schafer, 2004).

If we denote a spectrum with a mathematical equation s(f), then we can express the function as follows;

$$S(f) = E(f)D(f)$$

$$|S(f)| = |E(f)|\,|D(f)| \qquad \textit{(taking magnitudes)}$$

$$Log|S(f)| = Log|E(f)| + Log|D(f)| \qquad \textit{(taking Log)}$$

$$s(f) = e(f) + d(f) \qquad \textit{(taking IFFT)}$$

s(f) is the cepstrum, e(f) is the spectral envelope and d(f) is the remaining spectral details.

www.manaraa.com

Extracting the spectral envelope covers the entire frequency range of the signal, but in human perceptual experiments, the human ear acts as a filter that focuses on certain frequencies and ignores other frequencies. Designing artifacts to include this non-linearity in human ear perception generally improves recognition performance (Jung et al., 2017). This idea has led to the introduction of frequency filters that are positioned at certain frequencies on the spectral signal. These are called mel-frequency filterbanks and the non-linear scale is called the mel-scale. This scale is non-linear because there are shorter distances between mel filterbanks at the lower frequency range than at the high-frequency range. Applying the mel-filterbanks on the spectrum signal yields the mel-spectrum.

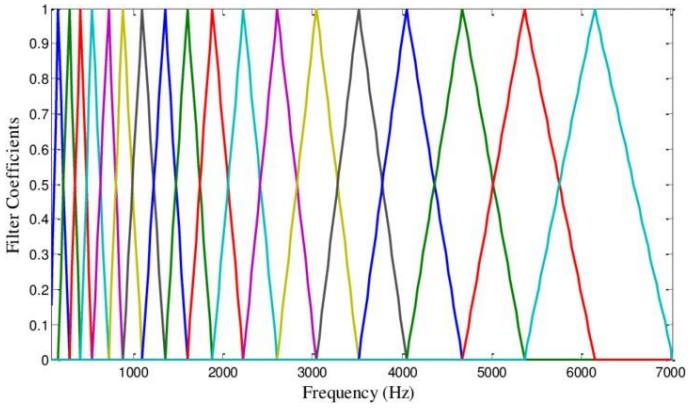


Figure 12: Mel-Scale Filterbank

As with extracting the spectral envelope, we can apply the log scale over the mel-spectrum to obtain the mel-spectrum in the logarithm domain.

$$Log\ M(f) = mel - spectrum\ in \log domain$$

$$Log\ M(f) = Log\ C(f) + Log\ G(f) \qquad \textit{(taking Log)}$$

$$m(f) = c(f) + g(f) \qquad \textit{(taking IFFT)}$$

The cepstral coefficient c(f) obtained from taking inverse fast Fourier transform (IFFT) on the mel-spectrum in the log domain is referred to as the Mel-Frequency Cepstral Coefficient (MFCC).

www.manaraa.com

4.4.2 CODETERMINATION COMPONENT

The codetermination component leverages the powerful capabilities of deep learning technologies. In this component, the objective is to separate the speech made by adults from those made by children (speaker age group recognition) using the token clusters. Recall that the token clusters contain attributes or features that are most effective for a given research objective. Hence the token clusters are fed into two codeterminators – a combination of convolutional neural network (CNN) and a specific version of recurrent neural network called the long short-term memory (LSTM) network. This combination is used because extant literature in speech classification research utilizes it very often and it has been shown to be effective (Lee et al., 2018; Zhao et al., 2019). The idea of codetermination is borrowed from the Codetermination Theory (McCain, 1980) which is a corporate governance theory where workers and management co-operate together in decision making especially with respect to the representation of workers on the board of directors in the company. In this scenario, the CNN and LSTM are the codeterminators that decide which speech was made by an adult and which was made by a child.

The speech that has been classified as adult speakers are termed the *filtrate* because there is no further use for them in the design pipeline. However, the speech that has been recognized as coming from a child speaker is termed *Affectationate* and is used in the last component of the design architecture.

www.manaraa.com

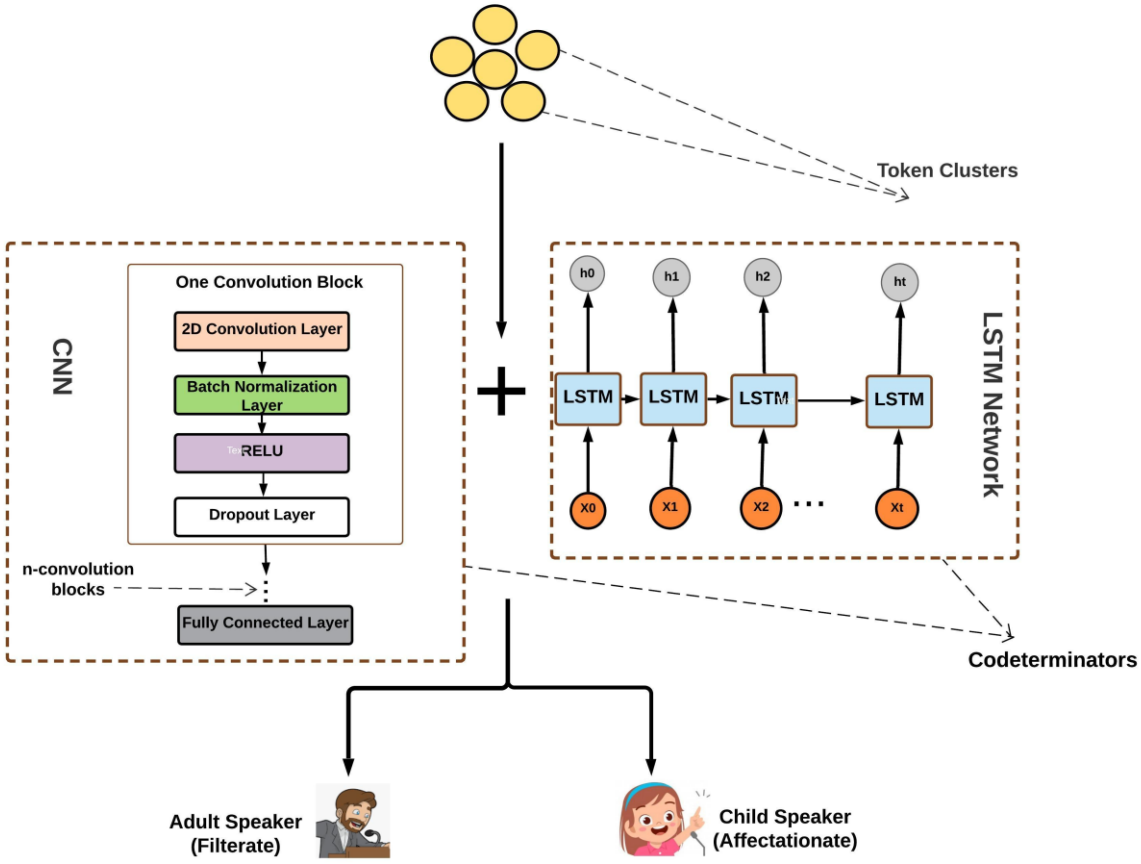


Figure 14: Codetermination Component

Long Short-Term Memory (LSTM) NETWORKS

Long short-term memory networks are a variation of the recurrent neural networks (RNN) that was developed to solve some of the original problems of RNN such as the vanishing gradient problem and short-term memory (Chen et al., 2019). The vanishing gradient problem is one that is encountered in most deep neural networks where the weight changes learned and propagated either forward or backward become small that it results in insignificant differences from the weights learned in the previous pass. Layers that receive this very small weight increment from the previous weight do not learn.

www.manaraa.com

A layer is a building block of a deep learning model that receives input, transforms it using non-linear functions, and passes it as output to the next layer. RNNs are known to use previous contextual information to make predictions such as in their application with language modeling; however, the difficulty arises when the contextual information required to make a current prediction was learned a while ago, possibly by a layer far away from the current layer. For example, RNN could easily predict the correct word 'garage' from this sample, '*the car is parked in the garage.'* This is because the words 'car' and 'parked' were learned not too long ago by layers that are close enough to the layer required to learn 'garage', hence the contextual information is still available to the current layer. However, when the context has been learned further back, in those scenarios, RNN suffers from correct prediction. For Instance, a paragraph begins and ends with, *'I lived in Spain for the first 20 years of my life...I speak Spanish fluently.'* Because of the distance from when the word 'Spain' occurred to when the model needs to predict the language the writer speaks, RNN will struggle to predict 'Spanish'. This problem of RNN is known as long-term dependencies. LSTM attempts to solve this problem by having a feature called the cell state that runs across the LSTM cells from the previous step to the current step. It is very easy for information to flow unmodified along the cell state, however, when needed, information on the cell state can be modified using components of the LSTM cell called the gates.

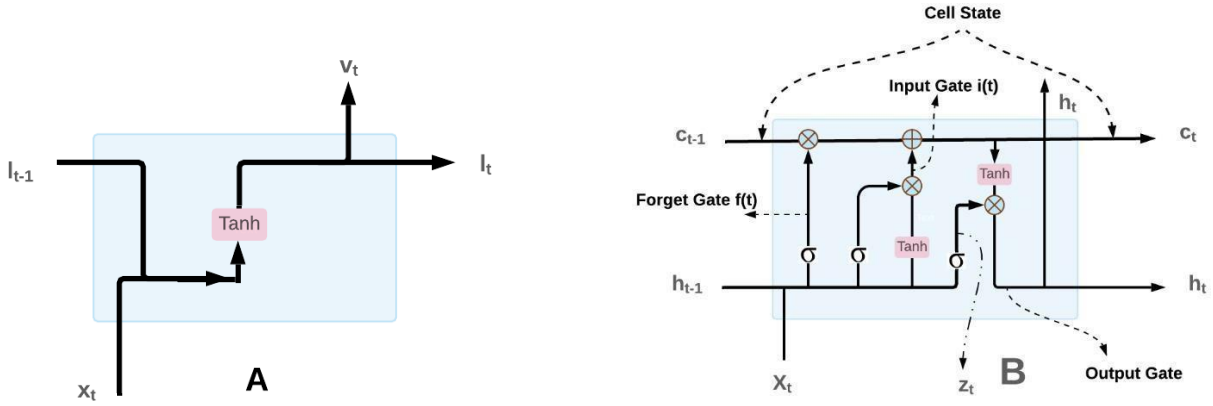


Figure 15: (A) Vanilla RNN Cell (B) LSTM Cell

www.manaraa.com

LSTM has three gates: the forget gate, the input gate, and the output gate. The forget gate determines how an LSTM cell handles the cell state. The forget gate vector function f(t) is given as

$$f(t) = \sigma\left(w_{fo} * [h_{t-1}, x_t] + b_{fo}\right) \quad (1)$$

where w_{fo} is the weight vector and b_{fo} is the bias vector, x_t is the new input on layer, *t,* and h_{t-1} is the output of the previous LSTM cell. The bias vector allows for the adjustment of the output f(t) so that the model can fit the data as best as possible.

Because the forget gate is activated with a sigmoid function σ, its output is 0 or 1, where zero means to forget the original information carried forward in the cell state, and 1 means to retain the information carried forward from the previous LSTM cell. While the forget gate decides on forgetting or retaining previous information, the input gate actually implements this decision. The input gate comprises two vector functions: one activated by a sigmoid activation and the second activated with tanh activation. Activation functions are used to convert layer outputs to 0s, 1s, or -1s. The equation for both functions respectively is;

$$i_1(t) = \sigma\left(w_{i1} * [h_{t-1}, x_t] + b_{i1}\right) \quad (2)$$

$$i_2(t) = tanh\left(w_{i2} * [h_{t-1}, x_t] + b_{i2}\right) \quad (3)$$

$$i(t) = (i_1(t) * i_2(t)) \quad (4)$$

The input gate vector i(t) and the forget gate vector f(t) are used to update the value of the cell state. The first part of this update is the product of the forget gate f(t) and the previous information on the cell state C_{t-1}. This operation determines if previous information is kept (f(t) =1 so that 1 * C_{t-1} = C_{t-1}) or forgotten (f(t) = 0 so that 0* C_{t-1} = 0). The second part of the update adds the input gate vector i(t) to the result of the first part.

$$c_t = (f(t) * ct_{-1}) + i(t) \quad (5)$$

This new cell state c_t is carried forward to the next cell. The output of the current cell also consists of two parts. In the first part, the new cell state c_t is passed through a tanh activation (output of 1 or -1), and in the second part, a sigmoid activation is taken over x_t and h_{t-1} to produce z_t.

$$z_t = \sigma\left(w_z * [h_{t-1}, x_t] + b_z\right) \quad (6)$$

www.manaraa.com

$$h_t = \tanh(c_t) * zt \quad (7)$$

h_t is the output gate vector.

The long-term dependency problem is an important motivation in developing the working memory recurrent network (WMRN) proposed in this study. As some of the speech files are as short as a single word while others are very long sentences, the WMRN being proposed must be able to store emotional information learned at the beginning of long sentences and ensure that they are remembered at the end.

4.4.3 AFFECTATION COMPONENT

From child speech dataset, we build a new token cluster using component criterion. The component criterion that is used to abstract the token clusters from the codetermination component is for the purposes of speaker age group recognition. This component criterion is for child emotion recognition and the token clusters for this objective could vary from the previous objective of speaker age group recognition at least theoretically. It is possible though that there could be some sort of *token cluster sharing* between the two objectives. That is, some token clusters used in speaker age group recognition could also be effective for child emotion recognition. Examples of speech features that are found in the literature to be effective for several speech syntheses and analyses using artificial intelligence are the Mel-frequency cepstral coefficient (MFCC) features (Qawaqneh et al., 2017), ZCR, and Energy.

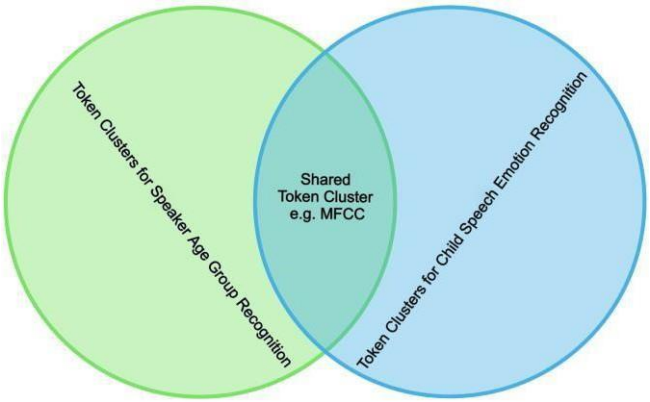


Figure 16: Token Cluster Sharing

www.manaraa.com

The affectation component uses convolutional neural networks and the working memory recurrent network to perform emotion recognition. We will test other configurations of the deep learning model, such as using a fully convolutional neural network and full RNN.

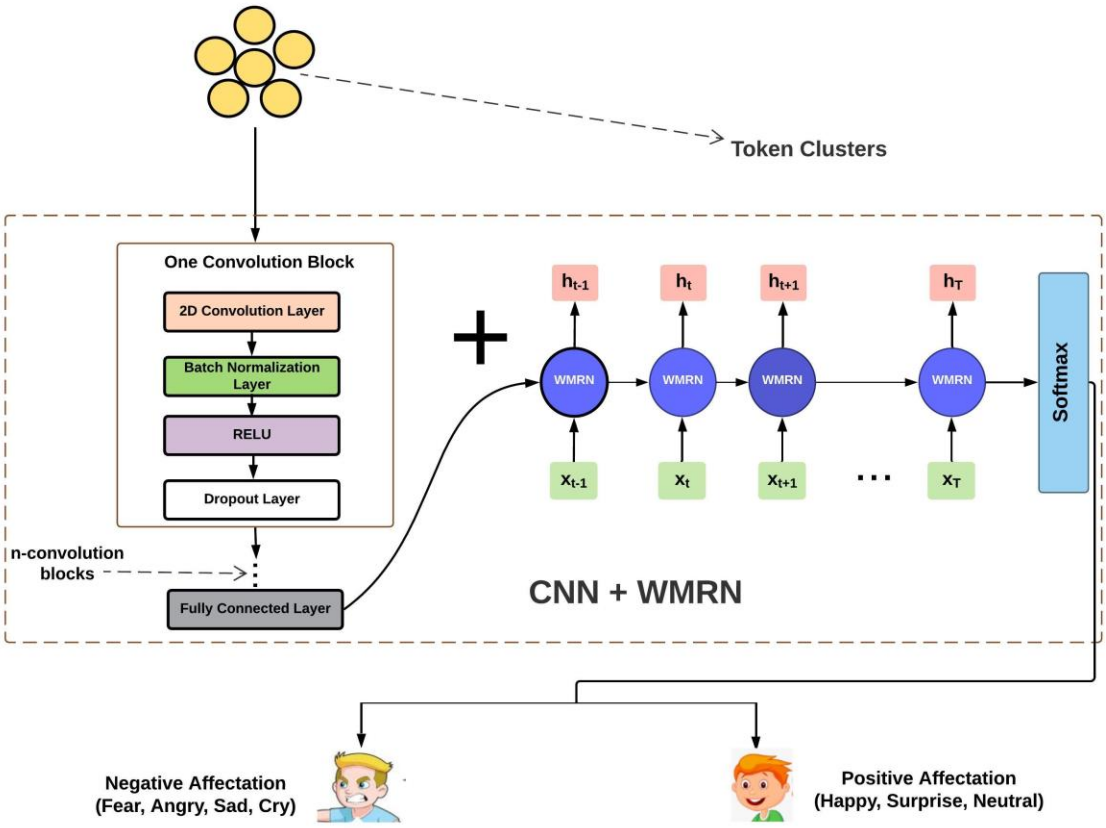


Figure 17: Affectation Component

The positive emotions, such as happy, surprise, and calm, are termed positive affectation, and the negative emotions such as fear, anger, sad, and crying are termed negative affectation.

4.4.4 Proposed Working Memory Recurrent Network (WMRN)

In this section, we propose a new deep learning unit for child speech emotion recognition that is based on Baddeley's theory of Working Memory also known as Baddeley's model of Working Memory (A. D. Baddeley & Hitch, 1974). We use this theory to develop the proposed deep learning units because we believe that they are a good representation of how humans learn. Furthermore, Baddeley's theory explains short-term memory and the models developed from it are capable of overcoming the long-term dependency problems of recurrent neural networks. Though this model

www.manaraa.com

was proposed in 1974, it has received several revisions and updates since (A. Baddeley, 1992, 1996; A. D. Baddeley, 1986; A. D. Baddeley et al., 2011).

Baddeley's Model of Working Memory

Working memory refers to the capacity to hold information briefly in memory while performing other mental operations on the information (Mirsky et al., 1995). Working memory is the memory that allows us to hold information that is still being processed. It allows us to retain partial outcomes when performing mathematical operations mentally and also to follow along when listening to a lecture, and call the attention of the lecturer to explain a seeming discrepancy between what was just stated and what was stated before. Working memory is very crucial to learning, cognitive development, and education (Cowan, 2014). Working memory is often misconstrued as short-term memory but there is a fundamental difference between the two. While short-term memory stores information for a short period, working memory holds information temporarily for processing and manipulation purposes. Working memory is usually regarded as the counterpart of long-term memory. Before Baddeley's model of working memory, the dominating school of thought holds that working memory is a singular component of the brain that is situated at the prefrontal cortex. However, Baddeley's model shows that the working memory consists of three main modules that work together: the Phonological Loop, the Visuo-Spatial Sketchpad, and the Central Executive. Hence this model is also often described as the multicomponent model.

www.manaraa.com

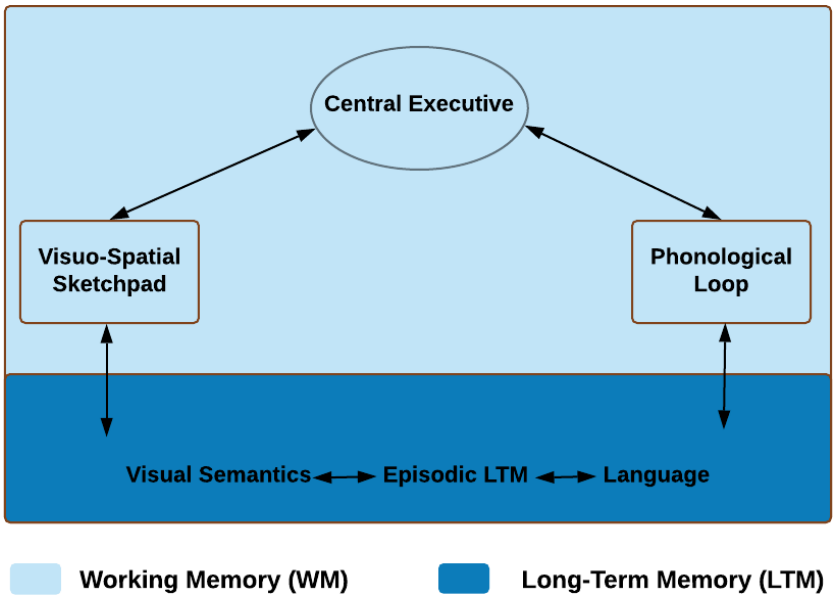


Figure 18: Baddeley's Model of Working Memory (A. Baddeley, 2012)

The **Phonological Loop** is a store with limited storage that holds phonological and aural information in the working memory. It consists of the phonological store which holds verbal information for a limited amount of time and the articulatory control process which manipulates and rehearses the verbal information in the phonological store. Over time, older information is forgotten either by interference theory or trace decay theory. The interference theory basically says that older information (memories) is forgotten because of interference and disruption caused by other information (memories) while trace decay theory says that information automatically decays with time without any need for interference. Baddeley's model favors the interference theory (A. D. Baddeley & Mehrabian, 1976), hence older information could be displaced or overwritten in the phonological loop (A. Baddeley, 2012).

The **Visuo-spatial Sketchpad** is a memory store that stores visual and/or spatial information. Digital systems do not differentiate between visual, aural, spatial, numerical, or phonological information as does the human brain but treat all digital information as binary digits of zeros and ones (Gregg, 1998), hence the visuo-spatial sketchpad and phonological loop are modeled as the storage of the working memory. We model them as the cell state for holding digital information.

The **Central Executive** is the decision-maker of the working memory. It highlights what needs

www.manaraa.com

additional explanation or attention, switches between tasks, shares time during multitasking, and drops irrelevant information (A. Baddeley, 2012). It is the most complex part of the working memory and its functions are still being progressively understood. It also works with the episodic buffer to commit information to long-term memory. The central executive plays the role of the forget gate in dropping irrelevant information and the update gate in regulating information moving to the long-term memory (Kropotov, 2016). We model the central executive as the update gate and the forget gate.

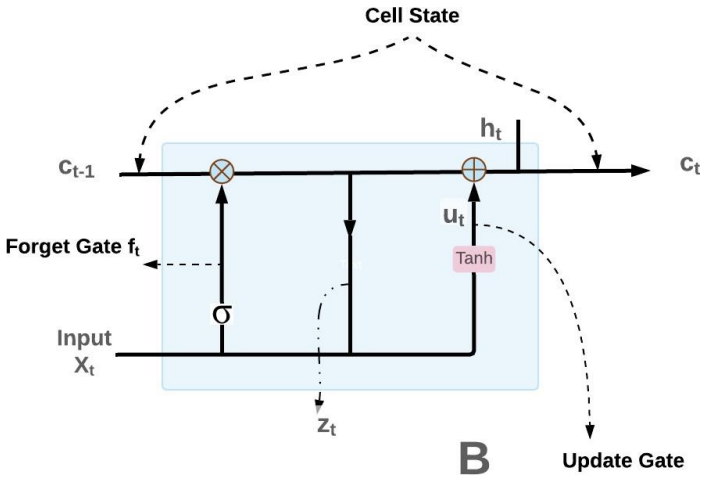


Figure 19: Proposed Working Memory Recurrent Network (WMRN) Cell

Input X_t is new information that flows into the cell. The central executive takes this new information and decides if it is worth keeping. To implement this, we pass the input through a sigmoid activation which reduces the value of the signal to 0 or 1. Zero means the information can be discarded while one means the information should be stored. The update gate takes the new input and previously stored information through a tanh activation and updates the cell state. If necessary, output h_t shows the result of the computation performed by the unit. The signal $c_t = h_t$ carries forward to the next unit.

www.manaraa.com

$$f_t = \sigma\left(w_f * [x_t] + b_f\right) \quad (8)$$

$$z_t = \sigma\left(w_z * [c_{t-1}, f_t] + b_z\right) \quad (9)$$

$$u_t = tanh\left(w_u * [z_t, x_t] + b_u\right) \quad (10)$$

$$h_t = c_t = u_t + z_t \quad (11)$$

w_f, w_z, and w_u are the weights, and b_f, b_z, and b_u are the bias vectors. These four equations implement the working memory recurrent network (WMRN) cell.

Why Develop the Working Memory Recurrent Network (WMRN)?

The RNN has some major shortcomings and one of them is the lack of a cell state (Dinesh, 2019; Stackexchange, 2018). The cell state is what enables a recurrent neural network to carry forward and load information learned not only from the immediate previous events but from much earlier events. Although the original recurrent neural network can model sequential events by connecting them in the hidden state using time propagation, every prior step is taken into consideration by the time propagation. Information from the previous cell must interact with the input in the current cell before the cell determines what must be carried forward. Hence, events from earlier time steps are not retained. For ease of understanding, this issue could loosely be compared with two people involved in a dialogic conversation and each person is starting to think from the scratch for every utterance made by the other person. This will make a very awkward situation because every utterance is understood based on the understanding from previous utterances and there is a building up the occurs as the conversation proceeds. Previous utterances are not thrown away because upcoming ones will use them as building blocks in the entire dialogue no matter where they occur in the previous conversations. The WMRN addresses this problem by creating a cell state which is in essence a larger memory that could store and load events that are not always from the immediate past but earlier past events. The forget gate enables these learned past events to be transmitted along the network without modification while the update gate could update learned events with inputs from the current layer.

RNN is further challenged by the chain rule that is applied in the hidden state. Although in general, RNN is considered to be very successful, it is plagued by the vanishing gradient problem because every previous step is considered in the calculation of the backpropagation, and gradient from the

www.manaraa.com

backpropagation computation is unable to reach those early events because they are near zero or less than zero. In the WMRN, the cell state acts as an information superhighway for the gradients to flow better to capture events that happened early in the sequence, hence solving the long-term dependency problem with RNN.

Solving the long-term dependency problem is very crucial in speech emotion recognition. In some speech data, an utterance could take very long and it is important that in analyzing speech with long utterances, the emotions learned at the beginning of the utterance are remembered all the way to the end of the utterance. As a result, RNN will perform poorly in speech emotion recognition tasks that involve recognizing emotions from long utterances. The working memory recurrent network (WMRN) is very suitable in such cases and also in principle with every problem that requires learning long-term dependencies such as demand forecasting, stock price prediction, credit fraud detection, and solving time series problems.

4.5 INSTANTIATION OF FRAMEWORK

The framework is instantiated using a web browser at the front end and streamlit at the back end. Though there is an option to use a different platform and deploy the framework as a traditional application, however, we opted to use the web-based application because of the flexibility it offers to users. The burden of deploying the application in each client machine is avoided and also the operating system (OS) platform dependency is removed as the application runs on a browser irrespective of the operating system. Furthermore, updates made in the application and bug fixes will be made at a central point where the application is deployed rather than performing updates on each individual machine. The web application option is also chosen because it is more easily adaptable to mobile applications. The only downside is that when deployed as a web application, it cannot be used offline. The user must be online to use it and this limits its use in climes that have extensive internet penetration.

The streamlit backend uses python and librosa to process the uploaded audio file and pass the output to the saved model that is loaded on streamlit for prediction. The artifact is currently hosted on a local machine but could be uploaded on a web hosting service so that users who wish to test the artifact can access it.

المنارة للاستشارات
www.manaraa.com

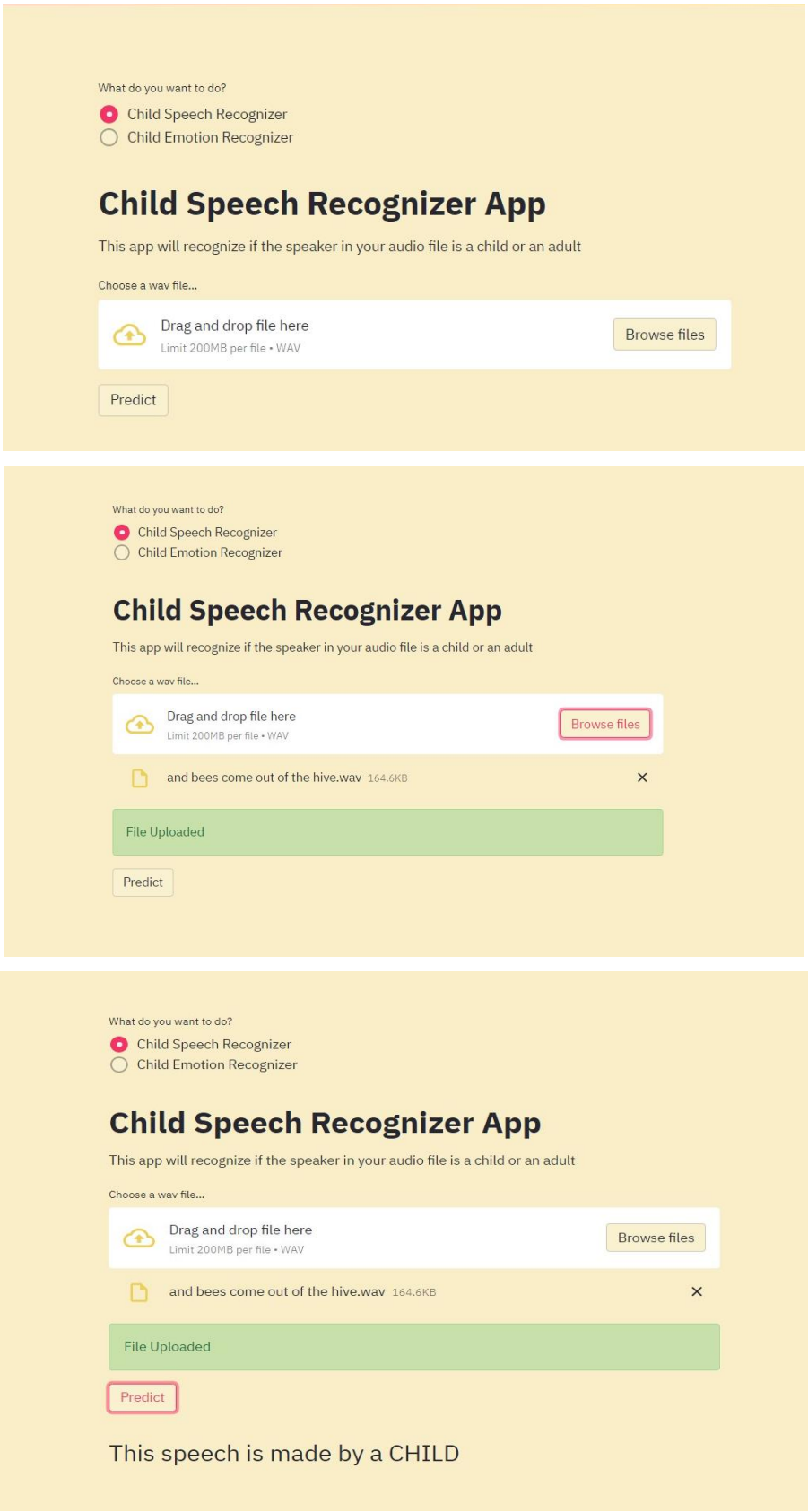
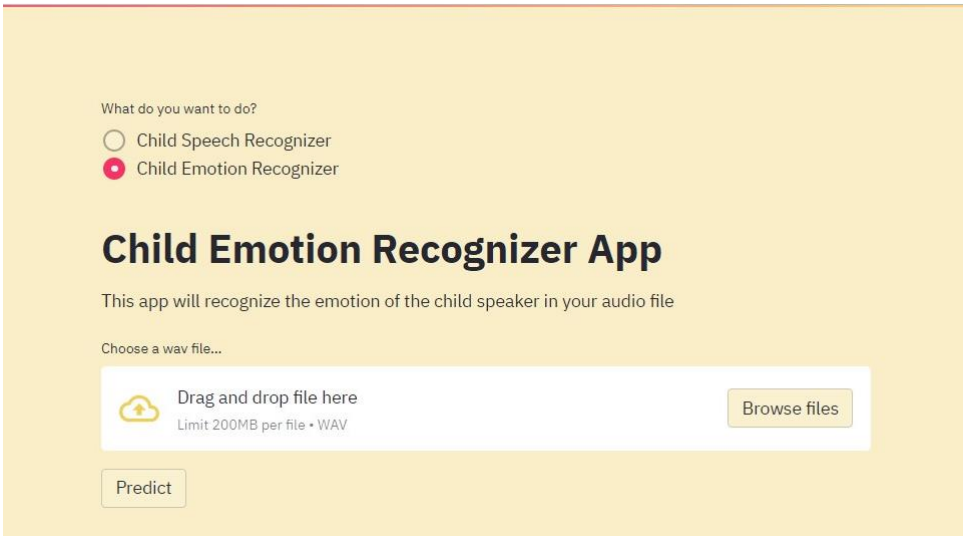


Figure 43: Child Speech Recognizer Web Interface

www.manaraa.com

In the framework, the codetermination component is matched with the child speech recognizer in the instantiation and the affectation component is matched with the child emotion recognizer in the instantiation. It is better to use a term that people can easily understand rather than using the term codetermination and affectation which can only be understood by someone who has read the entire dissertation. For the child speech recognizer, the user wants to know if the speaker in a speech audio file is an adult or a child and will use the radio button to select the child speech recognizer option. The codetermination components are activated with that selection. The user will use the browse button to navigate to the location of the file and select a .wav audio file. When the file is uploaded, the file uploaded message is displayed and the file name and file size is also displayed. When the user clicks 'Predict' the system analyzes the audio file and predicts the speaker as a child or an adult.



www.manaraa.com

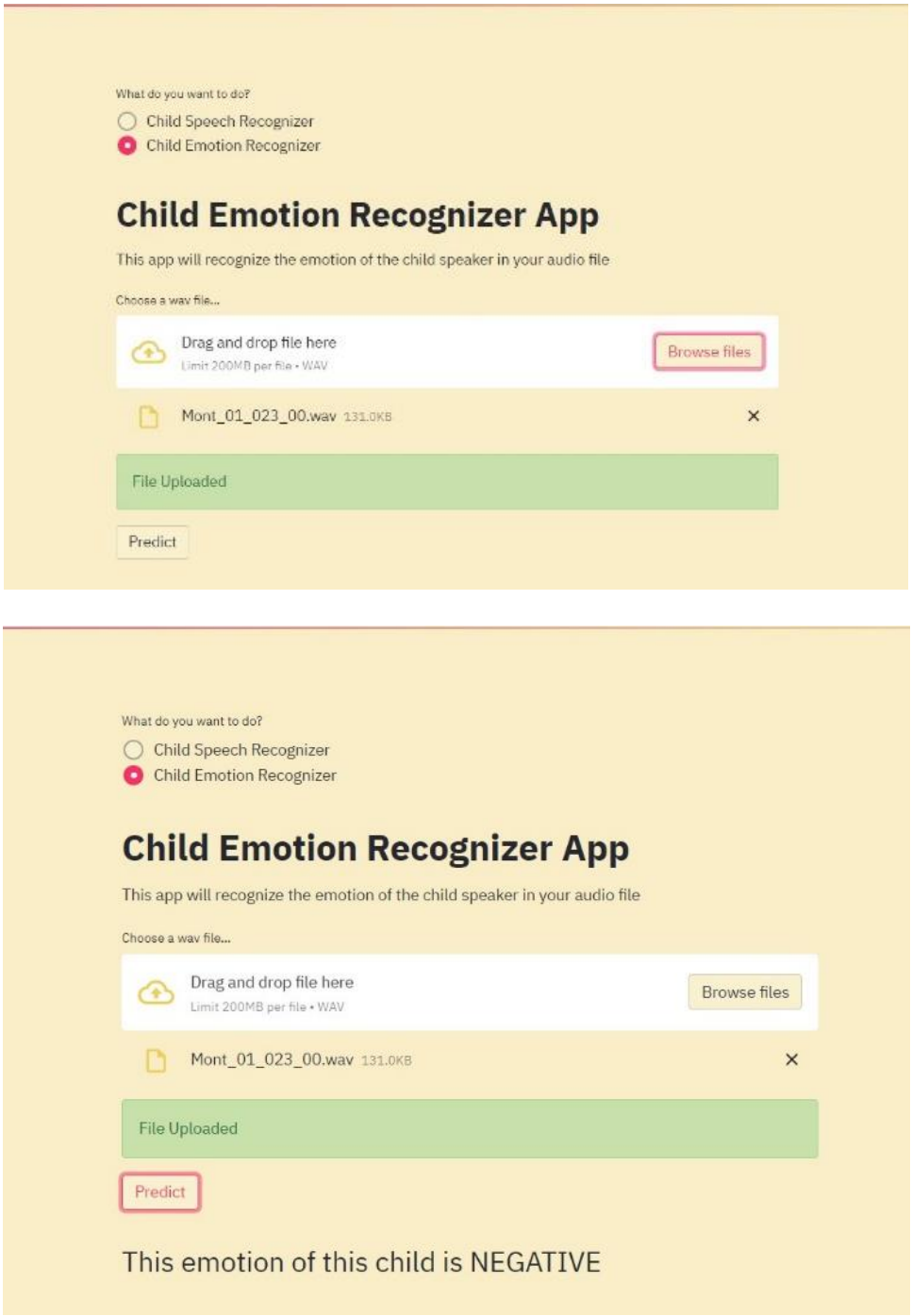


Figure 44: Child Emotion Recognizer Web Interface

www.manaraa.com

When the child emotion recognizer radio button is selected, the affectation component of the framework is loaded and active. The user will use the browser button to navigate to the location of the speech audio file that has been recognized as belonging to a child speaker. When the file is uploaded, the file name, the file size, and a message are displayed showing that the file is successfully uploaded. On clicking the 'Predict' button, at the backend, librosa will read the audio file as an array and a duration of 0.5 seconds is taken from the file with an offset of 0.1 seconds. The dimension of the data is adjusted to fit the input of the deep learning model which will then make a prediction and the prediction is displayed on the interface.

www.manaraa.com

CHAPTER 5

EVALUATION

In this chapter, we present experiments that we have conducted to evaluate the performance of the proposed artifacts to meet the research objectives.

5.1 Experiment 1: Identification of the Best Features for Speaker Age Group Recognition

Description

In addressing RQ1, the objective of this experiment is to determine which speech feature is best to use for speaker age group recognition. This task is very crucial as it forms the foundation upon which other experiments can be performed. Its purpose is to recognize that the speaker in a given speech data is a child from a mixture of speech datasets containing speech from people of various age groups. This is a complex process as several other characteristics are embedded in a speech signal, making speaker age group recognition a very challenging problem (Qawaqneh et al., 2017). In essence, the experiment is to determine the features extracted from the speech signal that will yield the best recognition of a child speaker from an adult speaker. The features are extracted from the speech using Python and Librosa library. The extracted features are fed into the CNN + LSTM network. The CNN + LSTM is used because of its popularity and performance from the extant literature. Speech features used for the token clusters include MFCC, ZCR, and Energy (Palo et al., 2017), and others.

Three datasets are used for the experiment. They are the English Children Dataset (Kennedy et al., 2017) and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset (Livingstone & Russo, 2018), and FAU Aibo Emotional Speech corpus (Steidl, 2009). Other studies that have used these datasets include (Cunningham et al., 2018; Sönmez & Varol, 2020; Yadav & Vishwakarma, 2020). These three datasets are used because the English Children dataset and the FAU Aibo Emotion Speech corpus contain only child speech and the RAVDESS dataset contains only adult speech. In the experiment, we merge these three datasets and we randomize the merge, resulting in one batch of dataset containing adult speakers and child speakers that is fed to the system to recognize child speakers.

www.manaraa.com

English Children Dataset – This dataset consists of 556 spontaneous speech utterances from eleven (11) children who took part in the collection of the data. They include five female children and six male children all of whom were residents in the UK at the time of data collection. Even though the individual ages of the children were not disclosed, their average age is 4.9 years and they were all elementary school children. They have proficiency in speaking English that is appropriate to their ages. The dataset includes single-word utterances and multi-word utterances. The single-word utterances were a reciting of numbers from 1 to 10 and the multi-word utterances were simple sentences from a picture book such as 'the dog is in front of the horse'. All audio files were recorded using the lossless .WAV audio format with a minimum sampling rate of 44kHz. Each speech sample is recorded with three instruments – a studio-grade microphone, a portable microphone, and an Aldebaran NAO robot.

RAVDESS Dataset – The RAVDESS dataset is a widely used database of emotional speech and song. Only the speech portion of the dataset was used for this experiment. It consists of 1440 audio files. The dataset has 24 professional adult actors (12 males and 12 females) who acted lexical statements using North American English. The Age range of the 24 actors is 21 – 33 years and the average age is 26 years.The emotional expressions acted in the speech dataset were calm, happy, sad, angry, fearful, surprise, and disgust. Each utterance was recorded at dual levels of emotional intensity (normal and strong) and a neutral expression. The audio files were recorded in 16bit .WAV format with a sampling rate of 48kHz. Like the English Children dataset, the utterances were also simple sentences such as 'kids are talking by the door'.

FAU AIBO emotional speech corpus - The FAU AIBO emotional speech corpus is used for this experiment. FAU is the acronym for Friedrich–Alexander University which is a public university in Erlangen and Nuremberg in Germany. This dataset contains the recording of 51 children who were interacting with a Sony pet robot. The gender distribution of the children is 21 male and 30 female and the age range is from 10-13 years. The dataset was recorded in two different schools and contains 9959 training speech files and 8257 testing files with an average speech length of 1.7 seconds.

www.manaraa.com

Experimental Setup

For this experiment, we create the three datasets as individual data frames and labeled the speech files from the English Children dataset and FAU Aibo dataset with 'C' denoting child and all the files from the RAVDESS dataset with 'A' denoting adult. In other to maintain a class balance between adult classes and child classes, we used about 1000 data samples from the FAU AIBO dataset. The three datasets were merged into one. The merging of the three datasets was randomized to ensure a uniform random distribution of the speech files. Mel Frequency Cepstral Coefficient (MFCC), Spectral Roll Off, Zero Crossing Rate (ZCR), Energy, Spectral Centroid, Chroma, and Spectral Bandwidth were abstracted from the speech files and fed into the deep learning model. These speech features are the most popular speech features in speaker age recognition research. We build the deep learning model using Keras with TensorFlow backend. For this experiment, our model is trained using the RMSprop optimizer with learning rate set to 0.001, and batch size set to 16. A dropout of 0.5 was used and RELU activation was used on all the layers of the network and kernel initializer was set to 'he_uniform'. We opted for max-pooling layers in between the convolutional layers to downsample the spatial dimension of the network and to minimize the size and computation in the network. Our hardware is a single core NVidia Tesla T4 GPU card with 16GB of RAM and our disk size is 110GB. The CPU is an Intel Xeon (R) running at 2.2GHz with two threads per core. We ran the entire model on GPU only.

Experiment 1 – Results

Accuracy

The accuracy data from the experiments to determine the best performing speech feature for recognizing child speakers from a mix of child and adult speakers is presented in Figure 20 and Table 2. Accuracy for all the experiments is defined as;

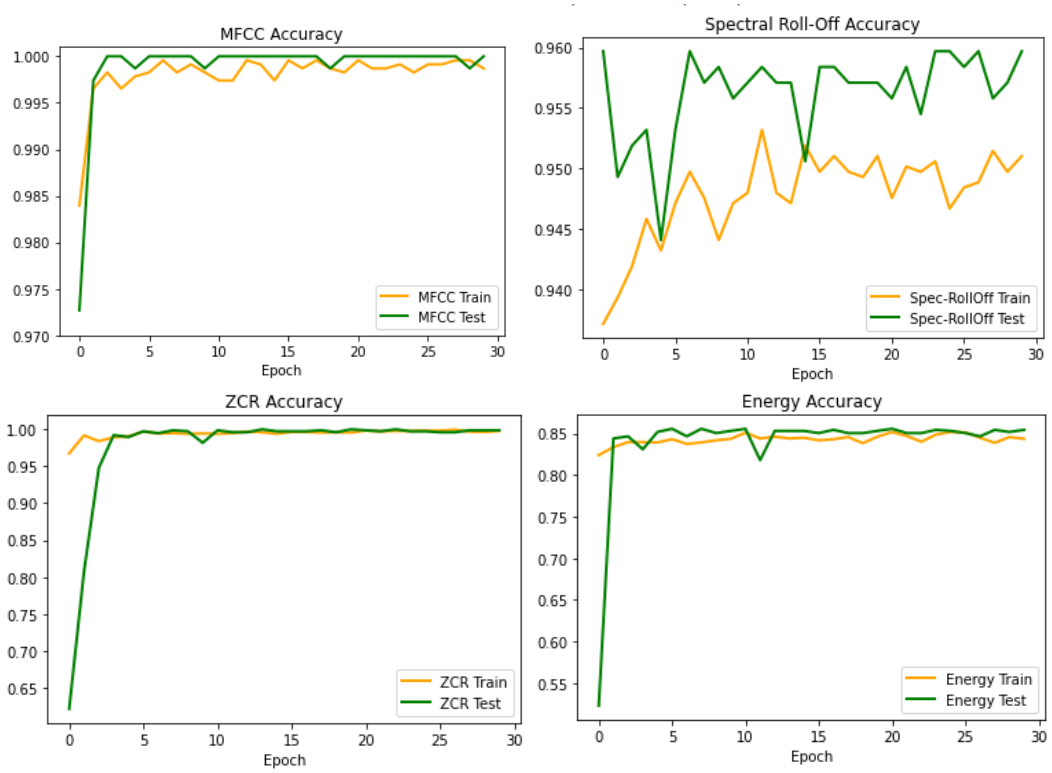
$$Accuracy = Number\ of\ Correct\ Predictions\ /Total\ Number\ of\ Predictions$$

$$Accuracy = (TP + TN)/\ (TP + TN + FP + FN)$$

Where TP is true positives, TN is true negatives, FP is false positives and FN is false negatives. The total number of data samples used in the experiment is 3076 and the classes are fairly balanced with child samples numbering 1636 and adult samples 1440. Figure 20 shows the training and

www.manaraa.com

testing accuracy plot for all the speech features. The Y-axis is the accuracy values and the X-axis is the epochs. The accuracy plot for MFCC, ZCR, Spectral Centroid, and Spectral Bandwidth was reasonably smooth through the entire training process while the accuracy plot for the Spectral roll-off was jumpy and jagged. The reason could be that the model was having difficulties in learning patterns from the Spectral roll-off data. Thirty epochs were chosen because the model is able to converge at this point. Convergence is often denoted by the plateau in the accuracy plot. From the results obtained, it is evident that the Mel frequency cepstral coefficient (MFCC) is the best-performing speech feature for our speaker age recognition deep learning model, achieving a 100% testing accuracy on the 25% hold out and 99.99 on the 4-fold cross-validation. The chroma features was the worst performing feature achieving 71.78% accuracy on the test dataset and 71. 29% on cross-validation.



www.manaraa.com

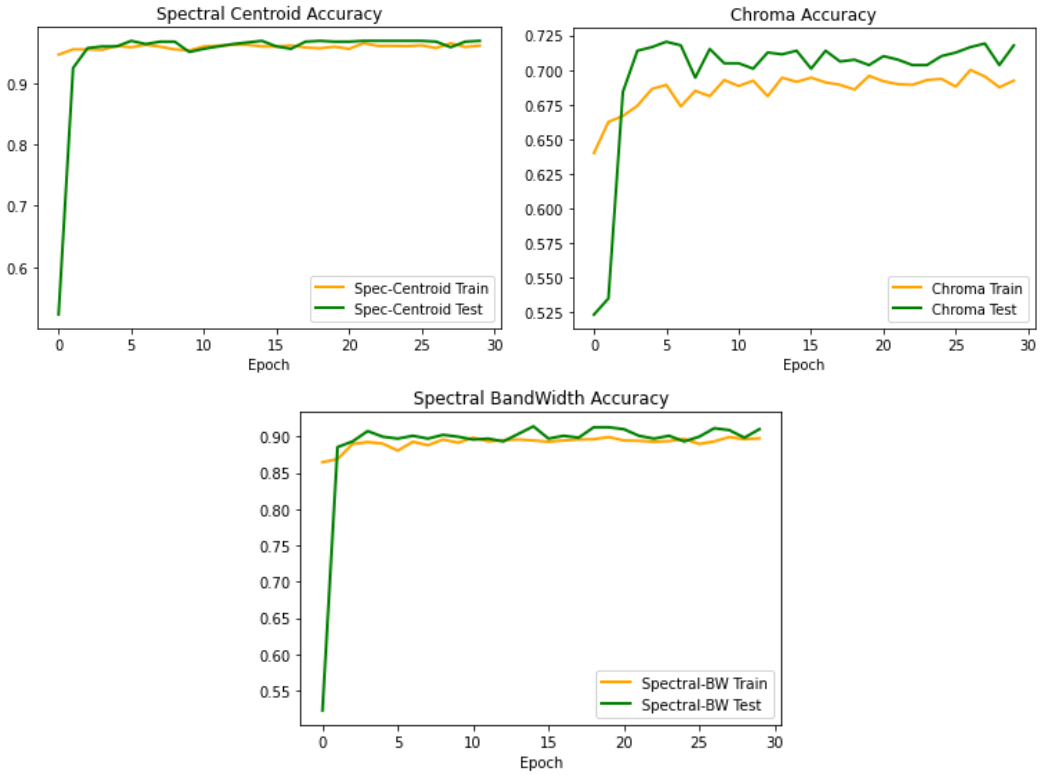


Figure 20: Accuracy Plot of Speech Features for Speaker Age Group Recognition

S/N	Speech Feature	4-fold Cross Validation (%)
1	MFCC	**99.99**
2	Spectral Roll Off	95.25
3	ZCR	99.84
4	Energy	84.87
5	Spectral Centroid	96.61
6	Chroma	71.29
7	Spectral Bandwidth	90.21

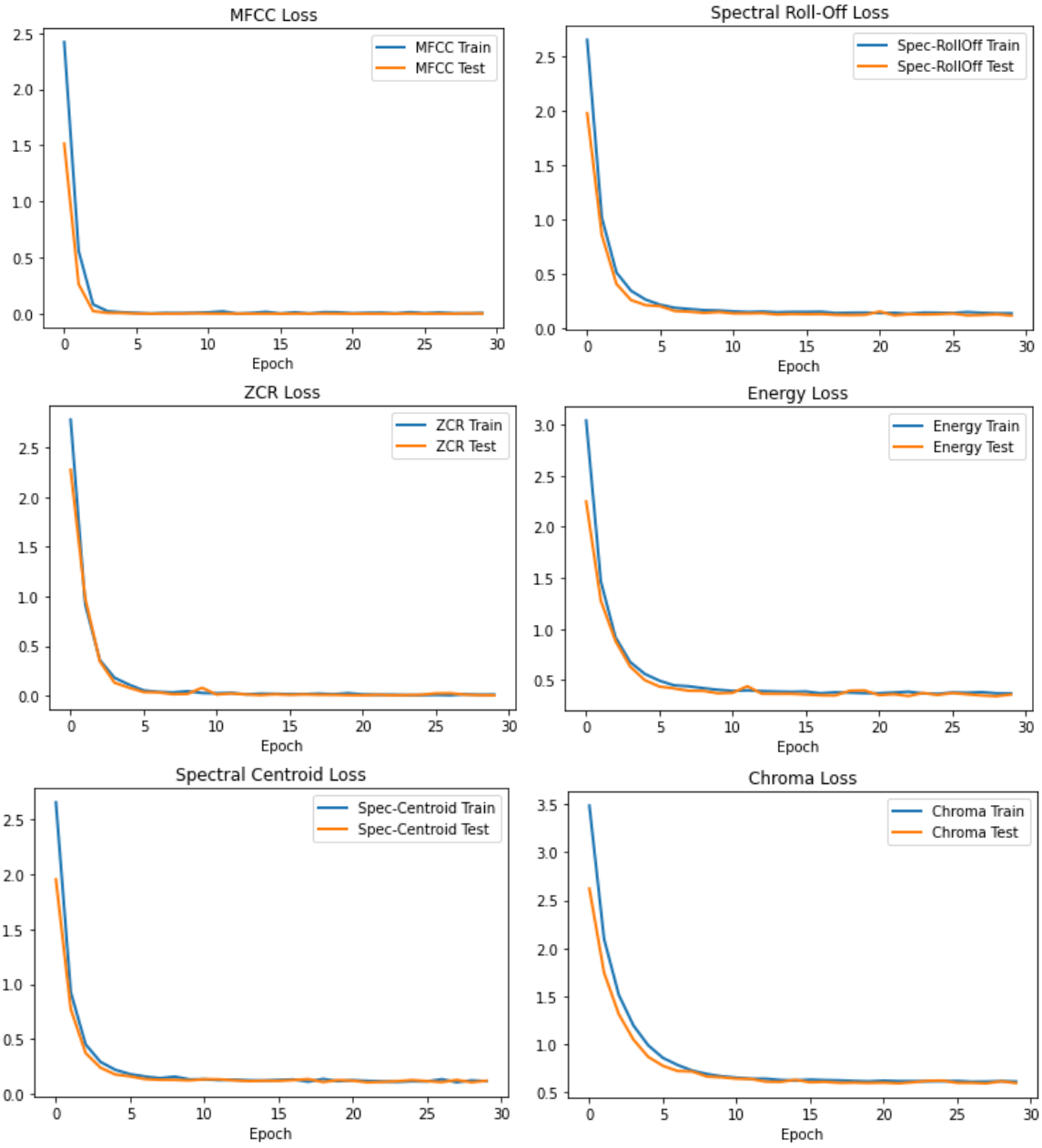
Table 2: Speech Features for Speaker Age Group Recognition with Training and Testing Accuracies

Loss

Fundamentally, the loss function measures how well a deep learning model performs with respect to predicting the expected outcome. When the loss function is used as a metric for performance, the objective is to diminish the loss function's value with respect to the model's parameters by updating the weight vector values through each forward pass. A lower score implies better performance and the model optimizer will attempt to reduce the loss values to a perfect score of zero. The results obtained from the experiments when using the loss function as a measurement

www.manaraa.com

of performance are shown in Table 3 and Figure 21. The model trained with Mel frequency cepstral coefficient (MFCC) features gave the best performance for speaker age recognition while the model trained with chroma features gave the worst performance. The loss value at epoch 1 for MFCC is the least for all the features tested while the loss value at epoch 1 is largest for the chroma features at 3.0 for the line plot of training loss and 2.6 for that of the testing loss.



www.manaraa.com

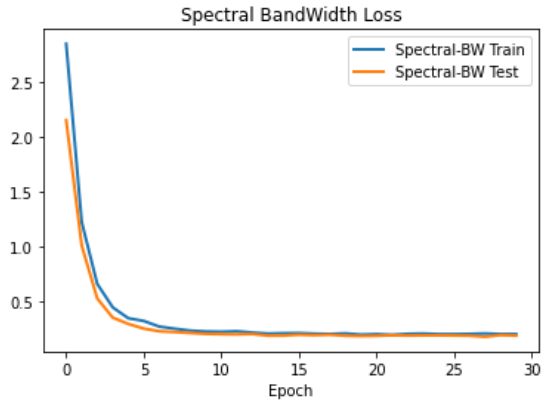


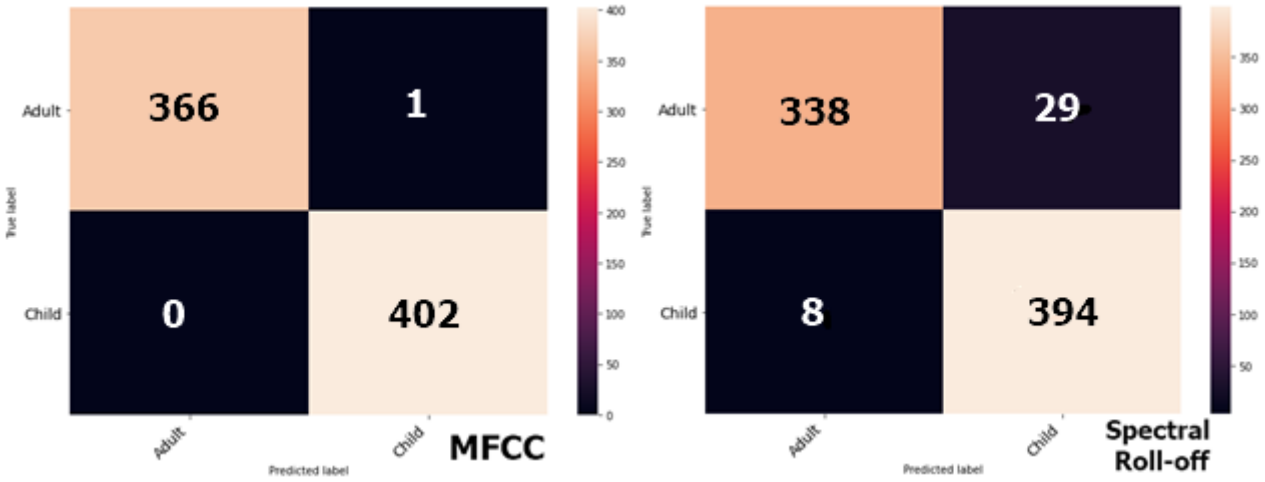
Figure 21: Loss Plot of Speech Features

S/N	Speech Feature	4-fold Cross Validation Loss
1	MFCC	**0.0017**
2	Spectral Roll Off	0.1249
3	ZCR	0.0057
4	Energy	0.3622
5	Spectral Centroid	0.1315
6	Chroma	0.6054
7	Spectral Bandwidth	0.2024

Table 3: Speech Features with Training and Testing Losses

Confusion Matrix

Confusion matrix is a technique that is often employed to access the performance of a model. It summarizes the classification performance of a prediction algorithm.



www.manaraa.com

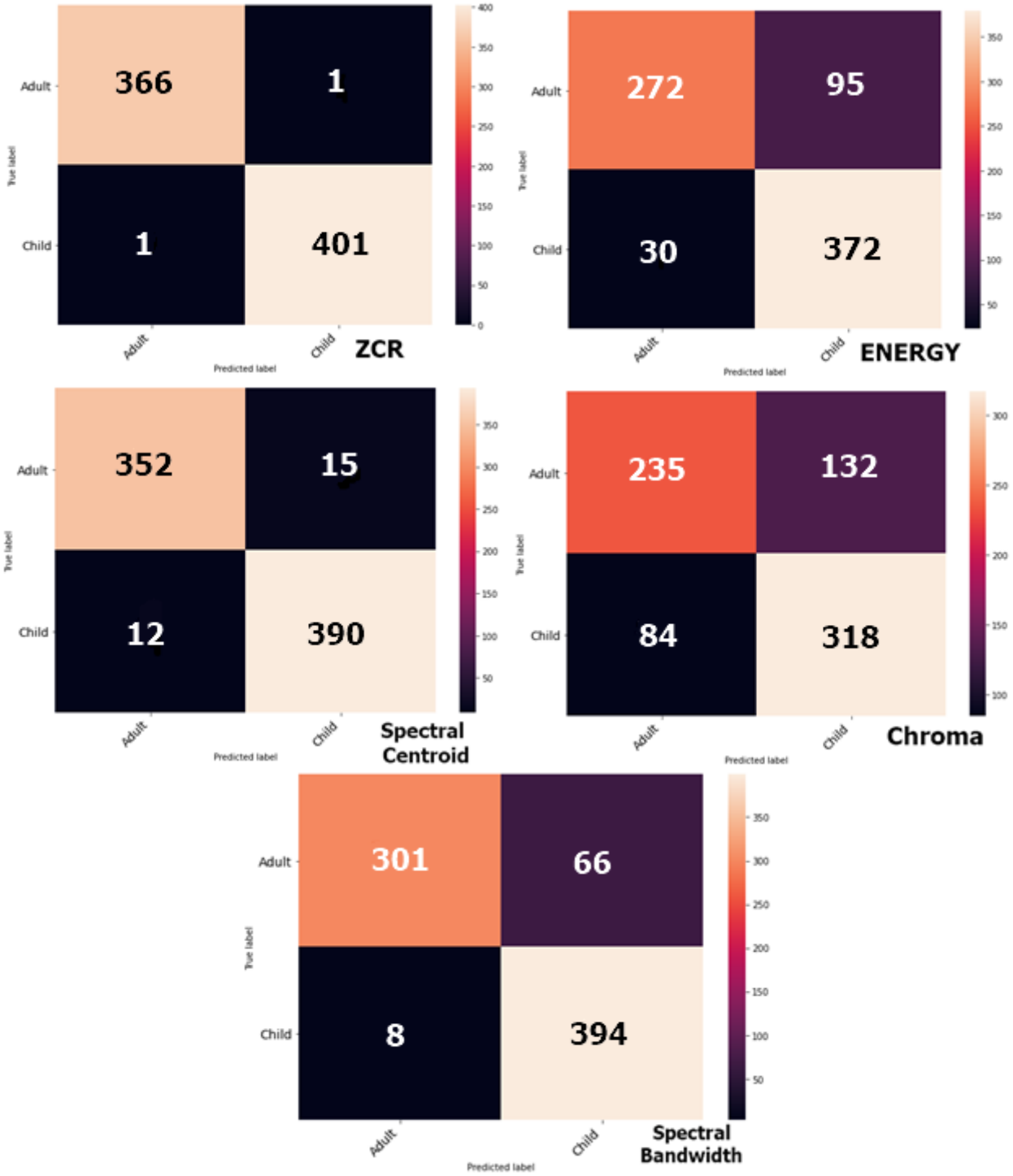


Figure 22: Confusion Matrix of Speech Features

It has the ability to show how the prediction model is 'confused' about making predictions. It overcomes some limitations of classification accuracy by showing the class that the algorithm makes the most errors in and the class where it is performing well. This information cannot be obtained by looking at accuracy alone.

www.manaraa.com

Figure 22 shows the confusion matrix for all the speech features used in the experiments. The model trained with Mel frequency cepstral coefficient (MFCC) feature achieved the best prediction of all the classes for the test dataset. The adult classes were all correctly classified with zero misclassification while the child classes had one misclassification. The ZCR features also performed well with one misclassification each for the child and adult classes. The chroma feature was the worst performing feature among all the features tested, with several examples being misclassified from both the child and adult classes.

Precision, Recall and F-Measure

Precision, Recall, and F-measure is some of the most widely used performance metrics for artificial intelligence models. Precision measures the predicted number of a class that actually belongs to that class, recall computes the predicted number of a class out of the total number of that class, and f-measure balances out both precision and recall. Mathematically;

$$Precision = True\ Positives/(True\ Postives + False\ Positives)$$
$$Recall = True\ Positives/(True\ Positives + False\ Negatives)$$
$$F - measure = 2 * Precision * Recall/(Precision + Recall)$$

Features	Classes	Precision	Recall	F-Measure	Support
MFCC	Adult	0.99	0.99	0.99	367
	Child	0.99	0.99	0.99	402
Spectral Roll Off	Adult	0.98	0.92	0.95	367
	Child	0.93	0.98	0.95	402
ZCR	Adult	0.99	0.99	0.99	367
	Child	0.99	0.99	0.99	402
Energy	Adult	0.90	0.74	0.81	367
	Child	0.80	0.92	0.86	402
Spectral Centroid	Adult	0.97	0.96	0.96	367
	Child	0.96	0.97	0.96	402
Chroma	Adult	0.74	0.64	0.68	367
	Child	0.71	0.79	0.74	402
Spectral Bandwidth	Adult	0.97	0.82	0.89	367
	Child	0.86	0.98	0.90	402

Table 4: Precision, Recall and F-Score for Speech Features for Speaker Age Group Recognition

www.manaraa.com

Table 4 presents the precision, recall, and f-measure obtained for the speech features. From the results, MFCC and ZCR have near-perfect scores of 0.99 for the three measures. The model trained with chroma features gave the worst performance. These results are consistent with the results obtained from using accuracy and loss plots as metrics for measuring performance.

5.2 Experiment 2: Emotion Recognition

In addressing RQ2, the objective of this experiment is to determine the best features for child speech emotion recognition. In chapter one, the point was made that audio features that work well for speaker age group recognition do not necessarily work well for child emotion recognition. Furthermore, researchers have noted that systems that show good performance for adult speech emotion recognition do not perform well when used for child emotion recognition (Palo et al., 2017). Hence, we cannot directly import the well-researched adult speech emotion recognition solutions into this work.

Dataset:

The FAU AIBO emotional speech corpus is used for this experiment. This dataset contains the recording of 51 children who were interacting with a pet robot called AIBO. The FAU AIBO corpus contains the chunk set and the word set. The word files and the chunk files make up the total files. When a chunk has a word that at least 3 of the 5 labelers agree on the emotion, that word is split out of the chunk and becomes part of the FAU AIBO word set. The word set label is larger than the chunk set label because one chunk could have more than one word where at least 3 of 5 labelers agree on the emotion of the word. The chunk set is made up of those files that contain at least one word that made it into the word set. The dataset specifies that the chunk set is preferable for emotion recognition and we used only the chunk set in our experiment. The total chunk set files used were 4543 audio files.

The network architecture is the affection component of the framework. Different speech features are tested. Metrics used to access performance are accuracy, loss plot, precision, recall, and f-measure, and confusion matrix.

Experimental Setup

The experiment is divided into two major runs. In the first run, we used the four classes in the FAU

www.manaraa.com

AIBO emotional speech corpus. The class labels are neutral, emphatic, negative (comprising of angry, touchy, and reprimanding), and joyful. In the second run, the data samples were relabeled into two classes of positive and negative emotions. For each run, MFCC, ZCR, Energy, and other features were abstracted from the speech files and fed separately into the deep learning model. The model was built with python. For this experiment, our model is trained using the RMSprop optimizer with learning rate set to 0.001, and batch size set to 16. A dropout of 0.5 was used and RELU activation was used on all the layers of the network. We opted for max-pooling layers in between the convolutional layers to downsample the spatial dimension of the network and to minimize the size and computation in the network. This is the most commonly used pooling algorithm (Badshah et al., 2019; Zhao et al., 2019). Our hardware is a single core NVidia Tesla T4 GPU card with 16GB of RAM and our disk size is 110GB. The CPU is an Intel Xeon (R) running at 2.2GHz with two threads per core.

Experiment 2 Results – Four Classes

Accuracy

The results obtained while using accuracy as the metric for evaluating performance is shown in Figure 23 and 24. Figure 23 is the accuracy over epochs plot of the training and testing accuracy for all the speech features. Fifty epochs was chosen for this experiment because the model was able to converge at this point and additional epochs did not result in significant performance improvement. The green line is the accuracy plot on the test set while the orange line represents the accuracy plot on the training set. Figure 24 shows the collated plot of testing and training accuracies of speech features. The model trained with Zero crossing rate (ZCR) features gave the best performance for child speech emotion recognition, with 59.02% average accuracy on the test sets followed by the MFCC-trained model which achieved 49.53% average accuracy on the test sets. The worst performing model is the model trained with chroma features which reached an accuracy of 42.81%. While these numbers may not seem short, they are very consistent with the results obtained in the original paper where the FAU AIBO emotional speech corpus was first collated and used (Steidl, 2009).

www.manaraa.com

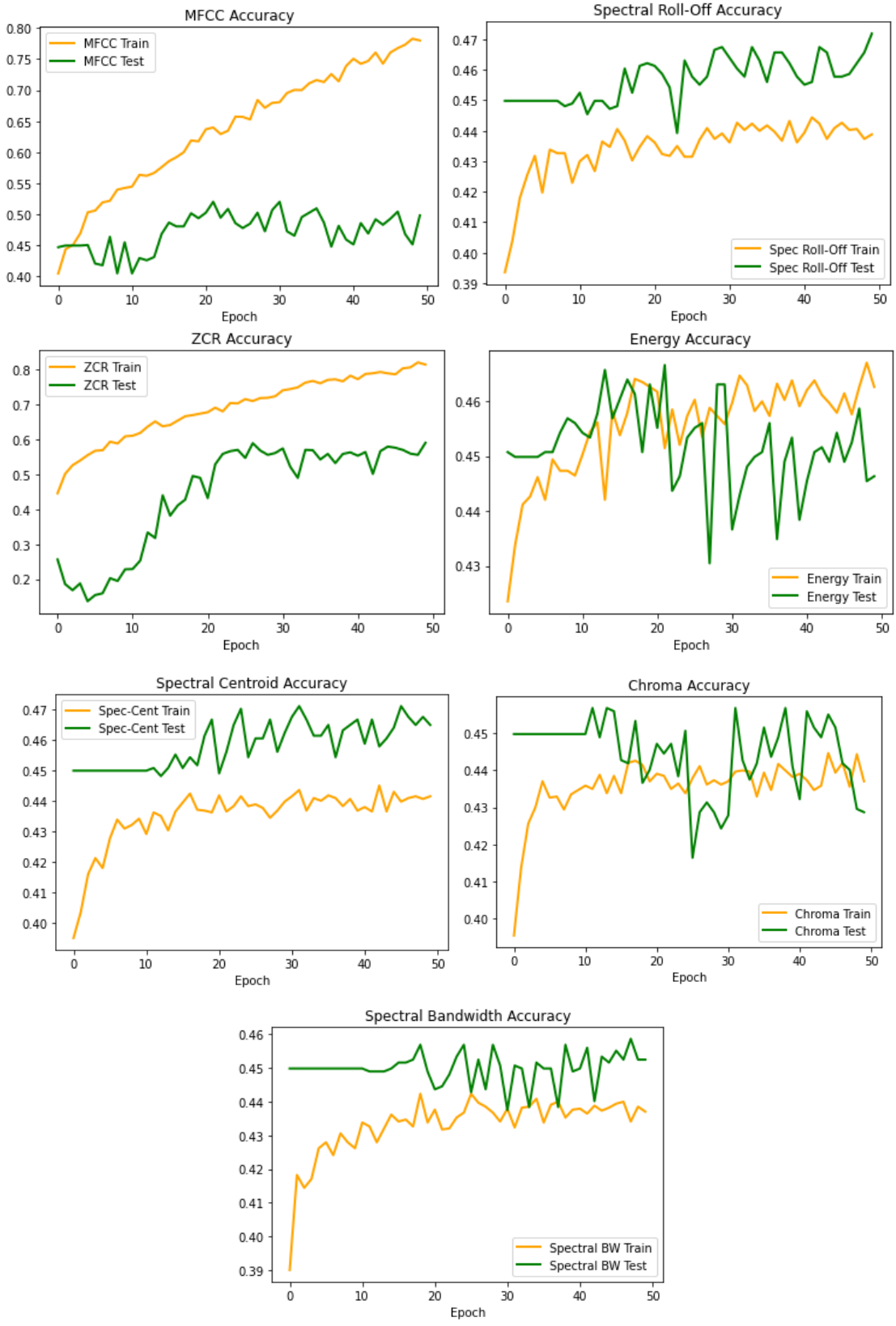


Figure 23: Accuracy Plot of Speech Features for Child Emotion Recognition

www.manaraa.com

The best result the author obtained using energy features was 50.90% and on MFCC features, he obtained 45.50% accuracy on the frame-level analysis. Other researchers that have used this dataset also obtained similar results (Cummins et al., 2017; Deng et al., 2014; Jalal et al., 2019). Jalal et al., 2019 report an unweighted accuracy of 61.80% on the multiclass classification, Deng et al., 2014 report a recall of 51.60%, and Cummins report 39.60%. The model trained with MFCC and ZCR features tend to overfit the training set however, other models trained with the rest of the speech features did not have the overfitting problem.

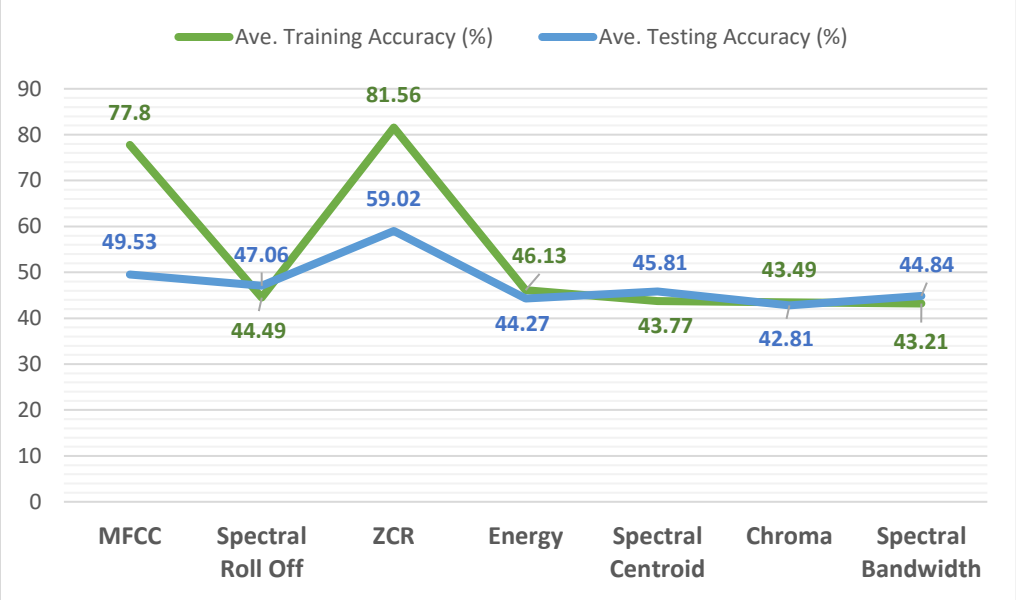


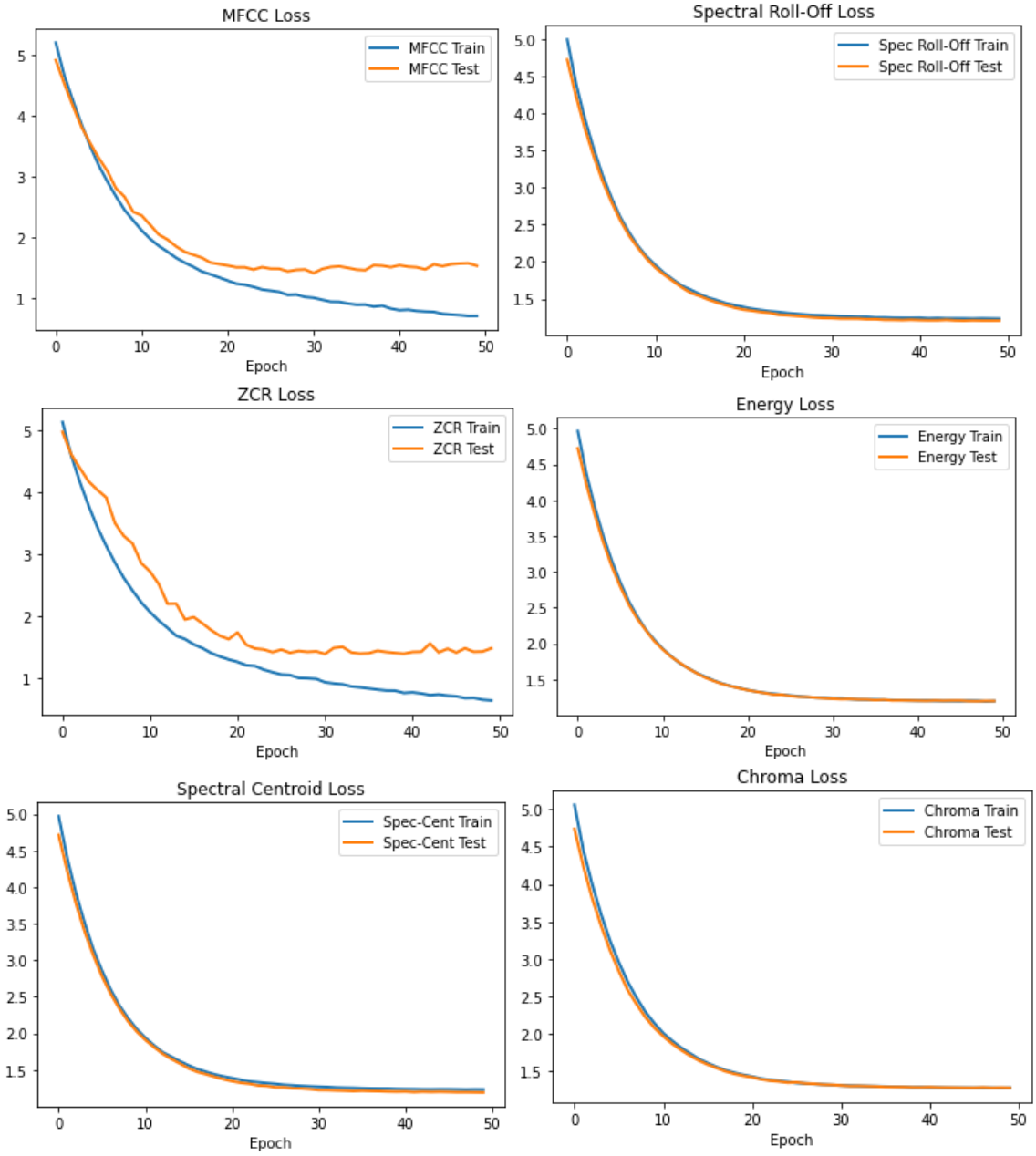
Figure 24: Speech Features for Child Emotion Recognition with Training and Testing Accuracies

Loss Plot

The results obtained for the loss function are shown in Figures 25 and 26. Each plot represents the speech features. The Y-axis is the loss values while the X-axis is the training epochs. Generally, loss plots start at high loss values in the early epochs and steadily decrease towards zero as the model learns the data and makes fewer mistakes. The blue line is the training loss and the orange line is the testing loss. The close fit between the two plots shows that the train-test split is ideal for the dataset. A large gap between the train and test loss plot is usually due to model complexity, regularization issues, and unbalanced test-train split. The model trained with Energy and Spectral Centroid features results in the best outcomes as they were able to minimize the loss in the network more than the other models. In general, the losses were very similar for the MFCC model and the ZCR model. For these two features, the loss plot for the train and test set were pulling

www.manaraa.com

apart as the epoch was increasing. A technique called early stopping (Brownlee, 2018; Keras, 2021) could be applied to stop the training process earlier than the specified number of epochs. However early stopping is usually applied when the model performance stops improving. In this case, the loss value was still improving with training epochs and early stopping could not be applied. The loss plots for spectral roll-off, energy, chroma, and spectral bandwidth were similarly shaped and the least performing model is the model trained with MFCC features.



www.manaraa.com

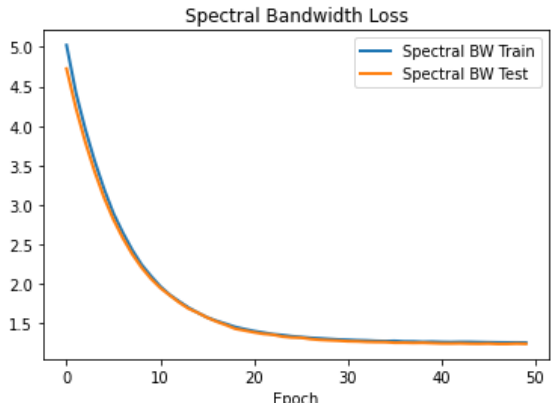


Figure 25: Loss Plot of Speech Features for Child Emotion Recognition

In Figure 26, the collated loss values for the speech features are shown. The Y-axis shows the loss and the X-axis is the speech features. The training and testing loss values for spectral roll-off, energy, spectral centroid, chroma, and spectral bandwidth are very near which is a sign of good model performance. The MFCC and the ZCR loss values are more spaced apart. Spectral centroid obtained the least average loss value of 1.20 on the test set which makes it the best performing features using the loss function minimization metrics. Although the ZCR obtained a lower average loss score of 0.674 on the test sets, priority is given to the average testing loss values as they are more indicative of how the model will perform on field data.

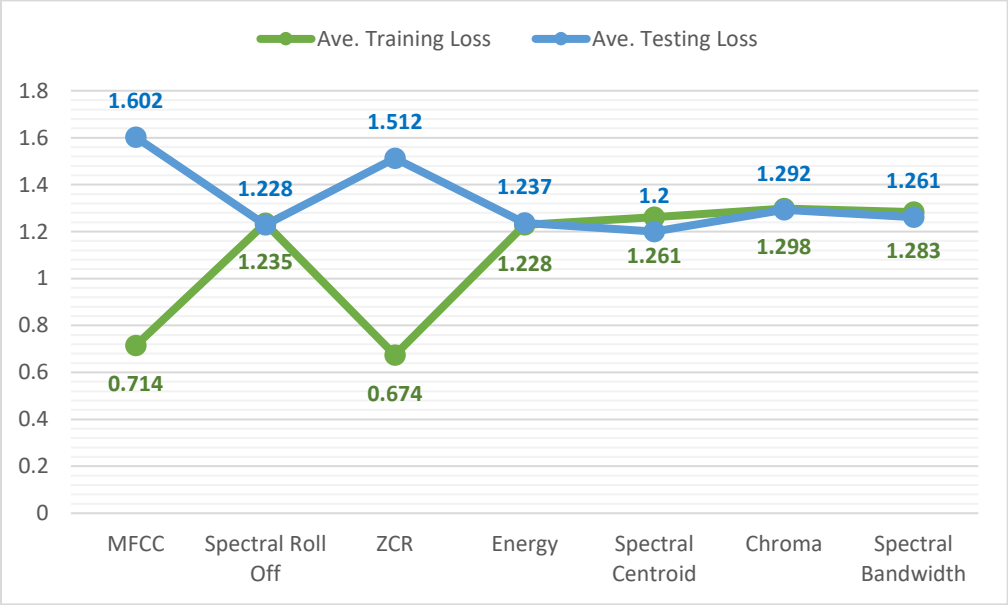
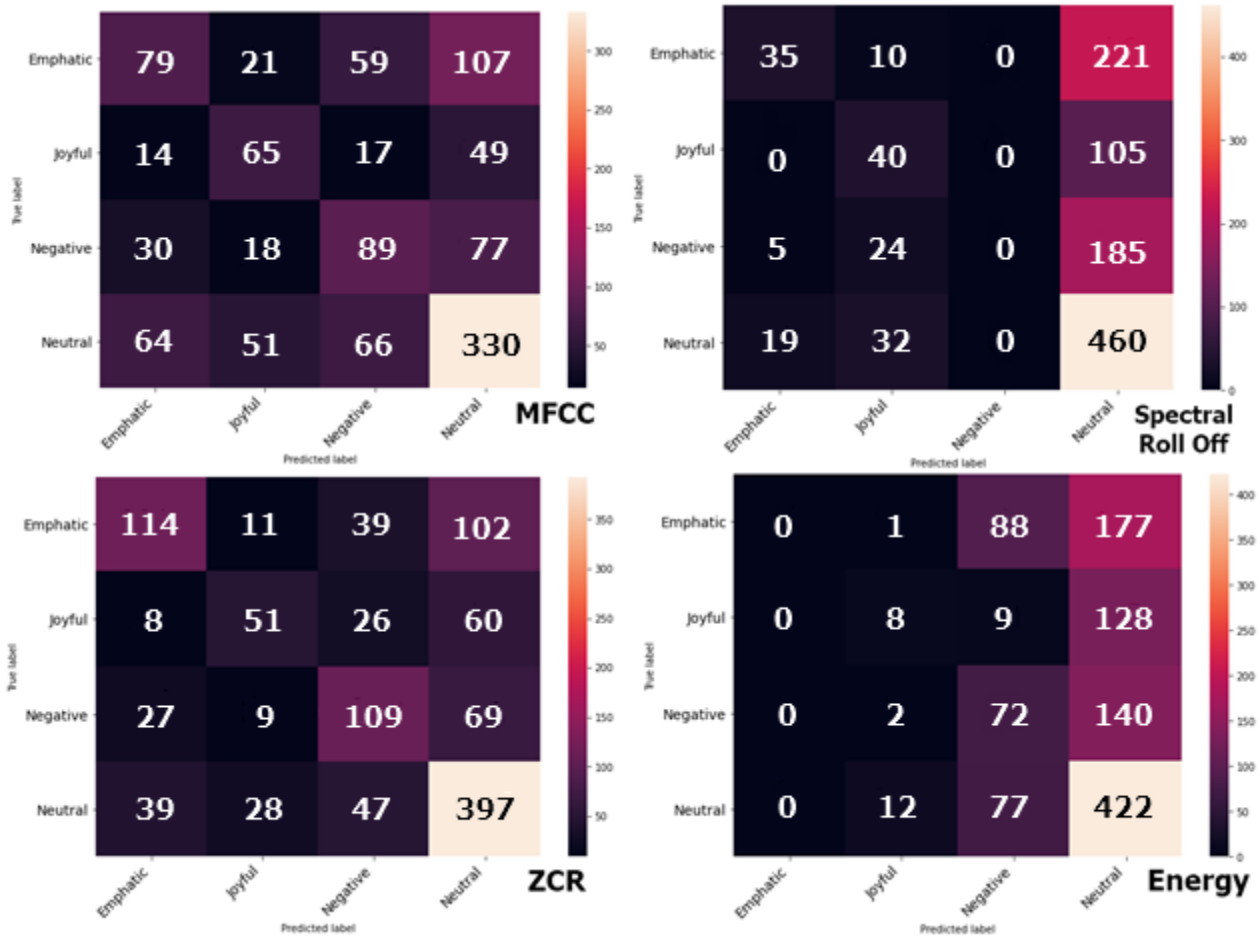


Figure 26: Speech Features for Child Emotion Recognition with Training and Testing Losses

www.manaraa.com

Confusion Matrix

The result for the confusion matrix is presented in Figure 27. With the confusion matrix, one can see exactly which classes the models are performing well on and the classes they are struggling to recognize. Counting the number of instances that were correctly classified by totaling the numbers in the diagonal of each matrix, the ZCR model obtained the best overall results based on the confusion matrix while the chroma obtained the worst results. When considering each individual class, the model trained with ZCR features was also able to correctly recognize more negative emotion data samples than any other model. This result is consistent with the results obtained by using accuracy as the metric for measuring performance.



www.manaraa.com

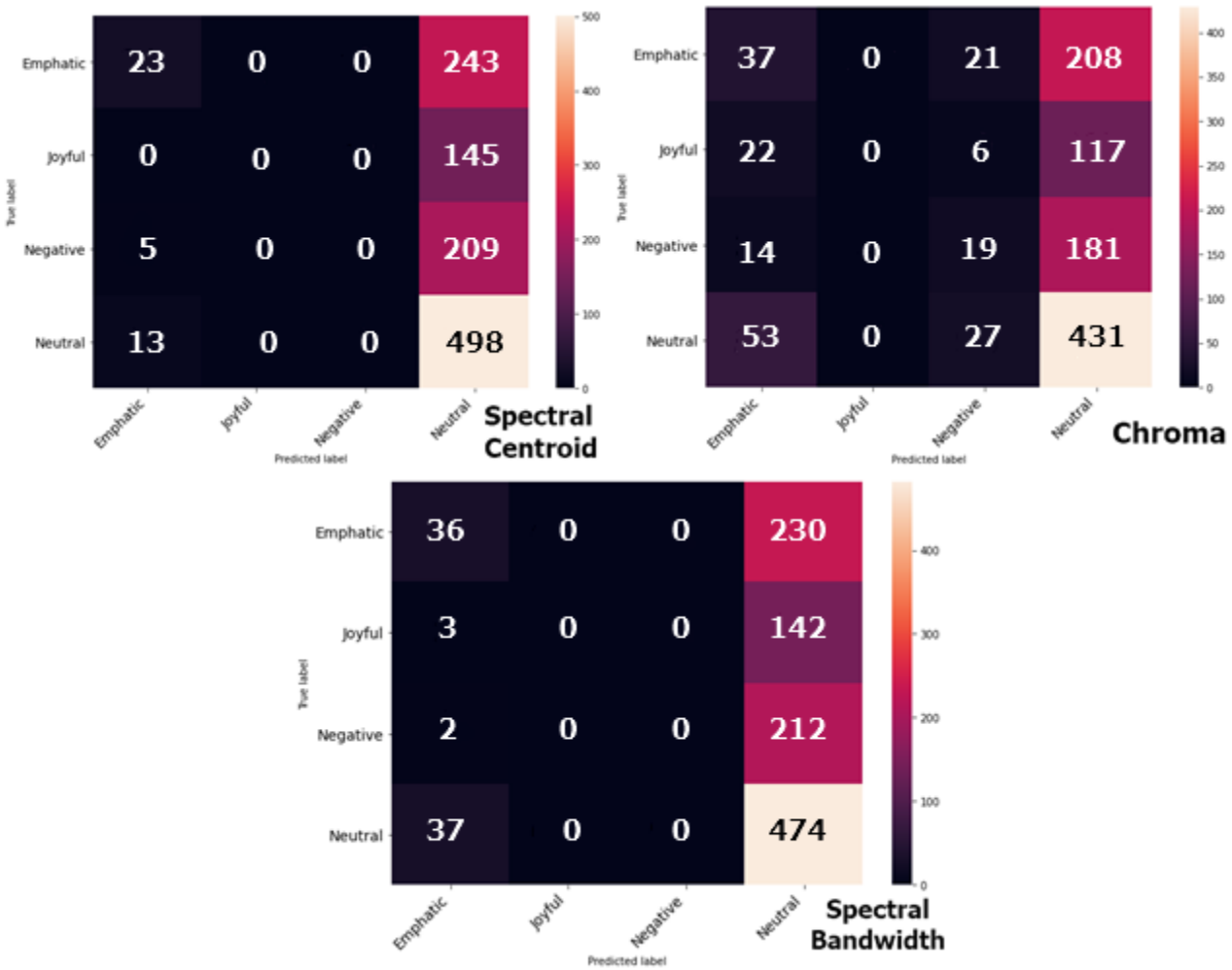


Figure 27: Confusion Matrix for Speech Features for Child Speech Emotion Recognition

Experiment 2 Results – Two Classes

In this experiment, the FAU AIBO emotional speech corpus was relabeled into two classes of positive and negative emotion. The neutral and joyful emotions were labeled as positive emotion and the emphatic emotion was included in the negative emotion. This labeling is consistent with previous research that has relabeled the FAU AIBO corpus into positive and negative emotions (Deng et al., 2014), mapping the neutral emotion as positive and the emphatic emotion as negative. This is the most reasonable way to map the emotions for a positive-negative binary classification. The negative samples were 1959 in total and the positive samples were 2584 in total.

www.manaraa.com

Accuracy

The results obtained for the accuracy evaluation of performance is shown in Figure 28 and 29. The Y-axis represents the accuracy values and the X-axis represents the number of epochs. The test accuracy is colored green while training accuracy is colored orange. Fifty epochs was chosen as the model tends to converge at this point. This is indicated by the plateauing of the test set accuracy plot. The model trained with Zero crossing rate (ZCR) features gave the best performance for child speech emotion recognition, with 70.47% accuracy on the test set followed by the Energy features-trained model achieved 65.65% accuracy on the test set. The worst performing model is the model trained with chroma features which reached an accuracy of 57.54% on the test set.

www.manaraa.com

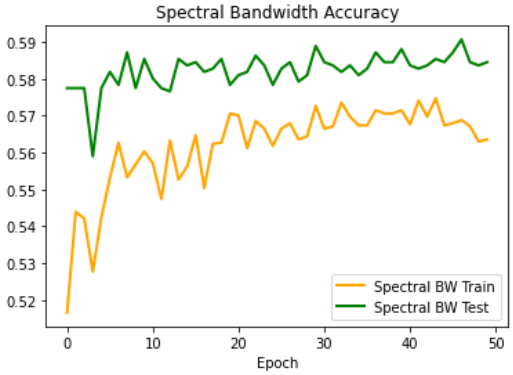


Figure 28: Accuracy Plot of Speech Features for Child Emotion Recognition

As with the four classes experiment, the models trained with ZCR and MFCC tend to overfit the training set. Attempts made to address overfitting by adjusting some training parameters did not result in any significant improvement. However, the model trained with the spectral roll-off, energy, spectral centroid, chroma, and spectral bandwidth did not overfit.

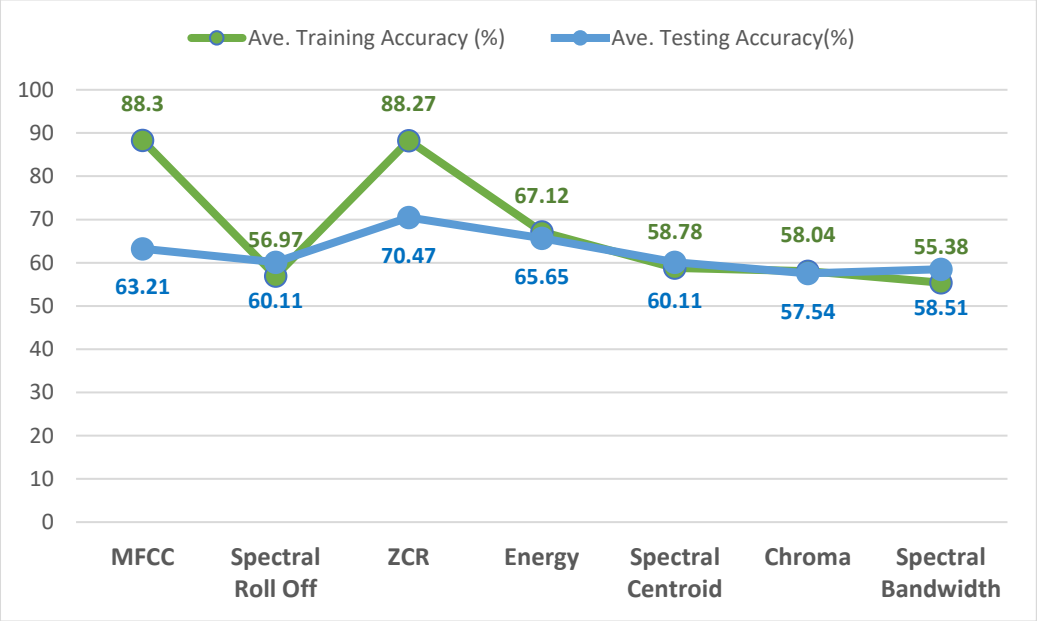


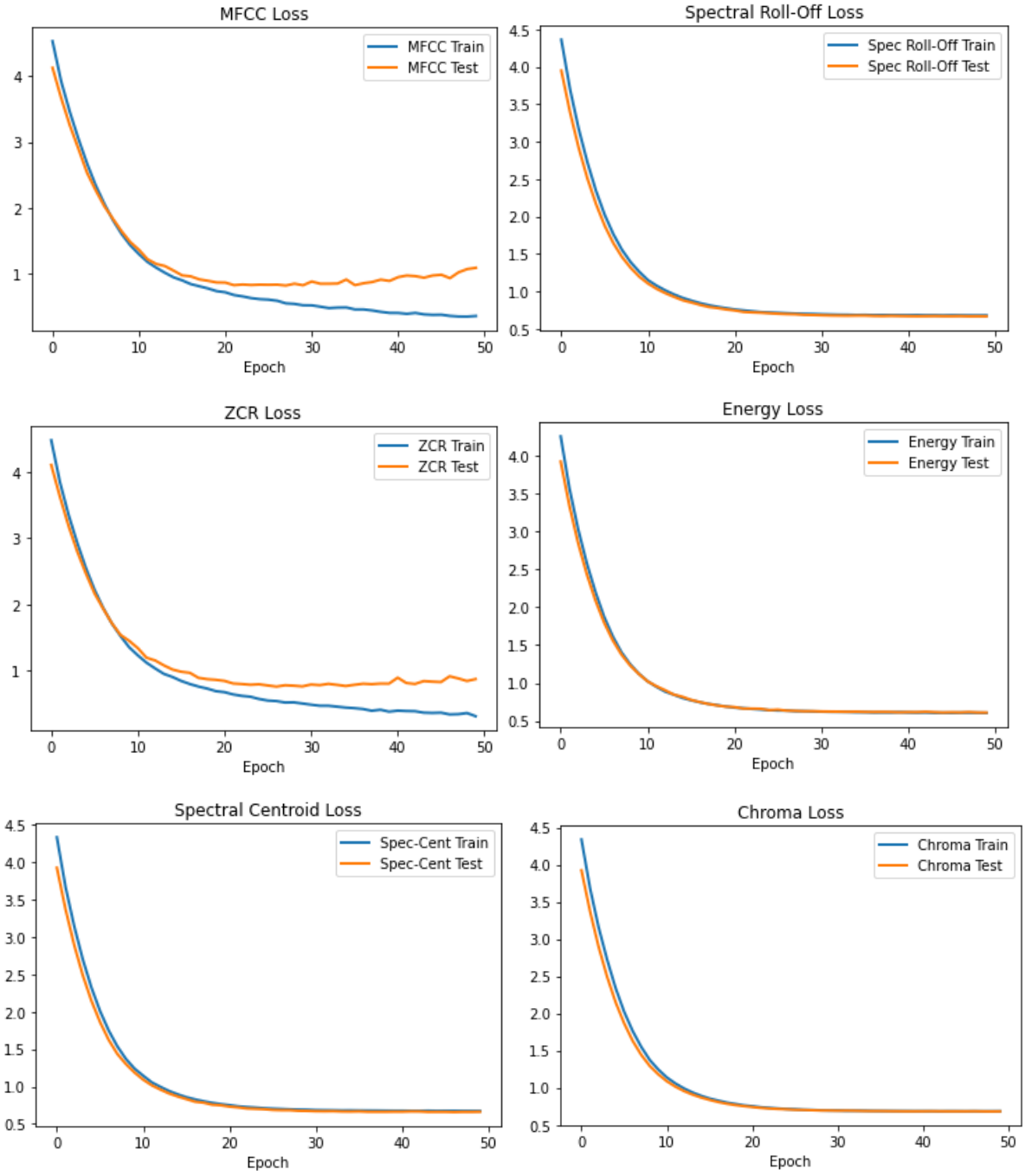
Figure 29: Speech Features for Child Emotion Recognition with Training and Testing Accuracies

Loss Plot

In general, the results obtained for minimizing the loss function for the two-class experiment are similar to those obtained from the four-class experiment. They are shown in Figures 30 and 31. The model trained with energy features showed slightly superior performance by minimizing the loss values in the test set better than the other features. The model behavior with the MFCC and

www.manaraa.com

ZCR features was similar to the results obtained previously. For both features, the gap between the train set and test set loss values were widening with increasing epochs. It seems that ZCR and MFCC converge faster and do not requires as many epochs as the other features. The MFCC and the ZCR features also gave the worst performance in minimizing the loss in the network on the test set.



www.manaraa.com

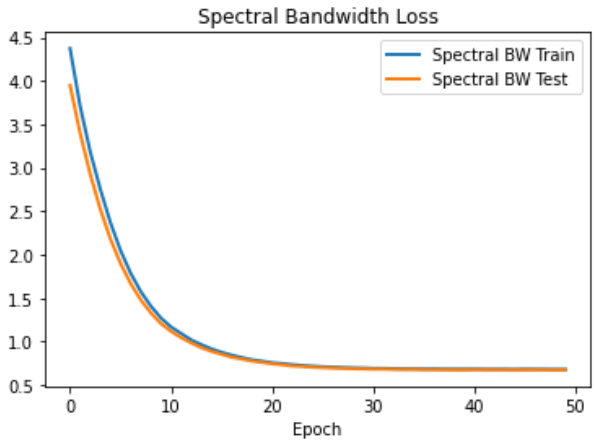


Figure 30: Loss Plot of Speech Features for Child Emotion Recognition

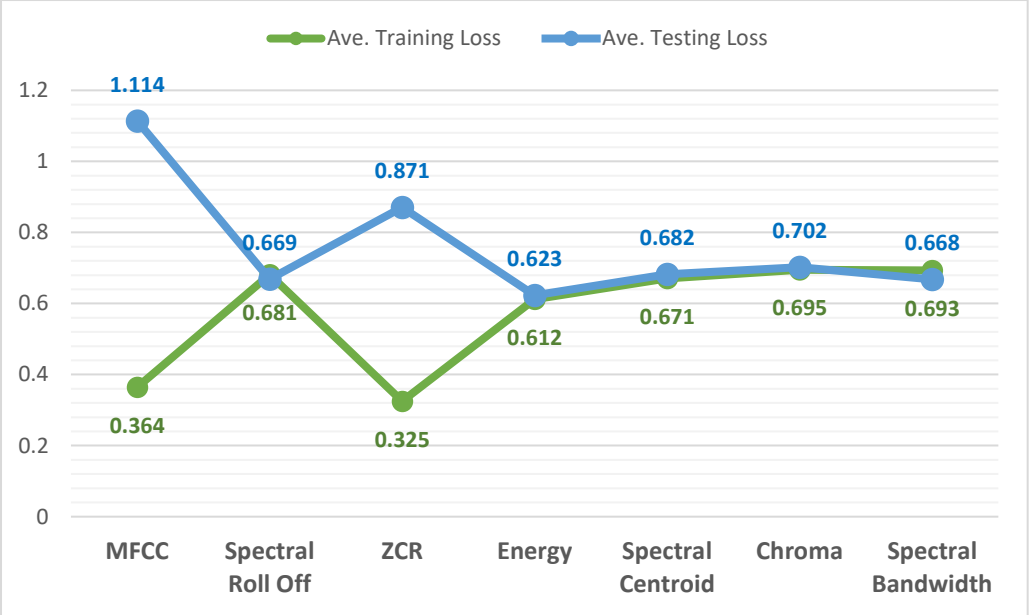


Figure 31: Speech Features for Child Emotion Recognition with Training and Testing Losses

Precision, Recall and F-Measure

The results for the precision, recall, and f-measure metrics are shown in Table 6. The ZCR features achieved the best precision of 0.76 of the positive samples among all the features used. This implies that 76% of data samples that were classified as belonging to the negative emotion class are truly negative emotion class. The chroma gave the worst performance for precision, correctly classifying 58% of the positive emotion class. The spectral bandwidth performed best in the recall metric, achieving a 0.97 score on the positive emotion class. The model was able to classify 97%

www.manaraa.com

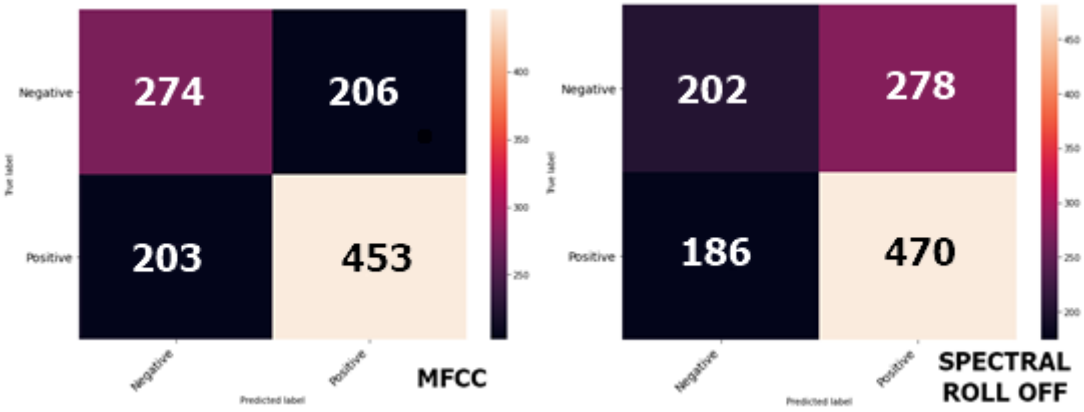
of the positive emotion samples. Spectral bandwidth managed the best performance on the f-measure metric, achieving a score of 0.73 while the spectral roll-off and MFCC obtained the least f-measure score of 0.67 and 0.69 respectively on the positive emotion class.

Features	Classes	Precision	Recall	F-Measure	Support
MFCC	Negative	0.57	0.57	0.57	480
	Positive	0.69	0.69	0.69	656
Spectral Roll Off	Negative	0.52	0.42	0.46	480
	Positive	0.63	0.71	0.67	656
ZCR	Negative	0.62	0.70	0.66	480
	Positive	0.76	0.68	0.72	656
Energy	Negative	0.62	0.42	0.50	480
	Positive	0.65	0.81	0.72	656
Spectral Centroid	Negative	0.53	0.25	0.34	480
	Positive	0.60	0.84	0.70	656
Chroma	Negative	0.56	0.08	0.14	480
	Positive	0.58	0.96	0.72	656
Spectral Bandwidth	Negative	0.59	0.06	0.11	480
	Positive	0.59	0.97	0.73	656

Table 6: Precision, Recall and F-Score for Speech Features for Child Speech Emotion Recognition

Confusion Matrix

The result for the confusion matrix for the two-class experiment is presented in Figure 32. The ZCR showed superior performance by correctly classifying the greatest number of negative emotion samples as negative. 339 of the negative samples were correctly classified as negative while 205 samples were incorrectly classified as positive. The MFCC correctly classified 274 negative emotion samples as negative while misclassifying 203 samples as positive. The worst performing features are the spectral centroid and the chroma.



www.manaraa.com

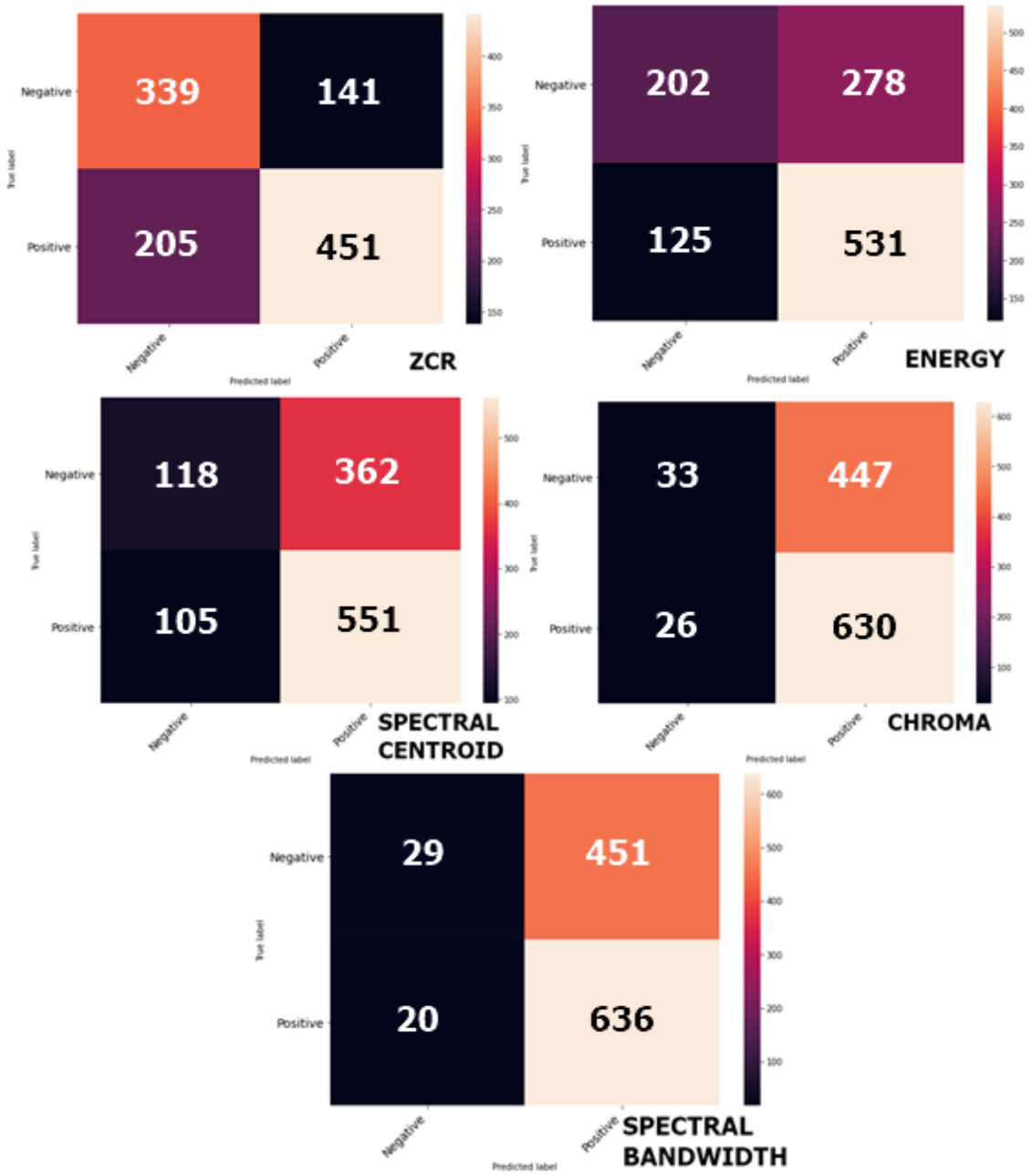


Figure 32: Confusion Matrix for Speech Features for Child Speech Emotion Recognition

5.3 Experiment 3: Spectrogram Versus Audio Features

Spectrograms have been used for adult speech emotion recognition (Ma et al., 2018; Satt et al., 2017). However, there is little research on how spectrograms perform for child speech emotion recognition. To answer RQ3, we propose to conduct this experiment using the FAU AIBO emotional speech corpus described earlier. The speech corpus is converted to spectrogram using fast Fourier Transform (FFT). Some researchers split the audio signal into equal frame lengths of about 3 seconds each before converting to spectrograms (Satt et al., 2017), this method tends to introduce some

www.manaraa.com

performance errors because the frames receive the same emotional label as the entire speech which is not always true in reality. Furthermore, frames that do not measure up to 3 seconds get padded with zeros to make it up (Ma et al., 2018). In this experiment, we use variable-length spectrograms because the network is capable of handling variable-length input. Metrics to be used to access performance are accuracy, loss plot, and confusion matrix.

Experimental Setup

This experiment is set up similarly to experiment 2. This is because the goal is to test how spectrogram performs for child speech emotion recognition and compare the performance with audio features. The experiment is divided into two runs. In the first run, we used the four classes that are available in the FAU AIBO emotional speech corpus. The class labels are neutral, emphatic, negative, and joyful. The dataset contains 4543 audio samples. In the second run, the data samples were relabeled into two classes of positive and negative emotions.

A sampling rate of 48000 was used for all the audio samples. The length of the fast Fourier transform (FFT) window was 2048 and the hop length, which is the number of samples between sequential frames is 512. The 'hann' window was used for signal smoothing. The librosa audio library was used to convert the audio samples to spectrograms.

The experiment is conducted using the affection component of the framework. The model is trained using the RMSprop optimizer with learning rate set to 0.001, and batch size set to 16. A dropout of 0.5 was used and RELU activation was used on all the layers of the network. We opted for max-pooling layers in between the convolutional layers to downsample the spatial dimension of the network and to minimize the size and computation in the network. Our hardware is a single core NVidia Tesla T4 GPU card with 16GB of RAM and our disk size is 110GB. The CPU is an Intel Xeon (R) running at 2.2GHz with two threads per core.

Experiment Results

www.manaraa.com

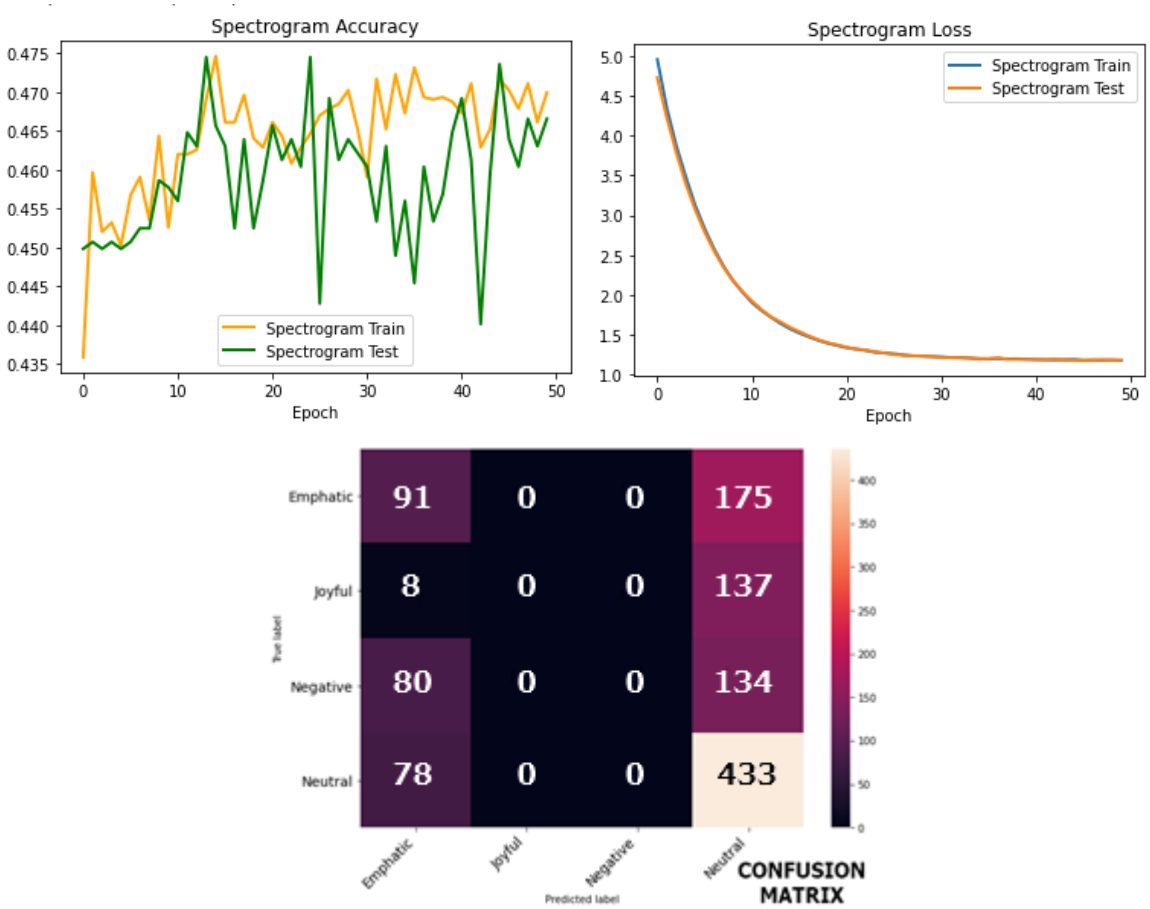


Figure 33: Spectrogram Accuracy, Loss and Confusion Matrix for Four-Class experiment

The results obtained from the four-class spectrogram experiments are shown in Figure 33. The figure includes the accuracy plot, the loss plot, and the confusion matrix. The model achieved an average accuracy of 46.13% on the 4-fold cross-validation. Comparing this performance with the result shown in Figure 34, it is evident that the ZCR, MFCC, and spectral roll-off outperformed the spectrogram for child speech emotion recognition while the spectrogram achieved better scores than the spectral centroid, chroma, and spectral bandwidth. The average network loss value is 1.169 which is the lowest score comparing with the results obtained for the speech features. Hence the spectrogram-trained model was able to minimize the network loss better than any of the speech features. This network however did not perform well on the negative emotion samples based on the confusion matrix. It consistently classified the negative emotion samples as neutral emotion and emphatic emotion.

www.manaraa.com

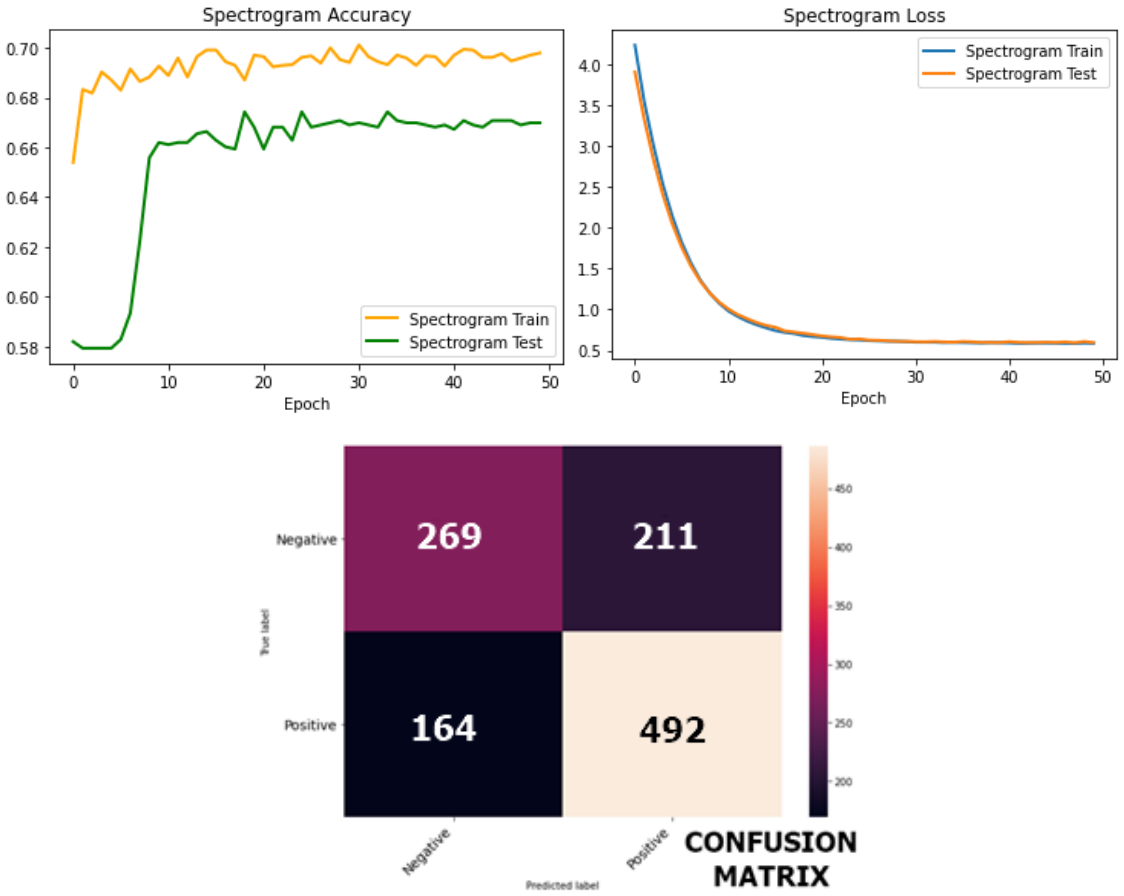


Figure 34: Spectrogram Accuracy, Loss and Confusion Matrix for Two-Class experiment

Figure 34 and Table 7 show the results obtained for the two-class spectrogram experiments. An average accuracy score of 66.48% was while the average loss at the end of the experiment was 0.5991. On the accuracy metric, the spectrogram was outperformed only by the ZCR audio feature. The spectrogram though was able to minimize network loss better than all the speech features with energy feature having the next lowest loss value. The spectrogram model did better in predicting the positive class than the negative class. MFCC, spectral roll-off, ZCR, and energy had better precision than spectrogram on the negative emotion class. All the speech features performed better than the spectrogram on recall and f-measure.

www.manaraa.com

	Precision	Recall	F-Measure	Support
Negative	0.62	0.56	0.59	480
Positive	0.70	0.75	0.72	656
Micro avg	0.65	0.65	0.65	1136
Macro avg	0.66	0.66	0.66	1136
Weighted avg	0.67	0.67	0.67	1136
Samples avg	0.66	0.66	0.66	1136

Table 7: Spectrogram Precision, Recall and F-measure for Two-Class Experiment

5.4 Experiment 4: Experiment with the Proposed Framework

To answer RQ4 requires training the affectation component of the proposed framework with raw speech signals. Using raw speech signals boycott the need to develop speech features and also the need to use spectrograms. This is also called end-to-end recognition. Using raw speech signals enables us to build real-time speech emotion recognition systems that could recognize a speaker's emotions on the go. For this experiment, we use the unprocessed raw speech files from the FAU AIBO emotional corpus. The speech files with their corresponding emotional labels will be used to train the network. Metrics to measure the performance of the solution include accuracy, loss plot, and precision, recall, and f-measure.

Experiment Setup

4543 audio samples were used in the experiment. Each of the samples was offset with a value of 0.1 seconds, meaning that the first 0.1 seconds in each speech file was not included in the analysis. The purpose is to ensure that noise and other jitters at the beginning of the audio signal are removed. The audio files were all converted to mono as there are no significant differences in performance in using mono or stereo audio; however, using mono reduces the computation in the network. The model is trained using the RMSprop optimizer with learning rate set to 0.001, and batch size set to 16. A dropout of 0.5 was used and RELU activation was used on all the layers of the network. We opted for max-pooling layers in between the convolutional layers to downsample the spatial dimension of the network and to minimize the size and computation in the network. Our hardware is a single core NVidia Tesla T4 GPU card with 16GB of RAM and our disk size is 110GB. The CPU is an Intel Xeon (R) running at 2.2GHz with two threads per core.

www.manaraa.com

The complete duration of each audio file in the dataset was read into the network apart from the 0.1-seconds offset. This resulted in a very large dimension of 4543 rows X 90,000 column matrix for the audio data to be used to train and test the model. The columnar data depends on the duration of the audio and the row data is the audio samples. While training the model with the high dimension audio data, the RAM buffer was overwhelmed and the system shut down severally. To reduce the dimension of the audio data, a duration of 0.5 seconds was set for all the audio samples. Hence, only 0.5 seconds was read into the network for all audio samples that have a duration longer than 0.5 seconds. This resulted in audio data of 4543 X 8003 matrix dimensions which the hardware was able to handle. The four-class experiment and the two-class experiments were then performed.

Experiment Results

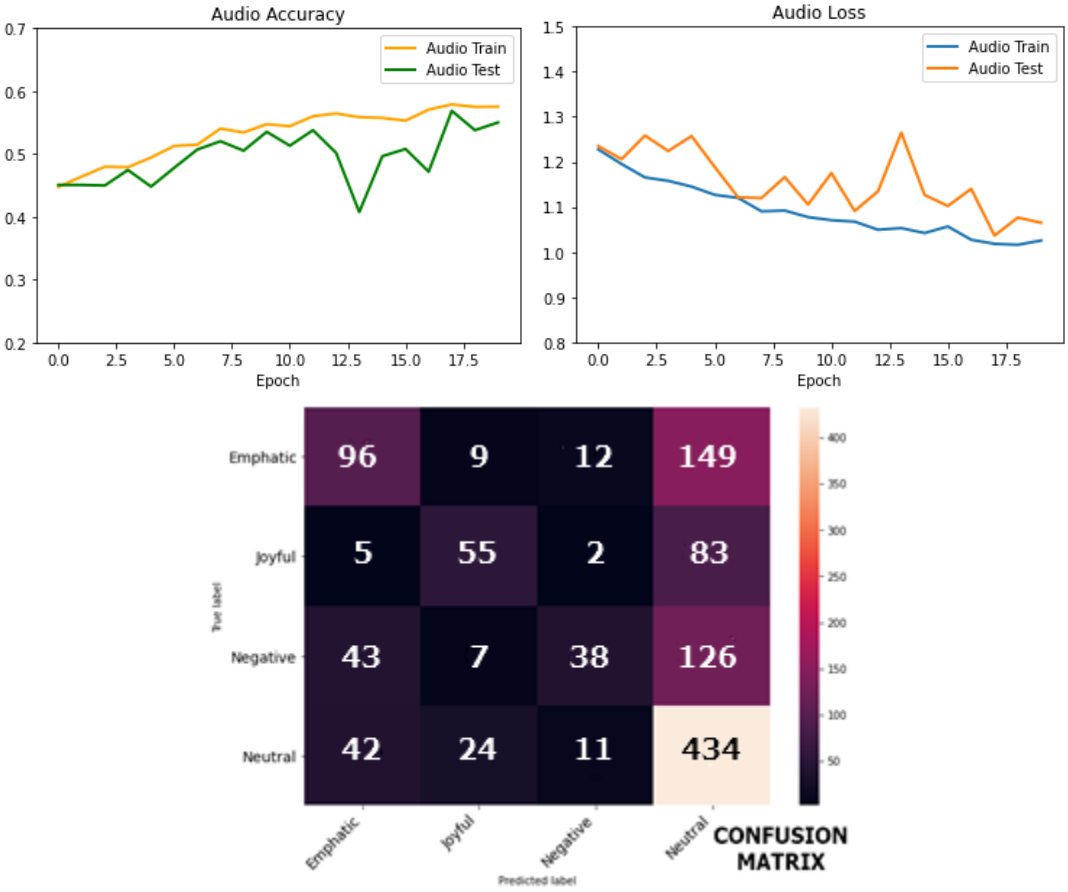


Figure 35: Four-Class Accuracy, Loss and Confusion Matrix for End-to-End Recognition

The result for the experiments is shown in Figure 35 and Table 8. Average accuracy of 54.83% was obtained for the four-class experiment and the network was able to minimize to an average of

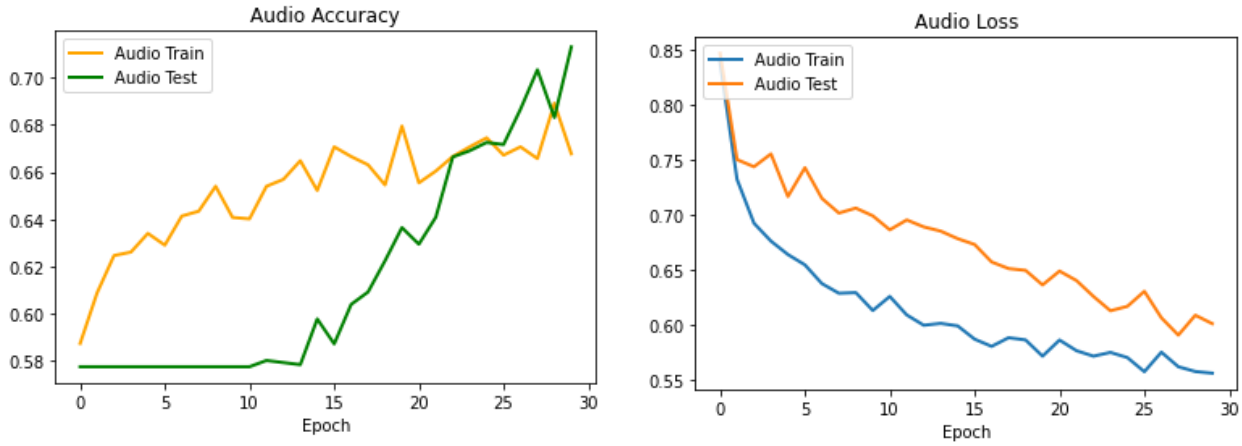
www.manaraa.com

1.069. The network did classify many of the negative samples as emphatic and neural from the confusion matrix. It did better in classifying the neutral emotion class than the other classes. This also reflects in the precision, recall, and f-measure metrics where the neutral class has the best score for the recall and f-measure followed by the negative emotion class.

	Precision	Recall	F-Measure	Support
Emphatic	0.61	0.23	0.33	266
Joyful	0.82	0.12	0.21	145
Negative	0.70	0.06	0.11	214
Neutral	0.61	0.61	0.61	511
Micro avg	0.62	0.33	0.43	1136
Macro avg	0.70	0.25	0.37	1136
Weighted avg	0.66	0.35	0.46	1136
Samples avg	0.33	0.34	0.33	1136

Table 8: Four-Class Precision, Recall and F-Measure for End-to-End Experiment

The results for the two-class experiment are presented in figure 36 and Table 9. The average accuracy score obtained is 71.14% while the average loss value for the network is 0.621. Towards the final epochs, the network performed better on the test data than on the train data which is the goal for most artificial intelligence models even though it is often difficult to achieve.



www.manaraa.com

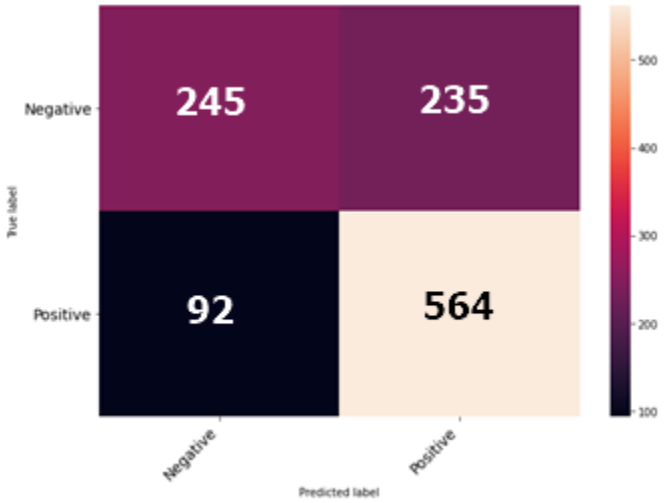


Figure 36: Two-Class Accuracy, Loss and Confusion Matrix for End-to-End Recognition

	Precision	Recall	F-Measure	Support
Negative	0.73	0.51	0.60	480
Positive	0.71	0.86	0.78	656
Micro avg	0.70	0.70	0.70	1136
Macro avg	0.71	0.69	0.70	1136
Weighted avg	0.71	0.71	0.71	1136
Samples avg	0.70	0.71	0.70	1136

Table 9: Two-Class Precision, Recall and F-Measure for End-to-End Experiment

5.5 Experiment 5: Comparative Performance Analysis with other Approaches

In this experiment, we compare the performance of the affectation component, which performs the child emotion recognition, with other standard deep neural networks such as the CNN and LSTM configurations and speech features. In this experiment, we use CNN only and CNN+LSTM to replace the working memory recurrent network in the affection component. The experimental result obtained from RQ 4 is compared with those obtained from using CNN and LSTM configurations. This allows us to evaluate the performance of the affection component and the working memory recurrent network (WMRN). The dataset to be used in this experiment is the FAU AIBO emotional speech dataset. Metrics for benchmarking performance will include accuracy, loss plot, precision, recall, f-measure, and confusion matrix.

Experiment Setup

The experimental setup was very similar to the setup for experiment four. The processing of the

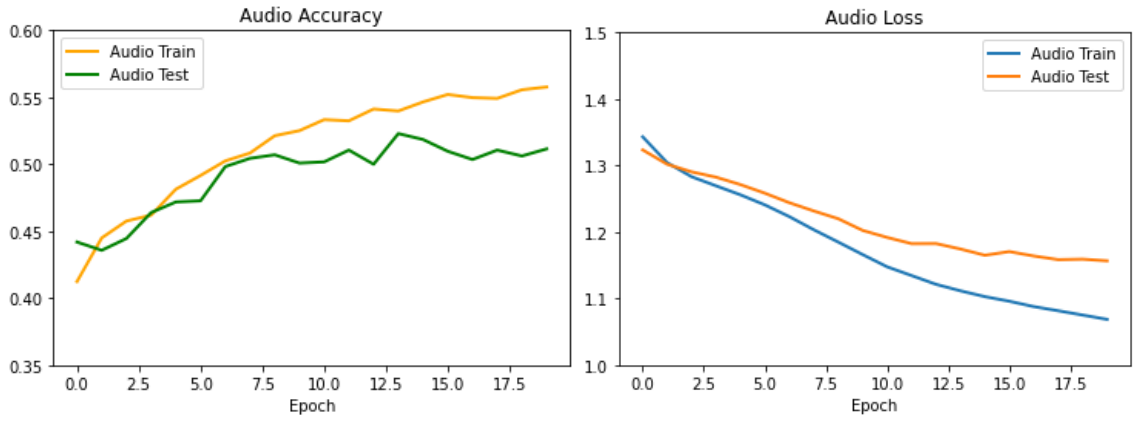
www.manaraa.com

audio data was the same. The first 0.1 seconds in each audio was removed to minimize noise. 0.5 seconds was read from each audio file and used in the analysis and they were converted to 'mono' as before. 4543 audio samples were used in the experiment.

The model is trained using the SGD optimizer with learning rate set to 0.0005, and batch size set to 16. Other optimizers were used and the SGD optimizer gave the best results. The learning rate was also tuned with different values and 0.0005 was the final selection. A dropout of 0.25 was used and RELU activation was used on all the layers of the network. We opted for max-pooling layers in between the convolutional layers to downsample the spatial dimension of the network and to minimize the size and computation in the network. Our hardware is a single core NVidia Tesla T4 GPU card with 16GB of RAM and our disk size is 110GB. The CPU is an Intel Xeon (R) running at 2.2GHz with two threads per core. For the CNN + LSTM experiment, the return sequence was set to true, the model was initialized with ones and the merge mode was set to concat.

Experiment results

CNN+LSTM – the results for the CNN+LSTM model are presented in figure 37 for the four-class experiment and figure 38 for the two-class experiment. In the four-class experiment, the model achieved an average accuracy of 50.87%, and the average network loss was 1.189. The model struggled with identifying negative emotion samples and misclassified many of them as emphatic or neutral emotions.



www.manaraa.com

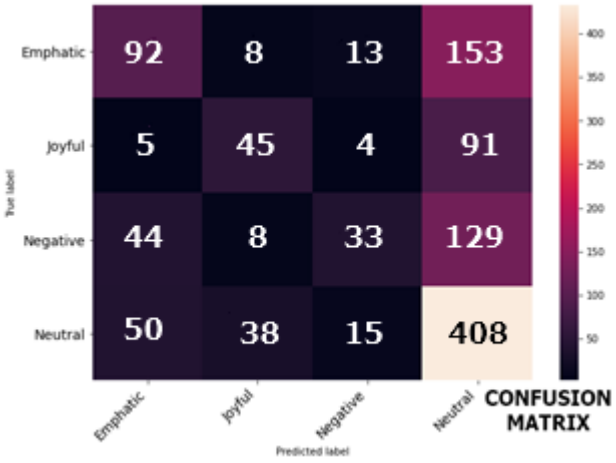


Figure 37: Four-Class Accuracy, Loss and Confusion Matrix for CNN+LSTM Configuration

In the two-class experiment, an average accuracy of 68.72% was obtained on the test sets and an average loss value of 0.603 on the test sets. Based on the confusion matrix, the model performed better in recognizing the positive emotion class than the negative emotion class. Nearly half of the negative emotion samples were misclassified as positive.

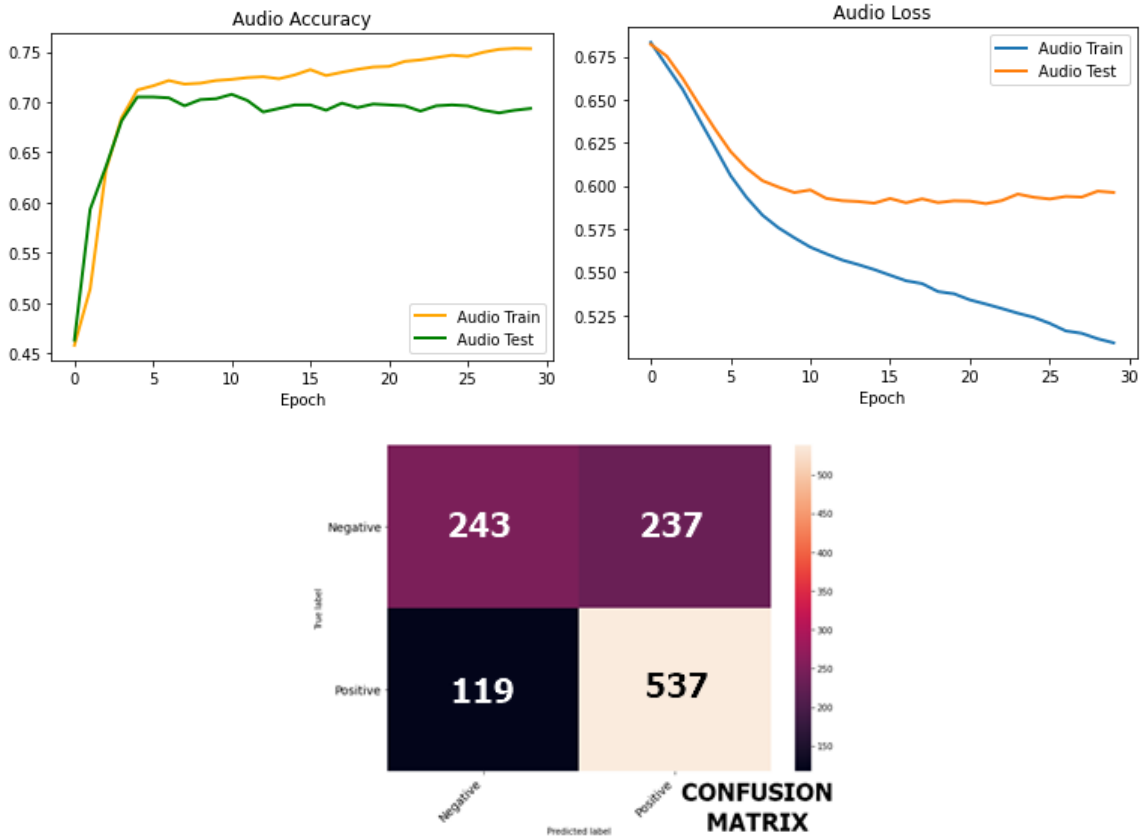


Figure 38: Two-Class Accuracy, Loss and Confusion Matrix for CNN+LSTM Configuration

www.manaraa.com

CNN – The CNN-only model outperformed the CNN+LSTM model on the FAU AIBO dataset for the four-class experiment on the accuracy metrics. It obtained 52.06% average accuracy and an average loss value of 1.652. This model also did better in recognizing the negative emotion class better than the CNN+LSTM model. It was able to correctly classify 58 negative emotion samples from the test data as negative while the CNN+LSTM model classified 38 samples correctly. Both models performed best on the neutral emotion data samples.

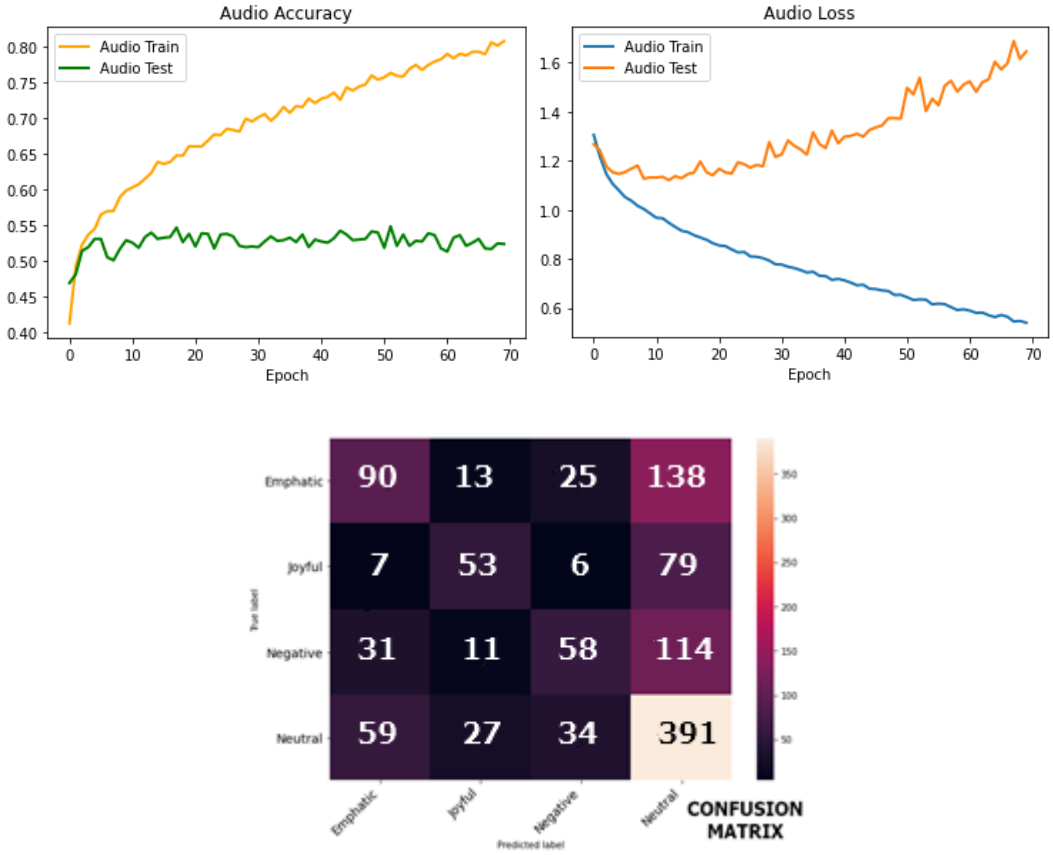


Figure 39: Four-Class Accuracy, Loss and Confusion Matrix for CNN only Configuration

The results for the two-class experiment on the CNN-only model are shown in Figure 40. The model obtained an average accuracy score of 67.23% on the test sets and an average loss of 0.39 on the test sets. The CNN+LSTM model shows a better accuracy performance for the two-class experiment, however narrowing down specifically to the negative samples, we find that the CNN only network performed better by classifying 275 negative emotion samples in the test data correctly as against 243 from the CNN+LSTM model.

www.manaraa.com

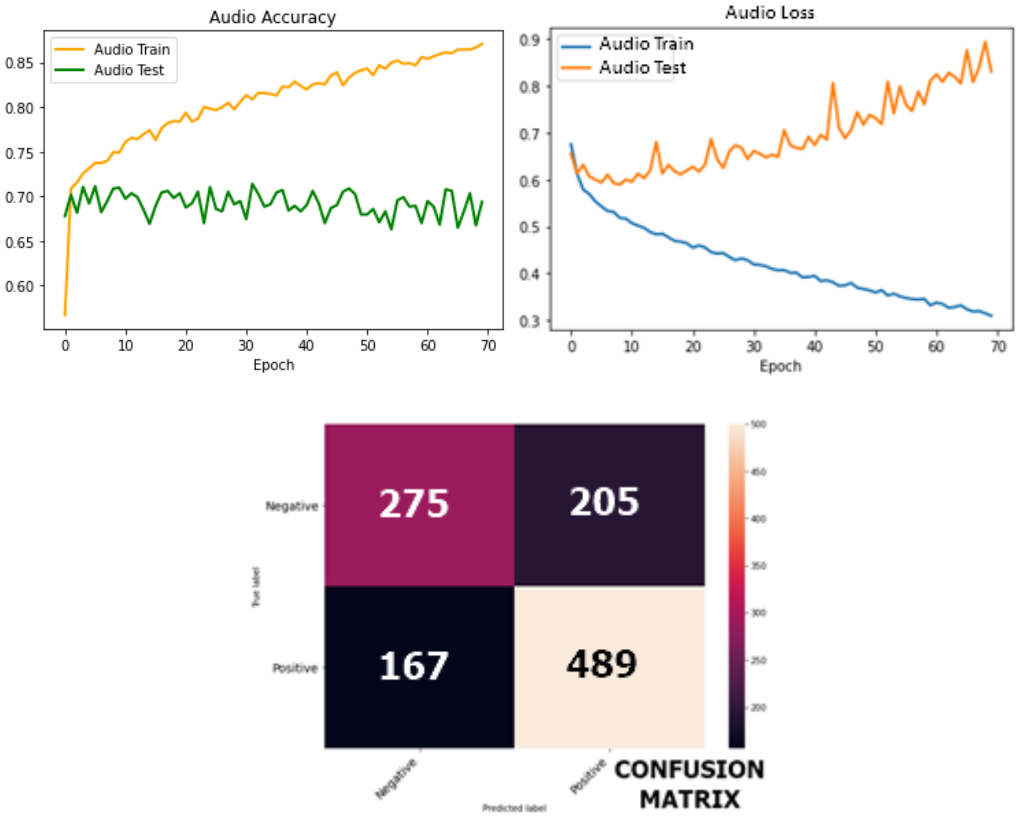


Figure 40: Two-Class Accuracy, Loss and Confusion Matrix for CNN only Configuration

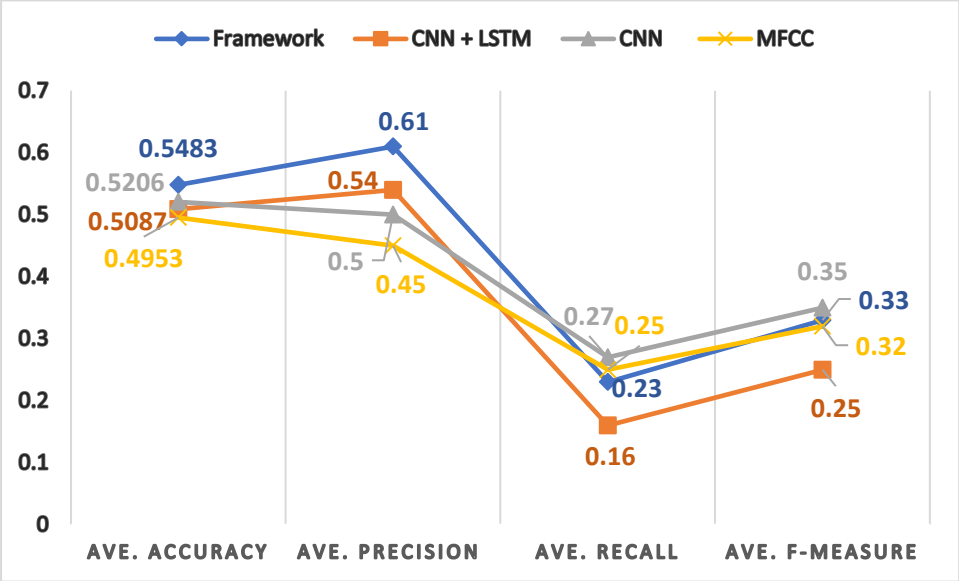


Figure 41: Performance comparison with Framework for the Four-class experiment.

Figure 41 and 42 compares the performance of the CNN+LSTM, CNN and MFCC speech feature and our framework on the accuracy and precision, recall and f-measure metrics. Figure 41 is for

www.manaraa.com

the four-class experiment while Figure 42 is for the two-class experiment. The precision, recall, and f-measure shown in Figure 41 is for the negative emotion class while Figure 42 is the positive emotion class. MFCC is used here as a representative speech feature because it is the most popular speech feature for speech emotion recognition in the literature. On the accuracy metrics, our framework performed better than all the other approaches on the two-class and four-class experiments. Accuracy is an important metric in evaluating artificial intelligence solutions and in cases where the classes are balanced such as our experiments, accuracy alone is a sufficient metric. The framework did not give better scores on the recall and f-measure for the four-class experiment and precision for the two-class experiment. We think that the nature of the dataset did not lend itself to the strengths of the framework. For example, the average speech length in the dataset is 1.7 seconds while one of the strengths of the framework is to address long-term dependency which is best demonstrated with longer speech duration. Despite this, the framework showed superior performance in precision on the four-class experiment and recall and f-measure in the two-class experiment. Most importantly, on the accuracy metric, the framework outperformed the other approaches in the four-class and two-class experiments.

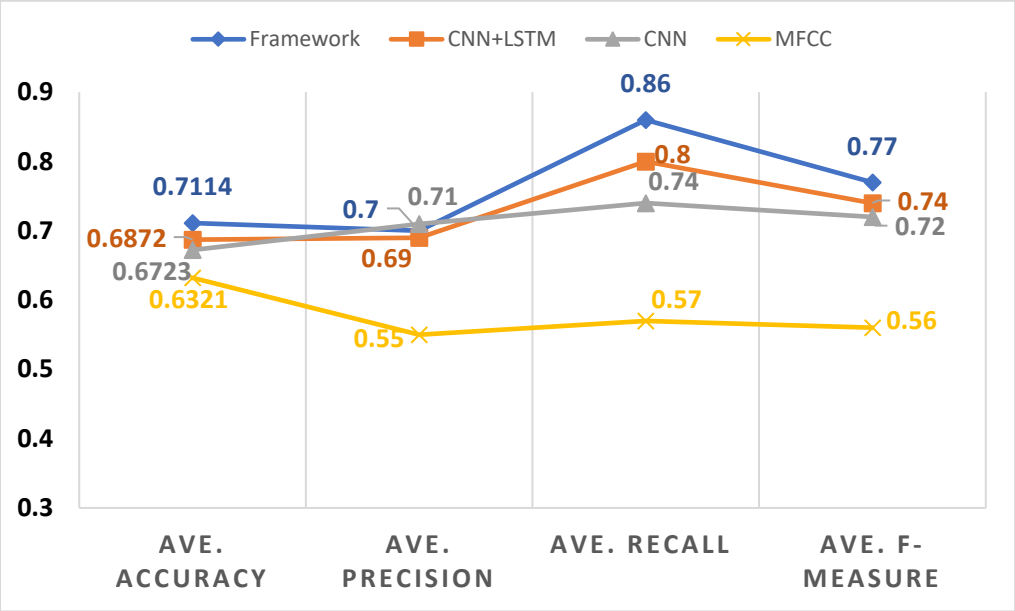


Figure 42: Performance Comparison with Framework for the Two-class experiment

www.manaraa.com

CHAPTER 6

CONCLUSION

6.1 Summary of Research and Research Findings

This dissertation first analyzes key speech features (Mel frequency cepstral coefficient (MFCC), spectral roll-off, zero-crossing rate (ZCR), spectral centroid, energy, spectral bandwidth, and chroma) for speaker age group recognition using the codetermination component of the proposed framework. This is the first crucial step used to sort the data that contains both adult and child speakers and to ensure that only speech made by children is being analyzed. An example of such a scenario is a recorded conversation between a child and a pediatrician where it may be necessary to sift and separate the pediatricians' statements from that of the child before analysis is done.

Furthermore, we explored the best features for speaker age group recognition as it is important to ensure that we are analyzing only speech that is made by children. The world health organization's definition of children as a person that is less than eighteen years old was used as the benchmark and we ensured that all the data used for child speech emotion recognition were created from persons under the age of eighteen. We did note a caveat to this definition as certain cultures and climes define a child differently and we contextualized our definition and scope to the western, educated, industrialized, rich, and democratic (WEIRD) societies. In our analysis, after testing several speech features, we were able to find that the feature that worked best for child speaker recognition. It was highlighted that speech emotion recognition for adults is different from children as previous research has identified that recognition systems designed for adults tend to fail with children (Palo et al., 2017). Hence, we proceed to examine the speech feature that works best for child speech emotion recognition and we ran experiments to determine that. Deep learning models allow us to perform end-to-end speech emotion recognition without needing to extract spectrograms or audio features, consequently, we used the working memory recurrent network of the affectation component of the framework to perform end to end child speech emotion recognition using raw audio. To benchmark the performance of our model, we compared it with the convolutional neural network (CNN), long short-term memory (LSTM), and MFCC, a popular speech feature.

www.manaraa.com

6.2 Answers to Research Questions

What speech features work best for identifying a child speaker from a mix of speech signals containing adult and child speakers?

To answer this question, the RAVDESS dataset, FAU AIBO emotional speech corpus, and the English Children dataset were used. The first dataset contains adult speech and the last two contain children's speech. Seven most frequently used speech features were extracted from the audio samples and used to train a CNN + LSTM deep learning model for speaker age group recognition. The speech features are Mel frequency cepstral coefficient (MFCC), spectral roll-off, zero-crossing rate (ZCR), energy, spectral centroid, chroma, and spectral bandwidth. MFCC worked best for identifying a child speaker. It correctly classified all child samples in the test data as a child and all adult samples as an adult with zero misclassification. The ZCR feature also correctly classified all child samples in the test data as a child, however, it misclassified one adult sample as a child. From our experiments, MFCC and ZCR are the two best features for identifying a child speaker.

What speech features work best for recognizing the emotion of a child speaker?

For this question, we used the FAU AIBO emotional speech corpus on the affection component of our framework. 4543 data samples were used and seven speech features were tested. Based on the results obtained, the ZCR features showed superior performance and worked best in recognizing the class of emotion of a child speaker. It obtained the best results using the accuracy, loss, and precision metrics for performance measurement. 76% of the data samples in the test data that is classified as negative emotion class was truly negative emotion class. This number is the highest among all the features tested. The ZCR was also the only speech feature that exceeded the 70% mark in classification accuracy for the two-class experiments.

What are the advantages of using spectrogram over audio features in child speech emotion recognition models?

To answer this question, we used the FAU AIBO emotional speech corpus and we converted the 4543 audio samples to spectrograms using the short-term fast Fourier transform. The experiment was conducted on the affection component of the framework. We trained the model using the spectrogram and the results obtained were compared with the result obtained using speech

www.manaraa.com

features. Overall, there were no advantages in using spectrograms over audio features for child speech emotion recognition. We found that the spectrogram performance was subpar to that of the ZCR, MFCC, energy, and spectral roll-off for child emotion recognition. The speech features have better precision scores than the spectrogram on the negative emotion class meaning that they made fewer mistakes in recognizing the negative emotion class and were better in predicting them correctly.

Can an effective child speech emotion recognition framework be developed using deep learning informed by theories of working memory?

For this question, we used the affection component of the framework and performed an end-to-end child speech emotion recognition. In this process, neither speech features or spectrograms were used but rather we trained our deep learning model directly with raw audio. This technique eliminates the steps in the workflow and results in a solution that will generally perform faster for online real-time processing. The model was trained using audio data labeled with four classes of emotions (neutral, negative, joyful, and emphatic) in the first experiment and was later retrained using data labeled with two classes of negative and positive emotions. The model performed well in identifying negative emotional class and the full results have been discussed in the previous chapters. The end-to-end recognition using our deep learning model overcomes the challenges of experimenting with several speech features and the associated problem of high dimensionality in the resulting feature data and allows child speech emotion recognition to be performed directly from the speech recordings. It is also able to overcome long-term dependency problems with some deep learning models such as the recurrent neural network.

How does the proposed framework perform compared to other approaches?

The proposed framework was evaluated using accuracy, loss, precision, recall, and f-measure and confusion matrix. The results obtained were shown and discussed in the previous chapter. To further evaluate the framework, we compared its performance with CNN, CNN + LSTM models, and MFCC features which are standard solutions used for speech recognition research. We found that our model was able to hold up with these standard models and outperform them on many occasions. Our framework showed better performance in the accuracy score for both the four-class

www.manaraa.com

and two-class experiments. In the four-class experiment, it also performed better than the others on the precision. It did struggle on the recall for the four-class experiment but so did the other approaches, leading to a generally low score on the recall metrics. On the two-class experiment, the framework showed better performance on the recall and f-measure than the other approaches. It recognized the positive and the negative classes better than the CNN, CNN + LSTM, and the MFCC.

6.3 Implications of the Study

Theoretical Implications

In this work, we developed a framework that is undergirded with theoretical groundwork. Baddeley's theory of working memory was the key theoretical model that guided the development of our framework. As far as we know, this theory has been used to study multitasking (Covre et al., 2019), mathematical learning (Passolunghi & Costa, 2019), listening (Sakai, 2018), and others. These types of studies reside in the domain of clinical psychology. They are typically aligned with the hypothetico-deductive mode of scientific study that uses theoretical backgrounds from extant literature to postulate testable hypotheses, identify, operationalize and manipulate experiment variables and dependencies, perform empirical studies based on experimentation and use the findings of the study to develop theories or enrich existing theories while contributing to the body of knowledge. This work precludes this mode of scientific inquiry and paves the way for Baddeley's theory of working memory to be used differently, in studying large quantities of data as is common in deep learning research. To the best of our knowledge, this research is the maiden use of Baddeley's theory of working memory, a clinical psychological theory in information systems research, an effort at cross-disciplinary research that could appeal to people from both sides of the aisle.

We made an assumption about Baddeley's theory of working memory that given that this theory offers one of the best explanations of how the working memory of the human brain works, it is going to be very effective in solving long-term dependency problems in artificial neural networks. The results that we obtained have demonstrated the validity of our assumption and the theory. We propose in principle that Baddeley's model will be effective for other deep learning applications that require learning long-term dependencies such as stock market prediction, fraud detection, and demand forecasting.

www.manaraa.com

Gregor, 2006 explains the *Theory of Explaining* as the theory of understanding that emphasizes "on showing others how the world may be viewed in a certain way, with the aim of bringing about an altered understanding of how things are or why they are as they are." The cognitive theories of emotion discussed in this work offer a unique explanation of why acted speech is not ideal for building speech emotion recognition systems. In chapter three, we detailed the cognitive theories of emotion such as the Schacter theory and Lazarus Cognitive mediational theory that maintain that an external stimulus precedes the cognitive appraisal of the stimulus before emotion is experienced. Beyond the cognitive theories, the early theories of emotion such as the James-Lange theory and the Canon-Bard theory form the bedrock of recent theories that explain human emotions. James-Lange theory proposes that emotional occurrences are due to physiological changes that are going on in the organs of the body due to an external stimuli. For example, when a person sees a wild animal in the dark, the heart begins to race and the individual is frightened. According to James-Lange theory, the racing of the heart is what causes the fright and not the fright causing the racing of the heart. James-Lange theory holds that the physiological changes cause the emotional experience. Cannon-Bard's theory, on the other hand, offers that physiological changes and emotional reactions occur independently and simultaneously. For example, thinking about an upcoming job interview may leave a person feeling worried and also increase his heartbeat rate. Cannon-Bard theory holds that both events occur at the same time due to an external stimulus, which in this case is the upcoming job interview. These theories despite their disparities agree that a stimulus must precede an emotional response, and as such, they invalidate the use of scripted/acted speech for speech emotion recognition as there is no external stimulus that evokes the emotions but they are fabricated or conjured by the actors. While previous researchers have raised this problem with acted speech, none to our knowledge has explained the problem in the light of the major theories and psychology of emotions and how their provisions abrogate the use of acted emotions as real emotions.

Practical Implications

While the apprehension of artificially intelligent machines taking over human jobs persist, AI technologies such as speech emotion recognition will continue to find applications that have the capacity to change business landscapes, reform the ways we socially interact, and in general affect the way we live on a daily basis.

www.manaraa.com

Potential applications of the proposed child speech emotion recognition framework abound, including child abuse detection, child learning robot, child therapy services, educational games, and estimating the level of engagement of children in entertainment games. The issue of child abuse has prompted Social Services, Child Protective Services, Department of Health and Human Services, and many other governmental and non-governmental agencies involved with child welfare to use conventional proxies such as school teachers, neighbors, sports coaches, etc., as a means of identifying and reporting child abuse. However, the efficacy of these proxies is declining due to minimized social interactions, hence making the need for other methods of child abuse detection that do not require proxies such as the use of artificial intelligence more compelling. Child abuse detection research conventionally resides in the social works, medical, and public health research domains. Typically, social services rely on the family members, neighbors, and school teachers, and the public in general to report suspected child abuse or maltreatment to local child protective services. However, since the outbreak of the covid-19 pandemic, interactions among neighbors, school teachers, and children have been severely limited with some climes still using virtual classrooms as the mode of learning for school children. In the medical field, pediatricians and other medical researchers use screening tools (Crichton et al., 2016; Louwers et al., 2014). These screening tools are usually in the form of a checklist designed to establish the levels of concern for child abuse on a per-child basis. Some other research uses radiological scanning technologies (Aertsen, 2017; Barber et al., 2013) such as sonography, scintigraphy, computed tomography, and magnetic resonance imaging for the detection of child abuse that resulted in a physical or inflicted injury.

Child speech emotion recognition has a strong potential application in child abuse detection. Researchers have shown that there is a strong correlation between child maltreatment and abuse and a child feeling or displaying negative emotions such as aggression, anger, fear, depression, sadness, and withdrawal (Hennessy et al., 1994; Says, 2016; Signs of Child Abuse, 2015). The framework developed in this study can be used to detect the emotions of a child from the speech. Negative emotions could be flagged and a human follow-up could ascertain the cause. Alternatively, the framework could be integrated with a text analytics system that can recognize and flag named entities to perform multimodal detection using speech and text to detect child abuse.

Children learning robots is growing at an astonishing pace in the last decade and it is seen by many as offering benefits in child pedagogy. Learning robots perform different functions and specific

www.manaraa.com

robots have specific niches such as teaching coding, teaching STEM, history, mathematics, physics, engineering, and science in general (Johnson, 2003). However, learning robots have also been associated with providing motivation for learning, critical thinking, self-esteem, teamwork, creativity, and imagination (Iberdrola, 2021). A popular learning robot named Miko 2 uses artificial intelligence to learn a child's preference, remember faces and recognize the child's voice. The framework developed in this study can be integrated with a child learning robot to better understand the emotion of the child and adapt behavior and learning to the child's emotion. For example, if the framework detects negative emotions from the child, rather than serving fresh materials, the robot can do a recap of all previous activities that child has performed successfully or serve edutainment contents as a means of helping to improve the child's emotional attitude.

Child therapy services provide a haven for maladjusted children with behavioral or mental disorders, providing emotional support and strategies to learn how to cope with the problem. Child therapists must find a way to engage and connect with the child and they use lots of play-based tools and techniques to do this (Ray et al., 2013). This is often referred to as play therapy and one of the goals is to allow a child to identify the emotions the child is feeling. In some cases such as (Albert Knapp & Associates, 2016), the child cannot identify or name the emotions that he is feeling. The framework developed in this study could be useful in these scenarios to help the child identify his emotion. Therapists believe that the child being able to identify and name the emotion he is feeling helps in the therapy (Albert Knapp & Associates, 2016). In other cases where a therapist is helping multiple children at the same time, the framework developed in this study could be used to develop a decision support system for the child therapist that would help to monitor and identify the dominant emotion that that the child is experiencing during their moments playing alone. The therapist could use such feedbacks from the system to work through some solutions during one-on-one sessions.

The proposed child speech emotion recognition framework can also be applied to education games and entertainment games in estimating the level of engagement of the child with the game. This framework could be integrated with stand-alone game consoles such as the play station game console. Game developers could build emotional markers at different stage gates in the game that are expected to trigger certain emotional reactions. The emotional reactions from players would be detected at these stage gates and the collated emotional feedbacks could be used to improve

www.manaraa.com

future versions of the game to have the desired emotional reaction.

6.4 Limitations of the study

While running the experiments for end-to-end recognition, limitations were encountered in the hardware. The 16GB RAM that was used could not support training the deep learning model with the entire duration of the audio samples. The RAM buffer gets filled during the training and the system would restart. To resolve this problem, we used 0.5 seconds from each audio file in the dataset to train the network. This limitation could be overcome by a more robust deep learning system.

In this dissertation, speech features were analyzed in a stand-alone fashion. It is possible to combine different speech features together for emotion recognition. The challenge with doing this is that the combined speech features usually result in very high dimensional data that often requires applying some dimensionality reduction techniques which can impact the overall performance of the system.

We have also highlighted that the definition of a child has cultural and geographic boundaries and we have adopted the general definition of the western, industrial, educated, rich, and democratic (WEIRD) society as someone that has reached the age of eighteen. This definition has limitations because it is not accepted in all climes. Hence this research is limited to the WEIRD societies where this definition is generally accepted.

In performing experiment 1 that requires recognizing child speakers from adult speakers, three different data sources were used together. The English Children dataset and the RAVDESS were both English speech while the FAU AIBO dataset is in the German language. Although speech emotion recognition systems will work well irrespective of the language, the differences in the methods of collection of these datasets and the recording devices use to collect them may affect the results of experiment 1 in ways that cannot be accounted for.

The other limitation that was encountered in this work is the paucity of children's emotional speech datasets. There are very limited speech datasets with emotional labels where children are the main speakers. Using many datasets typically strengthens the findings in deep learning research and it is hoped that more datasets will be used in the future work.

6.5 Future Work

In the future, we hope to use other datasets to train the model and make it more robust to the

المنارة للاستشارات
www.manaraa.com

variabilities in field data. We plan to explore how the framework behaves when training is done with one corpus and the testing performed using a different corpus. It is also possible to train the model with multiple datasets. In the current work, only the FAU AIBO split into train and test sets were used for child emotion detection. This type of experimentation tests the capacity of the framework to generalize across different corpora. In this work, anger, touchy and reprimanding were bundled together in the dataset as the negative emotion class. We hope that future datasets that will be used in this work will be able to treat each negative emotion as a different class.

It is in the plan for future work to continue to refine the framework and improve its performance, especially on the negative emotion samples. One of the ways this could be done would be to incorporate other types of mode within the framework such as semantic features and textual analysis and using an ensemble technique to aggregate results and potentially improve the overall performance of the system. Furthermore, we plan to deploy the model as a mobile application to allow interested organizations to use the solution.

www.manaraa.com

REFERENCES

Antipov, G., Baccouche, M., Berrani, S.-A., & Dugelay, J.-L. (2017). Effective training of convolutional neural networks for face-based gender and age prediction. *Pattern Recognition*, *72*, 15–26. https://doi.org/10.1016/j.patcog.2017.06.031

Assunção, G., Menezes, P., & Perdigão, F. (2019). Importance of speaker specific speech features for emotion recognition. *2019 5th Experiment International Conference (Exp.at'19)*, 266–267. https://doi.org/10.1109/EXPAT.2019.8876534

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559.

Baddeley, A. (1996). Exploring the central executive. *The Quarterly Journal of Experimental Psychology Section A*, *49*(1), 5–28.

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*, 1–29.

Baddeley, A. D. (1986). Working memory oxford. *England: Oxford Uni*.

Baddeley, A. D., Allen, R. J., & Hitch, G. J. (2011). Binding in visual working memory: The role of the episodic buffer. *Neuropsychologia*, *49*(6), 1393–1400.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In *Psychology of learning and motivation* (Vol. 8, pp. 47–89). Elsevier.

Baddeley, A. D., & Mehrabian, A. (1976). *The psychology of memory*. Citeseer.

Chae, M., Kim, T.-H., Shin, Y. H., Kim, J.-W., & Lee, S.-Y. (2018). End-to-end Multimodal Emotion and Gender Recognition with Dynamic Joint Loss Weights. *ArXiv:1809.00758 [Cs, Eess, Stat]*. http://arxiv.org/abs/1809.00758

www.manaraa.com

Cowan, N. (2014). Working Memory Underpins Cognitive Development, Learning, and Education. *Educational Psychology Review*, *26*(2), 197–223. https://doi.org/10.1007/s10648-013-9246-y

Grzybowska, J., & Kacprzak, S. (2016). *Speaker Age Classification and Regression Using i-Vectors*. 1402–1406. https://doi.org/10.21437/Interspeech.2016-1118

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75–105. JSTOR. https://doi.org/10.2307/25148625

Hixon, T. J., Weismer, G., & Hoit, J. D. (2020). *Preclinical speech science: Anatomy, physiology, acoustics, and perception* (Third edition). Plural Publishing.

Jung, Y., Kim, Y., Lim, H., & Kim, H. (2017). Linear-scale filterbank for deep neural network- based voice activity detection. *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 1–5. https://doi.org/10.1109/ICSDA.2017.8384446

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., Senft, E., & Belpaeme, T. (2017). Child speech recognition in human-robot interaction: Evaluations and recommendations. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 82–90.

Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, *7*, 117327–117345.

Kropotov, J. D. (2016). *Functional neuromarkers for psychiatry: Applications for diagnosis and treatment*. Academic Press.

www.manaraa.com

Li, Y., Zhao, T., & Kawahara, T. (2019). Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. *Interspeech 2019*, 2803–2807. https://doi.org/10.21437/Interspeech.2019-2594

Livingstone, S. R., & Russo, F. A. (2018). 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.

Ma, X., Wu, Z., Jia, J., Xu, M., Meng, H., & Cai, L. (2018). Emotion Recognition from Variable-Length Speech Segments Using Deep Learning on Spectrograms. *Interspeech 2018*, 3683–3687. https://doi.org/10.21437/Interspeech.2018-2228

Ma, X., Wu, Z., Jia, J., Xu, M., Meng, H., & Cai, L. (2017). Speech Emotion Recognition with Emotion-Pair Based Framework Considering Emotion Distribution Information in Dimensional Emotion Space. *Interspeech 2017*, 1238–1242. https://doi.org/10.21437/Interspeech.2017-619

Mao, S., Ching, P. C., & Lee, T. (2019). Deep Learning of Segment-Level Feature Representation with Multiple Instance Learning for Utterance-Level Speech Emotion Recognition. *Interspeech 2019*, 1686–1690. https://doi.org/10.21437/Interspeech.2019-1968 March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, *15*(4), 251–266. https://doi.org/10.1016/0167-9236(94)00041-2

Markitantov, M., & Verkholyak, O. (2019). Automatic Recognition of Speaker Age and Gender Based on Deep Neural Networks. In A. A. Salah, A. Karpov, & R. Potapova (Eds.), *Speech*

www.manaraa.com

and Computer (pp. 327–336). Springer International Publishing. https://doi.org/10.1007/978-3-030-26061-3_34

McCain, R. A. (1980). A Theory of Codetermination. *Zeitschrift Für Nationalökonomie / Journal of Economics*, *40*(1/2), 65–90. JSTOR.

Mena, M. E. (2012). *EMOTION RECOGNITION FROM SPEECH SIGNALS*. 116.

Mirsky, A. F., Ingraham, L. J., & Kugelmass, S. (1995). Neuropsychological assessment of attention and its pathology in the Israeli cohort. *Schizophrenia Bulletin*, *21*(2), 193–204.

Navyasri, M., RajeswarRao, R., DaveeduRaju, A., & Ramakrishnamurthy, M. (2018). Robust Features for Emotion Recognition from Speech by Using Gaussian Mixture Model Classification. In S. C. Satapathy & A. Joshi (Eds.), *Information and Communication Technology for Intelligent Systems (ICTIS 2017)—Volume 2* (pp. 437–444). Springer International Publishing. https://doi.org/10.1007/978-3-319-63645-0_50

OHCHR. (1990). *OHCHR | Convention on the Rights of the Child*. https://www.ohchr.org/en/professionalinterest/pages/crc.aspx

Oppenheim, A. V., & Schafer, R. W. (2004). From frequency to quefrency: A history of the cepstrum. *IEEE Signal Processing Magazine*, *21*(5), 95–106.

Palo, H. K., Mohanty, M. N., & Chandra, M. (2018, January 1). *Speech Emotion Analysis of Different Age Groups Using Clustering Techniques* [Article]. International Journal of Information Retrieval Research (IJIRR). www.igi-global.com/article/speech-emotion-analysis-of-different-age-groups-using-clustering-techniques/193250

Palo, H. K., Mohanty, M. N., & Chandra, M. (2017). *Emotion Analysis from Speech of Different Age Groups*. 283–287. https://doi.org/10.15439/2017R21

www.manaraa.com

Peffers, K., Rothenberger, M., Tuunanen, T., & Vaezi, R. (2012). Design Science Research Evaluation. In K. Peffers, M. Rothenberger, & B. Kuechler (Eds.), *Design Science Research in Information Systems. Advances in Theory and Practice* (pp. 398–410). Springer. https://doi.org/10.1007/978-3-642-29863-9_29

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Qawaqneh, Z., Mallouh, A. A., & Barkana, B. D. (2017). Deep neural network framework and transformed MFCCs for speaker's age and gender classification. *Knowledge-Based Systems*, *115*, 5–14. https://doi.org/10.1016/j.knosys.2016.10.008

Satt, A., Rozenberg, S., & Hoory, R. (2017). Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. *Interspeech 2017*, 1089–1093. https://doi.org/10.21437/Interspeech.2017-200

Schon, D. A. (1984). *The reflective practitioner: How professionals think in action* (Vol. 5126). Basic books.

Sharan, R. V., & Moir, T. J. (2015). Noise robust audio surveillance using reduced spectrogram image feature and one-against-all SVM. *Neurocomputing*, *158*, 90–99. https://doi.org/10.1016/j.neucom.2015.02.001

Simon, H. A. (1969). *The Sciences of the Artificial. 1969, 1981*. MIT Press.

Stone, S., Steiner, P., & Birkholz, P. (2017). A Time-Warping Pitch Tracking Algorithm Considering Fast f0 Changes. *INTERSPEECH*, 419–423.

www.manaraa.com

Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: A review. *International Journal of Speech Technology*, *21*(1), 93– 120. https://doi.org/10.1007/s10772-018-9491-z

Syed, Z. S., Schroeter, J., Sidorov, K., & Marshall, D. (2018). Computational Paralinguistics: Automatic Assessment of Emotions, Mood and Behavioural State from Acoustics of Speech. *Interspeech 2018*, 511–515. https://doi.org/10.21437/Interspeech.2018-2019

Tang, D., Zeng, J., & Li, M. (2018). An End-to-End Deep Learning Framework for Speech Emotion Recognition of Atypical Individuals. *Interspeech 2018*, 162–166. https://doi.org/10.21437/Interspeech.2018-2581

Verma, D., & Mukhopadhyay, D. (2016). Age driven automatic speech emotion recognition system. *2016 International Conference on Computing, Communication and Automation (ICCCA)*, 1005–1010. https://doi.org/10.1109/CCAA.2016.7813862

Yenigalla, P., Kumar, A., Tripathi, S., Singh, C., Kar, S., & Vepa, J. (2018). Speech Emotion Recognition Using Spectrogram and Phoneme Embedding. *Interspeech 2018*, 3688– 3692. https://doi.org/10.21437/Interspeech.2018-1811

References

Abhang, P. A., Gawali, B. W., & Mehrotra, S. C. (2016). *Introduction to EEG- and speech-based emotion recognition*. Elsevier/AP, Academic Press is an imprint of Elsevier.

Barber, I., Perez-Rossello, J. M., Wilson, C. R., Silvera, M. V., & Kleinman, P. K. (2013). Prevalence and relevance of pediatric spinal fractures in suspected child abuse. *Pediatric Radiology*, *43*(11), 1507–1515. https://doi.org/10.1007/s00247-013-2726-x

www.manaraa.com

Barsade, S. G. (2002). The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly*, *47*(4), 644–675.

Baxter, M. G., & Murray, E. A. (2002). The amygdala and reward. *Nature Reviews Neuroscience*, *3*(7), 563–573.

Becker, D. B., Needleman, H. L., & Kotelchuck, M. (1978). Child abuse and dentistry: Orofacial trauma and its recognition by dentists. *The Journal of the American Dental Association*, *97*(1), 24–28. https://doi.org/10.14219/jada.archive.1978.0447

Benger, J. R., & Pearce, A. V. (2002). Simple intervention to improve detection of child abuse in emergency departments. *BMJ*, *324*(7340), 780–782. https://doi.org/10.1136/bmj.324.7340.780

Bergland, C. (2017). How do neuroplasticity and neurogenesis rewire your brain. *An Article Published in Psychology Today*.

Bock, K., & Levelt, W. (1994). *Grammatical encoding*.

Bourassa, C., Lavergne, C., Damant, D., Lessard, G., & Turcotte, P. (2006). Awareness and detection of the co-occurrence of interparental violence and child abuse: Child welfare worker's perspective. *Children and Youth Services Review*, *28*(11), 1312–1328. https://doi.org/10.1016/j.childyouth.2006.02.002

Campo, D., Quintero, O. L., & Bastidas, M. (2016). Multiresolution analysis (discrete wavelet transform) through Daubechies family for emotion recognition in speech. *Journal of Physics: Conference Series*, *705*(1), 012034.

Caridakis, G., Castellano, G., Kessous, L., Raouzaiou, A., Malatesta, L., Asteriadis, S., & Karpouzis, K. (2007). Multimodal emotion recognition from expressive faces, body gestures and speech. *IFIP International Conference on Artificial Intelligence Applications and Innovations*, 375–388.

www.manaraa.com

Chae, M., Kim, T.-H., Shin, Y. H., Kim, J.-W., & Lee, S.-Y. (2018). End-to-end Multimodal Emotion and Gender Recognition with Dynamic Joint Loss Weights. *ArXiv:1809.00758 [Cs, Eess, Stat]*. http://arxiv.org/abs/1809.00758

Chatziagapi, A., Paraskevopoulos, G., Sgouropoulos, D., Pantazopoulos, G., Nikandrou, M., Giannakopoulos, T., Katsamanis, A., Potamianos, A., & Narayanan, S. (2019). Data Augmentation Using GANs for Speech Emotion Recognition. *Interspeech 2019*, 171–175. https://doi.org/10.21437/Interspeech.2019-2561

Chaudhari, S. J., & Kagalkar, R. M. (2015). *Automatic Speaker Age Estimation and Gender Dependent Emotion Recognition*. https://doi.org/10.5120/20644-3383

Chen, O. T.-C., Gu, J. J., Lu, P.-T., & Ke, J.-Y. (2012). Emotion-inspired age and gender recognition systems. *2012 IEEE 55th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 662–665. https://doi.org/10.1109/MWSCAS.2012.6292107

Cho, J., Pappagari, R., Kulkarni, P., Villalba, J., Carmiel, Y., & Dehak, N. (2018). Deep Neural Networks for Emotion Recognition Combining Audio and Transcripts. *Interspeech 2018*, 247–251. https://doi.org/10.21437/Interspeech.2018-2466

Choudhury, A. R., Ghosh, A., Pandey, R., & Barman, S. (2018). Emotion Recognition from Speech Signals using Excitation Source and Spectral Features. *2018 IEEE Applied Signal Processing Conference (ASPCON)*, 257–261. https://doi.org/10.1109/ASPCON.2018.8748626

Crichton, K. G., Cooper, J. N., Minneci, P. C., Groner, J. I., Thackeray, J. D., & Deans, K. J. (2016). A national survey on the use of screening tools to detect physical child abuse. *Pediatric Surgery International*, *32*(8), 815–818. https://doi.org/10.1007/s00383-016-3916-z

Daneshfar, F., Kabudian, S. J., & Neekabadi, A. (2020). Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality

www.manaraa.com

reduction, and Gaussian elliptical basis function network classifier. *Applied Acoustics*, *166*, 107360. https://doi.org/10.1016/j.apacoust.2020.107360

Demircan, S., & Kahramanli, H. (2018). Application of fuzzy C-means clustering algorithm to spectral features for emotion classification from speech. *Neural Computing and Applications*, *29*(8), 59–66.

Drubach, L. A., Johnston, P. R., Newton, A. W., Perez-Rossello, J. M., Grant, F. D., & Kleinman, P. K. (2010). Skeletal Trauma in Child Abuse: Detection with 18F-NaF PET. *Radiology*, *255*(1), 173–181. https://doi.org/10.1148/radiol.09091368

Drubach, L. A., Sapp, Mark. V., Laffin, S., & Kleinman, P. K. (2008). Fluorine-18 NaF PET imaging of child abuse. *Pediatric Radiology*, *38*(7), 776–779. https://doi.org/10.1007/s00247-008-0885-y

Fatima, S. N., & Erzin, E. (2017). Cross-Subject Continuous Emotion Recognition Using Speech and Body Motion in Dyadic Interactions. *Interspeech 2017*, 1731–1735. https://doi.org/10.21437/Interspeech.2017-1413

Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Networks*, *92*, 60–68. https://doi.org/10.1016/j.neunet.2017.02.013

Flaherty, E. G., Perez-Rossello, J. M., Levine, M. A., Hennrikus, W. L., Neglect, and the A. A. O. P. C. O. C. A. A., Radiology, S. O., Endocrinology, S. O., Orthopaedics, S. O., & Radiology, the S. F. P. (2014). Evaluating Children With Fractures for Child Physical Abuse. *Pediatrics*, *133*(2), e477–e489. https://doi.org/10.1542/peds.2013-3793

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language*, 27–52.

Gamage, K. W., Sethu, V., & Ambikairajah, E. (2017). Salience based lexical features for emotion recognition. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5830–5834. https://doi.org/10.1109/ICASSP.2017.7953274

www.manaraa.com

Gangamohan, P., Kadiri, S. R., Gangashetty, S. V., & Yegnanarayana, B. (2014). *Excitation Source Features for Discrimination of Anger and Happy Emotions*. 5.

Garrett, M. F. (1980). The limits of accommodation: Arguments for independent processing levels in sentence production. *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen, and Hand*, 263–271.

Gideon, J., Schatten, H. T., McInnis, M. G., & Provost, E. M. (2019). Emotion Recognition from Natural Phone Conversations in Individuals with and without Recent Suicidal Ideation. *Interspeech 2019*, 3282–3286. https://doi.org/10.21437/Interspeech.2019-1830

Goel, N. K., Sarma, M., Kushwah, T. S., Agrawal, D. K., Iqbal, Z., & Chauhan, S. (2018). *Extracting speaker's gender, accent, age and emotional state from speech*. 2.

Gordon, M., & Ladefoged, P. (2001). Phonation types: A cross-linguistic overview. *Journal of Phonetics*, *29*(4), 383–406.

Guo, L., Wang, L., Dang, J., Zhang, L., Guan, H., & Li, X. (2018). Speech Emotion Recognition by Combining Amplitude and Phase Information Using Convolutional Neural Network. *Interspeech 2018*, 1611–1615. https://doi.org/10.21437/Interspeech.2018-2156

Han, J., Zhang, Z., Schmitt, M., Ren, Z., Ringeval, F., & Schuller, B. (2018). Bags in Bag: Generating Context-Aware Bags for Tracking Emotions from Speech. *Interspeech 2018*, 3082–3086. https://doi.org/10.21437/Interspeech.2018-996

Han, K., Yu, D., & Tashev, I. (2014). *Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine*. 5.

Hayek, S. N., Wibbenmeyer, L. A., Kealey, L. D. H., Williams, I. M., Oral, R., Onwuameze, O., Light, T. D., Latenser, B. A., Lewis, R. W., & Kealey, G. P. (2009). The Efficacy of Hair and Urine Toxicology Screening on the Detection of Child Abuse by Burning. *Journal of Burn Care & Research*, *30*(4), 587–592. https://doi.org/10.1097/BCR.0b013e3181abfd30

www.manaraa.com

Hazarika, D., Poria, S., Zadeh, A., Cambria, E., Morency, L.-P., & Zimmermann, R. (2018). Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2122–2132. https://doi.org/10.18653/v1/N18-1193

Huang, J., Li, Y., Tao, J., & Lian, Z. (2018). Speech Emotion Recognition from Variable-Length Inputs with Triplet Loss Function. *Interspeech 2018*, 3673–3677. https://doi.org/10.21437/Interspeech.2018-1432

Ito, M. (2002). Japanese politeness and suprasegmentals-a study based on natural speech materials. *Speech Prosody 2002, International Conference*.

Jaspan, T., Narborough, G., Punt, J. A. G., & Lowe, J. (1992). Cerebral contusional tears as a marker of child abuse—Detection by cranial sonography. *Pediatric Radiology*, *22*(4), 237–245. https://doi.org/10.1007/BF02019848

Jaudes, P. K. (1984). Comparison of Radiography and Radionuclide Bone Scanning in the Detection of Child Abuse. *Pediatrics*, *73*(2), 166–168.

Jeon, J. H., Xia, R., & Liu, Y. (2011). Sentence level emotion recognition based on decisions from subsentence segments. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4940–4943.

Kaczor, K., Clyde Pierce, M., Makoroff, K., & Corey, T. S. (2006). Bruising and Physical Child Abuse. *Clinical Pediatric Emergency Medicine*, *7*(3), 153–160. https://doi.org/10.1016/j.cpem.2006.06.007

Kadiri, S. R., Gangamohan, P., Gangashetty, S. V., Alku, P., & Yegnanarayana, B. (2020). Excitation Features of Speech for Emotion Recognition Using Neutral Speech as Reference. *Circuits, Systems, and Signal Processing*. https://doi.org/10.1007/s00034-020-01377-y

www.manaraa.com

Kaya, H., Salah, A. A., Karpov, A., Frolova, O., Grigorev, A., & Lyakso, E. (2017). Emotion, age, and gender classification in children's speech by humans and machines. *Computer Speech & Language*, *46*, 268–283. https://doi.org/10.1016/j.csl.2017.06.002

Kensinger, E. A., & Corkin, S. (2003). Effect of negative emotional content on working memory and long-term memory. *Emotion*, *3*(4), 378.

Kensinger, E. A., & Schacter, D. L. (2007). Remembering the specific visual details of presented objects: Neuroimaging evidence for effects of emotion. *Neuropsychologia*, *45*(13), 2951–2962.

Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Mahjoub, M. A., & Cleder, C. (2019). Automatic Speech Emotion Recognition Using Machine Learning. *Social Media and Machine Learning*. https://doi.org/10.5772/intechopen.84856

Khorram, S., Aldeneh, Z., Dimitriadis, D., McInnis, M., & Provost, E. M. (2017). Capturing Long-Term Temporal Dependencies with Convolutional Networks for Continuous Emotion Recognition. *Interspeech 2017*, 1253–1257. https://doi.org/10.21437/Interspeech.2017-548

Koolagudi, S. G., Devliyal, S., Chawla, B., Barthwal, A., & Rao, K. S. (2012). Recognition of Emotions from Speech using Excitation Source Features. *Procedia Engineering*, *38*, 3409–3417. https://doi.org/10.1016/j.proeng.2012.06.394

Koolagudi, S. G., Kumar, N., & Rao, K. S. (2011). Speech Emotion Recognition Using Segmental Level Prosodic Analysis. *2011 International Conference on Devices and Communications (ICDeCom)*, 1–5. https://doi.org/10.1109/ICDECOM.2011.5738536

Ladefoged, P. (1971). *Preliminaries to linguistic phonetics*. University of Chicago Press.

Lalitha, S., Mudupu, A., Nandyala, B. V., & Munagala, R. (2015). Speech emotion recognition using DWT. *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 1–4.

www.manaraa.com

Latif, S., Rana, R., Khalifa, S., Jurdak, R., & Epps, J. (2019). Direct Modelling of Speech Emotion from Raw Speech. *Interspeech 2019*, 3920–3924. https://doi.org/10.21437/Interspeech.2019-3252

Le, D., Aldeneh, Z., & Provost, E. M. (2017). Discretized Continuous Speech Emotion Recognition with Multi-Task Deep Recurrent Neural Network. *Interspeech 2017*, 1108–1112. https://doi.org/10.21437/Interspeech.2017-94

LeDoux, J. E. (1992). *Emotion and the amygdala.*

Lee, B., & Cho, K.-H. (2016). Brain-inspired speech segmentation for automatic speech recognition using the speech envelope as a temporal reference. *Scientific Reports*, *6*(1), 37647. https://doi.org/10.1038/srep37647

Lee, C. W., Song, K. Y., Jeong, J., & Choi, W. Y. (2018). Convolutional attention networks for multimodal emotion recognition from speech and text data. *ACL 2018*, 28.

Lee, J., & Tashev, I. (2015). *High-Level Feature Representation Using Recurrent Neural Network for Speech Emotion Recognition*. 4.

Levelt, W. J. (1993). *Speaking: From intention to articulation* (Vol. 1). MIT press.

Leventhal, H. (1974). *Emotions: A basic problem for social psychology.*

Li, J.-L., & Lee, C.-C. (2018). Encoding Individual Acoustic Features Using Dyad-Augmented Deep Variational Representations for Dialog-level Emotion Recognition. *Interspeech 2018*, 3102–3106. https://doi.org/10.21437/Interspeech.2018-1455

Li, X., & Akagi, M. (2018). A Three-Layer Emotion Perception Model for Valence and Arousal-Based Detection from Multilingual Speech. *Interspeech 2018*, 3643–3647. https://doi.org/10.21437/Interspeech.2018-1820

Li, Y., Zhao, T., & Kawahara, T. (2019). Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. *Interspeech 2019*, 2803–2807. https://doi.org/10.21437/Interspeech.2019-2594

www.manaraa.com

Lindsley, D. B. (1970). The role of nonspecific reticulo-thalamo-cortical systems in emotion. In *Physiological correlates of emotion* (pp. 147–184). Academic Press New York.

Liu, Z., Xu, A., Guo, Y., Mahmud, J. U., Liu, H., & Akkiraju, R. (2018). Seemo: A computational approach to see emotions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12.

Louwers, E. C. F. M., Affourtit, M. J., Moll, H. A., Koning, H. J. de, & Korfage, I. J. (2010). Screening for child abuse at emergency departments: A systematic review. *Archives of Disease in Childhood*, *95*(3), 214–218. https://doi.org/10.1136/adc.2008.151654

Louwers, E. C. F. M., Korfage, I. J., Affourtit, M. J., Ruige, M., van den Elzen, A. P. M., de Koning, H. J., & Moll, H. A. (2014). Accuracy of a screening instrument to identify potential child abuse in emergency departments. *Child Abuse & Neglect*, *38*(7), 1275–1281. https://doi.org/10.1016/j.chiabu.2013.11.005

Louwers, E. C. F. M., Korfage, I. J., Affourtit, M. J., Scheewe, D. J. H., Merwe, M. H. van de, Vooijs-Moulaert, A.-F. S. R., Elzen, A. P. M. van den, Jongejan, M. H. T. M., Ruige, M., Manaï, B. H. A. N., Looman, C. W. N., Bosschaart, A. N., Teeuw, A. H., Moll, H. A., & Koning, H. J. de. (2012). Effects of Systematic Screening and Detection of Child Abuse in Emergency Departments. *Pediatrics*, *130*(3), 457–464. https://doi.org/10.1542/peds.2011-3527

Louwers, E. C. F. M., Korfage, I. J., Affourtit, M. J., Scheewe, D. J. H., Merwe, M. H. van de, Vooijs-Moulaert, F. A. F. S. R., Woltering, C. M. C., Jongejan, M. H. T. M., Ruige, M., Moll, H. A., & Koning, H. J. D. (2011). Detection of child abuse in emergency departments: A multi-centre study. *Archives of Disease in Childhood*, *96*(5), 422–425. https://doi.org/10.1136/adc.2010.202358

www.manaraa.com

Louwers, E. C., Korfage, I. J., Affourtit, M. J., De Koning, H. J., & Moll, H. A. (2012). Facilitators and barriers to screening for child abuse in the emergency department. *BMC Pediatrics*, *12*(1), 167. https://doi.org/10.1186/1471-2431-12-167

Luo, D., Zou, Y., & Huang, D. (2018). Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition. *Interspeech 2018*, 152–156. https://doi.org/10.21437/Interspeech.2018-1832

Ma, X., Wu, Z., Jia, J., Xu, M., Meng, H., & Cai, L. (2018). Emotion Recognition from Variable-Length Speech Segments Using Deep Learning on Spectrograms. *Interspeech 2018*, 3683–3687. https://doi.org/10.21437/Interspeech.2018-2228

Mandelstam, S. A., Cook, D., Fitzgerald, M., & Ditchfield, M. R. (2003). Complementary use of radiological skeletal survey and bone scintigraphy in detection of bony injuries in suspected child abuse. *Archives of Disease in Childhood*, *88*(5), 387–390. https://doi.org/10.1136/adc.88.5.387

Mao, S., Ching, P. C., & Lee, T. (2019). Deep Learning of Segment-Level Feature Representation with Multiple Instance Learning for Utterance-Level Speech Emotion Recognition. *Interspeech 2019*, 1686–1690. https://doi.org/10.21437/Interspeech.2019-1968

Marañon, G. (1924). *Contribution a l'etude de l'action emotive de l'adrenaline*. Revue Française d'Endocrinologie.

Marchand, J., Deneyer, M., & Vandenplas, Y. (2012). Educational paper. *European Journal of Pediatrics*, *171*(1), 17–23. https://doi.org/10.1007/s00431-011-1616-1

Mary, L. (2019). *Extraction of Prosody for Automatic Speaker, Language, Emotion and Speech Recognition*. Springer International Publishing. https://doi.org/10.1007/978-3-319-91171-7

Mehrabian, A., & Russell, J. A. (1974). *An approach to environmental psychology.* the MIT Press.

Murty, K. S. R., Yegnanarayana, B., & Joseph, M. A. (2009). Characterization of glottal activity from speech signals. *IEEE Signal Processing Letters*, *16*(6), 469–472.

www.manaraa.com

Neumann, M., & Vu, N. T. (2017). Attentive Convolutional Neural Network Based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech. *Interspeech 2017*, 1263–1267. https://doi.org/10.21437/Interspeech.2017-917

Palo, H. K., & Mohanty, M. N. (2020). Analysis of Speech Emotions Using Dynamics of Prosodic Parameters. In P. K. Mallick, V. E. Balas, A. K. Bhoi, & G.-S. Chae (Eds.), *Cognitive Informatics and Soft Computing* (Vol. 1040, pp. 333–340). Springer Singapore. https://doi.org/10.1007/978-981-15-1451-7_36

Palo, H. K., Mohanty, M. N., & Chandra, M. (2018, January 1). *Speech Emotion Analysis of Different Age Groups Using Clustering Techniques* [Article]. International Journal of Information Retrieval Research (IJIRR). www.igi-global.com/article/speech-emotion-analysis-of-different-age-groups-using-clustering-techniques/193250

Palo, H. K., Mohanty, M. N., & Chandra, M. (2017). *Emotion Analysis from Speech of Different Age Groups*. 283–287. https://doi.org/10.15439/2017R21

Parry, J., Palaz, D., Clarke, G., Lecomte, P., Mead, R., Berger, M., & Hofer, G. (2019). Analysis of Deep Learning Architectures for Cross-Corpus Speech Emotion Recognition. *Interspeech 2019*, 1656–1660. https://doi.org/10.21437/Interspeech.2019-2753

Pietrowicz, M., Hasegawa-Johnson, M., & Karahalios, K. G. (2017). Acoustic correlates for perceived effort levels in male and female acted voices. *The Journal of the Acoustical Society of America*, *142*(2), 792–811.

Pohjalainen, J., Fabien Ringeval, F., Zhang, Z., & Schuller, B. (2016). Spectral and cepstral audio noise reduction techniques in speech emotion recognition. *Proceedings of the 24th ACM International Conference on Multimedia*, 670–674.

Prasanna, S. R. M., & Govind, D. (2010). *Analysis of Excitation Source Information in Emotional Speech*. 4.

www.manaraa.com

Pravena, D., & Govind, D. (2017). Significance of incorporating excitation source parameters for improved emotion recognition from speech and electroglottographic signals. *International Journal of Speech Technology*, *20*(4), 787–797. https://doi.org/10.1007/s10772-017-9445-x

Qawaqneh, Z., Mallouh, A. A., & Barkana, B. D. (2017). Deep neural network framework and transformed MFCCs for speaker's age and gender classification. *Knowledge-Based Systems*, *115*, 5–14. https://doi.org/10.1016/j.knosys.2016.10.008

Rao, K. S., & Koolagudi, S. G. (2013a). *Emotion Recognition using Speech Features*. Springer New York. https://doi.org/10.1007/978-1-4614-5143-3

Rao, K. S., & Koolagudi, S. G. (2013b). *Robust Emotion Recognition using Spectral and Prosodic Features*. Springer New York. https://doi.org/10.1007/978-1-4614-6360-3

Regnaut, O., Jeu-Steenhouwer, M., Manaouil, C., & Gignon, M. (2015). Risk factors for child abuse: Levels of knowledge and difficulties in family medicine. A mixed method study. *BMC Research Notes*, *8*(1), 620. https://doi.org/10.1186/s13104-015-1607-9

Resonation. (1998). https://www.yorku.ca/earmstro/journey/resonation.html

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161.

Sarma, M., Ghahremani, P., Povey, D., Goel, N. K., Sarma, K. K., & Dehak, N. (2018). Emotion Identification from Raw Speech Signals Using DNNs. *Interspeech 2018*, 3097–3101. https://doi.org/10.21437/Interspeech.2018-1353

Sarma, M., Ghahremani, P., Povey, D., Goel, N. K., Sarma, K. K., & Dehak, N. (2019). Improving Emotion Identification Using Phone Posteriors in Raw Speech Waveform Based DNN. *Interspeech 2019*, 3925–3929. https://doi.org/10.21437/Interspeech.2019-2093

www.manaraa.com

Sato, Y., Yuh, W. T., Smith, W. L., Alexander, R. C., Kao, S. C., & Ellerbroek, C. J. (1989). Head injury in child abuse: Evaluation with MR imaging. *Radiology*, *173*(3), 653–657. https://doi.org/10.1148/radiology.173.3.2813768

Satt, A., Rozenberg, S., & Hoory, R. (2017). Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. *Interspeech 2017*, 1089–1093. https://doi.org/10.21437/Interspeech.2017-200

Schols, M. W., de Ruiter, C., & Öry, F. G. (2013). How do public child healthcare professionals and primary school teachers identify and handle child abuse cases? A qualitative study. *BMC Public Health*, *13*(1), 807. https://doi.org/10.1186/1471-2458-13-807

Sebastian, J., & Pierucci, P. (2019). Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts. *Interspeech 2019*, 51–55. https://doi.org/10.21437/Interspeech.2019-3201

Shaqra, F. A., Duwairi, R., & Al-Ayyoub, M. (2019). Recognizing Emotion from Speech Based on Age and Gender Using Hierarchical Models. *Procedia Computer Science*, *151*, 37–44. https://doi.org/10.1016/j.procs.2019.04.009

Sittig, J. S., Uiterwaal, C. S. P. M., Moons, K. G. M., Russel, I. M. B., Nievelstein, R. A. J., Nieuwenhuis, E. E. S., & Putte, E. M. van de. (2016). Value of systematic detection of physical child abuse at emergency rooms: A cross-sectional diagnostic accuracy study. *BMJ Open*, *6*(3), e010788. https://doi.org/10.1136/bmjopen-2015-010788

Smeekens, A. E. F. N., Henten, D. M. B., Sittig, J. S., Russel, I. M. B., Cate, O. T. J. ten, Turner, N. M., & Putte, E. M. van de. (2011). Successful e-learning programme on the detection of child abuse in Emergency Departments: A randomised controlled trial. *Archives of Disease in Childhood*, *96*(4), 330–334. https://doi.org/10.1136/adc.2010.190801

www.manaraa.com

Söderkvist, S., Ohlén, K., & Dimberg, U. (2018). How the Experience of Emotion is Modulated by Facial Feedback. *Journal of Nonverbal Behavior*, *42*(1), 129–151. https://doi.org/10.1007/s10919-017-0264-1

Strongman, K. T. (2003). *The psychology of emotion: From everyday life to theory* (5th ed). Wiley.

Sun, Y., Wen, G., & Wang, J. (2015). Weighted spectral features based on local Hu moments for speech emotion recognition. *Biomedical Signal Processing and Control*, *18*, 80–90.

Syed, Z. S., Schroeter, J., Sidorov, K., & Marshall, D. (2018). Computational Paralinguistics: Automatic Assessment of Emotions, Mood and Behavioural State from Acoustics of Speech. *Interspeech 2018*, 511–515. https://doi.org/10.21437/Interspeech.2018-2019

Tan, Z.-H., & Lindberg, B. (2010). Low-complexity variable frame rate analysis for speech recognition and voice activity detection. *IEEE Journal of Selected Topics in Signal Processing*, *4*(5), 798–807.

Tang, D., Zeng, J., & Li, M. (2018). An End-to-End Deep Learning Framework for Speech Emotion Recognition of Atypical Individuals. *Interspeech 2018*, 162–166. https://doi.org/10.21437/Interspeech.2018-2581

Tarantino, L., Garner, P. N., & Lazaridis, A. (2019). Self-Attention for Speech Emotion Recognition. *Interspeech 2019*, 2578–2582. https://doi.org/10.21437/Interspeech.2019-2822

Teeuw, A. H., Kraan, R. B. J., Rijn, R. R. van, Bossuyt, P. M. M., & Heymans, H. S. A. (2019). Screening for child abuse using a checklist and physical examinations in the emergency department led to the detection of more cases. *Acta Paediatrica*, *108*(2), 300–313. https://doi.org/10.1111/apa.14495

Tiyyagura, G., Gawel, M., Koziel, J. R., Asnes, A., & Bechtel, K. (2015). Barriers and Facilitators to Detecting Child Abuse and Neglect in General Emergency Departments. *Annals of Emergency Medicine*, *66*(5), 447–454. https://doi.org/10.1016/j.annemergmed.2015.06.020

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent

www.manaraa.com

network. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5200–5204. https://doi.org/10.1109/ICASSP.2016.7472669

Verma, D., & Mukhopadhyay, D. (2016). Age driven automatic speech emotion recognition system. *2016 International Conference on Computing, Communication and Automation (ICCCA)*, 1005–1010. https://doi.org/10.1109/CCAA.2016.7813862

Vuppala, A. K., Yadav, J., Chakrabarti, S., & Rao, K. S. (2012). Vowel onset point detection for low bit rate coded speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(6), 1894–1903.

Walker, P. L., Cook, D. C., & Lambert, P. M. (1997). Skeletal Evidence for Child Abuse: A Physical Anthropological Perspective. *Journal of Forensic Sciences*, *42*(2), 14098J. https://doi.org/10.1520/JFS14098J

Walsh, K., Bridgstock, R., Farrell, A., Rassafiani, M., & Schweitzer, R. (2008). Case, teacher and school characteristics influencing teachers' detection and reporting of child physical abuse and neglect: Results from an Australian survey. *Child Abuse & Neglect*, *32*(10), 983–993. https://doi.org/10.1016/j.chiabu.2008.03.002

Wang, B., Liakata, M., Ni, H., Lyons, T., Nevado-Holgado, A. J., & Saunders, K. (2019). A Path Signature Approach for Speech Emotion Recognition. *Interspeech 2019*, 1661–1665. https://doi.org/10.21437/Interspeech.2019-2624

Wang, Y., Du, S., & Zhan, Y. (2008). Adaptive and Optimal Classification of Speech Emotion Recognition. *2008 Fourth International Conference on Natural Computation*, *5*, 407–411. https://doi.org/10.1109/ICNC.2008.713

Wang, Z.-Q., & Tashev, I. (2017). Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5150–5154. https://doi.org/10.1109/ICASSP.2017.7953138

www.manaraa.com

Wills, R., Ritchie, M., & Wilson, M. (2008). Improving detection and quality of assessment of child abuse and partner abuse is achievable with a formal organisational change approach. *Journal of Paediatrics and Child Health*, *44*(3), 92–98. https://doi.org/10.1111/j.1440-1754.2007.01276.x

Yadav, I. C., Kumar, A., Shahnawazuddin, S., & Pradhan, G. (2018). Non-Uniform Spectral Smoothing for Robust Children's Speech Recognition. *Interspeech 2018*, 1601–1605. https://doi.org/10.21437/Interspeech.2018-1828

Yeh, J.-H., Pao, T.-L., Lin, C.-Y., Tsai, Y.-W., & Chen, Y.-T. (2011). Segment-based emotion recognition from continuous Mandarin Chinese speech. *Computers in Human Behavior*, *27*(5), 1545–1552.

Yenigalla, P., Kumar, A., Tripathi, S., Singh, C., Kar, S., & Vepa, J. (2018). Speech Emotion Recognition Using Spectrogram and Phoneme Embedding. *Interspeech 2018*, 3688–3692. https://doi.org/10.21437/Interspeech.2018-1811

Yogesh, C. K., Hariharan, M., Ngadiran, R., Adom, A. H., Yaacob, S., & Polat, K. (2017). Hybrid BBO_PSO and higher order spectral features for emotion and stress recognition from natural speech. *Applied Soft Computing*, *56*, 217–232.

Youmans, D. C., Don, S., Hildebolt, C., Shackelford, G. D., Luker, G. D., & McAlister, W. H. (1998). Skeletal surveys for child abuse: Comparison of interpretation using digitized images and screen-film radiographs. *American Journal of Roentgenology*, *171*(5), 1415–1419. https://doi.org/10.2214/ajr.171.5.9798889

Yücesoy, E. (2020). Speaker age and gender classification using GMM supervector and NAP channel compensation method. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-020-02045-4

Zhang, L., Wang, L., Dang, J., Guo, L., & Yu, Q. (2018). Gender-Aware CNN-BLSTM for Speech Emotion Recognition. In V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, & I. Maglogiannis (Eds.),

www.manaraa.com

Artificial Neural Networks and Machine Learning – ICANN 2018 (pp. 782–790). Springer International Publishing. https://doi.org/10.1007/978-3-030-01418-6_76

Zhao, H., Ye, N., & Wang, R. (2020). Speech emotion recognition based on hierarchical attributes using feature nets. *International Journal of Parallel, Emergent and Distributed Systems*, *35*(3), 354–364. https://doi.org/10.1080/17445760.2019.1626854

Zhao, J., Chen, S., Liang, J., & Jin, Q. (2019). Speech Emotion Recognition in Dyadic Dialogues with Attentive Interaction Modeling. *Interspeech 2019*, 1671–1675. https://doi.org/10.21437/Interspeech.2019-2103

Zheng, W. Q., Yu, J. S., & Zou, Y. X. (2015). An experimental study of speech emotion recognition based on deep convolutional neural networks. *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 827–831.

Zhou, Y., Sun, Y., Zhang, J., & Yan, Y. (2009). Speech Emotion Recognition Using Both Spectral and Prosodic Features. *2009 International Conference on Information Engineering and Computer Science*, 1–4. https://doi.org/10.1109/ICIECS.2009.5362730

Zvarevashe, K., & Olugbara, O. (2020). Ensemble Learning of Hybrid Acoustic Features for Speech Emotion Recognition. *Algorithms*, *13*(3), 70. https://doi.org/10.3390/a13030070

Zveglic, E. A. (2014). Chapter 7.5—Speech and singing. In L. Chaitow, D. Bradley, & C. Gilbert (Eds.), *Recognizing and Treating Breathing Disorders (Second Edition)* (pp. 203–214). Churchill Livingstone. https://doi.org/10.1016/B978-0-7020-4980-4.00019-8

www.manaraa.com

References

Abdelwahab, M. (2019). *Domain Adaptation for Speech Based Emotion Recognition* [Ph.D., The University of Texas at Dallas]. https://search.proquest.com/pqdtglobal/docview/2355976165/abstract/E46F7FD1C9174097PQ/65

Al-Talabani, A. (2015). *Automatic Speech Emotion Recognition- Feature Space Dimensionality and Classification Challenges* [Doctoral, University of Buckingham]. http://bear.buckingham.ac.uk/101/

Bays, J. (1990). Substance Abuse and Child Abuse: Impact of Addiction on the Child. *Pediatric Clinics of North America*, *37*(4), 881–904. https://doi.org/10.1016/S0031-3955(16)36941-3

Benasich, A. A., & Brooks-Gunn, J. (1996). Maternal attitudes and knowledge of child-rearing: Associations with family and child outcomes. *Child Development*, *67*(3), 1186–1205.

Burkhardt, F., Eckert, M., Johannsen, W., & Stegmann, J. (2010). A Database of Age and Gender Annotated Telephone Speech. *LREC*.

Busso, C., Bulut, M., Narayanan, S., Gratch, J., & Marsella, S. (2013). Toward effective automatic recognition systems of emotion in speech. *Social Emotions in Nature and Artifact: Emotions in Human and Human-Computer Interaction, J. Gratch and S. Marsella, Eds*, 110–127.

Douglas-Cowie, E., Devillers, L., Martin, J.-C., Cowie, R., Savvidou, S., Abrilian, S., & Cox, C. (2005). Multimodal databases of everyday emotion: Facing up to complexity. *Ninth European Conference on Speech Communication and Technology*.

Fatima, S. N., & Erzin, E. (2017). Cross-Subject Continuous Emotion Recognition Using Speech and Body Motion in Dyadic Interactions. *Interspeech 2017*, 1731–1735. https://doi.org/10.21437/Interspeech.2017-1413

Fischer, K. E., Kittleson, M., Ogletree, R., Welshimer, K., Woehlke, P., & Benshoff, J. (2000). The relationship of parental alcoholism and family dysfunction to stress among college students. *Journal of American College Health*, *48*(4), 151–156.

Gunn, V. L., Hickson, G. B., & Cooper, W. O. (2005). Factors affecting pediatricians' reporting of suspected child maltreatment. *Ambulatory Pediatrics*, *5*(2), 96–101.

Jones, R., Flaherty, E. G., Binns, H. J., Price, L. L., Slora, E., Abney, D., Harris, D. L., Christoffel, K. K., & Sege, R. D. (2008). Clinicians' description of factors influencing their reporting of suspected child abuse: Report of the Child Abuse Reporting Experience Study Research Group. *Pediatrics*, *122*(2), 259–266.

Kempe, C. H., Silverman, F. N., Steele, B. F., Droegemueller, W., & Silver, H. K. (1962). The Battered-Child Syndrome. *JAMA*, *181*(1), 17–24. https://doi.org/10.1001/jama.1962.03050270019004

المنارة للاستشارات
www.manaraa.com

Kuczkowski, K. M. (2004). Marijuana in pregnancy. *ANNALS-ACADEMY OF MEDICINE SINGAPORE*, *33*, 336–339.

Kuczkowski, Krzysztof M. (2007). The effects of drug abuse on pregnancy. *Current Opinion in Obstetrics and Gynecology*, *19*(6), 578–585.

Law, P. (1974). Law 93-247. *The Child Abuse and Prevention Treatment Act Of*.

Lazenbatt, A., & Freeman, R. (2006). Recognizing and reporting child physical abuse: A survey of primary healthcare professionals. *Journal of Advanced Nursing*, *56*(3), 227–236.

Lotfian, R. (2018). *Machine Learning Solutions for Emotional Speech: Exploiting the Information of Individual Annotations* [Ph.D., The University of Texas at Dallas]. https://search.proquest.com/pqdtglobal/docview/2355978737/abstract/E46F7FD1C9174097PQ/275

Lotfian, R., & Busso, C. (2017). Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*.

Ma, X., Wu, Z., Jia, J., Xu, M., Meng, H., & Cai, L. (2018). Emotion Recognition from Variable-Length Speech Segments Using Deep Learning on Spectrograms. *Interspeech 2018*, 3683–3687. https://doi.org/10.21437/Interspeech.2018-2228

Marchand, J., Deneyer, M., & Vandenplas, Y. (2012). Detection, diagnosis, and prevention of child abuse: The role of the pediatrician. *European Journal of Pediatrics*, *171*(1), 17–23. https://doi.org/10.1007/s00431-011-1616-1

Markus, L., & Sandison, B. (2020). *Australia's children. Australian Institute of Health and Welfare 2020.* https://nla.gov.au/nla.obj-2497069928

Pelletier, G., & Handy, L. C. (1986). Family Dysfunction and the Psychological Impact of Child Sexual Abuse. *The Canadian Journal of Psychiatry*, *31*(5), 407–412. https://doi.org/10.1177/070674378603100504

Qawaqneh, Z., Mallouh, A. A., & Barkana, B. D. (2017). Deep neural network framework and transformed MFCCs for speaker's age and gender classification. *Knowledge-Based Systems*, *115*, 5–14. https://doi.org/10.1016/j.knosys.2016.10.008

Raphael, L. J., Borden, G. J., & Harris, K. S. (2011). *Speech science primer: Physiology, acoustics, and perception of speech* (6th ed). Wolters Kluwer Health/Lippincott Williams & Wilkins.

Rayno, G. (2020, July 20). *Enhanced Child Protection Measures Now Law*. InDepthNH.Org. http://indepthnh.org/2020/07/20/enhanced-child-protection-measures-now-law/

Regnaut, O., Jeu-Steenhouwer, M., Manaouil, C., & Gignon, M. (2015). Risk factors for child abuse: Levels of knowledge and difficulties in family medicine. A mixed method study. *BMC Research Notes*, *8*(1), 620. https://doi.org/10.1186/s13104-015-1607-9

المنارة للاستشارات

www.manaraa.com

Sahu, S., Gupta, R., Sivaraman, G., AbdAlmageed, W., & Espy-Wilson, C. (2017). Adversarial Auto-Encoders for Speech Based Emotion Recognition. *Interspeech 2017*, 1243–1247. https://doi.org/10.21437/Interspeech.2017-1421

Sarma, M., Ghahremani, P., Povey, D., Goel, N. K., Sarma, K. K., & Dehak, N. (2019). Improving Emotion Identification Using Phone Posteriors in Raw Speech Waveform Based DNN. *Interspeech 2019*, 3925–3929. https://doi.org/10.21437/Interspeech.2019-2093

Satt, A., Rozenberg, S., & Hoory, R. (2017). Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. *Interspeech 2017*, 1089–1093. https://doi.org/10.21437/Interspeech.2017-200

Schweitzer, R. D., Buckley, L., Harnett, P., & Loxton, N. J. (2006). Predictors of failure by medical practitioners to report suspected child abuse in Queensland, Australia. *Australian Health Review*, *30*(3), 298–304.

Stedt, E. V. (2019). Child Maltreatment 2017. *U.S. Department of Health & Human Services Administration for Children and Families Administration on Children, Youth and Families Children's Bureau*, 261.

Stedt, E. V. (2020). Child Maltreatment 2018. *U.S. Department of Health & Human Services Administration for Children and Families Administration on Children, Youth and Families Children's Bureau*, 274.

Teeuw, A. H., Kraan, R. B. J., Rijn, R. R. van, Bossuyt, P. M. M., & Heymans, H. S. A. (2019). Screening for child abuse using a checklist and physical examinations in the emergency department led to the detection of more cases. *Acta Paediatrica*, *108*(2), 300–313. https://doi.org/10.1111/apa.14495

Tiyyagura, G., Gawel, M., Koziel, J. R., Asnes, A., & Bechtel, K. (2015). Barriers and Facilitators to Detecting Child Abuse and Neglect in General Emergency Departments. *Annals of Emergency Medicine*, *66*(5), 447–454. https://doi.org/10.1016/j.annemergmed.2015.06.020

Wheeler, S. M., Williams, L., Beauchesne, P., & Dupras, T. L. (2013). Shattered lives and broken childhoods: Evidence of physical child abuse in ancient Egypt. *International Journal of Paleopathology*, *3*(2), 71–82.

Wu, Y.-T., Chen, H.-Y., Liao, Y.-H., Kuo, L.-W., & Lee, C.-C. (2017). Modeling Perceivers Neural-Responses Using Lobe-Dependent Convolutional Neural Network to Improve Speech Emotion Recognition. *Interspeech 2017*, 3261–3265. https://doi.org/10.21437/Interspeech.2017-562

Yu, H., Tan, Z.-H., Ma, Z., & Guo, J. (2017). Adversarial Network Bottleneck Features for Noise Robust Speaker Verification. *Interspeech 2017*, 1492–1496. https://doi.org/10.21437/Interspeech.2017-883

www.manaraa.com

www.manaraa.com