SPEECH EMOTION RECOGNITION AND DEPRESSION PREDICTION USING DEEP NEURAL NETWORKS

BY

MUHAMMAD FAHREZA ALGHIFARI

A thesis submitted in fulfilment of the requirement for the degree of Master of Science (Computer and Information Engineering)

Kulliyyah of Engineering
International Islamic University Malaysia

JULY 2021

ABSTRACT

Speech signals contain ample information from which computers can gain insight into the user's state, including emotion recognition and depression prediction. The applications are numerous, from customer service to suicide prevention due to depression. In this research, we propose several deep-learning-based methodologies to detect emotion, as well as depression. Deep neural networks variations such as deep feedforward networks and convolutional networks were used. The deep learning model training, multi-languages emotion and depression database have been utilized, using well-known databases such as the Berlin Emotion Database and DAIC-WOZ Depression Dataset. For speech emotion recognition, the algorithm yields an accuracy of 80.5% across 4 languages, English, German, French and Italian. For depression detection, the current algorithm obtains an accuracy of 60.1% tested with the DAIC-WOZ dataset. This research has also created the Sorrow Analysis Dataset – an English depression audio dataset that contains 64 individuals samples of depressed and not-depressed. Further testing achieved an average accuracy of 97% with 5-fold validation using 1-dimensional convolutional networks. Finally, a prototype currently in development with Skymind Xpress.ai is presented, outlining the design and possible applications in the real world. It has been shown that the model is capable of performing both training and inference on a Raspberry Pi 3B+.

خلاصة البحث

تحتوي إشارات الكلام على معلومات وافرة يمكن لأجهزة الكمبيوتر من خلالها اكتساب نظرة ثاقبة على حالة المستخدم مثل معرفة المشاعر أو التنبؤ بحالات الاكتئاب. التطبيقات الممكنة لهذا المجال متعددة، يمكن أن تدرج من عملاء خدمة الزبائن إلى منع حالات الإنتحار الناتجة عن الإكتئاب. في هذا البحث، تم تقديم عدة منهجيات بالإعتماد على التعلم العميق لكشف المشاعر بالإضافة للإكتئاب. عدد من شبكات التعلم العميق مثل شبكات الإدخال المباشر و شبكات المسح الإلتفافي استخدمت في هذا البحث. لتدريب شبكة التعلم العميق، تم استخدام قواعد بيانات متعددة اللغات عن المشاعر و الإكتئاب، منها قواعد البيانات العروفة Berlin Emotion Database و DAIC-WOZ Depression Dataset.للتعرف على عاطفة الكلام ، الخوارزمية المطورة انتجت دقة تصل إلى 80.5٪ عبر 4 لغات ، الإنجليزية ، الألمانية ، الفرنسية والإيطالية. بالنسبة لاكتشاف الاكتئاب ، تحصل الخوارزمية الحالية على دقة تبلغ 60.1٪ تم اختبارها باستخدام مجموعة بيانات .DAIC-WOZ وقد أنشأ هذا البحث أيضًا مجموعة بيانات تحليل الحزن - وهي مجموعة بيانات صوتية باللغة الإنجليزية للاكتئاب تحتوي على عينات من الاكتئاب وغير المكتئب من 64 فردًا. حققت الاختبارات الإضافية متوسط دقة بنسبة 97 في المائة مع تحقق 5 أضعاف باستخدام شبكة تلافيفية أحادية البعد. أخيرًا ، يتم تقديم نموذج أولي قيد التطوير حاليًا مع Skymind Xpress.ai، يوضح التصميم بالإضافة إلى التطبيقات الممكنة في العالم الحقيقي. الشكبة المطورة أظهرت إمكانية تدريبها و تشغيلها على وحدة Raspberry Pi 3B.

APPROVAL PAGE

I certify that I have supervised and read this study and that, in my opinion, it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a thesis for the degree of Master of Science (Computer and Information Engineering).

……..
Teddy Surya Gunawan
Supervisor

……..
Mimi Aminah binti Wan Nordin
Co-Supervisor

……..
Nik Nur Wahidah binti Nik Hashim
Co-Supervisor

I certify that I have read this study and that, in my opinion, it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a thesis for the degree of Master of Science (Computer and Information Engineering).

……..
Othman O. Khalifa
Internal Examiner

……..
Kamarul Hawari Ghazali
External Examiner

This thesis was submitted to the Department of Electrical and Computer Engineering and is accepted as a fulfilment of the requirement for the degree of Master of Science (Computer and Information Engineering).

……..
Mohamed Hadi Habaebi
Head, Department of Electrical and Computer Engineering

This thesis was submitted to the Kulliyyah of Engineering and is accepted as a fulfilment of the requirement for the degree of Master of Science (Computer and Information Engineering).

……..
Sany Izan Ihsan
Dean, Kulliyyah of Engineering

DECLARATION

I hereby declare that this thesis is the result of my own investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Muhammad Fahreza Alghifari

Signature ..... Date .....

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

SPEECH EMOTION RECOGNITION AND DEPRESSION PREDICTION USING DEEP NEURAL NETWORKS

I declare that the copyright holders of this thesis are jointly owned by the student and IIUM.

Affirmed by Muhammad Fahreza Alghifari

……..…….. ……..
Signature Date

ACKNOWLEDGEMENT

Firstly, I would like to thank Allah Almighty for the opportunity and strength to complete my master's thesis.

It is my utmost pleasure to dedicate this work to my dear parents and my family, who granted me the gift of their unwavering belief in my ability to accomplish this goal: thank you for your support and patience.

A special thanks to Professor Teddy for his continuous support, encouragement, and leadership, and for that, I will be forever grateful. The thesis would not be complete without the support from my two co-supervisors, Dr. Mimi and Dr. Nik. They have supplemented the areas where I was lacking. I would also like to mention Brother Asif, who has helped me in data collection and moral support. Thank you.

I would like to thank IIUM Counselling and Career Unit for their help and feedback in developing the research. I hope the research I've completed will be useful for the center. I would also like to thank IIUM RMC for supporting the research team to participate in exhibitions. I hope the achievements obtained made you proud.

I would like to thank Mr. Eduardo and Skymind Xpress.Ai for the internship opportunity. I've received valuable guidance and experience in building industrial-level deep learning models and allowing me to use the company server to train my depression model.

Finally, I wish to express my appreciation and thanks to those who provided their time, effort, and support for this project. To the members of my dissertation committee, thank you for sticking with me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
CNN	Convolutional Neural Network
DNN	Deep Neural Network
LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepstral Coefficients
ML	Machine Learning
SER	Speech Emotion Recognition

LIST OF SYMBOLS

c_i	MFCC Coefficient
$Mel(f)$`	Mel-Scale Frequency
$\psi((x)t)$	TEO Coefficient
X_t	Input of Neural Network at time *t*
S_t	Hidden State of Neural Network at time *t*
O_t	Output of Neural Network at time *t*

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND OF THE STUDY

Speech, which is a part of human's daily life interactions, is full of information. Words from daily conversations can then be linguistically processed for data mining. Automatic speech recognition is one field that deals with capturing speech signals and transcribing them to text, which machines can further process.

But what is under the literal sense of words is often the intentions behind it. The words "I am okay" can be interpreted in many ways depending on how the speaker conveys it. A normal or even positive statement could be a warning to stay away if the speaker is angry or even be a secret cry of help if the speaker is depressed. Hence, the user's state, or to an extent, "the user's emotions," is important when considering human-to-computer interface implementations. This field has developed into affective computing, which can be understood as a field to enable intelligent systems to recognize, feel, infer and interpret human emotions (Poria, Cambria, Bajpai, & Hussain, 2017).

Speech emotion recognition (SER) was described by (Schuller, 2018) as the idea of lending machines emotional intelligence able to recognize human emotion and to synthesize emotion and emotional behavior. Scholars have proposed multiple ways of classifying emotions, but generally, they include happy, angry, sad, fearful, surprise, and neutral emotions. The list of emotions in several databases can be viewed in the literature review – database section. Regardless of the variations, the main purpose of SER is to extract emotional state. The research topic has been worked on for over two decades (Akçay & Oğuz, 2020), and only recently gained much attention due to the rise

of deep learning. The review by (Gunawan, Alghifari, Morshidi, & Kartiwi, 2018) observes the growing numbers of publications related to SER in IEEE, from less than 50 papers in the early 2000s to more than 150 papers in the late 2020s.

While the idea of speech emotion recognition may sound grand, the concept behind it is quite simple. Speech emotion recognition is nothing but the pattern recognition system (Ingale & Chaudhari, 2012), as the general flow is identical to that of pattern recognition in face recognition or signature recognition. The challenge comes from constructing a system that can recognize these emotional patterns, from choosing the correct features to constructing a robust system that is accurate and low cost in processing time and resources. The possible speech emotion recognition applications are numerous, such as in the call center industry to monitor a customer's satisfaction, e-learning to detect whether the student is frustrated, or even in the horror entertainment industry.

Another focus of this research is the detection of depression from speech signals. Depression is defined by Hooley et al. (Hooley, Butcher, Nock, & Mineka, 2017) in Abnormal Psychology as "an emotional state characterized by extraordinary sadness and dejection." They explain that depression is a mood disorder and is unusually severe or prolonged and impair their ability to function when dealing with responsibilities, showing that there are major depressive disorder and dysthymic disorder, which is milder.

The effect of depression is disastrous. Depression is one of the leading causes of suicides in the modern world, according to the World Health Organization (WHO), which is around 800 thousand each year. Their report estimated that the total number of people suffering from depression in 2015 exceeds 300 million, depression being the largest contributor to global disability (World Health Organization, 2017).

In (Bhaduri, Chakraborty, & Ghosh, 2016), the author explains that suicide occurs when a human being is under 'abnormal' mental conditions. These mental health problems are heavily influenced by emotions, which can be discerned from perturbation in conversations or speech. There has been extensive research in detecting stress and suicidal tendencies from speech signals from a recent study (Yingthawornsuk, 2016) and (Venek, Scherer, Morency, Rizzo, & Pestian, 2017) where the latter has highlighted that suicidal behavior often remains untreated or undetected. Those who are aware of their condition rarely seek professional care or help.

It is worth mentioning that there several other mediums to which emotion recognition and depression prediction can be performed. For example, there is much research on detecting depression from video analysis (Mantri, Patil, Agrawal, & Wadhai, 2015), image processing (He, Jiang, & Sahli, 2019), EEG signals (Guo, Zhang, & Pang, 2017), or even social media feeds (Deshpande & Rao, 2017). However, we have found the speech to be the least intrusive (does not require a video cam setup), can be easily implemented anywhere (does not require special setup like EEG), and has a solid use case discussed in Chapter 5.2. Therefore, considering all the above, emotion recognition and depression detection are the focus of this research.

1.2 PROBLEM STATEMENT

This research has two problem statements.

1. Speech emotion recognition is currently a research hotspot with a growing number of paper submissions each year, yet a gold standard has not been found due to various factors to consider, such as different speech features, databases, and classifiers. SER has potential applications in real life, to which this research aims to be implanted in counseling services and call centers to detect emotional instability.

2. Depression is one of the leading causes of suicide. This problem is also occurring in Malaysia. Every day, 20 of the 68 people who call the Befrienders Kuala Lumpur hotline have suicidal thoughts. These numbers continue to rise, from 21,256 in 2015 to 24,821 in 2016 – a 16% increase. Speaking in terms of calls - 5,739 people have tried to reach out by phone in 2015 and now to a sum of 7,446 who called in 2016 had suicidal intentions (Pillay, 2017; Radhi, 2018). Although there is existing research in depression detection, very few of them are focused on Malaysia and, far as this research knows, have no working implementation. This research would like to encourage more research on the field by proposing a deep learning approach for depression detection and a prototype.

1.3 RESEARCH OBJECTIVES

The study aimed to achieve the following objectives:

1- To utilize available emotion and depression speech databases across many languages and develop a speech depression audio database.

2- To design a robust deep learning methodology that includes deep neural network configuration and training parameters.

3- To evaluate the proposed algorithm's performance in terms of accuracy and computational time benchmarked with other research in the same topic.

4- To design a prototype that implements the system proposed in this research

1.4 SIGNIFICANCE OF THE RESEARCH

The significance of this research is threefold. From the SER part, we hope to achieve a robust emotion recognition system that can be used for real-life applications, such as customer service.

From the depression detection side, this research's outcome is hoped to assist suicide prevention hotlines in dealing with suicidal callers suffering from depression, which may be difficult to be detected in normal ways. The research is also hoped to encourage more future researchers to investigate ways to help depressed individuals. Finally, this research is also hoped to be an awareness campaign on mental health, as any individual can suffer from it yet may not get the deserved help.

The prototype development is a working collaboration between IIUM and Skymind XpressAi. Aside from a tangible proof-of-concept that implements the theories in this research, the project is hoped to encourage more collaboration between industry and the academic field.

1.5 RESEARCH SCOPE AND LIMITATIONS

The research will primarily focus on detecting emotions and depression solely through speech signals using a speech processing deep learning approach. The primary target of this research was to be conducted in Malaysia. The justification of this method is elaborated more in Chapter 2, Literature Review.

Several limitations need to be addressed. In Chapter 4, the data collection was intended to have multiple languages, including the Malay language, but due to COVID-19, which limits physical interactions, the data collection needed to be performed online, and only English was collected. In Chapter 5, the prototype is discussed; however, further testing was not conducted due to time constraints, and only the overview design is discussed.

1.6 ORGANIZATION

The rest of the thesis is organized as follows. Chapter 2 is the literature review which discusses common speech features, classifiers, and databases related to speech emotion recognition and depression detection. Chapter 3 elaborates research conducted on speech emotion recognition which encompasses multi-language datasets. Chapter 4 focuses on depression detection in three phases, which include data collection. Chapter 5 elaborates on the prototype developed, which implements speech emotion recognition and depression prediction on an edge device. Chapter 6 concludes the thesis, summarizing all the thesis content, milestones achieved, and future research directions and suggestions.

CHAPTER TWO

LITERATURE REVIEW

2.1 INTRODUCTION

The concept of speech emotion recognition and speech depression detection are interrelated in that they both derive insights through speech. Hence, in this chapter, an overview of commonly used speech recognition components is discussed, alongside speech emotion recognition and depression detection trends.

In general, a typical speech emotion recognition system is shown in Figure 2.1.

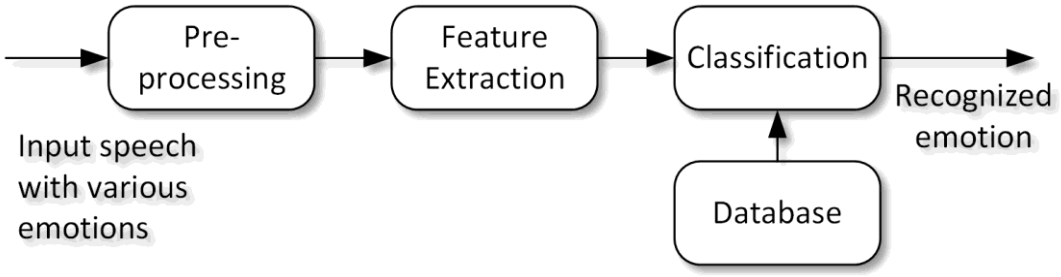

Figure 2.1 General Speech Emotion Recognition Algorithm

The input, a speech signal, is passed through a pre-processing step, typically involving denoising and silence removal. The cleaned audio is then passed to a feature extractor.

As in speech processing, the feature extraction marks the start of an SER system. It includes selecting the features appropriate for emotion recognition. Next, these features are processed by a classifier. These classifiers are trained by referring to an emotion database. The system will then be put into testing by cross-checking with the same database. The processed data obtained will be the determinant of the decision, typically in accuracy and processing time.

A similar approach is often be taken to detect depression through speech. However, instead of using an emotion database, a depression database that contains audio data of people clinically diagnosed as depressed is utilized. The classifier output can either be in the form of simple binary depressed or not a prediction, positive-negative-neutral emotion, or even predicting the level of depression (Kiss & Vicsi, 2017b) depending on the training dataset used.

2.2 SPEECH FEATURES

Speech features used in depression prediction are typically the same as those in speech emotion recognition. According to (Mantri, Agrawal, Dorle, Patil, & Wadhai, 2013), the commonly used features can be differentiated into prosodic, glottal, spectral, cepstral, and TEO (Teager energy Operator)-based features. Most researches in depression prediction use the combination of all those features, such as conducted in (Kiss & Vicsi, 2017a) and (Kiss & Vicsi, 2017b). According to (El Ayadi, Kamel, & Karray, 2011), speech features can be divided into four, as shown in Figure 2.2.

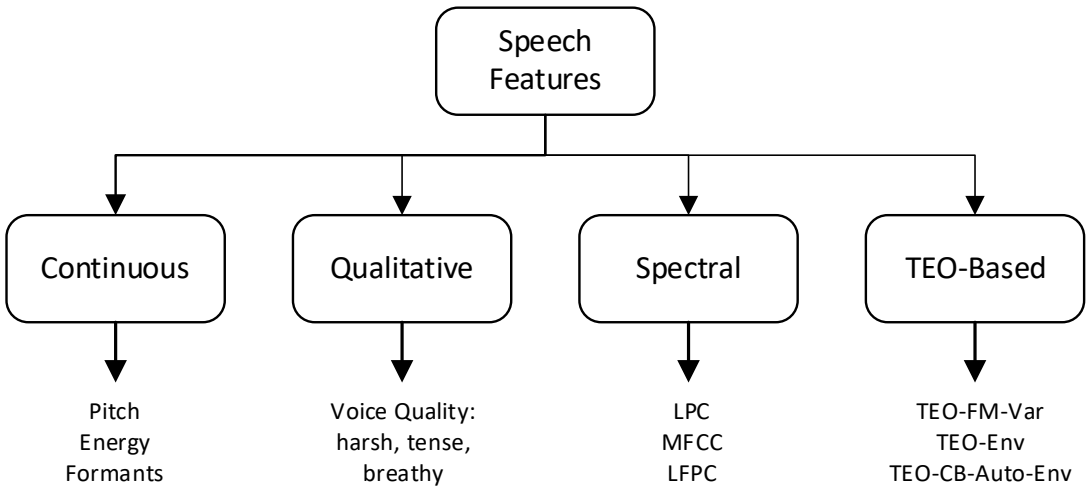

Figure 2.2 Categories of Speech Features

The following are 3 commonly used features in speech processing, including speech emotion recognition and depression detection.

2.2.1 LPCC

Linear predictive coding (LPC) is a digital method for encoding an analog signal (Dixit, Vidwans, & Sharma, 2016). LPC works because it predicts the next value of a signal based on the information it has received in the past, forming a linear pattern. LPC's main objective is to obtain a set of predictor coefficients to minimize the mean squared error, E_m. The formula used to obtain the LPC coefficients is:

$$E_m = \sum_n e_m^2[n] = \sum_n \left(x_m[n] - \sum_{j=1}^{p} a_j x_m[n-j]\right)^2 \qquad (2.1)$$

where $x_m[n]$ is a frame of the speech signal and p the order of the LPC analysis.

LPC encoding generally gives a good quality speech at a lower bitrate and supplies pinpoint approximations of speech parameters.

2.2.2 MFCC

The Mel-frequency cepstral coefficient (MFCC) is one of the most popular features extraction in speech processing. It represents the speech signals where a feature called the cepstrum of a windowed short-time signal is derived from the FFT of that signal. Afterward, the signal goes to the Mel-frequency scale's frequency axis using a log-based transform and then decorrelated using a Modified Discrete Cosine Transform (X. Huang, Acero, & Hon, 2001).

The researchers (Konar & Chakraborty, 2014) noted that the difference between MFCC and the normal MFC is that in MFCC, the frequency bands are tuned to the Mel-scale, which is adjusted to the human’s auditory hearing.

The process of extracting the MFCCs can be derived from the definition given above.