SPEECH EMOTION RECOGNITION USING SPECTROGRAMS AND CONVOLUTIONAL NEURAL NETWORKS

BY

TAIBA MAJID

A dissertation submitted in fulfilment of the requirement for the degree of Master of Science (Communication Engineering)

Kulliyyah of Engineering
International Islamic University Malaysia

APRIL 2021

ABSTRACT

Speech Emotion Recognition (SER) is the task of recognising the emotional aspects of speech irrespective of the semantic contents. Recognising these human speech emotions have gained much importance in recent years in order to improve both the naturalness and efficiency of Human-Machine Interactions (HCI). Deep Learning techniques have proved to be best suited for emotion recognition over traditional techniques because of their advantages like fast and scalable, all-purpose parameter fitting and infinitely flexible function. Nevertheless, there is no common consensus on how to measure or categorise emotions as they are subjective. The crucial aspect of SER system is selecting the speech emotion corpora (database), recognition of various features inherited in speech and a flexible model for the classification of those features. Therefore, this research proposes a different architecture of Deep Learning technique - Convolution Neural Networks (CNNs) known as Deep Stride Convolutional Neural Network (DSCNN) using the plain nets strategy to learn discriminative features and then classify them. The main objective is to formulate an optimum model by taking a smaller number of convolutional layers and eliminate the pooling-layers to increase computational stability. This elimination tends to increase the accuracy and decrease the computational time of speech emotion recognition (SER) system. Instead of pooling layers, notable strides have been used for the necessary dimension reduction. CNN and DSCNN are trained on three databases; a German database Berlin Emotional Database (Emo-DB), an English database Surrey Audio-Visual Expressed Emotion (SAVEE) and Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC), a Hindi database. The speech signals of three databases are converted to clean spectrograms by applying STFT on them after preprocessing. For the evaluation process, four emotions angry, happy, neutral, and sad have been considered. Besides, F1 scores have been calculated for all the considered emotions of all databases. Evaluation results show that the proposed architecture of both CNN and DSCNN outperform the-state-of-art models in terms of validation accuracy. The proposed architecture of CNN improves the accuracy of absolute 6.37%, 9.72% and 5.22% for EmoDB, SAVEE database and IITKGP-SEHSC database respectively. In comparison, as DSCNN architecture improves the performance by absolute 6.37%, 10.72% and 7.22% for EmoDB, SAVEE database and IITKGP-SEHSC database respectively compared to the best existing models. Furthermore, the proposed DSCNN architecture performs better for the three examining databases than proposed CNN architecture in terms of computational time. The computational time difference is found to be 60 seconds, 58 seconds and 56 seconds for EmoDB, SAVEE database and IITKGP-SEHSC respectively on 300 epochs. This study has set new benchmarks for all the three databases for upcoming work, which proves the effectiveness and significance of the proposed SER techniques. Future work is warranted to examine the capability of CNN and DSCNN for the voice-based identification of gender and image/video-based emotion recognition.

خلاصة البحث

التعرف على المشاعر الكلام (SER) هي مهمة التعرف على الجوانب العاطفية للكلام بغض النظر عن المحتويات الدلالية. اكتسب التعرف على مشاعر الكلام البشرية أهمية كبيرة في السنوات الأخيرة ، وذلك من أجل تحسين طبيعية وكفاءة التفاعلات بين الإنسان والآلة (HCI). أثبتت تقنيات التعلم العميق أنها الأنسب للتعرف على المشاعر مقارنة بالتقنيات التقليدية بسبب سرعتها وقابليتها للتوسع ،إمكانية تركيب المعلمات لجميع الأغراض ودعم الوظائف المرنة بلا حدود. ومع ذلك ، لا يوجد إجماع مشترك حول كيفية قياس العواطف أو تصنيفها لأنها ذاتية. تمثل الجوانب الحاسمة لنظام SER في اختيار هيئة عاطفة الكلام (قاعدة بيانات) ، والتعرف على الميزات المختلفة الموروثة في الكلام وتصنيف تلك الميزات من خلال نموذج مرن. لذلك ، يقترح هذا البحث بنية مختلفة لتقنية التعلم العميق - الشبكات العصبية الالتفافية (CNNs) المعروفة باسم الشبكة العصبية التلافيفية ذات الخطوة العميقة (DSCNN) وذلك باستخدام استراتيجية الشبكات البسيطة لتعلم الميزات التمييزية ثم تصنيفها. الهدف الرئيسي هو تصميم نموذج مناسب عن طريق أخذ عدد أقل من الطبقات التلافيفية وأيضًا عن طريق التخلص من طبقات التجميع لزيادة الاستقرار الحسابي. يميل هذا الحذف إلى زيادة الدقة وتقليل الوقت الحسابي لنظام التعرف على مشاعر الكلام (SER). بدلاً من طبقات التجميع ، تم استخدام خطوات خاصة لتقليل الأبعاد الضرورية. تم تدريب CNN و DSCNN على ثلاث قواعد بيانات ؛ يتم تدريب CNN و DSCNN على ثلاث قواعد بيانات ؛ قاعدة بيانات برلين العاطفية الألمانية (Emo-DB) ، قاعدة بيانات باللغة الإنجليزية باسم "Surrey للمشاعر المعبر عنها بالصوت والصورة (SAVEE)" وقاعدة بيانات هندية للمعهد الهندي للتكنولوجيا في خراجبور تسمى "مجموعة الكلام الهندية المحاكاة للعاطفة" (IITKGP-SEHSC). يتم تحويل إشارات الكلام لقواعد البيانات الثلاث إلى مخططات طيفية نظيفة عن طريق تطبيق STFT على الإشارات ، بعد المعالجة المسبقة. بالنسبة لعملية التقييم ، تم تبني أربعة مشاعر غاضبة وسعيدة ومحايدة وحزينة. بالإضافة إلى ذلك ، تم حساب درجات F1 لجميع المشاعر المدروسة لجميع قواعد البيانات. تظهر نتائج التقييم أن البنية المقترحة لكل من CNN و DSCNN تتفوق على أحدث النماذج من حيث دقة التحقق. تعمل البنية المقترحة لـ CNN على تحسين الدقة المطلقة 6.37٪ و 9.72٪ و 5.22٪ لقاعدة بيانات EmoDB و SAVEE وقاعدة بيانات IITKGP-SEHSC على التوالي. بينما تعمل بنية DSCNN على تحسين الأداء بنسبة 6.37٪ و 10.72٪ و 7.22٪ لقاعدة بيانات EmoDB وقاعدة بيانات SAVEE وقاعدة بيانات IITKGP-SEHSC على التوالي وذلك بالمقارنة مع أفضل النماذج الموجودة.علاوة على ذلك ، تعمل بنية DCNN المقترحة بشكل أفضل لقواعد بيانات الفحص الثلاث. مقارنة ، بنية CNN المقترحة من حيث الوقت الحسابي. تم العثور على فارق الوقت الحسابي ليكون 60 ثانية و 58 ثانية و 56 ثانية ل EmoDB وقاعدة بيانات SAVEE و IITKGP-SEHSC على التوالي في 300 فترة. وضعت هذه الدراسة معايير جديدة لجميع قواعد البيانات الثلاثة للأعمال القادمة ، مما يثبت فعالية وأهمية تقنيات SER المقترحة. العمل المستقبلي له ما يبرره لفحص قدرة CNN و DSCNN على التحديد الصوتي للجنس والتعرف على المشاعر القائمة على الصورة / الفيديو.

DECLARATION

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

Taiba Majid

Signature .... Date ....

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

SPEECH EMOTION RECOGNITION USING SPECTROGRAMS AND CONVOLUTIONAL NEURAL NETWORKS

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

Copyright © 2021 Taiba Majid and International Islamic University Malaysia. All rights reserved.

No part of this unpublished research may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission of the copyright holder except as provided below

1. Any material contained in or derived from this unpublished research may be used by others in their writing with due acknowledgement.

2. IIUM or its library will have the right to make and transmit copies (print or electronic) for institutional and academic purposes.

3. The IIUM library will have the right to make, store in a retrieved system and supply copies of this unpublished research if requested by other universities and research libraries.

By signing this form, I acknowledged that I have read and understand the IIUM Intellectual Property Right and Commercialisation policy.

Affirmed by Taiba Majid

……..……………………..   ………………………..

Signature   Date

ACKNOWLEDGEMENTS

First and foremost, praises to Almighty ALLAH for his showers of blessings throughout my research work to complete the research successfully.

I would like to express my deep and sincere gratitude to my research supervisor Prof. Dr. Teddy Surya Gunawan, for providing invaluable guidance throughout this research. His dynamism, vision and motivation have deeply inspired me. It was a great privilege and honour to work and study under his guidance. I am incredibly grateful to my co-supervisor Dr. Hasmah Mansor, for helping me keep perspective on where my research fits into the more outstanding picture and helping me in strengthening my research proposal and dissertation.

I am extending my heartfelt thanks to my parents and grandparents for their encouragement, prayers and sacrifices for educating and preparing me for my future. I am indebted to my brother Faisal Majid Wani for uncovering the potential in me and for acceptance and patience during the discussion I had with him on my research work. A special thanks go to my senior colleague Syed Asif Ahmad Qadri and my friends Farheen Fayaz and Rufaida Riyaz for supporting continuously and lending help whenever I needed. Also, I express my thanks to all other friends for their unprecedented support and valuable prayers. May Allah bless them all.

I would like to express my gratitude to the Malaysian Ministry of Education (MOE), which has provided research funding through the Fundamental Research Grant, FRGS19-076-0684.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

p	Dropout ratio
z	Normalised data
ψ	Teager Energy Operator
μ	Mean
$\boldsymbol{\sigma}$	Standard Deviation

LIST OF ABBREVIATIONS

SER	Speech Emotion Recognition
CNN	Convolutional Neural Network
DSCNN	Deep Stride Convolutional Neural Network
Emo-DB	Berlin Emotional Database
SAVEE	Surry Audio-Visual Expressed Emotion
IITKGP-SEHSC	Indian Institute of Technology Kharagpur-Simulated Emotion Hindi Speech Corpus
CL	Convolutional Layers
FC	Fully Connected Layers

CHAPTER ONE

INTRODUCTION

1.1. BACKGROUND OF STUDY

As humans, we tend to refer to speech to be the foremost natural manner to convey ourselves. We rely so much on it that we have the propensity to acknowledge its significance when alternative ways of communication like text messages or emails should be utilised. It is nothing unexpected that emoticons have acquired rudimentary in instant messages because these instant messages could be misjudged, and we might want to pass the emotion alongside the content as we do in speech. Both paralinguistic, as well as linguistic information, are contained in the data-rich signal, i.e., speech. Classical Automatic Speech Recognition (ASR) system studied less about some of the essential paralinguistic information which is passed on by speech like gender, personality, emotion, aim and state of mind (Li et al., 2016). The human mind utilises all phonetic and paralinguistic data to comprehend the hidden importance of the utterances and has efficacious correspondence (Hook et al., 2019). The superiority of communication gets badly affected if there is any meagreness in the cognisance of paralinguistic features. There have been some arguments regarding children who cannot comprehend the emotional conditions of the speaker evolve substandard social skills. In certain instances, they manifest psychopathological manifestations (Chatterjee et al., 2015). This accentuates the significance of perceiving the emotional conditions of speech ineffective communication. Therefore, creating coherent and human-like communication machines that comprehend paralinguistic data, for example, emotion is essential (Schuller, 2018).

Emotion recognition has been the subject of exploration for quite a long time. The fundamental structure of research in emotion recognition was formed from the detection of emotions from facial expressions (Smith & Rossit, 2018). Emotion recognition from speech signal has been studied to a great extent during recent times. In the Human-Computer Interaction (HCI), emotions play an essential role (Min Chen et al., 2017). Recently, Speech Emotion Recognition (SER) system has become one of the essential key elements in HCI as it aims to examine the emotional states of human beings through the speech signals. Speech Emotion Recognition is still a very challenging task for which how to extract practical, emotional features is an open question.

SER system is described as the ensemble of the techniques which processes the speech signals and simultaneously detect the emotions present in them by the classification process A SER system requires a classifier, a supervised learning construct, which is programmed in a way to recognise any emotions in new speech signals. A system like that which is supervised for all activities introduces the need for labelled data that has emotions embedded in it. However, before any processing can be done on the data to extract the features, it needs pre-processing. For this reason, the sampling rate across all the databases should be consistent. Further, all audio utterances can be converted into spectrograms (Prasomphan, 2015). The variation of energy at different frequencies across time is displayed as an image known as the spectrogram. The classification process essentially requires features. They help in the reduction of raw data into the most critical characteristics only. Finally, whether it suffices to use acoustic features for modelling emotions or if it is necessary to combine them with other types of features such as linguistic, discourse information, or facial features.

The performance of classifiers can be said to depend mainly on the techniques of feature extraction and also those features that are considered salient for a specific emotion (Yala et al., 2017). If additional features can be incorporated from other modalities such as linguistic or visual, it can strengthen the classifiers. However, this depends upon the significance and accessibility. These features are then allowed to pass to the classification system which has a wide range of classifiers at its disposal. From amongst the numerous machine learning algorithms, all have been examined to classify emotions according to their acoustic correlation in speech utterances. Hidden Markov models (HMM), Gaussian mixture models (GMM), Nearest Neighbourhood classifiers, linear discriminant classifiers, Artificial Neural Networks (ANN) and Support Vector Machines (SVM) are some examples that have been widely used to classify emotions based on their acoustic features of interest (Khalil et al., 2019). The feature extraction techniques used by these classifiers is the determining factor for the performance of these classifiers. Till date, numerous acoustic features and classifiers have been put through experimentation to test their credibility, but the accuracy still needs to be improved. More recently, classifiers that incorporate deep learning have become standard such as Deep Neural Network (DNN), Deep Belief Network (DBM), Recurrent Neural Network (RNN) and Long Short- Term Memory (LSTM) and Convolution Neural Network (CNN).

Now, talking about advantages, this is a known fact that deep learning methods for SER have several advantages over traditional methods, which includes their capability to detect the complex structure and features without the need for manual feature extraction and tuning; tendency toward extraction of low-level features from the given raw data, and ability to deal with un-labelled data. Taking inspiration from the success of deep learning methods, this research has formulated different network architecture of Convolution Neural Networks (CNN) with 3 convolutional layers and Deep Stride Convolution Neural Networks (DSCNN) with 6 convolution layers with different filter size and kernels in order to increase the computational stability. The detailed description of the methodology is explained in the subsequent section. Two experiments are carried out, one using Convolutional Neural Networks and the other using Deep Stride Convolutional Neural Networks. Both are trained on the spectrograms generated from the Berlin Emotional Database (EMO-DB), Surrey Audio-Visual Expressed Emotion (SAVEE) and Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC).

1.2. PROBLEM STATEMENT

When considering SER, the most distinctive role that it plays is in the domain of man-machine interactions, which can be best described by any computer tutorial application, wherein the emotions of the user determine the kind of response the system is going to provide. The functioning of SER Systems is based on extracting features from a speech which thereby help in predicting emotions. Nevertheless, the system does not remain unencountered by various difficulties faced by researchers that include the selection of appropriate speech features, the robustness of speaking styles, speaking rates and the way emotions are expressed in different cultures and environments. Problems of other sorts may include extraction of discriminative, robust, and silent features from speech. Some level of feat has been achieved in the field though, using state-of-art feature extraction methods like Mel frequency cepstral coefficients (MFCC), Linear Predictive cepstral coefficients (LPCC), and Teager Energy Operator (TEO). However, there remains a load of uncertainty on the accuracy of the results. This, in turn, makes the results far from applicable in real practice. There have been advances in Deep Learning methods to provide better emotion recognition. This research, which stands inspired by the success of Deep Learning methods, uses Convolution Neural Networks (CNN) architecture to extract silent discriminative features from spectrograms and reducing the computational complexity of the presented SER model. Different aspects of the feature extraction, content representation and classification are analysed and discussed in the context of SER.

1.3. RESEARCH OBJECTIVES

The main objective of this study is to learn high-level features from Spectrograms generated from the speech signals, then implementing to the deep learning methods Convolution Neural Networks (CNNs) and Deep Stride Convolutional Neural Networks (DSCNN) for the extraction and classification of features.

Other objectives are:

1. To formulate efficient spectrograms to represent speech signals from three different databases: Berlin Emotional Database (EMO-DB), Surrey Audio-Visual Expressed Emotion (SAVEE) and Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC).
2. To formulate an optimum model architecture of Convolutional Neural Networks (CNN) and Deep Stride Convolution Neural Network (DSCNN) and training parameters for Speech Emotion Recognition (SER) setup.

3. To analyse performance measure in terms validation and training accuraries, time, confusion matrix and F1 score with respect to various configurations of epochs and iterations.
4. To evaluate and benchmark the existing state of the art SER algorithms in terms of computational efficiency and accuracy.

1.4. RESEARCH SCOPE

Speech, which is essentially a signal is information-rich that contains paralinguistic as well as linguistic information. Considering in-depth the paralinguistic information, the "emotion", which is conveyed in part by those speech developing machines which might well be capable of understanding such paralinguistic information. To exemplify, consider the fact that emotions facilitate human-machine communication by making the communication more natural and thus more precise. Now, this research intends to thoroughly investigate the efficacy of Convolution Neural Networks (CNN) in recognition of speech emotions by developing an optimum model architecture of CNN- a Deep Stride Convolutional Neural Network (DSCNN) using the plain nets strategy to learn discriminative features from the spectrogram. The input features of the network are taken in the form of clean spectrograms of speech signals. The databases under consideration are three types, Berlin Emotional Database (EMO-DB), Surrey Audio-Visual Expressed Emotion (SAVEE) and Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC). This work aspires to realise the potential of spectrograms as well as establishing their suitability in other speech and acoustic recognition tasks. The inclination of this research is towards the reduction of complexity in the computations of the proposed SER framework.

1.5. METHODOLOGY

The proposed framework endeavours to use a discriminative Convolutional Neural Networks (CNNs) for feature learning schemes using spectrograms produced from speech signals. The proposed architecture of Deep Stride Convolutional Neural Networks (DSCNN) has input layers, convolutional layers, and fully connected layers followed by a SoftMax classifier. Both the models are trained on three different databases Berlin Emotional Database (EMO-DB), Surrey Audio-Visual Expressed Emotion (SAVEE) and Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC). As already mentioned, spectrograms are put into use; they tend to hold rich information. Also, extraction and application of such information when the transformation of the audio speech signal to text or phonemes takes place is highly unlikely. This capability lets the spectrogram improve speech recognition emotion. Therefore, the main idea is to study high-level discriminative features from speech signals making the use of CNN and DSCNN architecture highly imperative. The primary strategy is depicted in Figure 1.1.

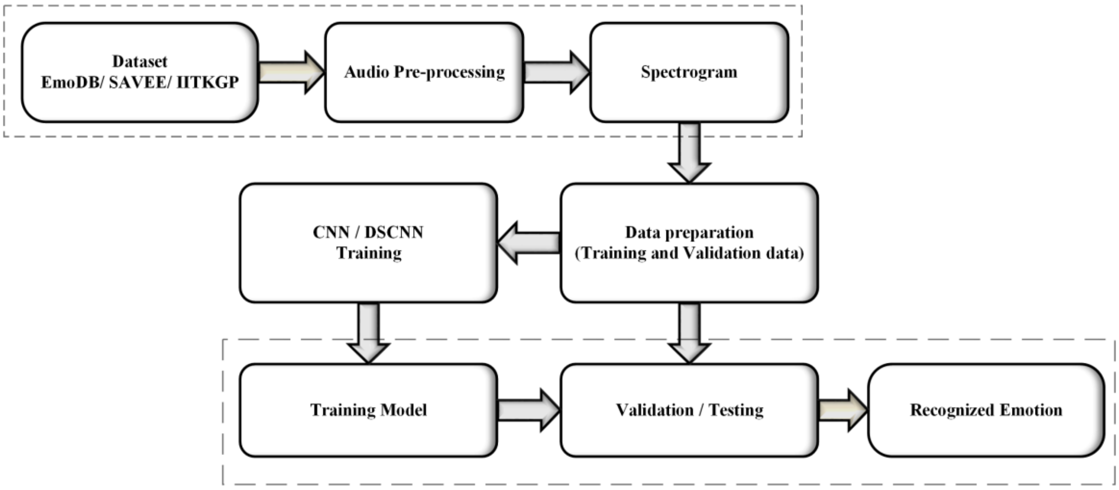

Figure 1.1 Proposed Algorithm using Spectrograms, Convolutional Neural Networks and Deep Stride Convolutional Neural Networks.

The above figure shows the proposed overall system diagram. The initial step is to prepare the dataset and get it sorted in terms of emotions. Three different datasets, EmoDB, SAVEE and IITKGP, are utilised for this task. The audio files are read in a loop. The audio preprocessing block reads the audio file, removes noise if any performs audio normalisation and echo cancellation. The spectrogram module extracts the Short-Time Fourier Transform (STFT) of the audio signal and plots the spectrogram image. All the images are saved in the image (.jpg,.png) format and .mat format so that the code can read it and perform testing and validation. This concludes the preprocessing. Next step is to divide the dataset into Training and validation/testing. In this module, the data is divided into a ratio of about 68.75% for training and 31.25% for Validation.

Next stage is the training stage where CNN/DSCNN configuration is built like its layer's configuration, pooling, connected layers and its training configurations. Then the data is trained, which gives training and validation accuracy. Using the trained module, the emotions in the testing set are classified further; confusion matrix, precision, recall and F1 scores are calculated and displayed to measure the overall accuracy of the developed system.

1.6 THESIS ORGANIZATION

The rest of the report is organised as follows. Chapter 2 is the literature review and discusses the main parameters and research conducted related to Speech Emotion Recognition and various Deep Learning techniques. Chapter 3 is the design and implementation of the research. The results and discussion are elaborated in Chapter 4, featuring the network optimisation. Finally, the conclusion and future work are given in Chapter 5.

CHAPTER TWO

LITERATURE REVIEW

2.1. SPEECH EMOTION RECOGNITION SYSTEM

In the area of speech processing, one of the most arduous tasks for the researchers still is speech emotion recognition. The emotional speech recognition is of most interest while studying human-computer recognition. It implies that the system must bear the capability of understanding the emotions of the user, which will define the actions of the system accordingly. Various tasks such as speech to text conversion, feature extraction, feature selection and classification of those features to identify the emotions must be performed by a well-developed framework that includes all these modules (El Ayadi et al., 2011). The task of classification of features is yet another challenging work, and it involves the training of various emotional models to perform the classification appropriately.

Now comes the second aspect of emotional speech recognition; database used for training models. It involves a challenging task of selecting only the features which happen to be salient to depict the emotions accurately. On merging all the above modules in the desired way provides us with an application that can recognise a user's emotions and further provide it as an input to the system to respond appropriately. At the point when we take a superior view, it may be isolated into a few fields as depicted in Figure 2.1. The enhancement of the classification process can be attributed to a better understanding of emotions.