SPEECH RECOGNITION BASED ON SPECTROGRAMS BY USING DEEP LEARNING

ROY EDUARDO AGUILAR LEON

UNIVERSITI TEKNOLOGI MALAYSIA

Replace this page with form PSZ 19:16 (Pind. 1/07), which can be obtained from SPS or your faculty.

*Replace this page with the Cooperation Declaration form, which can be obtained from SPS or your faculty. This page is **OPTIONAL** when your research is done in collaboration with other institutions that requires their consent to publish the finding in this document.]*

SPEECH RECOGNITION BASED ON SPECTROGRAMS BY USING DEEP LEARNING

ROY EDUARDO AGUILAR LEON

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Engineering (Computer and Microelectronic Systems)

Faculty of Electrical Engineering
Universiti Teknologi Malaysia

JUNE 2018

To my beloved parents and siblings.

ACKNOWLEDGEMENT

I express my gratitude to my supervisor Asoc. Prof. Muhammad Mun'im bin Ahmad Zabidi for his valuable motivation and support throughout this research. His supervision was of great importance and it led me in the most appropiate way to successfully complete the requirements of this project.

I am sincerely grateful to all my lecturers, both in the Deparments of Mechatronics and Microelectronics Engineering, for their teachings, advices, guidance and constructive comments, this project would not have been possible without the valuable knowledge imparted by them. Special thanks to Prof. Dr. Usman Ullah Sheikh, Prof. Dr. Mohamed Khalil Bin Mohd Hani and Prof. Dr. Johari Halim Shah Osman.

I am also thankful to the support of my classmates and friends especially whom I was lucky to meet here in the university, their support encouraged me to go ahead and their companionship made me feel like home.

My gratitude to this diverse country, Malaysia, and its people for having added specially value to my educational peryod.

Finally, I am deeply indebted to the the support and affection of my parents who have made all this possible.

ABSTRACT

Speech Recognition is widely being used and it has become part of our day to day. Several massive and popular applications have taken its use to another level. Most of the existing systems use machine learning techniques such as artificial neural networks or fuzzy logic, whereas others may just be based in a comparative analysis of the sound signals with a large lookup tables that contain possible realizations of voice commands. These models base their speech recognition algorithms on the analysis or comparison of the analog acoustic signal itself. The sound has particular characteristics that can not be seen through the representation of its propagation wave in time. This project proposes speech recognition through an innovative model that analyzes the graphic representation of the acustic signal, its spectrogram. Therefore the model does not classify the speech through its acoustic signal but its graphical representation. This leads the research to an approximation of the problem through the use of image classification techniques. Image clasification was considered a task only the humans can do, with the devoloping of machine learning techniques this perception has drastically changed. This project covers several techniques and shows the potential of Deep Learning for objects classification and within this field presents the convolutional neural networks as the most suitable algorithim for the classifcation of spectrograms. As a method to clearly illustrate the efficacy of the proposed model, the used alorithim was trained with two self-obtained datasets. Several experiments were conducted to make a detailed comparison of the system throughput and its levels of accuracy.

ABSTRAK

Pengecaman pertuturan digunakan secara meluas dan telah menjadi sebahagian daripada hari- hari kita. Beberapa aplikasi yang masif dan popular telah mengambil kegunaannya ke tahap yang lain. Kebanyakan sistem yang sedia ada menggunakan teknik pembelajaran mesin seperti rangkaian saraf buatan atau logik kabur, sementara yang lain hanya berdasarkan analisis perbandingan isyarat bunyi dengan jadual carian yang besar yang mengandungi pernyataan perintah suara yang mungkin. Model-model ini berdasarkan algoritma pengenalan pertuturan mereka terhadap analisis atau perbandingan isyarat akustik analog itu sendiri. Suara ini mempunyai ciri-ciri tertentu yang tidak boleh dilihat melalui perwakilan gelombang rambatan dalam masa. Projek ini mencadangkan pengecaman pertuturan melalui model inovatif yang menganalisis perwakilan grafik isyarat akustik, iaitu spektrumnya. Oleh itu, model ini tidak mengklasifikasikan pertuturan melalui isyarat akustiknya tetapi perwakilan grafiknya. Ini membawa penyelidikan kepada penghampiran masalah melalui penggunaan teknik klasifikasi imej. Klasifikasi imej dianggap sebagai tugas yang hanya dapat dilakukan oleh manusia, walau bagaimanapun, dengan pembangunan teknik pembelajaran mesin, persepsi ini berubah secara drastik. Projek ini merangkumi beberapa teknik dan menunjukkan potensi "pembelajaran dalam" untuk klasifikasi objek dan dalam lingkungan disiplin ini membentangkan rangkaian neural konvulusi sebagai algoritma yang paling sesuai untuk klasifikasi spektrogram. Sebagai kaedah untuk menggambarkan kecekapan model yang dicadangkan dengan jelas, algoritma yang digunakan telah dilatih dengan dua set data yang diperoleh sendiri. Beberapa eksperimen telah dijalankan untuk membuat perbandingan terperinci mengenai kadar celus sistem dan tahap-tahap ketepatannya.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADT	-	Automatic Drum Transcription
AE	-	AutoEncoder
AGI	-	Artificial General Intelligence
AI	-	Artificial Intelligence
AMT	-	Automatic Music Transcription
ANN	-	Artificial Neural Network
ARNN	-	Anticipation Recurrent Neural Network
BILSTM	-	Bidirectional Long Short-Term Memory
BPTT	-	Back-Propagation Through Time
BRNN	-	Bidirectional Recurrent Neural Network
CDBN	-	Convolutional Deep Belief Networks
CEC	-	Constant Error Carousel
CLNN	-	ConditionaL Neural Networks
CNN	-	Convolutional Neural Network
ConvNet	-	Convolutional Neural Network
CRBM	-	Conditional Restricted Boltzmann Machine
CRNN	-	Convolutional Recurrent Neural Network
DAE	-	Denoising AutoEncoder or Deep AutoEncoder
DBM	-	Deep Boltzmann Machine
DBN	-	Deep Belief Network
DeconvNet	-	DeConvolutional Neural Network
DL	-	Deep Learning
DNN	-	Deep Neural Network
DSN	-	Deep Stacking Network

DWT	-	Discrete Wavelet Transform
ELM	-	Extreme Learning Machine
FC	-	Fully Connected
FCN	-	Fully Convolutional Network
FC-CNN	-	Fully Convolutional Convolutional Neural Network
FC-LSTM	-	Fully Connected Long Short-Term Memory
GAN	-	Generative Adversarial Network
GBRCN	-	Gradient-Boosting Random Convolutional Network
GFNN	-	Gradient Frequency Neural Networks
GLCM	-	Gray Level Co-occurrence Matrix
HAN	-	Hierarchical Attention Network
HHDS	-	HipHop Dataset
LSTM	-	Long Short-Term Memory
MCLNN	-	Masked ConditionaL Neural Networks
MER	-	Music Emotion Recognition
ML	-	Machine Learning
MLM	-	Music Language Models
MLP	-	Multi-Layer Perceptron
MRS	-	Music Recommender System
MSDAE	-	Modified Sparse Denoising Autoencoder
MSE	-	Mean Squared Error
MSR	-	Music Style Recognition
NN	-	Neural Network
NNMODFF	-	Neural Network based Multi-Onset Detection Fun Fusion
ODF	-	Onset Detection Function
PNN	-	Probabilistic Neural Network
PReLU	-	Parametric Rectified Linear Unit
RANSAC	-	RANdom SAmple Consensus
RBM	-	Restricted Boltzmann Machine

ReLU	-	Rectified Linear Unit
RICNN	-	Rotation Invariant Convolutional Neural Network
RNN	-	Recurrent Neural Network
RTRL	-	Real-Time Recurrent Learning
SAE	-	Stacked AE
SDAE	-	Stacked DAE
SGD	-	Stochastic Gradient Descent
SVM	-	Support Vector Machine
SVD	-	Singing Voice Detection
SVS	-	Singing Voice Separation
VAD	-	Voice Activity Detection
VAE	-	Variational AutoEncoder
VPNN	-	Vector Product Neural Network
WPE	-	Weighted Prediction Error

LIST OF SYMBOLS

KB	-	Kilobyte
S	-	Sigmoid Function
H	-	Unitary Step Function
α	-	Sigmoid Function
λ	-	Learning Rate of the Feedforward Network
σ	-	Activatoin Function of a Neuron
w	-	Weight of the Neuron connections
$\ominus$	-	Erosion Operator
$\oplus$	-	Dilation Operation
$\circ$	-	Opening Operator
$\bullet$	-	Closing Operator
$*$	-	Convolution Operator
$\cup$	-	Union of sets
$\cap$	-	Intersection of sets
$\subseteq$	-	Subset
$\subset$	-	Proper subset / strict subset
$\supseteq$	-	Superset
$\supset$	-	Proper Superset / Strict Superset
$\in$	-	Element of
$\ni$	-	Proper Superset / Strict Superset
$\sim$	-	Is Row Equivalent to

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Problem Background

Doctors can classify between a good blood sample, and a bad one. Photographers can classify if their latest shot was beautiful, or not. Musicians can classify what sounds good and what does not in a piece of music.

The ability to classify well takes many hours of training. Persons get it wrong over and over again, until eventually they get it right, that is the normal and appropriate method of learning. With the development of artificial intelligence techniques it is possible to replicate these learning processes for training machines to recognize objects without interfering or programming the decision parameters to be used in the classification.

There were different attempts to improve the image classification models. In the 80s and early 90s, researchears tried a similar approach. Think about the features that makes up an image, and hand code detectors for each of them, but there is so much variety out there, no two apples look the same, thus the results for classifying images ended up with models with very low accuracy. This was considered a task only the humans could do until new artificial intelligence-based models were introduced. They are capable of classifying several objects with an accuracy of over 95%, better than humans.

An AI-based model mainly needs a set of data for “teaching” the system the features of the objects and build a model able to recognize and classify those objects.

With a quality data set, machine learning algorithms can classify just as well, if not better than humans can. This technology has been tested and it is currently being applied in many instances in daily life (industrial process, home, offices, airports, etc). One of the strengths of machine learning is image classification.

Because of the high accuracy the AI models provide for image classification, they could be used not only for classifying captured image, video or pictures. Based on image classification it could be built models to predict the weather, to read the external parameters from optical sensors, to read the temperature, etc. Their applicability goes beyond and could be also applied for sound recognition.

Sound could be classified through images by using a graphical representation of the acoustic wave, transforming the wave into an image, unlike it was previously done, by analysing the wave itself and its physical behaviour.

Speech recognition has gotten so much better in the past few decades. In the 50s the general consensus amongst computer scientists was that speech signals needed to first be split into little phonetic units, then those units could be grouped into words but even thought this seemed like it would work well, this approach did not give us good results.

The first ever speech recognizer was called Audrey by Bell labs in 1952. It could only recognize spoken numbers between 1 and 9 and it was built with analog electronic circuits. The renowned scientist at Bell Labs, John Pierce banned speech recognition research because the results were not promising enough.

A small group of visionaries at a newly formed team called were against popular opinion and created a system called Harpy. Harpy used fifteen thousand interconnected nodes and each represents all the possible utterances within the domain. They used a brute-force search algorithm to match the speech to the right nodes, and thus to get the text. This approach was slightly better but then IBM invented something called the Hidden Markov Model (HMMs) [1].

HMM represented utterances as states and probabilistic. it was predicted what a word was by given the phonemes it was made up of. When words like “You” were pronounced, they could have different durations like “you” or “yoouu”. HMM captured the plasticity of words by using a probabilistic approach.

The HMM pretty much maintained its position as king of the speech recognition throughout the 80s and 90s as researches improved them more and more. Some researches kept trying the artificial neural network models but they did not get good results. Geoffrey Hinton kept on trying out neural networks until all of a sudden; started outperforming everything [2]. The key was to give it more data and computer power, and this is Deep Learning. Nowadays, these Deep Neural Nets are how services like Siri, Echo, Alexa, Google now hear people speak.

Applications which use speech recognition are becoming part of people’s routines and are helping to interconnect in better way humans with systems. Because of its flexibility, software based nature, voice recognition technology is quite versatile in terms of the possibility of the applications it could be used. Identifying and authenticating users through the qualities of the voice is contactless, fast and simple to deploy in different situations where a audio sensor is available. Several examples could easily illustrate the strength of the speech recognition systems.

- Protect your data and bank account with your voice – Voice biometrics may be paired with facial recognition to bolster a multi factor system. Combined with face biometrics, voice recognition could add a higher security level as well as a built-in liveness detection test. Every person has a different voice frequency and pattern, it is a unique personal identification, as the finger prints are.
- Purchasing goods and services with the sound of your voice - One of the most popular and mainstream applications of biometrics is mobile payments, voice recognition has brought is way into a highly efficient and competitive area. Several companies are leading the voice authentication to mobile commerce, aiming to bring suitable security for monetary transactions without using the physical card and dispensing with a password but the voice of the user.
- Solving crimes with speech recognition – When it comes to a crime scene, the

cliché is to find any fingerprints or swab for DNA samples, but those are not the only traces left behind by criminals. In many cases there not any evidence but audio samples, in this forensics situations acoustic signals are the only accessible data available, voice biometrics could be deployed to great effect.

- An AI-based hands free assistant that recognizes who you are – Nowadays, this is probably the most popular application being introduced by the electronic devices and gadgets providers. The AI assistants like Siri, Alexa, Echo, Google, etc. capable to discriminate your voice from the others. In those voice-activated devices only authorized users could activate them. The user is able to ask the devices to perform tasks, answer questions or even tell jokes. The inclusion of speech recognition on electronic devices has also opened the possibility of voice-based unlock systems allowing users who prefer to go hands free to use a biometric security feature of their own.
- Controlling devices or procedures by voice commands - There are many applications where voice recognition can drastically enhance the performance of the systems by letting the operators uses their voice to take actions or control any procedure instead of using their hands which in turn may simultaneously be controlling other processes. This gives the operators an extra actuator, their voice.

1.2 Problem Statement

Although there are many applications for speech recognition, many of them are focuses on a major language, there is not a robust one able to process simple voice commands and which is reliable enough to control electronic devices efficiently in other languages except English.

Besides the above mentioned, there is not deep previous research of speech recognition based on the graphic analysis of the acoustic wave.

This research will find a model for speech recognition through the spectrogram analysis by using machine-learning techniques.

Usually the representation of an acoustic signal is related to a diagram of the wave amplitude versus the time of its propagation, the well-known oscillating plots in the music equalizers displays. For interpretation purposes, a better approach is to use the relation between the wave amplitude, frequency and its time of propagation: the spectrograms. They provide detailed information of the acoustic signals that is the reason why it is important to determine the best representation of them, thus the model can define stronger features and can extract detailed information of the generated images.

The goal pursued by this research is to classify spectrograms, thus it will be used image classification techniques. It is necessary to find the best model to process this sort of pictograms and set up an optimal parameterized system which results in a system with an adequate accuracy.

A fundamental part of the research is to find a quality dataset; good results will depend mostly on that. Because there are not many datasets available, the samples to be used will be self-collected. An efficient and automated method to collect the maximum amount of data has to be designed.

The execution of this research will be carried out by using informatics tools. There is specialized software with built-in libraries, which facilitate the implementation of neural networks. Several programs are to be used to determine the one which is more suitable and provides better accuracy for the spectrogram classification.

1.3 Objectives

Throughout this project, there are four objectives to be achieved.

1. To determine the most appropriate Deep Learning algorithm for speech recognition based on the analysis of the spectrograms of the acoustic signals.

2. To design a system for speech recognition based on spectrograms by using Deep Learning.
3. To evaluate and analyze the effectiveness and accuracy of the model compared to similar and traditional speech recognition techniques.
4. To obtain a quality dataset for training and validating the neural network model.

1.4 Scope of Work

The scope of this project is the fundamental task that needs to be carried out in order to achieve the project objectives. The Scope of this project covers:

1. The different machine learning techniques will be investigated.
2. An algorithm within Deep Learning will be chosen and an optimal configuration of it will be set.
3. The research will be delimited for a set of 12 words.
4. A quality dataset will be obtained, processed and converted into spectrogram files for training and testing the model, the process of collecting data will be automated.
5. It will use 2 different dataset to enrich the results and perform a better comparison of the effectiveness of the model, Spanish and English.
6. The automation of the data collections and the application of machine learning algorithms will be developed in Python 3.62 and Matlab.
7. The results are obtained by using machine learning techniques and because there is not a suitable benchmark for this project, other models will be applied which will allow doing an effective comparisons.

1.5 Organization

This report consists of five chapters, which includes introduction, literature review, methodology, results and conclusions.

Chapter one introduces the overall information regarding to this project. A general background of the data classification, speech recognition and machine learning techniques. It also presents the motivation for this research to be conducted by showing interesting examples where this model could be used. This chapter also identifies the problem statement, objectives and scope of the work to be carried out.

Chapter two presents an overview of basic concepts to be used within this project: it includes Speech processing, spectrogram analysis, morphological image processing, artificial intelligence techniques, deep learning, and convolutional neural networks. Besides, it is being analyzed previous studies conducted in this field, consequently, an appropriate model could be develop based on the study and analysis on those previous works.

Chapter three introduces the proposed solution, an optimal algorithm within Deep Learning for classifying spectrograms, and thus perform the speech recognition. In this chapter it is also analyzed the data set acquisition and the automation of this process. The universe of data is delimited and also the words to be used in both languages which the model will be applied in. Subsequently, the methodology implemented throughout this project is presented in this chapter.

Chapter four presents the results of the project and the comparisons among the different models applied. This chapter includes imperative discussion on the validity, reliability and efficiency of the proposed model alongside with the depiction of graphics and data diagrams associated to the experiments done. Chapter five presents the conclusions obtained after the model and its parameterization has been carried out. In addition, it is given recommendations for future analysis in this field besides the future projections for this research and the works, which could be conducted, based on this project.

Chapter five presents the conclusions obtained after the model and its parameterization has been carried out. In addition, it is given recommendations for future analysis in this field besides the future projections for this research and the works, which could be conducted, based on this project.

REFERENCES

1. Huang, X., Baker, J. and Reddy, R. A historical perspective of speech recognition. *Communications of the ACM*, 2014. 57(1): 94–103.

2. Werbos, P. J. Neural networks for intelligent control, 2005. US Patent 6,882,992.

3. Rabiner, L. R. and Schafer, R. W. *Digital processing of speech signals*. vol. 100. Prentice-hall Englewood Cliffs, NJ. 1978.

4. Attenborough, K., Taherzadeh, S., Bass, H. E., Di, X., Raspet, R., Becker, G., Güdesen, A., Chrestman, A., Daigle, G. A., L'Esperance, A. *et al.* Benchmark cases for outdoor sound propagation models. *The Journal of the Acoustical Society of America*, 1995. 97(1): 173–191.

5. Kinsler, L. E., Frey, A. R., Coppens, A. B. and Sanders, J. V. Fundamentals of acoustics. *Fundamentals of Acoustics, 4th Edition, by Lawrence E. Kinsler, Austin R. Frey, Alan B. Coppens, James V. Sanders, pp. 560. ISBN 0-471-84789-5. Wiley-VCH, December 1999.*, 1999: 560.

6. ANSI, A. American National Standard Acoustical Terminology. *ANSI Sl*, 1994: 1–1994.

7. Flanagan, J. L. *Speech analysis synthesis and perception*. vol. 3. Springer Science & Business Media. 2013.

8. Russ, J. C. *The image processing handbook*. CRC press. 2016.

9. Gonzalez, R. C. and Richard, E. Woods.(2002). Digital Image Processing.

10. Comer, M. L. and Delp, E. J. Morphological operations for color image processing. *Journal of electronic imaging*, 1999. 8(3): 279–290.

11. Palomares, F. G., Serrá, J. A. M. and Martínez, E. A. Aplicación de la convolución de matrices al filtrado de imágenes. *Modelling in Science*

Education and Learning, 2016. 9(1): 97–108.

12. Vargas, M. G. F. and Cruz, E. A. A. Estudio del Efecto de las Máscaras de Convolución en Imágenes Mediante el Uso de la Transformada de Fourier. *Ingeniería e Investigación*, 2001. (48): 46–51.

13. Russell, S. J. and Norvig, P. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,. 2016.

14. Takeyas, B. L. Introducción a la inteligencia artificial, 2017.

15. Robert, C. Machine learning, a probabilistic perspective, 2014.

16. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural networks*, 2015. 61: 85–117.

17. Kingsbury, B. E., Morgan, N. and Greenberg, S. Robust speech recognition using the modulation spectrogram. *Speech communication*, 1998. 25(1-3): 117–132.

18. Satt, A., Rozenberg, S. and Hoory, R. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. *Proc. Interspeech 2017*, 2017: 1089–1093.

19. Uchida, S., Ide, S., Iwana, B. K. and Zhu, A. A further step to perfect accuracy by training CNN with larger data. *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on*. IEEE. 2016. 405–410.

20. Graves, A., Mohamed, A.-r. and Hinton, G. Speech recognition with deep recurrent neural networks. *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE. 2013. 6645–6649.

21. Sprengel, E., Jaggi, M., Kilcher, Y. and Hofmann, T. Audio based bird species identification using deep learning techniques. *LifeCLEF 2016*. 2016, EPFL-CONF-229232. 547–559.

22. Greenberg, S. and Kingsbury, B. E. The modulation spectrogram: In pursuit of an invariant representation of speech. *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*. IEEE. 1997, vol. 3. 1647–1650.

23. Goyal, M. Morphological image processing. *IJCST*, 2011. 2(4).

24. Bovik, A. C. *Handbook of image and video processing*. Academic press. 2012.

25. LeCun, Y., Bengio, Y. and Hinton, G. Deep learning. *nature*, 2015. 521(7553): 436.