SPEECH EMOTION RECOGNITION USING DEEP NEURAL NETWORKS

BY

SYED ASIF AHMAD QADRI

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

Kulliyyah of Engineering
International Islamic University Malaysia

AUGUST 2020

ABSTRACT

With the ever-increasing interest of research community in studying human-computer/human-human interactions, systems deducing and identifying emotional aspects of a speech signal has emerged as a hot research topic. Speech Emotion Recognition (SER) has brought the development of automated and intelligent analysis of human utterances to reality. Typically, a SER system focuses on extracting the features from speech signals such as pitch frequency, formant features, energy related and spectral features, tailing it with a classification quest to understand the underlying emotion. However, as of now there still exists a considerable amount of uncertainty arising from factors like, determining influencing features, development of hybrid algorithms, type and number of emotions and languages under consideration, etc. The key issues pivotal for successful SER system are driven by proper selection of proper emotional feature extraction techniques. In this research Mel- frequency Cepstral Coefficient (MFCC) and Teager Energy Operator (TEO) along with a new-fangled fusion of MFCC and TEO referred as Teager-MFCC (TMFCC) is examined over multilingual database consisting of English, German and Hindi languages. These datasets have been retrieved from authentic and widely adopted sources. The German corpus is the well-known Berlin Emo-DB, the Hindi corpus is Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC) and the English corpus is Toronto emotional speech set (TESS). Deep Neural Networks has been used for the classification of the different emotions considered viz., happy, sad, angry, and neutral. Evaluation results shows that MFCC with recognition rate of 87.8% outperforms TEO and TMFCC. With TEO and TMFCC configurations, the recognition rate has been found as 77.4% and 82.1% respectively. However, while considering energy-based emotions, contrasting results were fetched. TEO with recognition rate of 90.5% outperforms MFCC and TMFCC. With MFCC and TMFCC configurations, the recognition rate has been found as 83.7% and 86.7% respectively. The outcome of this research would assist information of a pragmatic emotional speech recognition implementation driven by wiser selection of underlying feature extraction techniques.

خلاصة البحث

مع الاهتمام المتزايد لمجتمع الأبحاث بدراسة التفاعلات بين الإنسان والحاسوب / الإنسان والإنسان، برزت الأنظمة التي تستنبط وتعرف الجوانب العاطفية لإشارة الكلام كموضوع بحثي ساخن. أدى التعرف على عاطفة الكلام (SER) إلى تطوير التحليل الآلي والذكاء للألفاظ البشرية إلى واقع ملموس. عادةً ما يركز نظام SER على استخلاص الميزات من إشارات الكلام مثل تردد النغمة وميزات التكوين والميزات الطيفية ذات الصلة بالطاقة، وتكييفها مع تصنيف سعري لفهم المشاعر الكامنة. ومع ذلك، حتى الآن لا يزال هناك قدر كبير من عدم اليقين الناشئ عن عوامل مثل تحديد الخصائص المؤثرة وتطوير الخوارزميات المختلطة ونوع وعدد المشاعر واللغات قيد النظر، إلخ. تتمحور القضايا الرئيسية المحورية لنظام SER الناجح عن طريق الاختيار الصحيح لتقنيات استخراج الميزة العاطفية المناسبة. في هذا البحث، يتم فحص معامل ميلتر Cepstral (MFCC) ومشغل الطاقة Teager (TEO) مع اندماج جديد fangled من MFCC و TEOيشار إليه Teager-MFCC (TMFCC)في قاعدة بيانات متعددة اللغات تتكون من اللغات الإنجليزية والألمانية والهندية.تم تجميع مجموعات البيانات هذه من مصادر أصلية ومعتمدة على نطاق واسع .مجموعة البينات الألمانية هي برلين Emo-DB المعروف جدًا ، و مجموعة البينات الهندية هي المعهد الهندي للتكنولوجيا خراغبور محاكاة اللغة الهندية (IITKGP-SEHSC) ومجموعة اللغة الإنجليزية هي مجموعة الكلام العاطفي في تورونتو.(TESS) تم استخدام الشبكات العصبية العميقة لتصنيف المشاعر المختلفة التي تم اعتبارها بمعنى: سعيدة وحزينة وغاضبة ومحايدة. تظهر نتائج التقييم أن MFCC مع معدل التعرف على 87.8 ٪ يتفوق TEO و .TMFCCمع تكوينات TEO وTMFCC، تم العثور على معدل التعرف بنسبة 77.4 ٪ و82.1 ٪ على نحو محترم. ومع ذلك، أثناء النظر في المشاعر القائمة على الطاقة، تم جلب نتائج متباينة TEO . مع معدل الاعتراف 90.5 ٪ يتفوق MFCCو .TMFCC مع تكوينات MFCC وTMFCC، تم العثور على معدل التعرف على 83.7٪ و86.7٪ على نحو متتابع. سوف تساعد نتائج هذا البحث في انشاء معلومات تنفيذ عملية للتعرف على العاطفة في الكلام بانتقاء الاختيار الأكثر حكمة بناء على تقنيات استخلاص الخصائص الأساسية للإشارات الصوتية.

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 dissertation for the degree of Master of Science (Computer and Information Engineering)

……..
Teddy Surya Gunawan
Supervisor

……..
Hasmah Mansor
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 dissertation for the degree of Master of Science (Computer and Information Engineering)

……..

Dr. Rashidah Funke Olanrewaju

……..

Dr. Nurul Fariza Zulkarnain

This dissertation 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 dissertation 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)

……..
Ahmad Faris Ismail
Dean, Kulliyyah of Engineering

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.

Syed Asif Ahmad Qadri

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

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

SPEECH EMOTION RECOGNITION USING DEEP NEURAL NETWORKS

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

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 Commercialization policy.

Affirmed by Syed Asif Ahmad Qadri

……..…….. ……..

Signature Date

ACKNOWLEDGEMENTS

All praises and countless thanks be to Almighty Allah whose grace and compassion conferred upon me the courage to remain steadfast with the tedious work of the research project.

I find no words to express my heartfelt thanks to my supervisor, guide and mentor, Prof. Teddy Surya Gunawan for extending his full support from the very first day of starting this project. He left no stone unturned to be liberal enough in offering his prudent guidance coupled with his oceanic knowledge and dexterous wisdom all along the way to get me through this milestone. In addition, I would like to thank my Co-supervisor, Dr. Hasmah Manzoor for providing me with her valuable guidance and caring support.

It would be ungrateful on my part if miss to acknowledge my colleagues Taiba Majid Wani and Mohammad Fahreza Ghifari for their unprecedented support and valuable suggestions. This project would not have been complete without the moral support and prayers of my family and friends. May Allah bless them all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

N_m	Number of tests inside each frame
Y_m	Resulting Signal
$W(m)$	Hamming Window
$X(m)$	Representation of Signal
$x(n)$	Speech Signal
$\psi[x(n)]$	Discrete time signal
A	Amplitude
$V(z)$	Transform function for vocal track

LIST OF ABBREVIATIONS

SER	Speech Emotion Recognition
DNN	Deep Neural Network
MFCC	Mel-frequency cepstral coefficients
LPCC	Linear Predictive Coding
TEO	Teager Energy Operator
HMM	Hidden Markov Model
GMM	Gaussian Mixture Model
SVM	Support Vector Machine
KNN	K-Nearest neighbor
ANN	Artificial Neural Network

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND OF THE STUDY

As time goes by and as we step into the future, intelligent or smart machines will gradually supplant and upgrade human abilities in numerous zones. The intelligence showed by machines or programming projects are regularly named as "Artificial Intelligence" which is a subfield of computer engineering. Artificial intelligence alongside machine learning is currently a potential distinct advantage throughout the entire existence of computing backed with solid information investigation. Study in the area of artificial intelligence has quickly affected the rise of keen advancements that has a huge impact on our everyday lives. The field of science, engineering, business and medicine has become more brilliant with forecast abilities to smoothen our lives in our everyday exercises. There are numerous areas in which the human capabilities will be enhanced or replaced by artificial machines in coming years. The speech input facility is the most user- friendly way, adopted by development of speech recognition based on sophisticated technologies.

Compared to many other biological signals (e.g., electrocardiogram), speech signals usually can be acquired more readily and economically. Therefore, most researchers are interested in speech emotion recognition (SER). The initial start of that is simple speech recognition dates back from the late fifties. Automatic Speech Recognition or simply Speech Recognition is the technology that manages methodologies and procedures to perceive the speech the speech signals. It was discovered that voice can be next mechanism for speaking with machines particularly with computer-based systems. A requirement for gathering emotions from spoken

utterances increment exponentially. Emotion Recognition manages the investigation of inducing emotions, techniques utilized for gleaning. Emotion can be perceived from facial expressions and speech signals. Different techniques have been evolved to detect the emotions such as neural networks, signal processing, computer vision and machine learning. Emotion Recognition is acquiring its prevalence in research (Furui, Kikuchi, Shinnaka, & Hori, 2004). The principal requirement of Emotion Recognition from Speech is laborious errand in Artificial Intelligence where for the computer systems the speech signals are alone an input. For this very reason Speech Emotion Recognition came into considerations of researchers and became a hot topic in the research area which attempts to presume the emotions from the speech signals. Emotions may differ based on culture, environment and individual face findings which leads to perplexing findings. Speech corpus is not sufficient to precisely assume the kind of emotion, also the paucity of speech database in many languages make Emotion Recognition a challenging problem (El Ayadi, Kamel, & Karray, 2011). However, with the advancement of new technologies like classification of emotions, has led a great impact in the recognition of emotions in speech signal. This classification of emotion is most important part in SER system of emotions using classifiers. There are several classifiers such as GMM, HMM, SVM, DNN etc. In this work we have used Deep Neural Network Classifier to distinguish the emotions. Deep Neural Network is a deep learning technique in which data models are formulated in such a way that are bounded to perform specific task (LeCun, Bengio, & Hinton, 2015). DNN is used for number of tasks such as classification tasks, pattern recognition, image recognition, decision making etc (Schmidhuber, 2015). DNN has been utilized and have come up with great favourable outcomes in recognizing emotions and made feature extraction process a simple task (Ngiam et al., 2011). Deep

neural networks are amazing class of machine learning algorithms actualized by stacking layers of neural networks along the profundity and width of littler architectures. Despite the extraordinary advancement made in artificial intelligence, there is still a long way from having the option to normally interface with machines, somewhat on the grounds that machines comprehend the emotion states. SER aims to recognize the underlying emotional state of a speaker from his/her voice. The area has received increasing research interest all through recent years.

1.2 PROBLEM STATEMENT

Human speech is one of the fundamental ways of conveying information between people. Spoken words are not the sole component of speech, but also the acoustics properties and emotions. Exchange of emotions can happen during conversation, in which emotional state of a speaker may easily trigger an interlocutor emotional state resulting in a change in the speech style or tone. Many primary emotions can be manifested in speech signals, such as disgust, fear, sadness, boredom, joy/happiness, and anger. Many acoustic features have been widely used in emotion recognition, including pitch, Teager energy operator, vocal tract, spectral, and duration. Many algorithms have been utilized as classifier, such as Hidden Markov Model (HMM), support vector machine (SVM), Gaussian mixture models (GMM), artificial neural networks (ANN). Although many acoustic features and classifiers have been experimented for speech emotion recognition, yet it is still unclear what features are effective for the task when accuracy and computational time are considered. It is hoped that the recent advances in Deep Neural Networks (DNN) could be exploited for better emotion recognition.

1.3 OBJECTIVES

The main objective of this project is to extract and analyse speech features from the speech files using MATLAB, then classifying the features extracted into SER using DNN. Side objectives that were pursued are:

- To investigate speech emotion database across many languages and formulate a common operational multilanguage repository.
- To design an optimum deep neural networks configuration and training parameters for Speech Emotion Recognition (SER) setup.
- To evaluate and benchmark existing state of art SER algorithms in terms of computational time and accuracy.

1.4 METHODOLOGY

As stated in (Yu & Deng, 2015), for speech recognition application deep feedforward neural network configuration could be utilized. The book discussed as well Restricted Boltzmann Machines (RBM), Deep Belief Network (DBN), and autoencoder to be used as classifier while using traditional acoustic features, such as MFCC. For speech emotion recognition, this research derived a strategy that used traditional audio features and use DNN as classifier. The optimum audio features and the DNN configuration were investigated properly. The strategy is depicted in Figure 1.1, respectively.

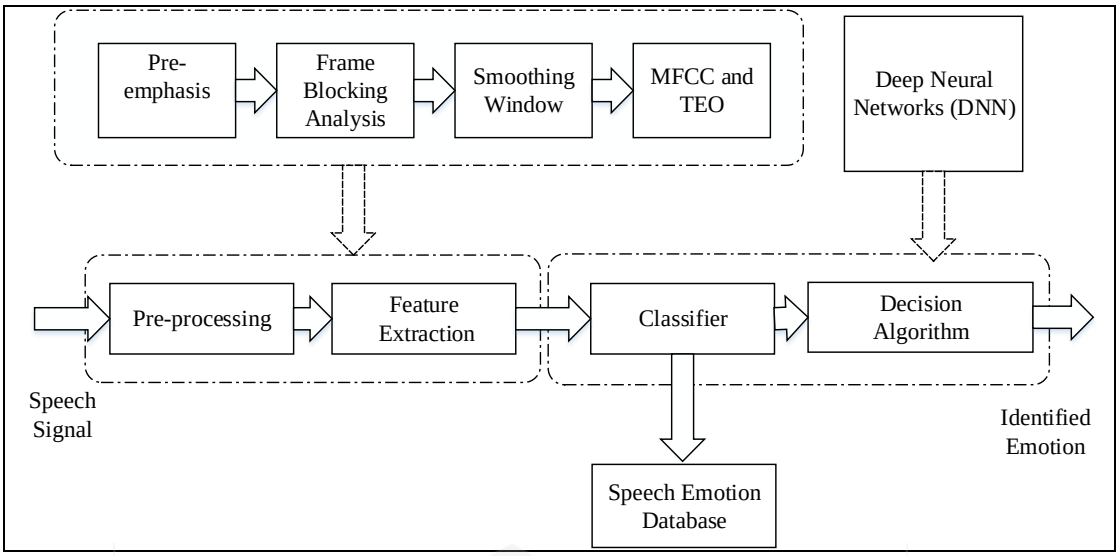

Figure 1.1 Proposed Algorithm using Traditional Acoustic Features and Deep Neural Networks

Figure 1.1 shows that the traditional acoustic features, such as MFCC, Teager Energy Operator and their fusion (TEO+MFCC) that were used as feature extraction techniques. The classifier and decision rule used deep feedforward neural network (DFNN) and other DNN configuration.

The performance of the existing state of art SER algorithm are compared to other algorithms in terms of recognition rate and computational time. Based on the result analysis, the algorithm is further optimized.

1.5 RESEARCH SCOPE

There are various features in speech that vary from speaker to speaker. In general, the accuracy of speaker-dependent emotion recognition is much higher than that of speaker-independent recognition. The SER system consists of four main steps. First is the collection of samples (databases). Followed by features vector which is formed by

extracting the features. As the next step there is the determination of features which are most relevant to differentiate each emotion.

This study considered English, Hindi, and German databases only and developed a new database which is a multilingual database. This research put forward a new algorithm using Traditional Acoustic Features and Deep Neural Networks. The experiment considered a total of 1342 utterances from all databases.

1.6 ORGANIZATION

The rest of the report is organized as follows. Chapter 2 is the literature review and discusses research conducted relating to SER and DNN. Chapter 3 is the methodology and implementation of the research. The results and discussion will be elaborated in Chapter 4 featuring the network optimization. Finally, the conclusion and future work will be in Chapter 5.

CHAPTER TWO

LITERATURE REVIEW

2.1 INTRODUCTION

Speech Emotion Recognition (SER) lies in the most continuously researched subject matter from several years in the field of speech. The initial start of the speech recognition goes back to the late fifties (El Ayadi et al., 2011). The increasing number of publications every year in SER is evident that this topic is quite a research hotspot. Figure 2.1 shows the rough estimation of IEEE publications that are related to SER. The data was analyzed from IEEE Explore.

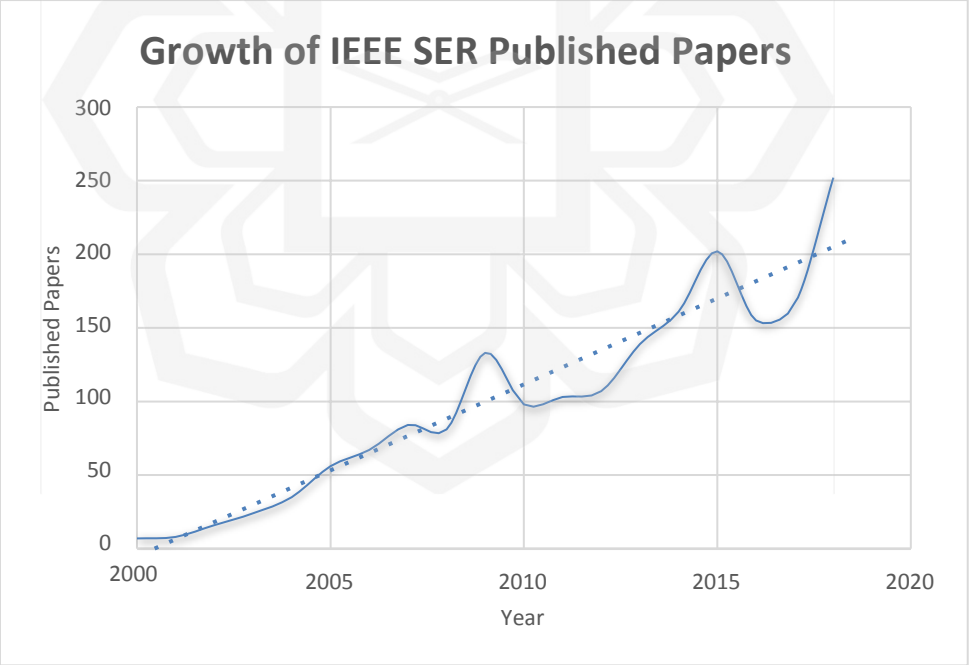

Figure 2.1 Growth of IEEE SER Publications from 2000 until 2019 (IEEE Xplore)

The aim of Speech Emotion Recognition (SER) system is to draw out the emotion from the undisclosed input speech (Joshi & Kaur, 2013). There is respective

emotional state for every individual, usually emotions are assembled into a universal category of sad, fear, surprise, happy, neutral as well as anger. Some other researchers have their own categories, for example the database utilized in (Takebe, Yamamoto, & Nakagawa, 2016) categorized emotions into ten types, namely anticipation, sadness, joy, disgust, anger, neutral and others. Even though the classification of emotion may contradict, but the objective of SER is same, which is to extract emotional state. In (Ingale & Chaudhari, 2012), it is claimed that SER is roughly a pattern recognition system.

SER has many applications in various sectors like, an auto caller in bank may be provided with SER which may abet in the detection the emotion of the customer and will help in the generation of custom responses which will be based on the result (Guo, Li, Wei, & Xu, 2017; Irastorza & Torres, 2016; Pappas, Androutsopoulos, & Papageorgiou, 2015). In the sector of education, SER with an e-learning portal could identify the user's emotion such as stress and frustration and persuade if the studying is beneficial or not and give suitable attainments (Kerkeni, Serrestou, Mbarki, Raoof, & Mahjoub, 2017). One more application of SER is in transportation, where in the coming years vehicles will have the ability of auto-driving, the system itself will take over the steering wheel if there is an ailing amount of emotion detected from the driver (J. S. K. Ooi, Ahmad, Harun, Chong, & Ali, 2017).

2.2 GENERAL FLOW OF SPEECH EMOTION RECOGNITION

Speech Emotion Recognition system is kindred as pattern recognition system. Speech emotion recognition system involves same steps as that of pattern recognition system. The SER system involves 5 main phases viz., Emotional speech input, Feature extraction, Feature selection, classification and recognized emotional output (Ingale &

Chaudhari, 2012). The structure of speech emotion recognition (SER) is shown as in Figure 2.2.

Figure 2.2 General Flow of Speech Emotion Recognition

The examination of the SER is grounded on the dimension of straightforwardness of the database, is utilized as a contribution to the SER framework. The huge issue in speech emotion recognition (SER) system is the need to recognize a great deal of essential emotions to be arranged by an automatic emotion recognizer. A fundamental arrangement of emotions comprises of 300 passionate states and hence it is hard to recognize extensive number of emotions. According to "Palette hypothesis", any emotion can be disintegrated into essential emotion like how any shading is a blend of some fundamental hues. Essential emotions are fear, joy, sadness, anger and disgust (Fernandez, 2004). Feature extraction is the most noteworthy phase of speech emotion recognition and contains numerous strategies. A portion of the parametric portrayals are, The Mel-recurrence cepstrum coefficients (MFCC), the linear cepstrum coefficients (LFCC), the linear prediction coefficients (LPC) and the reflection coefficients (RC).

2.3 AUDIO FEATURES

Feature processing for speech recognition is an efficient tool to be used for model building and recognition which can be done by extracting the information which is speaker dependent primarily. Many parameters are used in the determination of any

specific emotion present in the speech. As such the change in the parameters would represent the changes in the emotions. The challenge in the process stands in the fact that suitable features need to be extracted very carefully in order to characterize the emotions correctly. Speech signals are not steady in border sense either; hence we need to take up piece wise smaller speech signal fragments. These small speech signals are known as frames. Hence with each frame we consider the speech signal to be almost stationary. Some of the speech features which can be called as prosodic features and contain pitch and energy like characteristics can be extracted from every frame, thus known as local features.

The predominant subject in speech emotion recognition is the extraction of speech features that effectively characterize the emotional content of speech and simultaneously are independent of the verbal content and the speaker. Since, various speech features are scrutinized in speech emotion recognition, yet the researchers are unable to discover the best speech features for the task. Speech features may be classified among four classes: continuous features, qualitative features, spectral features, and TEO (Teager energy operator)-based features. Figure 2.3 demonstrates instances of features relating to each class.

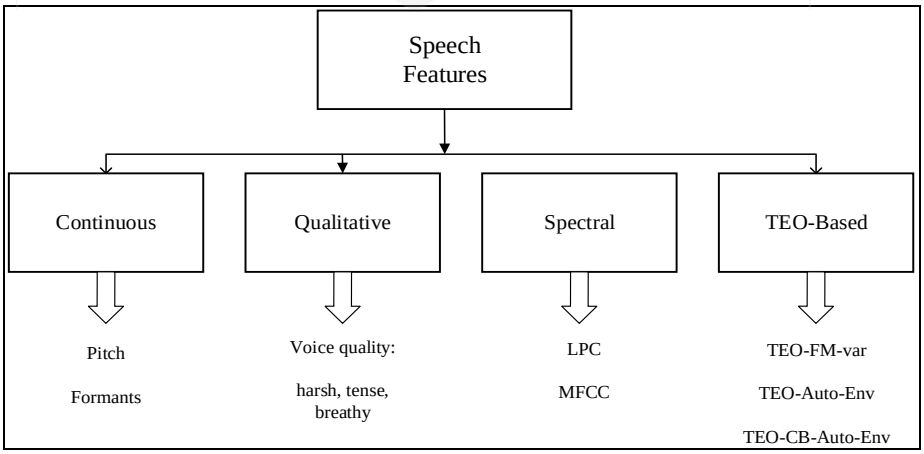

Figure 2.3 Categories of Speech Features