CHAPTER 1

INTRODUCTION

This chapter includes the following subtopics, namely: (1) Rationale; (2) Theoretical Framework; (3) Conceptual Framework/Paradigm; (4) Statement of the problem; (5) Hypothesis (Optional); (6) Assumption (Optional); (7) Scope and Delimitation; and (8) Importance of the study.

1.1 Rationale

The interest in the recognition of emotions and their benefits has been increased in the field of speech and language processing over the last decade [9][35][33]. Emotion recognition can improve the quality of services and even the quality of life. Speech Emotion Recognition (SER) is particularly useful in which the emotional state of the speaker plays an important role and finds numerous applications in automated speech services such as interactive voice recognition systems, in investigative application like lie detectors, in medical applications such as diagnosis of depression, etc. SER is also used in certain applications such as tutoring systems to detect the learner's state and adjust the presentation style according to the detected mood. SER useful in call-center systems for detecting consumer's states and enhancing the service quality. It can also be employed even in medicine as a diagnostic tool for detecting the emotional state of the patient during consultations or even for emotion recognition of autistic individuals [6][1][22][10].

Although recognition of emotions has benefits, there are still problems to develop a method to recognize emotion from the speech signal, because speaking styles and speaking rates of the speakers is different from person to person i.e., different for native speakers [16][17]. Besides, there are many issues related to the recognition of emotions, especially the selection of measurement and the results of evaluation methods, the selection of measurement hardware and software. Moreover, the issue of emotion recognition and evaluation remains complicated because of its interdisciplinary nature: emotion recognition and strength evaluation are the objects of psychology sciences [9].

Nowadays, most methods of emotion recognition detection are built based on an OpenSMILE toolkit [14][38]. It is a tool for automatic feature extraction from audio signals and for the classification of speech and music signals [11]. OpenSMILE has various standard feature sets for emotion recognition and is available as openSMILE configuration files, i.e *emobase* which have 988 features. Nevertheless, they only provide statistical feature values and cannot describe the emotional dynamic states of the speakers [14][6][42].

Speech dynamics are the systematic patterns in the sequencing of speech sounds and in their manifestation in the acoustic speech signal, which can be observed at any one or

more time scales. Visualizing the time-varying changes of features is important to speech emotion recognition [19]. It is a crucial factor in speech communication because it offers feedback information to the listeners.

A vital part of the emotion recognition system pipeline is a process where emotional speech is not constant across time, so it is necessary to know how to describe the speaker's emotional features to classify speech segments in different emotional classes. Physiological studies show that expressing emotion in speech has a beginning, a raising side a peak, and a falling side [22]. The speech segment should be long enough to capture the most possible information, but not very small, since it may lack an informative emotion area.

1.2 Statement of the Problem

Many works have been done for speech emotion recognition but most of them use speech segments ranging from a word to a couple of sentences [22][2], The speech segment should be long enough to capture the most possible information, but not very small, since it may lack an informative emotion area. Gao et.al [14] proposed prosodic and spectral features to establish emotion recognition that extracted using openSMILE. However, their method only provides the statistical features of the entire signal. This method cannot capture the dynamics of the speaker's emotional state from the signal. Moreover, their performance achieves 87.3% for EMO-DB and 79.4% for RAVDESS.

1.3 Objective

Based on the above analysis methods, our methods to tries catch the dynamic emotional states from voice segments of the signals and extract features from them. We assume that emotional state are in the voice segment. We also use discrete wavelet transform (DWT) to be able to select the level of DWT which have minimum noise level. So that, this research presents a speech emotion recognition approach aiming at improving the recognition correct rate identification for human emotions [14].

1.4 Hypotheses

This study using wavelet analysis has been successfully applied in speech emotion recognition and also in many other signal and image processing applications [36][39]. The wavelet can decompose a signal structure at all frequencies below its signal bandwidth so that this study can choose at which level the signal can be processed. This will help in identifying a band where the information related to emotion is concentrated. Furthermore, voiced speech segmentation is considered to contain human voice emotions [22].

1.5 Theoretical Framework

Emotion recognition is typically performed by measuring various human bodies. The most popular techniques are speech [35][9]. Speech emotion refers to the use of various methods to analyze vocal behavior as a marker of affection (e.g., emotions, moods, and stress), focusing on the nonverbal aspects of speech. The basic assumption is that there is a set of objectively measurable voice parameters that reflect the affective state of a person who is currently experiencing (or expressing for strategic purposes in social interaction). The most effective states involve physiological reactions which in turn modify different aspects of the voice production process. For example, the sympathetic arousal associated with an anger state often produces changes in respiration and an increase in muscle tension which influences the vibration of the vocal folds and vocal tract shapes. It also affects the acoustic characteristics of the speech, which in turn can be used to infer the respective state [30]. How emotions are expressed in the voice can be analyzed at three different levels:

1. The physiological level (e.g., describing nerve impulses or muscle innervation patterns of the major structures involved in the voice-production process)

2. The phonatory-articulatory level (e.g., describing the position or movement of the major structures such as the vocal folds)

3. The acoustic level (e.g., describing characteristics of the speech waveform emanating from the mouth)

Most of the current methods for measurement at the physiological and phonatory-articulatory levels are rather intrusive and require specialized equipment as well as a high level of expertise. In contrast, acoustic cues of vocal emotion expression may be obtained objectively, economically, and unobtrusively from speech recordings, and allow some inferences about voice production and physiological determinants.

Acoustic features, as one of the most popular effective features, mainly contain prosody features, voice quality features, and spectral features [27][18][14][30][7]. Pitch, loudness, and duration are commonly used as prosody features since they express the stress and intonation patterns of spoken language. Voice quality features, as the characteristic auditory coloring of an individual voice, are discriminative in expressing positive or negative emotions [27][18]. The widely used voice quality features are the first three formants (F1, F2, F3), spectral energy distribution, harmonics-to-noise-ratio, pitch irregularity (jitter), amplitude irregularity (shimmer), and so on. Meanwhile, the part of Spectral features are Linear Prediction Cepstral Coefficients (LPCC), Log Frequency Power Coefficients (LFPC), and Mel Frequency Cepstral Coefficients (MFCC) computed based on the short-term power spectrum of speech [2].

1.6 Conceptual Framework/Paradigm

This research, Speech Emotion Recognition, proposes a development speech emotion recognition system based on discrete wavelet transform (DWT) and voice segmentation combined with several existing features that represent emotion speech. Firstly, the speech dataset containing the silence area at the beginning and the end area are removed because they influence the emotional information of the signal. Then, speech signals are decomposed into discrete wavelet transform (DWT) to analyze and extract information from data because humans are very sensitive to low ear frequencies. This study decompose on low frequencies. Decompose signal is carried out based on the human speech frequency [37][15]. Based on the calculation, it shows that level 4 to level 6 provide better information about the speech signal.

Because each signal emotion dynamic and varied, the best way of segmenting the speech signal is using voiced speech segmentation. Since voiced segments contain information about emotions, for this reason this study select blocks of speech containing a few voiced segments. Nonetheless, the result voice segmentation provides different numbers of the segments for each speaker. Therefore, the classifier needs the same length of feature vector so that the maximum number of voice segments have to be selected from the whole dataset. To guarantee that the feature of each segment has equal length segment, the remaining segments needs to be filled with the available segments until maximal number of segments.

Then, feature extraction plays a crucial role in the overall performance of a speech recognition as well as speaker recognition system. This study uses the existing several features, i.e., ZCR, peak, energy, cepstrum, and Fourier coefficients [15][4][40]. In this part, these features in the voice segmentation to analyze emotional information of the speech signal are extracted and they are used for classification.

Artificial Neural Network (ANN) is used for the classification task in this work. The results are compared with other very popular classification techniques to establish the relevance of the feature set [27][18][17].

The methods applied in this study attempts to catch the dynamic emotional states from voice segments of the signals and extract features from them. In this study emotional states are assumed in the voice segment. Discrete wavelet transform (DWT) is also used to to select the most noiseless level of a signal. Thus, this study contributes to the development of a speech emotion recognition system based on discrete wavelet transform (DWT) and voice segmentation with a combination of a peak, ZCR, energy, cepstrum, and Fourier coefficient features. The block diagram of speech emotion recognition is shown in Figure 1.1.

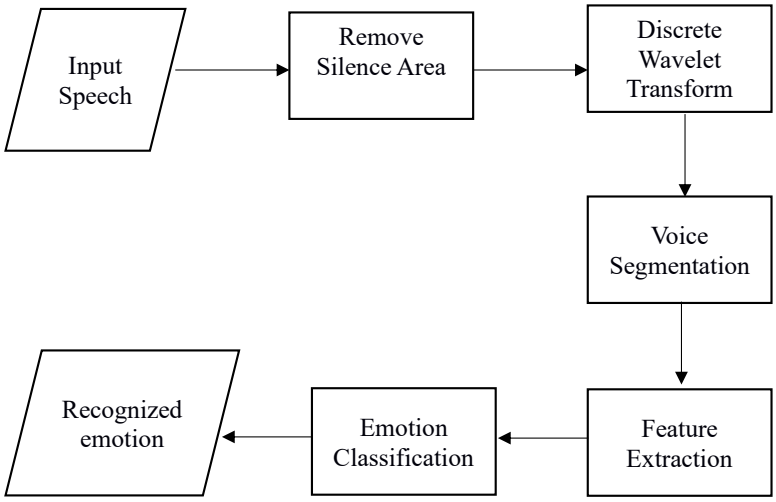

Figure 1.1: Block Diagram of speech emotion recognition

1.7 Scope and Delimitation

In this research, the scope and delimitation of this study are:

1. The dataset used for this study is RAVDESS where audio utterances with normal intensity.
2. Our datasets consist of only male and female adults.