DUAL-LEVEL SEGMENTATION METHOD FOR FEATURE EXTRACTION ENHANCEMENT STRATEGY IN SPEECH EMOTION RECOGNITION

NOOR AINA BINTI ZAIDAN

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Computer Science)

School of Computing
Faculty of Engineering
Universiti Teknologi Malaysia

JANUARY 2022

DEDICATION

This thesis is dedicated to my mother and father,
who taught me the best knowledge in this world, and always being there for me,
thank you for the blessed and magical Du'a, for the consistent support,
encouragement, and constant love that has sustained me throughout my life.

Also dedicated to my mother and father-in-law,
a very special thank you for providing a 'writing space' and for nurturing me
through the months of writing, for always asking for my well-being,
thank you for the constant motivation and blessing.

To all my siblings,
thank you for the motivation and constant support, for always being there hearing me out in ups and downs, thank you for making the world a happier place for me to be in, so I could gather all my strength, to finish what I've started.

Specially dedicated to my husband,
thank you for all the sacrifices, emotional encouragement, to keep believing in me long after I had lost belief in myself, and for sharing my wish to reach the goal of completing this precious journey, thank you for your endless support and motivation that helps me keep moving forward.

To my massive source of strength,
you inspire me every day to not give up, to keep me going no matter hard,
to help me grow and be better, thank you for the positive energy
you have showered me with, that push me beyond my capability,
this hard work is dedicated to you, little princess.

ACKNOWLEDGEMENT

Throughout the writing of this thesis, I have received a great deal of support and assistance. I wish to express my sincere appreciation to my main supervisor, Dr. Md. Sah Hj. Salam, for encouragement, guidance, critics, advice, motivation, and valuable guidance throughout my studies. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. Without the continued support and interest, this work would not have been the same as presented here.

My sincere appreciation also extends to all my colleagues, friends, and others who have aided in various events. Their views and tips are useful indeed. Thank you for the wonderful support. I am grateful to all my family members for their constant support.

My gratitude extends to The Ministry of Higher Education (MOHE) for the scholarship and funding opportunity to complete my postgraduate study at the School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia.

ABSTRACT

The speech segmentation approach could be one of the significant factors contributing to a Speech Emotion Recognition (SER) system's overall performance. An utterance may contain more than one perceived emotion, the boundaries between the changes of emotion in an utterance are challenging to determine. Speech segmented through the conventional fixed window did not correspond to the signal changes, due to the random segment point, an arbitrary segmented frame is produced, the segment boundary might be within the sentence or in-between emotional changes. This study introduced an improvement of segment-based segmentation on a fixed-window Relative Time Interval (RTI) by using Signal Change (SC) segmentation approach to discover the signal boundary concerning the signal transition. A segment-based feature extraction enhancement strategy using a dual-level segmentation method was proposed: RTI-SC segmentation utilizing the conventional approach. Instead of segmenting the whole utterance at the relative time interval, this study implements peak analysis to obtain segment boundaries defined by the maximum peak value within each temporary RTI segment. In peak selection, over-segmentation might occur due to connections with the input signal, impacting the boundary selection decision. Two approaches in finding the maximum peaks were implemented, firstly; peak selection by distance allocation, and secondly; peak selection by Maximum function. The substitution of the temporary RTI segment with the segment concerning signal change was intended to capture better high-level statistical-based features within the signal transition. The signal's prosodic, spectral, and wavelet properties were integrated to structure a fine feature set based on the proposed method. 36 low-level descriptors and 12 statistical features and their derivative were extracted on each segment resulted in a fixed vector dimension. Correlation-based Feature Subset Selection (CFS) with the Best First search method was applied for dimensionality reduction before Support Vector Machine (SVM) with Sequential Minimal Optimization (SMO) was implemented for classification. The performance of the feature fusion constructed from the proposed method was evaluated through speaker-dependent and speaker-independent tests on EMO-DB and RAVDESS databases. The result indicated that the prosodic and spectral feature derived from the dual-level segmentation method offered a higher recognition rate for most speaker-independent tasks with a significant improvement of the overall accuracy of 82.2% (150 features), the highest accuracy among other segmentation approaches used in this study. The proposed method outperformed the baseline approach in a single emotion assessment in both full dimensions and an optimized set. The highest accuracy for every emotion was mostly contributed by the proposed method. Using the EMO-DB database, accuracy was enhanced, specifically, happy (67.6%), anger (89%), fear (85.5%), disgust (79.3%), while neutral and sadness emotion obtained a similar accuracy with the baseline method (91%) and (93.5%) respectively. A 100% accuracy for boredom emotion (female speaker) was observed in the speaker-dependent test, the highest single emotion classified, reported in this study.

ABSTRAK

Pendekatan segmentasi pertuturan boleh menjadi salah satu faktor utama yang menyumbang kepada prestasi keseluruhan sistem Pengecaman Emosi Ucapan (SER). Satu ucapan mungkin mengandungi lebih dari satu jenis emosi, sempadan antara perubahan emosi dalam ucapan sukar untuk ditentukan. Ucapan yang disegmentasikan melalui cara konvensional tetingkap tetap tidak mengambil kira perubahan isyarat, menghasilkan titik segmen rawak dan keratan segmen secara rambang, sempadan segmen mungkin berada dalam ayat atau di antara perubahan emosi. Kajian ini memperkenalkan penambahbaikan segmentasi berasaskan segmen pada Selang Waktu Relatif (RTI) tetingkap tetap dengan menggunakan pendekatan segmentasi Perubahan Isyarat (SC) untuk menetapkan batas isyarat bagi segmentasi ucapan berdasarkan peralihan isyarat. Strategi penambahbaikan pengekstrakan ciri berasaskan segmen menggunakan kaedah segmentasi dua tingkat telah dicadangkan iaitu segmentasi RTI-SC yang menggabungkan pendekatan konvensional. Selain daripada pembahagian keseluruhan ucapan pada selang waktu relatif, kajian ini menggunakan kaedah analisis puncak untuk mendapatkan batas segmen yang ditentukan oleh nilai puncak maksimum dalam setiap segmen RTI sementara. Dalam pemilihan puncak, pembahagian berlebihan mungkin berlaku disebabkan oleh sambungan dengan isyarat input, yang memberi kesan kepada keputusan pemilihan sempadan. Dua pendekatan dalam mencari puncak maksimum telah dilaksanakan iaitu pertama; pemilihan puncak dengan peruntukan jarak dan kedua; pemilihan puncak oleh fungsi Maksimum. Penggantian segmen RTI sementara dengan segmen SC bertujuan untuk mendapatkan ciri emosi ucapan yang lebih baik melalui statistik tingkat tinggi yang diperoleh dari peralihan isyarat. Ciri prosodik, spektrum dan gelombang isyarat telah disatukan untuk menghasilkan set ciri emosi ucapan yang lebih baik berdasarkan kaedah yang dicadangkan. 36 deskriptor tahap rendah dan 12 ciri statistik dan terbitannya telah diekstrak pada setiap segmen menghasilkan dimensi vektor tetap. Pemilihan Ciri Subset berasaskan Korelasi (CFS) dengan kaedah carian Terbaik Pertama digunakan untuk pengurangan dimensi sebelum Mesin Vektor Sokongan (SVM) dengan Pengoptimuman Minimum Berurutan (SMO) dilaksanakan untuk klasifikasi. Prestasi gabungan ciri yang dibina dari kaedah yang dicadangkan telah dinilai melalui ujian penutur-bersandar dan bebas penutur-bersandar pada pangkalan data EMO-DB dan RAVDESS. Hasil kajian menunjukkan bahawa ciri prosodik dan spektrum yang diperolehi melalui kaedah segmentasi dua tingkat menawarkan kadar pengiktirafan yang lebih tinggi untuk kebanyakan ujian bebas penutur-bersandar dengan peningkatan yang ketara pada ketepatan keseluruhan 82.2% (150 ciri), ketepatan tertinggi antara pendekatan segmentasi lain yang digunakan dalam kajian ini. Kaedah yang dicadangkan mengatasi pendekatan asas dalam penilaian emosi tunggal dalam kedua-dua dimensi penuh dan optimum set. Ketepatan tertinggi untuk setiap emosi tunggal banyak disumbangkan oleh kaedah yang dicadangkan. Dengan menggunakan pangkalan data EMO-DB, ketepatan ditingkatkan, khususnya emosi gembira (67.6%), marah (89%), takut (85.5%), meluat (79.3%) sementara emosi neutral dan sedih memperoleh ketepatan yang setara dengan kaedah asas, masing-masing (91%) dan (93.5%). Ketepatan 100% untuk emosi bosan (penutur wanita) diperolehi dalam ujian penutur-bersandar, emosi tunggal tertinggi dikelaskan, dilaporkan dalam kajian ini.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	-	Artificial Intelligence
ASM	-	Acoustic Segment Model
ASR	-	Automatic Speech Recognition
ATIR	-	Absolute Time Intervals at Relative Positions
AVEC	-	Audio/ Visual Emotion Challenge
CFS	-	Correlation-based Feature Selection
DFT	-	Discrete Fourier Transform
DSP	-	Digital Signal Processing
EEG	-	Electroencephalogram
FFT	-	Fast Fourier Transform
GTI	-	Global Time Interval
HCI	-	Human-Computer Interaction
HSF	-	High-level Statistical Function
LLD	-	Low-level Descriptor
LP	-	Linear Prediction
LPCs	-	Linear Prediction Coefficient
MFCC	-	Mel Frequency Cepstral Coefficient
RSPA	-	Residual Sinusoidal Peak Amplitude
RTI	-	Relative Time Interval
SBS	-	Sequential Backward Selection
SC	-	Signal Change
SER	-	Speech Emotion Recognition
SFS	-	Sequential Forward Selection
SMO	-	Sequential Minimal Optimization
STFT	-	Short-time Fourier Transform
SVM	-	Support Vector Machine
WAV	-	Waveform Audio File Format
ZCR	-	Zero-Crossing Rate

CHAPTER 1

INTRODUCTION

1.1 Introduction

Humans have conversations almost every day to deliver and exchange information, emotion is certainly included in the discussion. Emotions shown by humans have a great impact on the decision-making process. A system with the ability to understand human speeches and emotions is anticipated to greatly contribute to more natural human-computer interaction (HCI). With such machine, messages can be delivered accurately, individual characters can be identified, the emotional state of humans can be classified and even stress levels can also be detected, thus the communication process between man and machine will work more effectively and could deliver great purposes to human life. To make a more human-like machine, a depth understanding of the emotional intelligent principle must be acquired so that the man-machine interaction could be improved by having machines capable to offer a natural and reliable conversation where the user's emotional state is considered.

Emotional state can be recognized through facial expression, speech, and commonly used physiological signals - electroencephalogram (EEG). Speech emotional recognition (SER) has its significance in today's technology. SER is the task of automatically recognizing human emotion and affective states from speech. Emotion recognition from speech signals is progressively developed and the interest in the methods of integrating emotion detection in the machines has been increasing a lot within the past two decades. Studies in the SER field have been widely carried out in the areas related to speech user interfaces and spoken language processing and has evolved to some extent where it can be the "next big thing" for the industry in developing further beneficial applications while improving the life quality (Schuller, 2018).

Human speech consists of a combination of sentences; words syllables and phonemes. There are two major components in a continuous speech signal, one part contains speech information (which can be further divided into voice and unvoiced speech), and the other part contains noise or silent properties in between the spoken word (Sakran *et al.*, 2017). A continuous speech signal can be segmented based on a phonemic, sub phonemic, syllabic, word level, syntagmatic level (Amirgaliyev *et al.*, 2017) depending on the segmentation algorithm employed.

Segmentation is an important signal pre-processing step in SER system design for the conversion of a single section of the signal to smaller segments before going through a feature extraction process. The segmentation method is implemented during the pre-processing phase to define the speech segment boundaries by splitting speech signals into several small frames and a feature vector is constructed from each segmented speech. Traditionally, speech labelling and segmentation were manually done depending on the linguistic information of the spoken utterance. Manual segmentation is at disadvantage compared to automatic segmentation because the result is inconsistent, time-consuming, and prone to error since it is implemented by trained phoneticians based on personal listening and visual judgment on required boundaries (Sharma and Mammone, 1996).

Automatic segmentation procedure is another preferred way to segment speech automatically according to the signal acoustic properties depending on the linguistic knowledge and it was broadly used in the Automatic Speech Recognition (ASR) system (Sakran *et al.*, 2017). The boundaries between the two standard signals can be identified in the same way when automated segmentation is implemented; repeated segmentation results for the entire signal can be detected. When the linguistic knowledge is not necessarily required, a 'blind' speech segmentation procedure is implemented which allows a speech sample to be segmented into several frames (Sharma and Mammone, 1996), (Schuller and Rigoll, 2006). The initial step of blind segmentation is entirely based on the signal's acoustic characteristics (Sakran *et al.*, 2017), due to the limited linguistic knowledge, finding a starting and endpoint of speech boundary concerning the emotional content is a challenging task. Segment boundary could be located using the endpoint detection to differentiate the silence and voice part and the emotion information is measured within the whole dialogue rather than part of the sentence. The purpose of endpoint detection is to find the beginning and the end of meaningful partitions. The following criteria are used to assess the efficacy of segmentation algorithms: precision in establishing segment boundaries, robustness, noise resistance, and executing time (Amirgaliyev *et al.*, 2017). It is an important procedure in the machine learning domain to discover the knowledge, patterns and avoid the predictive model from learning on unrelated features.

Speech features can be extracted based on low-level descriptor (LLD) – local, and high-level statistical function (HFS) - global approach. Local features define the temporal dynamics in the prosody and global feature highlights the statistical value (Rao, Koolagudi and Vempada, 2013). Mean, standard deviation, max, min, kurtosis, skewness, and median, are some global statistical features mostly used in SER (Wen *et al.*, 2017). Global statistic features could be useful to reduce computation as it produces smaller and fixed dimensionality details compared to local features extracted from each frame (Badshah *et al.*, 2019). The main idea of feature extraction is to obtain a set of desired information that represents the properties of the original data (Giannakopoulos and Pikrakis, 2014a).

Since the past decade, the search for the optimal speech feature set to represent emotion and the extraction strategy has been actively pursued, (Bitouk *et al.*, 2010), (A. Ingale and Chaudhari, 2012), (Sezgin *et al.*, 2012). Feature extraction strategy has been a current challenging research topic, due to the data insufficient problem. A frequent number of researches involves in-depth studies on extraction strategies among various types of feature groups that lead to better recognition accuracy were previously reported (Rao *et al.*, 2010), (Kishore and Satish, 2013), (Gharsellaoui *et al.*, 2015), (Jing *et al.*, 2018), (Guo *et al.*, 2019). In the research literature, some studies proved that feature integration is effective in classifying emotion, but these different types of feature representations are usually diverse as it is structured from various type of feature, so a basic challenge is how to effectively integrate the diversity information for better recognition performance. Multiple features are merely concatenated into a single high-dimensional feature vector and fed into a final classifier which has difficulty in joining learning fundamental correlations between different acoustic feature representations (Jiang *et al.*, 2019).

Ayadi, Kamel and Karray, (2011) stated, prosody continuous features like energy and pitch greatly represent the emotional information of an utterance. According to Origlia, Galatà and Ludusan, (2010) and Koolagudi and Rao, (2012), global prosodic features are usually used in the emotion recognition task. Most of the earlier researches works were mainly focused on prosodic features alone such as pitch/ fundamental frequency (f0), intensity, duration, energy, and MFCC (Anagnostopoulos *et al.*, 2012), (Origlia *et al.*, 2010), (Yutai *et al.*, 2009), and some only focus on spectral feature alone like MFCC (Bitouk *et al.*, 2010), (Bhaykar *et al.*, 2013).

Single emotion features lead to inconsistency in recognition with a lower recognition rate, a combination of multiple features that are capable to describe emotional information is needed in generating the optimal feature set. Soon afterward, researchers were actively conducting studies on the integration of a few feature categories: prosodic, spectral, and voice quality features by combining them to maximize the rate of emotional recognition, resulting in a robust feature set (Bozkurt and Erzin, 2009), (Zhou *et al.*, 2010), (A. Ingale and Chaudhari, 2012), (Safdarkhani *et al.*, 2012), (Seehapoch and Wongthanavasu, 2013), (Gharsellaoui *et al.*, 2015). (Watile *et al.*, 2017).

A feature set constructed with a high-dimensional feature vector usually elevates the computation complexity. An extensive study on a predictive model with a fine feature set structure is crucial. The efficiency of the emotion recognition process is heavily influenced by the quality of segmentation results that contribute to the good selection of required features. Appropriate segmentation approach, feature extraction strategy, and selection algorithm of data attributes are necessary for irrelevant data removal procedure and feature dimension reduction to improve learning performance by lowering computational complexity and providing a good decision-making process with shorter processing time. There are still more potential features extraction strategies that have not been studied and there is still room for improvement.

1.2 Research Background

Emotional expression may appear across several sentences, or on any word in speech. Since emotion is not highly dependent on the spoken words or the linguistic content, an utterance may contain a possible mixture of perceived emotion and the boundaries between the changes of emotion are difficult to determine, making it hard for the SER system to define the dominant emotion. According to physiological and psychological studies, expressing emotion in speech has a beginning, a rising side, a peak, and a declining side (Ekman, 2003).

Speech boundary could be defined by the temporal dynamics of the signal, based on the extracted feature. The technique of identifying the presence of voiced speech among other unvoiced speech and silence regions is known as endpoint detection, speech detection, or voice activity detection. The system’s accuracy is influenced by the performance of the endpoint detection algorithm, eliminating the voice and noise frames in a dynamic environment makes it easier to model speech (Berkehan and Kaya, 2020).

Defining the basic unit of the segmented speech in a continuous speech that best represents single emotion is one of the ongoing challenges; the segmented speech should be long enough to define single emotion and short enough to isolate the presence of other emotions in that utterance (Batliner *et al.*, 2010), (Guo *et al.*, 2019). Small segments may be providing insufficient informative peak area, while longer segments subsequently expressed emotions may affect each other (Mansoorizadeh and Charkari, 2007). As stated in (Lee and Cho, 2016), the frame size of 25ms could potentially wipe the dynamic properties in a speech signal due to the rapid changes of spectral characteristics. Research findings stated a speech segment longer than 0.25 seconds carries enough emotional information (Provost, 2013), (Sahoo *et al.,* 2019).

Looking at the progress of studies on segmentation approaches in the SER domain, it can be argued that the subtopic of segmentation is still under discussion. Based on current related research, finding the right segmentation approach has been one of the remaining challenges that need to be sought after. Several automatic segmentation approaches have been proposed with the idea of segmenting the signal into smaller frames under supervised and unsupervised segmentation before executing the desired procedure. (Schuller and Rigoll, 2006) and (Zhang *et al.*, 2014), (Tzinis and Potamianos, 2017) implying a timing-levels in segment-based segmentation in the previous study referring to a relative time interval (RTI) approach and absolute time intervals at relative positions (ATIR) while (Yeh *et al.*, 2011), Huang *et al.*, (2019), (Atmaja and Akagi, 2019) implemented the unsupervised segmentation strategy based on signal change detection method in their research.

The implementation of fixed-window segmentation is still relevant in terms of the emotion classification ability, reliable result is still achieved in Zhang, Warisawa and Yamada, (2014), (Lee and Cho, 2016), Sahoo *et al.*, (2019). The speech signal is divided into segments of a fixed window with predefined window lengths, resulting in an individual speech sample. The feature extraction phase is implemented based on each segmented speech frame, to capture the distinctive temporal dynamics within the speech, a suitable segmentation approach is required.

Fixed window segmentation is less favourable in some studies because the segmentation point may be in the middle of a short phrase, the segmentation result is not optimal (Yeh *et al.*, 2011). Furthermore, if one partition contains two or more emotional expressions, the recognition result will be inaccurate. The segmentation result from fixed-length segment might not be optimal due to the segment points location, the segment boundary might be within the sentence or in-between emotional changes, the method might carry inadequate emotional information, thus leading to the inaccurate outcome. Other researchers support the use of signal change segmentation for better recognition accuracy compared to fixed window segmentation. Lee *et al.*, (2013) are concerned about the need to incorporate temporal information in acoustic feature sequences in determining emotional speech category, the use of fixed-window segmentation alone is still lacking to provide satisfactory results. The Acoustic Segment Model (ASM) approach is proposed to classify utterances by their acoustic feature sequences. The implementation of ASM approach is supported in (Zheng *et al.*, 2021), due to the potential use of acoustic information for performing SER tasks. Amirgaliyev, Hahn and Mussabayev, (2017) used pitch frequency analysis by observing the average number of zero transitions functions and the signal energy function to construct a speech parameterization. The speech signal is segmented using the parameterization result to isolate the segments with stable spectral properties. Huang *et al.,* (2019), implement signal change segmentation for silence detection, verbal/nonverbal segment detection, and prosodic-phrase segmentation procedures to obtain sound/speech segments. (Atmaja *et al.*, 2019) remove the silence part, considering silence brings unnecessary information, and use only the segmented speech part of the utterance for feature extraction.

The mutual proclamation about explicit features in speech signals that represent emotional information is uncertain and insufficient, it is a widespread challenge being faced by SER systems including the range of features that can distinguish individual emotion (Sahoo *et al.*, 2019), (Badshah *et al.*, 2019). Thus far, researchers are still experimenting and proposing new emotion-related features as indicated by (Jing *et al.*, 2018), old-fashioned acoustic features with traditional approaches still cannot promise satisfactory system performance due to the deficiency of discriminative acoustic features. Most existing research related to emotion recognition from speech focuses on basic emotion classification since the main feature for each basic emotion is still unclear make it hard to emphasize the most persuasive feature for classifying emotions.

Spectral and prosodic are among two features that well describe emotion. Speech energy, fundamental frequency, formant, and Mel-frequency Cepstral Coefficient (MFCC) are widely used in research literature because they can differentiate certain states of emotion effectively (B. A. Ingale & D. Chaudhari, 2012). Features from the spectral group alone also delivered satisfactory results using MFCC and Modulation spectral (MS) feature (Kerkeni *et al.*, 2018). It was further reported that the MFCC feature is often used and considered as the best representation of the voice signal's spectral property where human perception sensitivity towards frequency is considered. Aside from auditory suggestive features, MFCC was optimally merged with chosen prosodic and voice quality features to improve recognition accuracy (Gharsellaoui *et al.*, 2015).

As the research area expands, various fusion set has been proposed including the wavelet Sub-band Based Cepstral (SBC) that has been early introduced in (Sarikaya *et al.*, 1998) for the efficiency of recognizing emotion in a noisy environment. A comparative study by Kishore and Satish, (2013), evaluate the sensitivity of MFCC and wavelet features, SBC towards noisy data. The result shows, SBC parameters produce better recognition accuracy than MFCC and are proven to have less sensitivity towards noisy data. Chenchah and Lachiri, (2014) also proved that speech emotion recognition systems based on the wavelet packet energy and entropy features yield the best average result and are robust for both acted and spontaneous databases. Since wavelet features give better results in a noisy environment and are robust in both acted and spontaneous databases, the combination of wavelet, spectral, and prosodic features might further improve the recognition accuracy, a system that may well be withstanding environmental noise should be more practical and reliable used in a future application with real-time processing.

1.3 Problem Statement

The segmentation approach could be one of the major factors that contribute to the overall performance of an SER system. An utterance may contain more than one perceived emotion but the boundaries between the changes of emotion in an utterance are difficult to determine. Speech segmented through the conventional fixed window did not correspond to the signal changes, due to the random segment point. The segmentation approach at relative time interval in finding the segment boundaries might carry insufficient emotional information, as it produces arbitrary segmented frame, a refined segmentation method is required to isolate the boundaries between emotion change according to the signal transition, hence a better feature structure could be constructed when emotional information is well defined. In summary, some problems that need to be addressed in emotional recognition are:

Table 1.1 Problem to be addressed

Problem	**Description**
Fixed-length speech segment	• Segmentation results from fixed-length segment might not be optimal due to the segment points location, the segment boundary might be within the sentence or in-between emotional changes (Yeh *et al.*, 2011), and lacking temporal information in acoustic feature (Lee *et al.*, 2013), Deficient semantic functionality access to a sequence of associated patterns or interpretations (Amirgaliyev *et al.*, 2017), even segmentation approach is used with an unsatisfactory result on performing SER task (Zheng *et al.*, 2021).
Peak Detection	• As stated in Giannakopoulos and Pikrakis, (2014b), during signal change detection, over-segmentation will occur if a huge number of local maxima might be detected when a short-term feature vector is employed directly in the computation, unless a refined peak selection technique is implemented. In peak selection, the signal thresholding computation is difficult, due to connections with the input signal, which impact the boundary selection decision (Maka, 2020).
Insufficient feature	• Single emotion features lead to inconsistency in recognition with a lower recognition rate, a combination of multiple features is needed in generating the optimal feature set (Gharsellaoui *et al.*, 2015). Old-fashioned acoustic features with traditional approaches still cannot promise satisfactory system performance due to the deficiency of discriminative acoustic features (Jing *et al.*, 2018). Explicit features that represent emotional information are insufficient (Sahoo *et al.*, 2019), the optimal feature to differentiate distinct emotions is still in pursuit.

1.4 Research Questions

To address the problems mentioned, the solution to the following research questions should be sought:

a) Which segmentation method corresponds more to the emotional changes in an utterance and how to define a clear segment boundary in emotional speech signal?

b) Could the refined peak selection method avoid the over-segmentation problem and capture better emotional information within the signal transition?

c) Does integrating prosodic, spectral and wavelet statistical representation derived from feature extraction strategy through proposed segmentation approach provide better emotion recognition accuracy?

1.5 Research Aims

This research aims to identify a new speech segment boundary based on maximum peak selection implemented on the proposed method: a dual-level segmentation, the feature extraction strategy will be enhanced by selecting statistical features representation derived from a speech segmented which reflect the signal change, instead of the random segmented speech signal.

1.6 Research Objectives

Three main objectives to be achieved in constructing a robust statistical feature set for speech emotion recognition:

a) To propose a dual-level segmentation method for identifying the new segment boundary based on maximum peak selection.

b) To enhance feature extraction strategy and construct statistical representation based on hybrid features through the proposed segmentation approach.

c) To evaluate the performance of the statistical feature set derived from the proposed method.

1.7 Research Scope

The scope of the project has been determined to carry out this research are as follow:

a) Several emotional states are selected to be classified including the basic emotions: happiness, sadness, fear, surprise, anger, disgust, boredom, and neutral.

b) Low-level descriptor (LLDs): zero-crossing rate, energy, the entropy of energy, spectral centroid and spread, spectral flux, spectral roll-off, MFCCs 13 coefficient, 12 Chroma Vector, Harmonic, Fundamental frequency (F_0), SBC.

c) 12 high-level statistical functions (HSFs): min, max, mean, median, mode, standard deviation, variance, skewness, kurtosis, range, interquartile range, mean absolute deviation.

d) Feature combination of the prosodic, spectral, and wavelet group that carries the most emotional information will be constructed using RTI segmentation and signal change detection approach based on peak analysis.

e) Optimizing feature set using Correlation-based feature subset selection with best first search method.

f) SVM classifier with the SMO algorithm has been selected to classify those emotions using WEKA analysis tools.

g) The experiment will be conducted using Berlin Emotional Speech Database (EMO-DB) with acted emotional data and The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) with elicited emotional data.

1.8 Importance of Study

Improper speech segmentation algorithm might lead to lower recognition accuracy when emotional information captured within a segmented partition is carried along with the unnecessary information. A feature extraction strategy must be well structured; the selection of features is crucially important to advance the system performance with better recognition accuracy. This study focuses on the importance of the segmentation approach during the pre-processing phase in structuring the optimal feature extraction strategy. The performance is validated on speaker-dependent and speaker-independent tests using the state-of-the-art classifier. The efficiency of the statistical feature set constructed from the proposed dual-level segmentation method based on peak analysis has been analyzed to discover whether it captures better emotional information compared to the conventional approach and the emotional change in between signal transition is observed.

1.9 Thesis Organization

The details of the process flow for this thesis are structured in the following chapters accordingly for better reference. The remainder of this thesis is organized as follows:

Chapter 2 provides a further explanation of digital signal processing, speech emotion recognition, emotional model and database, background research, previous studies conducted by other researchers related to the segmentation approach, and feature extraction strategy are also presented. The methods used for analyzing speech signals from the emotional speech database are further discussed.

Chapter 3 will discuss on research framework and the methods used in executing the enhancement strategy of feature extraction based on the proposed method: RTI segmentation and signal change detection to classify emotion through speech signal. The whole methodologies chapter will have a general discussion on design and procedure, emotional data collection, software justification, segmentation approach, baseline feature extraction method, feature optimization, and classification technique used.

Chapter 4 will explain the detailed explanation of the whole study covering specific implementation tasks, experimental design, data analysis, and evaluation to accomplish the objective of the study. Each of the implementation phases will be explained in detail based on the research framework element for a better understanding of the research flow. The experimental setup, preliminary and comparative study on the experiment conducted, segmentation approach, feature extraction strategy, feature selection algorithm, and the optimal classification are presented.

The results of this study are presented in Chapter 5, the comparison of results from the baseline feature extracted using the RTI segmentation approach and the result after implementing signal change detection using peak analysis. Few experiments are conducted, the performance of the proposed dual-level segmentation has been observed. This study also highlights the potential of the proposed algorithm through framework design and detailed result analysis, focusing on reviewing other related models, and showing how this study is distinguished from others' work. Possible further works using the proposed method will be suggested at the end of the chapter.

Chapter 6 will highlight the research finding, achievements, and contributions of this study that might be useful to advance the industry. The advantage of using the proposed algorithm dual-level segmentation method and the summary of the study is further discussed.

REFERENCES

Ali, S. A., Zehra, S., Khan, M. and Wahab, F. (2013) ‘Development and Analysis of Speech Emotion Corpus Using Prosodic Features for Cross Linguistics’, *International Journal of Scientific & Engineering Research*, 4(1), pp. 1–8.

Álvarez, A., Cearreta, I., López, J. M., Arruti, A., Lazkano, E., Sierra, B. and Garay, N. (2006) ‘Feature Subset Selection Based on Evolutionary Algorithms for Automatic Emotion Recognition in Spoken Spanish and Standard Basque Language’, in, pp. 565–572.

Amirgaliyev, Y., Hahn, M. and Mussabayev, T. (2017) ‘The speech signal segmentation algorithm using pitch synchronous analysis’, *Open Computer Science*, 7(1), pp. 1–8.

Anagnostopoulos, C. N., Iliou, T. and Giannoukos, I. (2012) ‘Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011’, *Artificial Intelligence Review*, 43(2), pp. 155–177.

Atmaja, B. T. and Akagi, M. (2019) ‘Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model’, *Proceedings - 2019 IEEE International Conference on Signals and Systems, ICSigSys 2019*. IEEE, pp. 40–44.

Atmaja, B. T., Shirai, K. and Akagi, M. (2019) ‘Speech emotion recognition using speech feature and word embedding’, *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2019*, pp. 519–523.

Ayadi, M. El, Kamel, M. and Karray, F. (2011) ‘Survey on Speech Emotion Recognition: Features, classification schemes, and databases’, *Pattern Recognition*. Elsevier, 44(3), pp. 572–587.

Badshah, A. M., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Lee, M. Y., Kwon, S. and Baik, S. W. (2019) ‘Deep features-based speech emotion recognition for smart affective services’, *Multimedia Tools and Applications*. Multimedia Tools and Applications, 78(5), pp. 5571–5589.

Bahoura, M. and Pelletier, C. (2004) ‘Respiratory sounds classification using cepstral analysis and gaussian mixture models’, *Annual International Conference of the*

IEEE Engineering in Medicine and Biology - Proceedings, 26 I(February 2004), pp. 9–12.

Basharirad, B. and Moradhaseli, M. (2017) ‘Speech emotion recognition methods: A literature review’, *AIP Conference Proceedings*, 1891(January 2017).

Batliner, A., Seppi, D., Steidl, S. and Schuller, B. (2010) ‘Segmenting into adequate units for automatic recognition of emotion-related episodes: A speech-based approach’, *Advances in Human-Computer Interaction*, 2010.

Berkehan, M. and Kaya, O. (2020) ‘Speech emotion recognition : Emotional models , databases , features , preprocessing methods , supporting modalities , and classifiers’, 116(October 2019), pp. 56–76.

Bhargava, M. and Polzehl, T. (2013) ‘Improving Automatic Emotion Recognition from speech using Rhythm and Temporal feature’, *Icecit-2012*, pp. 139–147.

Bhaykar, M., Yadav, J. and Rao, K. S. (2013) ‘Speaker dependent, speaker independent and cross language emotion recognition from speech using GMM and HMM’, in *2013 National Conference on Communications, NCC 2013*.

Bitouk, D., Verma, R. and Nenkova, A. (2010) ‘Class-level Spectral Features for Emotion Recognition’, *Speech Communication*. Elsevier B.V., 52(7–8), pp. 613–625.

Blanton, S. (1915) The voice and the emotions, Quarterly Journal of Speech, 1:2, 154-172.

Boashash, B. and Barkat, B. (2001) ‘Introduction to Time-Frequency Signal Analysis’, in *Wavelet Transforms and Time-Frequency Signal Analysis*. Boston, MA: Birkhäuser Boston, pp. 321–380.

Bojanić, M., Delić, V. and Karpov, A. (2020) ‘Call redistribution for a call center based on speech emotion recognition’, *Applied Sciences (Switzerland)*, 10(13), pp. 6–8.

Bozkurt, E. and Erzin, E. (2009) ‘Improving Automatic Emotion Recognition from Speech Signals’, in *Interspeech 2009: 10th Annual Conference of the International Speech Communication Association*.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F. and Weiss, B. (2005) *A Database of German Emotional Speech*, *Interspeech*.

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S. and Narayanan, S. S. (2008) ‘IEMOCAP: Interactive emotional dyadic motion capture database’, *Language Resources and Evaluation*, 42(4), pp.

335–359.

Caponetti, L., Buscicchio, C. A. and Castellano, G. (2011) ‘Biologically inspired emotion recognition from speech’, pp. 1–10.

Caponetti, L., Buscicchio, C. and Castellano, G. (2011) ‘Biologically Inspired Emotion Recognition from Speech’, *EURASIP Journal on Advances in Signal Processing*. Springer Open Ltd, (1), p. 10.

Chandrasekar, P., Chapaneri, S. and Jayaswal, D. (2014) ‘Emotion Recognition from Speech using Discriminative Features’, *International Journal of Computer Applications*, 101(16), pp. 31–36.

Chenchah, F. and Lachiri, Z. (2014) ‘Speech emotion recognition in acted and spontaneous context’, *Procedia Computer Science*. Elsevier Masson SAS, 39(C), pp. 139–145.

Davletcharova, A., Sugathan, S., Abraham, B. and James, A. P. (2015) ‘Detection and Analysis of Emotion from Speech Signals’, *Procedia Computer Science*. Elsevier Masson SAS, 58, pp. 91–96.

Deb, S. and Dandapat, S. (2016) ‘Emotion Classification Using Residual Sinusoidal Peak Amplitude’.

Descartes, R. (1952). Descartes philosophical writings.

Dov, D., Talmon, R. and Cohen, I. (2017) ‘Multimodal Kernel Method for Activity Detection of Sound Sources’, *IEEE/ACM Transactions on Audio Speech and Language Processing*. IEEE, 25(6), pp. 1322–1334.

Ekman, P. (1992) ‘An argument for basic emotions’, *Cognition & Emotion*, 6(3–4), pp. 169–200.

Ekman, P. (2003) ‘Emotions revealed: Recognizing faces and feelings to improve communication and emotional life.’, *Times Books/Henry Holt and Co.*, 328(Suppl S5), p. 0405184.

Flake, G. W. and Lawrence, S. (2002) ‘Efficient SVM regression training with SMO’, *Machine Learning*, 46(1–3), pp. 271–290.

Frank, E., Hall, M. A. and Witten, I. H. (2017) ‘The WEKA workbench’, *Data Mining*, pp. 553–571.

Garg, U., Agarwal, S., Gupta, S., Dutt, R. and Singh, D. (2020) ‘Prediction of Emotions from the Audio Speech Signals using MFCC, MEL and Chroma’, *Proceedings - 2020 12th International Conference on Computational Intelligence and Communication Networks, CICN 2020*, pp. 87–91.

Gharsellaoui, S., Selouani, S. A. and Dahmane, A. O. (2015) 'Automatic emotion recognition using auditory and prosodic indicative features', in *Canadian Conference on Electrical and Computer Engineering*, pp. 1265–1270.

Giannakopoulos, T. and Pikrakis, A. (2014a) 'Audio Features', in *Introduction to Audio Analysis*, pp. 59–103.

Giannakopoulos, T. and Pikrakis, A. (2014b) *Audio Segmentation*, *Introduction to Audio Analysis*.

Guo, L., Wang, L., Dang, J., Liu, Z. and Guan, H. (2019) 'Exploration of Complementary Features for Speech Emotion Recognition Based on Kernel Extreme Learning Machine', *IEEE Access*. IEEE, 7, pp. 75798–75809.

Hall, M. A. (1999) 'Correlation-based Feature Selection for Machine Learning', (April).

Hall, M. A. and Smith, L. A. (1998) 'Practical Feature Subset Selection for Machine Learning', *proceedings of the 21st Australasian Computer Science Conference, ACSC'98, Perth, 4-6 February*, Volume 20, p. 586.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. (2009) 'The WEKA Data Mining Software: An Update', *SIGKDD Explorations*, 11(1).

Han, K., Yu, D. and Tashev, I. (2014) 'Speech Emotion Recognition using Deep Neural Network and Extreme Learning Machine', *Proceedings of Interspeech*, (September), pp. 223–227.

Han, Z., Lun, S. and Wang, J. (2012) 'Speech Emotion Recognition System Based on Integrating Feature and Improved HMM', *Proceedings of the 2nd International Conference on Computer Application and System Modeling*. Paris, France: Atlantis Press, pp. 571–574.

Hoque, M. E., Yeasin, M. and Louwerse, M. M. (2006) 'Robust recognition of emotion from speech', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4133 LNAI(May 2014), pp. 42–53.

Huahu, X., Jue, G. and Jian, Y. (2010) 'Application of speech emotion recognition in intelligent household robot', *Proceedings - International Conference on Artificial Intelligence and Computational Intelligence, AICI 2010*. IEEE, 1, pp. 537–541.

Huang, K.-Y., Wu, C.-H., Hong, Q.-B., Su, M.-H. and Chen, Y.-H. (2019) 'Verbal and

Nonverbal Speech', *Icassp*, pp. 5866–5870.

Iida, A., Iga, S., Higuchi, F., Campbell, N. and Yasumura, M. (2000) 'A Speech Synthesis System with Emotion for Assisting Communication', Proc. ITRW on Speech and Emotion, 2(2), pp. 63–70.

Iliev, A. (2009) *Emotion Recognition Using Glottal and Prosodic Features*. University of Miami.

Iliou, T. and Anagnostopoulos, C. (2009) 'Statistical Evaluation of Speech Features for Emotion Recognition', in *2009 Fourth International Conference on Digital Telecommunications*, pp. 121–126.

Ingale, A. and Chaudhari, D. (2012) 'Speech Emotion Recognition', *International Journal of Soft Computing and Engineering (IJSCE)*, 2(1), pp. 235–238.

Ingale, B. A. and Chaudhari, D. . (2012) 'Speech Emotion Recognition Using Hidden Markov Model and Support Vector Machine', *International Journal of Advanced Engineering Research and Studies (IJAERS)*, 1(3), pp. 316–318.

J.C.Platt (1999) 'Fast Training of Support Vector Machines Using Sequential Minimal Optimization', in *Advances in Kernel Methods, Support Vector Learning*, pp. 185–208.

Jagini, N. P. and Rao, R. R. (2017) 'Exploring emotion specific features for emotion recognition system using PCA approach', in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, pp. 58–62.

Ji, Y., Wei, J., Zhang, J., Fang, Q., Lu, W., Honda, K. and Lu, X. (2017) 'Speech Behavior Analysis by Articulatory Observations', *Procedia Computer Science*. Elsevier B.V., 111(2015), pp. 463–470.

Jiang, W., Wang, Z., Jin, J. S., Han, X. and Li, C. (2019) 'Speech emotion recognition with heterogeneous feature unification of deep neural network', *Sensors (Switzerland)*, 19(12), pp. 1–15.

Jin, Y., Song, P., Zheng, W. and Zhao, L. (2014) 'A feature selection and feature fusion combination method for speaker-independent speech emotion recognition', *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (1), pp. 4808–4812.

Jing, S., Mao, X. and Chen, L. (2018) 'Prominence features: Effective emotional features for speech emotion recognition', *Digital Signal Processing*. Elsevier Inc., 72, pp. 216–231.

Jones, C. and Jonsson, I.-M. (2008) 'Using Paralinguistic Cues in Speech to Recognise

Emotions in Older Car Drivers', in *Affect and Emotion in Human-Computer Interaction*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 229–240.

Kalamani, M., Valarmathy, S., Anitha, S. and Mohan, R. (2014) 'Review of Speech Segmentation Algorithms for Speech Recognition', 3(11), pp. 1572–1574.

Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Ali Mahjoub, M. and Cleder, C. (2019) 'Automatic Speech Emotion Recognition Using Machine Learning', *Social Media and Machine Learning [Working Title]*, (March).

Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K. and Mahjoub, M. A. (2018) 'Speech emotion recognition: Methods and cases study', *ICAART 2018 - Proceedings of the 10th International Conference on Agents and Artificial Intelligence*, 2(Icaart), pp. 175–182.

Kim, Y. and Provost, E. M. (2013) 'Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions', in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 3677–3681.

Kishore, K. V. K. and Satish, P. K. (2013) 'Emotion Recognition in Speech using MFCC and Wavelet Features', *Advance Computing Conference (IACC), 2013 IEEE 3rd International*, pp. 842–847.

Koduru, A., Valiveti, H. B. and Budati, A. K. (2020) 'Feature extraction algorithms to improve the speech emotion recognition rate', *International Journal of Speech Technology*. Springer US, 23(1), pp. 45–55.

Koolagudi, S. G. and Rao, K. S. (2012) 'Emotion recognition from speech: A review', *International Journal of Speech Technology*, pp. 99–117.

Lang, P. J. (1995). The emotion probe: Studies of motivation and attention. *American Psychologist*, *50*(5), 372–385.

Langari, S., Marvi, H. and Zahedi, M. (2020) 'Efficient speech emotion recognition using modified feature extraction', *Informatics in Medicine Unlocked*. Elsevier Ltd, 20, p. 100424.

Lee, B. and Cho, K. (2016) 'Brain-inspired speech segmentation for automatic speech recognition using the speech envelope as a temporal reference', *Nature Publishing Group*. Nature Publishing Group, (November), pp. 1–12.

Lee, H. Y., Hu, T. Y., Jing, H., Chang, Y. F., Tsao, Y., Kao, Y. C. and Pao, T. L. (2013) 'Ensemble of machine learning and acoustic segment model techniques for speech emotion and autism spectrum disorders recognition', *Proceedings* *of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, (August), pp. 215–219.

Lin, Y.-L. L. Y.-L. and Wei, G. W. G. (2005) 'Speech emotion recognition based on HMM and SVM', *2005 International Conference on Machine Learning and Cybernetics*, 8(August), pp. 18–21.

Livingstone, S. R. and Russo, F. A. (2018) 'The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north American english', *PLoS ONE*, 13(5).

Ma, D. and Saunders, M. (2018) *SVM using SMO vs PDCO : Support Vector Machines using Sequential Minimal Optimization vs Primal-Dual interior method for Convex Objectives*.

Maka, T. (2020) 'Influence of adaptive thresholding on peaks detection in audio data', *Multimedia Tools and Applications*. Multimedia Tools and Applications, 79(27–28), pp. 19329–19348.

Mansoorizadeh, M. and Charkari, N. M. (2007) 'Speech Emotion Recognition : Comparison of Speech Segmentation Approaches', (January 2007).

Martin, O., Kotsia, I., Macq, B. and Pitas, I. (2006) 'The eNTERFACE'05 Audio-Visual Emotion Database', in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, pp. 8–8.

Matin, R. and Valles, D. (2020) 'A Speech Emotion Recognition Solution-based on Support Vector Machine for Children with Autism Spectrum Disorder to Help Identify Human Emotions', in *2020 Intermountain Engineering, Technology and Computing (IETC)*. IEEE, pp. 1–6.

Mencattini, A., Martinelli, E., Costantini, G., Todisco, M., Basile, B., Bozzali, M. and Di Natale, C. (2014) 'Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure', *Knowledge-Based Systems*. Elsevier B.V., 63, pp. 68–81.

Mena, M. E. (2012) *Emotion Recognition from Speech Signals*. University of Ljubljana.

Mirsamadi, S., Barsoum, E. and Zhang, C. (2017a) 'Automatic speech emotion recognition using recurrent neural networks with local attention', in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2227–2231.

Mirsamadi, S., Barsoum, E. and Zhang, C. (2017b) 'Automatic speech emotion recognition using recurrent neural networks with local attention', in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 2227–2231.

Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America, 93*(2), 1097–1108.

Nema, B. M. and Abdul-Kareem, A. A. (2018) 'Preprocessing signal for Speech Emotion Recognition', *Al-Mustansiriyah Journal of Science*, 28(3), p. 157.

Niu, M., Tao, J., Liu, B. and Fan, C. (2019) 'Automatic depression level detection via ℓp-norm pooling', in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 4559–4563.

Nogueiras, A., Moreno, A., Bonafonte, A. and Mariño, J. B. (2001) 'Speech emotion recognition using hidden Markov models', *EUROSPEECH 2001 - SCANDINAVIA - 7th European Conference on Speech Communication and Technology*, pp. 2679–2682.

Origlia, A., Galatà, V. and Ludusan, B. (2010) 'Automatic classification of emotions via global and local prosodic features on a multilingual emotional database', *Proceedings of the 5th International Conference on Speech Prosody*.

Padmaja, J. N. and Rajeswarrao, R. (2017) 'Analysis And Identification Of Emotion Specific Features For Speaker Independent Emotion Recognition System Using Gaussian Mixture Models (GMMs)', 10(8), pp. 2491–25052.

Palo, H. K. and Mohanty, M. N. (2018) 'Wavelet based feature combination for recognition of emotions', *Ain Shams Engineering Journal*. Ain Shams University, 9(4), pp. 1799–1806.

Pampouchidou, A., Marias, K., Yang, F., Tsiknakis, M., Simantiraki, O., Fazlollahi, A., Pediaditis, M., Manousos, D., Roniotis, A., Giannakakis, G., Meriaudeau, F. and Simos, P. (2016) 'Depression Assessment by Fusing High and Low Level Features from Audio, Video, and Text', *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC '16*, pp. 27–34.

Pan, Y., Shen, P. and Shen, L. (2012) 'Speech emotion recognition using support vector machine', *International Journal of Smart Home*, 6(2), pp. 101–108.

Pfister, T. (2010) 'Emotion Detection from Speech', *Tomas.Pfister.Fi*.

Plannerer, B. (2005) ‘An Introduction to Speech Recognition’, *Munich, Germany*.

Polzehl, T., Schmitt, A., Metze, F., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N. and Kingsbury, B. (2010) ‘Approaching Multi-Lingual Emotion Recognition from Speech-On Language Dependency of Acoustic/Prosodic Features for Anger Recognition’, *Proceedings of International Conference on Speech Prosody*, (November), pp. 3–6.

Pradier, M. F. (2011) *Emotion Recognition from Speech Signals and Perception of Music*. University of Stuttgart.

Provost, E. M. (2013) ‘Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow’, *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. IEEE, pp. 3682–3686.

Rabiner, L. R. and Sambur, M. R. (1975) ‘An Algorithm for Determining the Endpoints of Isolated Utterances’, *Bell System Technical Journal*, 54(2), pp. 297–315.

Rabiner, L. R. and Schafer, R. W. (2007) ‘Introduction to Digital Speech Processing’, *Foundations and Trends® in Signal Processing*. Now Publishers Inc, 1(1–2), pp. 1–194.

Raghib, O., Sharma, E., Ahmad, T. and Alam, F. (2017) ‘Emotion analysis and speech signal processing’, in *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*. IEEE, pp. 2872–2875.

Ram, C. S. and Ponnusamy, R. (2016) ‘Assessment on Speech Emotion Recognition for Autism Spectrum Disorder children using Support Vector Machine’, *World Applied Sciences Journal*, 34(1), pp. 94–102.

Rao, K. S., Koolagudi, S. G. and Vempada, R. R. (2013) ‘Emotion recognition from speech using global and local prosodic features’, *International Journal of Speech Technology*, 16(2), pp. 143–160.

Rao, K. S., Reddy, R., Maity, S. and Koolagudi, S. G. (2010) ‘Characterization of Emotions Using the Dynamics of Prosodic Features’, *Speech Prosody 2010*, p. 100941.

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., Schmitt, M., Alisamir, S., Amiriparian, S., Eva-Maria Messner, Song, S., Liu, S., Zhao, Z., Mallol-Ragolta, A., Ren, Z., Soleymani, M. and Pantic, M. (2019) ‘AVEC

2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition', *AVEC 2019 - Proceedings of the 9th International Audio/Visual Emotion Challenge and Workshop, co-located with MM 2019,* (Avec), pp. 3–12.

Rohith, G. (2018) 'Support Vector Machine — Introduction to Machine Learning Algorithms', *Towards Data Science*, p. 1.

Rosão, C., Ribeiro, R. and De Matos, D. M. (2012) 'Influence of Peak Selection methods on onset detection', *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012,* (Ismir), pp. 517–522.

Safdarkhani, M. K., Mojaver, S. P., Atieghechi, S. and Riahi, M. S. (2012) 'Emotion Recognition of Speech Using ANN and GMM', *Australian Journal of Basic and Applied Sciences*, 6(9), pp. 45–57.

Sahoo, S., Kumar, P., Raman, B. and Roy, P. (2019) 'A Segment Level Approach to Speech Emotion Recognition using Transfer Learning A Segment Level Approach to Speech Emotion Recognition using Transfer Learning', (November).

Sakran, A. E., Abdou, S. M., Hamid, S. E. and Rashwan, M. (2017) 'A Review : Automatic Speech Segmentation', *International Journal of Computer Science and Mobile Computing*, 6(4), pp. 308–315.

Sarikaya, R., Pellom, B. L. and Hansen, J. H. L. (1998) 'Wavelet Packet Transform Features with Application to Speaker Identification', in *Third IEEE Nordic Signal Processing Symposium (NORSIG'98)*, pp. 81–84.

Satt, A., Rozenberg, S. and Hoory, R. (2017) 'Efficient emotion recognition from speech using deep learning on spectrograms', in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*.

Schaefer, A., Nils, F., Philippot, P. and Sanchez, X. (2010) 'Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers', *Cognition and Emotion*, 24(7), pp. 1153–1172.

Schlosberg, H. (1954). Three dimensions of emotion. *Psychological Review, 61*(2), 81–88.

Schuller, B., Arsic, D., Wallhoff, F. and Rigoll, G. (2006) 'Emotion recognition in the noise applying large acoustic feature sets', *Speech Prosody, Dresden*, pp. 276–289.

Schuller, B. and Rigoll, G. (2006) 'Timing levels in segment-based speech emotion recognition', in *INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing, INTERSPEECH 2006 - ICSLP*, pp. 1818–1821.

Schuller, B., Rigoll, G. and Lang, M. (2003) 'Hidden Markov model-based speech emotion recognition', *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, 1, pp. 401–404.

Schuller, B., Rigoll, G. and Lang, M. (2004) 'Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - belief network architecture', *Acoustics, Speech, and Signal Processing*, 1, pp. 577–580.

Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R. and Pantic, M. (2011) 'AVEC 2011 - The first international audio/visual emotion challenge', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6975 LNCS(PART 2), pp. 415–424.

Schuller, B. W. (2018) 'Speech Emotion Recognition two decades in a Nutshell', *Communications of the ACM*, 61(5), pp. 90–99.

Seehapoch, T. and Wongthanavasu, S. (2013) 'Speech Emotion Recognition Using Support Vector Machines', *5th International Conference on Knowledge and Smart Technology (KST)*.

Sezgin, M., Gunsel, B. and Kurt, G. (2012) 'Perceptual Audio Features for Emotion Detection', *EURASIP Journal on Audio, Speech, and Music Processing*. Springer Open Ltd, 2012(1), p. 16.

Shannon, C. E. (1948) 'A Mathematical Theory of Communication', *Bell System Technical Journal*, 27(4), pp. 623–656.

Sharma, M. and Mammone, R. (1996) 'Blind speech segmentation: automatic segmentation of speech without linguistic knowledge', *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, 2(d), pp. 1237–1240 vol.2.

Shen, P., Changjun, Z. and Chen, X. (2011) 'Automatic Speech Emotion Recognition using Support Vector Machine', *International Conference on Electronic and Mechanical Engineering and Information Technology*. Ieee, pp. 621–625.

Shirani, A. and Nilchi, A. R. N. (2016) 'Speech Emotion Recognition based on SVM as Both Feature Selector and Classifier', *International Journal of Image,* *Graphics and Signal Processing*, 8(4), pp. 39–45.

Shoiynbek, A., Kozhakhmet, K., Sultanova, N. and Zhumaliyeva, R. (2019) 'The robust spectral audio features for speech emotion recognition', *Applied Mathematics and Information Sciences*, 13(5), pp. 867–870.

Spasova, L. (2011) 'Paralinguistics As an Expression of Communicative Behaviour', *Trakia Journal of Sciences*, 9, pp. 204–209.

Strayer, D. L. and Johnston, W. A. (2001) 'Driven to Distraction: Dual-Task Studies of Simulated Driving and Conversing on a Cellular Telephone', *Psychological Science*, 12(6), pp. 462–466.

Sultana, S., Shahnaz, C., Fattah, S. a., Ahmmed, I., Zhu, W.-P. and Ahmad, M. O. (2014) 'Speech emotion recognition based on entropy of enhanced wavelet coefficients', *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 137–140.

Tang, J., Alelyani, S. and Liu, H. (2014) 'Feature selection for classification: A review', p. 37.

Tawari, A. and Trivedi, M. M. (2010) 'Speech emotion analysis: Exploring the role of context', *IEEE Transactions on Multimedia*, 12(6), pp. 502–509.

Tzinis, E. and Potamianos, A. (2017) 'Segment-based speech emotion recognition using recurrent neural networks', in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 190–195.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R. and Pantic, M. (2013a) 'AVEC 2013 - The continuous Audio/Visual Emotion and depression recognition challenge', in *AVEC 2013 - Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*. New York, NY, USA: ACM, pp. 3–10.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R. and Pantic, M. (2013b) 'AVEC 2013 - The continuous Audio/Visual Emotion and depression recognition challenge', in *AVEC 2013 - Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, pp. 3–10.

Ververidis, D. and Kotropoulos, C. (2006) 'Emotional speech recognition: Resources, features, and methods', *Speech Communication*, 48(9), pp. 1162–1181.

Vogt, T. and Andre, E. (2005) 'Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition', in *2005 IEEE*

International Conference on Multimedia and Expo. IEEE, pp. 474–477.

Wang, K., An, N., Li, B. N., Zhang, Y. and Li, L. (2015) ‘Speech emotion recognition using Fourier parameters’, *IEEE Transactions on Affective Computing*, 6(1), pp. 69–75.

Wang, Z. (2014) ‘Segment-based Fine-grained Emotion Detection for Chinese Text’, *Proceedings of the Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, (October), pp. 52–60.

Watile, A., Alagdeve, V. and Jain, S. (2017) ‘Emotion Recognition in Speech by MFCC and SVM’, in *International Journal of Science, Engineering and Technology Research (IJSETR)*, pp. 404–407.

Wen, G., Li, H., Huang, J., Li, D. and Xun, E. (2017) ‘Random Deep Belief Networks for Recognizing Emotions from Speech Signals’, *Computational Intelligence and Neuroscience*, 2017.

Williamson, J. R., Godoy, E., Cha, M., Schwarzentruber, A., Khorrami, P., Gwon, Y., Kung, H.-T., Dagli, C. and Quatieri, T. F. (2016) ‘Detecting Depression using Vocal, Facial and Semantic Communication Cues’, *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC ’16*, pp. 11–18.

Witten, I. H., Frank, E. and Geller, J. (2002) ‘Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations’, *SIGMOD Record*, 31(1), pp. 76–77.

Yeh, J. H., Pao, T. L., Lin, C. Y., Tsai, Y. W. and Chen, Y. Te (2011) ‘Segment-based emotion recognition from continuous Mandarin Chinese speech’, *Computers in Human Behavior*. Elsevier Ltd, 27(5), pp. 1545–1552.

Yutai, W., Bo, L., Qingfang, M. and Ping, L. (2009) ‘Emotional feature analysis and recognition in multilingual speech signal’, in *ICEMI 2009 - Proceedings of 9th International Conference on Electronic Measurement and Instruments*, pp. 41046–41050.

Zepf, S., Hernandez, J., Schmitt, A., Minker, W. and Picard, R. W. (2020) ‘Driver Emotion Recognition for Intelligent Vehicles: A Survey’, *ACM Computing Surveys*, 53(3).

Zhang, H., Warisawa, S. and Yamada, I. (2014) ‘An Approach for Emotion Recognition using Purely Segment-Level Acoustic Features’, *Keer2014, International Conference on Kansei Engineering and Emotion Research*.

Zhang, Shiqing, Zhang, Shiliang, Huang, T. and Gao, W. (2018) ‘Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching’, *IEEE Transactions on Multimedia*. IEEE, 20(6), pp. 1576–1590.

Zheng, S., Du, J., Zhou, H., Bai, X., Lee, C. and Li, S. (2021) ‘Speech Emotion Recognition Based on Acoustic Segment Model’, in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, pp. 1–5.

Zhou, Y., Li, J., Sun, Y., Zhang, J., Yan, Y. and Akagi, M. (2010) ‘A hybrid speech emotion recognition system based on spectral and prosodic features’, *IEICE Transactions on Information and Systems*, E93-D(10), pp. 2813–2821.

Zhou, Y., Sun, Y., Zhang, J. and Yan, Y. (2009) ‘Speech Emotion Recognition Using Both Spectral and Prosodic Features’, *2009 International Conference on Information Engineering and Computer Science*. Ieee, pp. 1–4.

LIST OF PUBLICATIONS

Indexed Journal

1. **Zaidan, N. A.** and Salam, M. S. H. (2015) 'A Review on Speech Emotion Features', *Jurnal Teknologi*, 75(2), pp. 1–6. **(Q3)**
2. **Zaidan, N. A.** and Salam, M. S. H. (2019) 'Emotional Speech Feature Selection Using End-Part Segmented Energy Feature', *Indonesian Journal of Electrical Engineering and Computer Science*, 15(3), p. 1374. **(Q3)**

Indexed Conference Proceedings

1. **Zaidan, N. A.** and Salam, M. S. (2016) 'MFCC Global Features Selection in Improving Speech Emotion Recognition Rate', in *Advances in Machine Learning and Signal Processing.* Springer, Cham, pp. 141–153. **(Q4)**