A STUDY ON STRESS AND EMOTIONS SPEECH RECOGNITION AND MODELING

by

Barlian Henryranu Prasetio

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

Course of Computer Science and Bio-informatics
Department of Materials and Informatics
Interdisciplinary Graduate School of Agriculture and Engineering
UNIVERSITY OF MIYAZAKI

September 2020

Declaration of Authorship

I, Barlian Henryranu Prasetio , declare that this thesis titled, 'A Study on Stress and Emotion Speech Recognition and Modeling' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: September 2020

Signed: Barlian Henryranu Prasetio

Advisor: Prof. Hiroki Tamura

PhD

abbrevation[pee-eych-dee]

An academic who has learned more and more to know about something but eventually they know nothing at all.

Miyazaki, September 2020

UNIVERSITY OF MIYAZAKI

Abstract

A Study on Stress and Emotions Speech Recognition and Modeling

by Barlian Henryranu Prasetio

In social life, the ability to recognize and make sense of the emotions, known as emotional awareness, makes us further understand what others telling and realize how our emotion affects others. In addition, emotional awareness makes people care about their emotional health, which also includes being able to solve problems by understanding emotions. Thus, it means that emotional awareness is not just for making sense of other's emotions but it is also to manage our emotions for a healthy life. Being emotionally healthy does not mean we are happy all the time but we have cared about our emotions. We can deal with emotions, whether they are positive or negative. Emotionally healthy, people still feel stress, anger, and sadness but they know how to manage them.

Stress and emotions trigger a particular biological response that causes hormones to surge throughout the body and it makes increasing the activity of the human physiological system. Thus, there are some physiological parameters that could be used to represent the emotional condition, such as brain activity, heart rate variability (HRV), Electrodermal Activity, respiratory system, pupil dilation, the facial, and the tone of voice (speech). Lately, speech analysis is the most interesting measurement method to recognize the stress and emotional condition due to its comfortableness and economic cost.

In this decade, stress and emotion recognition systems using speech analysis has been extremely studied. Most of them used a standard approach where feature extraction and classifier are the main components in recognizing the patterns. The effectiveness of feature representation is a crucial modal to make the system efficient. However, we should know that stress has diverse characteristics and different patterns for each individual. Along with these limitations, to make the system more robust and able to adapt in the real condition, more huge data training is required. Unfortunately, stress and emotion data are hard to be collected massively. To this end, some studies used the clustering approach to categorize stress and emotion speech data based on the similarity of their characteristics. Due to its effectiveness, the clustering approach becomes a popular method and widely used in many emotional-based applications. However, in some cases, emotion (e.g., stress) may change when triggered by an event during the speaking. Thus, the exploration with larger sets of contextual information becomes an important consideration to recognize the stress and emotion accurately. In this thesis,

we present a big framework of stress and emotion recognition and modeling in order to contribute to emotional awareness.

This thesis is organized as follows. Chapter 1 provides a research background, aims, scopes, contributions, and findings. The definitions of stress and emotion are discussed in Chapter 2. This chapter also describes the measurement method that could be used for identifying stress and emotions. Chapter 3 depicts the proposed system for stress and emotions speech recognition and modeling. The stress and emotions recognition system in binary and multi-class classification are presented in Chapter 4. The pre-processing phase and the clustering approach provided in Chapters 5 and 6, respectively. The clustering method is approached in an unsupervised and semi-supervised. Chapter 7 presents the stress and emotions speech prediction and modeling. We summarize the results of the thesis and discuss some problems connected with the findings in Chapter 8.

In this thesis, we introduce and evaluate three approaches of stress and emotion recognition using speech. The first approach is to develop and evaluate the stress and emotions speech recognition (SSR) system. In this approach, we explore the effectiveness of SSR in the classification tasks. The second approach is to develop and evaluate the stress and emotions speech clustering (SSC) system. The unsupervised and semi-supervised clustering methods are introduced. Moreover, we also discuss the pre-processing steps for this system. The third approach is to develop and evaluate the stress and emotions speech prediction and modeling (SSM) system. SSM analyzes the speech features and the prior emotional state for predicting the present emotional state and model their state transition.

This thesis contains several methods that are used in recognizing stress and emotion, such as Embedded Discriminant Analysis (EDA) for speech activity detection, Deep Time-delay Embedded Algorithm Clustering (DTEC) and Semi-Supervised Deep Time-Delay Embedded Clustering (SDTEC) for stress speech clustering, and Deep Time-delay Markov Network (DTMN) for prediction and modeling the stress and emotions.

The major finding and conclusion in this thesis is the emotional transition model. In general, males and females present a similar model of emotional transition. However, there are some fundamental differences between male and female emotional transition tendencies. Females tend to be more easily change their emotions, but they have a tendency longer in stress than males. After a stressful period, females tend to become sad, while males are easier to grow angry.

Acknowledgements

First and foremost, I would like to thank Allah SWT Almighty for giving me the age of blessing, happiness, and health. Praise the presence of Allah SWT that gives me the strength and opportunity to undertake this study and to persevere and complete it satisfactorily. Without his blessings, this achievement would not have been possible.

Blessings and greetings may always be devoted to our lord the Great Prophet Muhammad SAW, who brings and illuminates our conscience, becomes the light for all noble deeds.

I would like to express my deep and sincere gratitude to my supervisor, **Professor Hiroki Tamura**. His knowledge, support, enthusiasm, and academic guidance have been of great value to me.

I would also like to thank **Professor Koichi Tanno**. His kind advisement is most helpful during my studies.

Special thanks are due to my parents **Wiwik Sudarto Darsono**, **Siti Indahsyah**, **Warno Edy Supomo**, and **Haryati Usbandiyah**. They worked very hard to support the family and to raise me. Thanks for their unconditional love and support throughout my life.

Special thanks also addressed to my wife **Edita Rosana Widasari** and my daughter **Khayyira Quinza Prasetio**. Without them encouragement and understanding, it would have been impossible for me to finish my doctoral study.

Thank you to LDC that give me the SUSAS database scholarship that provides a database of emotional stress speech.

Thank you to the **University of Miyazaki (UoM)** to kind wisdom in giving tuition exemption during my study.

Thank you to the **University of Brawijaya (UB)** to kind support for me during my study.

My sincere thanks also go to **Dr. Lindsey R. Tate** and **Dr. Keiko Sakurai** for their much-appreciated advice and also to my Laboratory members of University of Miyazaki: **Takeshi Tomomizu**, and **Praveen Nuwantha Gunaratne**.

Thanks to my friends, **Sabriansyah R. Akbar**, **M. Chandra Saputra**, and **Bayu Priyambadha**, for all the help that has been given to my family.

Thanks also go to all member of Indonesian community (**PPI-Miyazaki**) for our togetherness during in Miyazaki.

List of Publications

This thesis titled, 'A Study on Stress and Emotion Speech Recognition and Modeling' is a unity of several publications, as follows:

Journal articles

1. Prasetio, B.H., Tamura, H., Tanno, K. Deep Time-delay Markov Network for Prediction and Modeling the Stress and Emotions State Transition. *Scientific Reports*. **xx**, xxxx (2020).

2. Prasetio, B.H., Tamura, H., Tanno, K. The Long Short Term Memory Based on I-Vector Extraction for Conversational Speech Gender Identification Approach. *Artificial Life and Robotics*. **25(2)**, 233–240 (2020).

3. Prasetio, B.H., Tamura, H., Tanno, K. Semi-Supervised Deep Time-Delay Embedded Clustering for Stress Speech Analysis. *Electronics*. **8(11)**, 1–13 (2019).

4. Prasetio, B.H., Tamura, H., Tanno, K. Generalized Discriminant Methods for Improved X-Vector Back-end Based Speech Stress Recognition. *IEEJ Transactions on Electronics, Information and Systems*. **139(11)**, 1341–1347 (2019).

Conference proceedings

1. Prasetio, B.H., Tamura, H., Tanno, K. Embedded Discriminant Analysis based Speech Activity Detection for Unsupervised Stress Speech Clustering. In International Conference on Imaging, Vision & Pattern Recognition (IVPR), Kitakyushu, Japan (2020).

2. Prasetio, B.H., Tamura, H., Tanno, K. A Deep Time-delay Embedded Algorithm for Unsupervised Stress Speech Clustering. In Proceeding of IEEE International Conference on Systems, Man, and Cybernetics (SMC), Bari, Italy, pp. 1193–1198 (2019).

3. Prasetio, B.H., Tamura, H., Tanno, K. A Study on Speaker Identification Approach by Feature Matching Algorithm using Pitch and Mel Frequency Cepstral Coefficients. In the International Conference on Artificial Life and Robotics (ICAROB), Beppu, Japan (2019).

4. Prasetio, B.H., Tamura, H., Tanno, K. Ensemble Support Vector Machine and Neural Network Method for Speech Stress Recognition. In International Workshop on Big Data and Information Security (IWBIS), Jakarta, Indonesia, pp. 57–62 (2018).

Contents

List of Figures

List of Tables

I dedicate this thesis to my homeland "Indonesia"

Kulihat ibu pertiwi
Sedang bersusah hati
Air matamu berlinang
Mas intanmu terkenang

Hutan gunung sawah lautan
Simpanan kekayaan
Kini ibu sedang susah
Merintih dan berdoa
Kulihat ibu pertiwi

Kami datang berbakti
Lihatlah putra-putrimu
Menggembirakan ibu

Ibu kami tetap cinta
Putramu yang setia
Menjaga harta pusaka
Untuk nusa dan bangsa

Chapter 1

Introduction

This Chapter presents the research background, aims, and scopes. The findings and contributions are also provided in this Chapter. An outline structure of this thesis report is also given.

1.1 Backgrounds

In simple, emotions are a class of feelings. Emotions are the psychological state that responses to significant internal and external events. Even though it is a normal reaction, emotional health is a fundamental factor in overall health. People who are emotionally healthy able to control their thoughts, emotions, and behaviors [1]. They are able to cope with life's challenges, keep problems in perspective, and bounce back from setbacks. They feel good about themselves and have good relationships. Being emotionally healthy does not mean we are happy all the time but we have cared about our emotions. We can deal with emotions, whether they are positive or negative. Emotionally healthy people still feel stress, anger, and sadness but they know how to manage them [2].

In social life, the ability to recognize and make sense of the emotions, known as emotional awareness, makes us further understand what others telling and realize how our emotion affects others [3]. In addition, emotional awareness makes people care about their emotional health [1], which also includes being able to solve problems by understanding emotions [4]. Thus, it means that emotional awareness is not just for making sense of other's emotions for social relationships but also to manage our emotions for a better life.

Modern life is full of emotional challenges. The pressure to succeed, following fear of everybody. Satisfaction in works can evoke volatile combinations of emotions. Emotions have energy that pushes up for expression. Minds and bodies respond by constricting the muscular and holding breath. Symptoms like anxiety, stress, and depression are on the rise in Japan. It might be caused by the way to deal with emotions, which are biological forces that should not be ignored. When the mind fights with the flow of

emotions, it puts stress on the mind and the body, creating psychological distress and symptoms. Some studies reported that there is a link between emotions that are closely associated with each other [5]. For instance, depression is a form of stress response [6, 7]. Stress can positively predict anxiety symptoms [8]. Certain negative emotions usually arise from stress [9]. Stress one of emotion that a normal reaction due to changes in environmental conditions [10, 11] or stimuli [12]. This situation triggers a particular biological response that causes hormones to surge throughout the body [13]. Therefore, there is increasing activity in the human physiological system [14].

Emotional stress has not only been linked to mental ills, but also to physical problems like heart disease, intestinal problems, headaches, insomnia, and autoimmune disorders. Most people feel emotions but unconscious that this is happening. Whereas, understanding a few about emotion can help greatly. Basic biology and anatomy explain that we cannot stop our emotions from being triggered, as they originate from the middle section of our brain that is not under conscious control. Therefore, emotional awareness is a crucial thing for a healthy life. By emotional awareness, we could manage emotions in healthy and not destructive to ourselves or others. Therefore, a notification system to recognize emotions is needed as an awareness form or effort for our health.

In order to recognize stress and emotions, there are some physiological parameters that reflect the increased efficiency of the body due to the emotional condition, such as brain activity [15]. This measurement is known as a non-invasive method because it placed the electrodes along the scalp to record the electrical activity of the brain. Stress also can be measured by another non-invasive method, such as heart rate variability (HRV) [16], Electrodermal activity [17], and respiratory system [18]. The free-contact or non-intrusive methods such as pupil dilation [19] and facial expression [20–23] could be used to measure the stress level. The tone of voice (speech) analysis [24–28] is the most interesting measurement method due to its comfortableness and economic cost.

In this decade, stress and emotion recognition systems using speech analysis has been extremely studied. Most of them used a standard approach where feature extraction and classifier are the main components in recognizing the patterns. The effectiveness of feature representation is a crucial modal to make the system efficient. An amount of stress and emotion database (e.g., Speech Under Simulated and Actual Stress (SUSAS) [29, 30], Belfast [31], France et.al [32], and Fischer et.al [33] has been provided. However, we should know that stress has diverse characteristics and different patterns for each individual. It is caused by various aspects such as characteristics, gender, experience background, and emotional tendencies [34]. Along with these limitations, to make the system more robust and able to adapt in the real condition, more huge data training is required. Unfortunately, stress and emotion data are hard to be collected massively.

To this end, some studies used the clustering approach to categorize stress and emotion

speech data based on the similarity of their characteristics [35–38]. An unsupervised algorithm defines its effective objective in a self-learning manner by computing the distance between data points in feature space [37, 39, 40]. In the past year, some researchers offered another approach to solve the problem of the curse of dimensionality by presenting a compact feature representation in the clustering assignment, known as deep clustering. Due to their effectiveness, deep clustering becomes a popular clustering method and widely used in many practical applications. However, in some cases, emotion (e.g., stress) may change when triggered by an event during speaking [41]. Thus, the exploration with larger sets of contextual information becomes an important consideration to recognize the stress and emotion accurately.
In this thesis, we propose an end-to-end stress and emotion recognition system using speech. The proposed system analyzes stress and emotions in terms of recognition and its state transition modeling approach, consist of three main parts, stress and emotions speech classification, clustering, and prediction and modeling.

1.2 Research Objectives

Conducting emotional awareness can be described in two steps. The first step is emotion recognition. It means that the system is able to recognize an time-series emotion class by its characteristics. In order to manage emotions, the second step is to model the state transition of emotions and recognize its patterns. To reach this objective, the proposed system is addressed as follows:

- develop the stress and emotions speech recognition (SSR) system and evaluate its effectiveness in the classification tasks.
- develop and evaluate the stress and emotions speech clustering (SSC) system. We use two approaches i.e. unsupervised and semi-supervised clustering. Since the clustering task uses unlabeled data with high noise, unknown speaker, and gender, we also discuss the pre-processing steps for SSC.
- develop and evaluate the stress and emotions speech prediction and modeling (SSM) system. We predict the present emotional state by analyzing the speech features and the prior emotional states.
- develop the emotional states transition model to recognize its patterns. Since males and females express emotion in different ways, we model the emotional states transition in different diagrams.

1.3 Research Scope

The thesis focused on natural speech i.e. speech with naturally expressed (not acted) stress and emotions. We used Speech under Stress and Actual Stress (SUSAS) database containing natural speech. Labeled and unlabeled speech data from SUSAS were used in all of the experiments. The stress and emotions speech classification, clustering, and prediction were classified into 5 different levels of stress i.e. high stress, low stress, neutral, soft, and angry. The number of data and their attribute is explained in more detail in Chapter 3.1.

1.4 Contributions and Findings

This thesis proposed and evaluated several new methods for stress and emotion identification. The effectiveness of the proposed methods were compared with existing classical approaches. The proposed methods include the following.

- Embedded Discriminant Analysis (EDA) for Speech Activity Detection
- Deep Time-delay Embedded Algorithm Clustering (DTEC) for Unsupervised Stress Speech analysis
- Semi-Supervised Deep Time-Delay Embedded Clustering (SDTEC) for Stress Speech Analysis
- Deep Time-delay Markov Network (DTMN) for Prediction and modeling the stress and emotions state transition

Moreover, this research led to the following major findings and conclusions:

- Generally, the differences between neutral and stress are the power will decrease while the frequency will increase.
- We hypothesize that the female tended to express their stress in soft (e.g., sadness) while the male tended to express their stress as anger.
- Generally, males and females generally present a similar emotional transition representation. However, there are some fundamental differences between males and females. Females have a tendency longer in stress than males but more easily change for other emotions. After a stressful period, females tend to become sad, while males are easier to grow angry.

1.5 Thesis Outline

We organized this thesis as follows:

Chapter 1 provides research background, aims and scopes. We also present the research's contributions and findings.

Chapter 2 introduces definitions of stress and emotion. It describes the expression of stress and emotions. The measurement method for identifying stress and emotions. This Chapter also provides stress and emotion speech-based applications.

Chapter 3 describes the proposed system for stress and emotions speech recognition and modeling. We also describe the speech database and the methods used in the pre-processing, classification, clustering, and prediction stages of this framework.

Chapter 4 presents the stress and emotions recognition system in two feature extraction techniques.

Chapter 5 explain in detail the pre-processing phase for stress and emotion speech clustering task. The pre-processing phase consists of speech activity detection, speaker identification, and gender identification.

Chapter 6 presents the stress and emotions speech clustering system in unsupervised and semi-supervised approach. In this chapter, we propose two new deep clustering methods.

Chapter 7 presents the stress and emotions speech prediction and modeling. In this chapter, we propose a new prediction and modeling method.

Chapter 8 summarizes the results of this thesis and discusses some of the problems connected with the findings.

Chapter 2

Stress and Emotions in Speech

In this chapter, we discuss stress and emotion definition, emotional effect in human life, and measurement methods of stress and emotion. We also present a general architecture of stress and emotion recognition systems. The applications that use stress and emotion speech recognition are also discussed.

2.1 What is stress and emotions

In neuroscience, emotions are biological states associated with the nervous system [42] brought on by neurophysiological changes variously associated with thoughts, feelings, behavioral responses, and a degree of pleasant or unpleasant [43, 44]. Emotion is often intertwined with mood, temperament, personality, disposition, creativity, and motivation [45].

Psychologists have used methods such as factor analysis to attempt to map emotion-related responses onto a more limited number of dimensions. Such methods attempt to boil emotions down to underlying dimensions that capture the similarities and differences between experiences [46]. For instance, Plutchik wheel emotion [47] and Ekman basic emotion [48–51]. These models uncovered by factor analysis are valence (how negative or positive the experience feels) and arousal (how energized or enervated the experience feels). Furthermore, Russell depicts on a 2D coordinate map [52]. This two-dimensional map has been theorized to capture one important component of emotion called core affect [53, 54]. Core affect is not theorized to be the only component to emotion, but to give the emotion its hedonic and felt energy. Russel's dimensional model, represented in Figure 2.1, is the most used with the dimensions valence and arousal.

Plutchik [47] found that complex emotions could arise from cultural conditioning or association combined with the basic emotions. For instance, certain emotions like anger, shame, and anxiety usually triggered by another emotion such as stress [9]. Based on Lazarus and Folkman's theory [55], stress and emotion depend on how an individual evaluates (appraises) transactions with the environments. During the appraisal, when

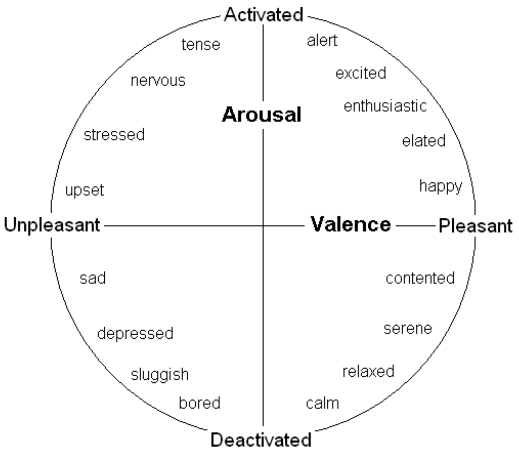


FIGURE 2.1: Russel's dimensional model of emotions

people find something significant to self is uncontrollable, they tend to feel high levels of stress. In a stressed condition, ones easy to misunderstand the intentions or what they would like to communicate and express an abnormal emotion as a reaction.
Stress is a non-specific response of the body in any claim [56]. Stress can also be defined as a condition that suppresses a person's psyche because of constraints or obstacles [57]. Naturally, the human body will exhibit physical, mental, and emotional responses to a potentially dangerous condition in an attempt to prevent injury.

2.2 Stress in human life

Stress can affect all aspects of our life, including trigger other emotions, social communication, behaviors, and mental health [58]. Everyone handles stress in different ways so that the symptoms of stress are also varied. The symptoms of stress can be vague and maybe has the same with other medical conditions. Therefore, it is important to have emotional awareness so that one can manage their behavior, healthy social relationships, personal emotional health, and able to make the best decisions for their life.

2.2.1 Stress and social communication

Communication is the act of sharing information between individuals. During speaking, humans share verbal information with other [59]. Verbal information is the use of words to transfer linguistic messages [60]. It includes sound and language in speaking. Human speech conveys not only linguistic messages but also non-verbal information. Non-verbal information is the additional messages that are delivered as clues to complete the meaning over spoken, such as emotional expression. The emotional content of speech can be perceived even when the message of the utterance is emotionally ambiguous, and even when the listener does not know the language. Emotion affects paralanguage elements

in a speech, which is how to speak like a tone of voice, loud or weak, speech speed, sound quality, intonation, and others [61].

2.2.2 Stress and behavior

Obviously, speech information can express emotions to reinforce the intent of what we have said [62]. In non-verbal communication, emotions can be grouped into two main categories: conscious and unconscious [63]. Expression of conscious emotion is easier to be recognizable, such as anger, sadness, and happiness. While unconscious emotions are very difficult to recognize, such as stress and depression. Conscious and unconscious emotions define the direction and intensity of the mental activity of the individual, determining the degree and direction of sensitivity, which is reflected in the external manifestations of behavior, due to the specifics of perception and reflection of the emotional impacts.
Unconscious emotion, such as stress, shows more specific symptoms in behavior [64]. Table 2.1 shows the common emotional stress effect on behavior [11].

2.2.3 Stress and health

In medical science, a stress response is considered as the increasing activity in the sympathetic branch of the autonomic system and the activation of the hypothalamic-pituitary-adrenal axis [12]. The main purpose of physiological changes under stress is increasing the efficiency of delivering energy (oxygen and glucose) to vital organs [65].
The human body responds to stress by releasing hormones that increase heart rates, breathing rates, and muscle tension [66] around the neck and the face [67]. These responses make the articulatory movements, airflow from the respiratory system, and timing of the vocal system physiology change [27, 68], such that speech characteristics [69] and facial expression [70] are also changed.
Thus, stress affects massive human physiological systems. Stress that's left unchecked can contribute to many health problems, such as high blood pressure, heart disease, obesity, and diabetes. Table 2.2 shows the common effect of stress on body and mental health.

2.3 Physiological parameters and stress relationship

The Sympathetic Nervous System (SNS) provokes the stress response in humans, carrying psychological, physiological and behavioral symptoms. From a physiological point of view, the increase of SNS activity changes the hormonal levels of the body and provokes reactions like sweat production, increased heart rate, and muscle activation. Respiration becomes faster and the blood pressure increases. As a consequence of changes in

TABLE 2.1: Common stress effects on behavior

On habits	On performance	On personality
Teeth grinding	Positive Effects:	Time urgency / Rushed Lifestyle
Hair pulling	Performance levels increase when stress management is effective. Stressors such as pressure and demands can facilitate better stress response and thus, higher levels of performance.	Aggressive, hostile, easily angered
Tremors or nervous tics		Hard-driving, unable to relax, cynical
Eating pattern		Polyphagia (multitasking)/2-things at one
Clumsiness	Negative Effects:	Rapid speech patterns
Alcoholism		Predictor of heart disease
Social withdrawal	When stress is perceived as uncontrollable or unmanageable, the person begins to experience a gradual to drastic decrease in performance levels, causing a decline in productivity and enthusiasm to respond to the stress.	Hopeless personality: Poor self-motivation, feel helpless, hopeless, give up
Impulse buying		Irrational-Illogical Personality: evaluators, do not perceive situations accurately, unrealistic expectations

TABLE 2.2: Common effects of stress

On the body	On mental
Headache	Anxiety
Muscle tension or pain	Restlessness
Chest pain	Lack of motivation or focus
Fatigue	Feeling overwhelmed
Change in sex drive	Irritability or anger
Stomach upset	Sadness or depression
Sleep problems	

the muscles which control the respiratory system and vocal tract, speech characteristics change too. Skin temperature decreases together with hands and feet temperature and

the Heart Rate Variability (HRV) decreases. Moreover, pupil diameter can vary. Articulatory movements, airflow from the respiratory system, and timing of the vocal system physiology are change. Table 2.3 shows a comparison of physiological parameters and their relation with stress [71].

TABLE 2.3: Physiological parameters and their relation to stress

Physiological parameter	Relation to stress
Blood pressure (BP)	Stress increases blood pressure depending on the experienced stress levels [72].
Electroencephalogram (EEG)	Alpha activity decreases in stress situations and Beta activity increases with mental workload (stress) [73].
Electrocardiogram (ECG)	Worldwide scientific research has shown that heart rate (HR) increases during stressful times [74].
Electrodermal activity (EA)	The amplitude of the electrodermal response was significantly correlated with subjective stress experience [75].
Respiration rate	when the stress level changes, the speed and depth of respiration system also change [76]
Pupil dilation (PD)	pupils are dilated more often under stress situations [77].
Electromyogram (EMG)	Stress provokes involuntary reactions on facial and Trapezius muscles [78].
Speech	Stress changes human vocal production [79].

As shown in Table 2.3, stress increases blood pressure but it is not an as good indicator to detect stress situations [71]. Alpha waves of EEG reflect a calm, open, and balanced psychological state, so Alpha activity decreases in stress situations [73]. Besides, the Beta activity of EEG reflects cognitive and emotional processes [80] so it increases with the mental workload and then with stress. HR is defined as the number of heartbeats per minute that frequently is used to analyze stress by computing the mean and standard deviation [81]. EA linearly related to arousal [82] and it has been widely used in stress and emotion detection [83]. EA and heart rate variability (HRV) were found to be the best correlates of real-time stress but less superior than EEG features in discriminating under cognitive load and relaxed states [84].

The possibility of estimating respiration rates is by an ECG signal [71] but the contribution of respiration signals to stress detection was far from being as evident as EA or HRV's contribution [83]. The high ability of PD features to discriminate between stress and relaxed situations was affirmed by Ren et al [85]. The LF/HF ratio of PD variability could effectively replace the LF/HF ratio of HRV in stress recognition [86]. EMG can sometimes be less selective than desirable because the electrical activity created by a muscle can be extended to the adjacent areas, and moreover, activities that are not related to emotions, such as speaking, can generate confusing EMG activity [71]. Under stress situations, stress changes human vocal production in pitch (fundamental

frequency) and in the speaking rate are usual, together with variations in features related to the energy and spectral characteristics of the glottal pulse [79]. The performance comparison of each physiological parameter has been summarized by [71], as shown in Table 2.4.

TABLE 2.4: The physiological parameters performance comparison

Reference	Target class	Number of classes	Signal	Accuracy (%)
Palanisamy et al.	Stress	2	HRV	93.75
			EA	70.83
			EMG	71.25
Wei et al.	Stress	2	EMG	97.8
			Resp	86.7
Ren et al.	Stress	2	PD	88.71
Dinges et al.	Stress	2	Facial	75-88
Demenko et al	Stress	2	Speech	84
Kurniawan et al.	Stress	2	Speech	92.6
			EA	80.72
Sharma et al.	Stress	2	EEG	98
Li et al.	Stress	5	ECG	96.4

Table 2.4 shows the non-invasive methods present better accuracy than non-intrusive methods. EEG is the best physiological parameter that correlates to stress. EA and HRV were found to be the best correlates of real-time stress but less superior than EEG features. However, as most of the physiological measurements, EEG, EA and HR are the contact method and it may makes uncomfortable. Contrarily, speech analysis has an interesting way because it can be easily measured in a completely contactless method. Furthermore, speech features are more efficient for stress detection than the selected EA features [74].

2.4 Architecture of stress and emotion recognition system

Stress and emotion recognition in speech is used the well-established architecture of pattern recognition [87] that was successfully applied in numerous speech and speaker recognition. Figure 2.2 illustrates a general architecture of stress and emotion recognition in speech.

The training process is an iterative procedure that usually has supervision. The speech samples with known class labels (known emotions) are first pre-processed to reduce the noise and remove the silence or unvoiced intervals. The labeled and pre-processed speech is then used to calculate sets of acoustic feature parameters characterizing each emotional class. In some cases, the features may undergo a process of data reduction (or redundancy removal) through an optimal selection of features. During the modeling stage, the characteristic features are used to derive class models in a form of estimated

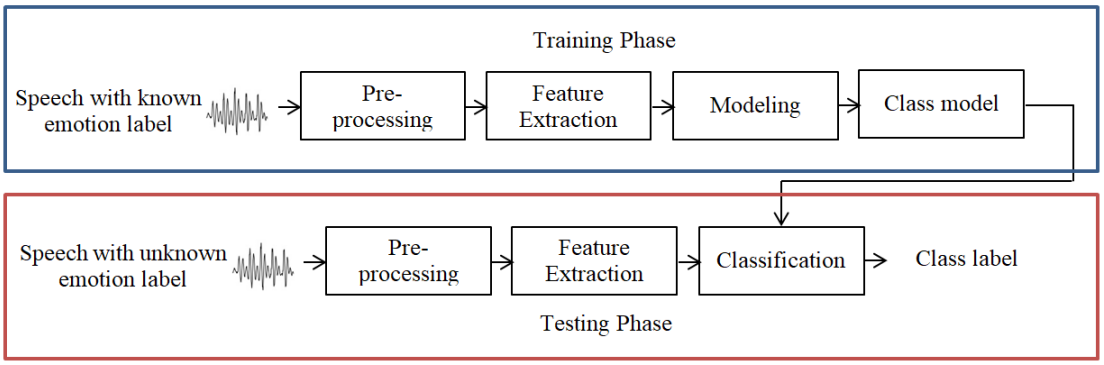


FIGURE 2.2: The general architecture of the stress and emotion speech recognition and classification system. The blue and red box denote training and testing (classification) phase, respectively.

parameters, such as probability density functions, statistically describing each class, or in a form of neural network structures with nodes represented by sets of constant weights derived from the training data.

In the testing phase, speech samples from unknown classes are subjected to pre-processing, feature extraction, and feature selection procedures, usually identical to those used during the training process. A classification method is then used to perform the pattern matching and decision-making process and produce the most probable emotional label for the examined speech sample. The classification is usually based on a pattern-matching approach that determines class that produces the highest probability values or activates certain nodes in a neural network structure.

2.4.1 Existing emotional speech database

The recognition accuracy of stress and emotion depends on the types of speech samples used in the process of statistical modeling of different classes of stress or emotion. The first type used emotions simulated by professional actors in a recording laboratory allowing experimental control but having low ecological validity. The second type of data represented natural vocal expressions recorded in the field or from reality media broadcasts. It provided high ecological validity, but it was difficult to determine the actual emotion felt by the speaker. The third type used experimentally induced emotional expressions in the laboratory. This approach provided a low level of control over emotional arousal and valence.

Apart from ecological validity, other essential factors characterizing the quality of emotional speech data are the size of the speech database (number of speech samples). A small dataset of data may cause the trained model low-robust to the natural environment. In contrast, a large set of ecological data is hard to be collected. Table 2.5 contains a list of several widely used databases of natural stress and emotion speech, described in terms of numbers of subjects, types of emotions, language [88].

TABLE 2.5: Natural stress and emotion speech database.

Database	Number of subjects	Type of emotions	Language
Belfast [31]	125 (31 male and 94 female)	Wide range (active positive emotion, active negative emotion, passive positive emotion, passive negative emotion)	English
SUSAS [29, 30]	32 (13 male and 19 female)	Wide range (neutral, high stress, low stress, angry, soft, etc)	English
France [32]	115 (67 male and 48 female)	Depression, neutrality, suicidal state	English
Fischer [33]	56 (unknown gender)	Anger, depression, neutrality	German

2.4.2 Existing techniques of feature extraction in emotional speech analysis

The concept of automatic emotion recognition was introduced when the use of statistical properties of speech in automatic emotion recognition. In recognizing emotion, the feature extraction technique is a vital part. There are some feature extraction techniques that are widely used for speech emotion recognition.

2.4.2.1 Prosodic features

The human being are able to more or less independently control phonation (source) with the larynx and articulation (filter) with the vocal track. Thus, we could assume speech sound are the response coming from a vocal-track system, where a sound source is fed into and filtered by the resonance characteristic of the vocal track. This kind of modeling by a linear system is called the source-filter theory of speech production.
The source-filter model has a linear character, where it conveying accurate linguistic content. The majority of the current approaches to emotional speech analysis rely on the assumption that the emotional state of a speaker affects in some way speech parameters assumed by the source-filter model. These parameters including the fundamental frequency, formats, and energy, or parameters derived from them, are the most often cited in the literature as characteristic features used in emotion recognition from speech. During phonation, the vocal folds vibrate. The number of cycles per second determines the frequency of the vibration, which is subjectively perceived as pitch or objectively measured as the fundamental frequency F_0. The sound is then modulated by the vocal tract configuration and the resonant frequencies of the vocal tract are known as formants. Figure 2.3 illustrate how to obtain the estimation of F_0.
The fundamental frequency F_0 of the vibration of the vocal folds is estimated simultaneously using the autocorrelation method in the time domain and the cepstral method

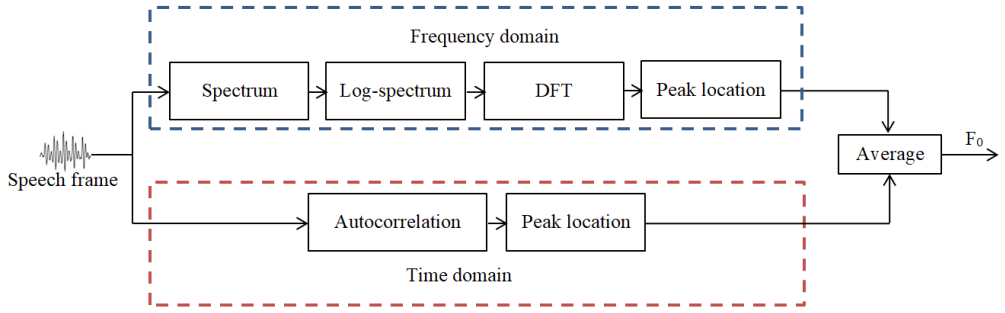


FIGURE 2.3: A flowchart of the fundamental frequency estimation method

in the frequency domain [89]. The average value of these two measurements provided the final estimate of F_0.

Then, the first three formant frequencies, F_1, F_2, and F_3, where estimated as the resonant frequencies of the vocal tract filter using the linear predictive coding (LPC) analysis [90]. The formants values are obtained by factoring the predictor polynomial and solving the roots of the polynomial to find the locations of the resonances representing the values of formants.

2.4.2.2 Mel Frequency Cepstral Coefficients (MFCC)

MFCC is a robust technique to capture the cepstral representations of speech information. In order to maintain the signal periodicity, a short-term analysis of the speech signal was performed to compute the MFCC feature [91]. The MFCC flow diagram is presented in Figure 2.4.

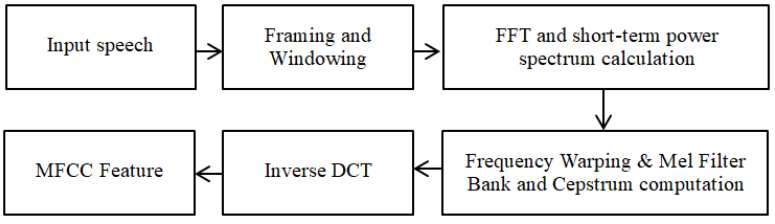


FIGURE 2.4: The MFCC flow diagram

The MFCC coefficients were calculated on each frame (framing). Then, the Hamming window algorithm was applied to each frame over the speech sample (windowing) to remove the discontinuities in the speech signal. We obtained the Mel-frequency-wrapped spectrum from the Fast Fourier Transform (FFT). Mel energy is computed by multiplying each filter bank with the power spectrum then adding the product to the coefficients. The Mel filter bank consists of triangle bandpass filters that span overlapping frequency bands. The width of each filter is set according to the Mel-frequency wrapping band, as exemplified in Table 2.6 [92]. The Mel-frequency scale expresses the important phonetic characteristics of the speech signal. A measured frequency f in Hz can be converted to

the Mel scale $mel(f)$ using Eq. 2.1. The Inverse Discrete Cosine Transform (IDCT) was applied to obtain the d-dimensional order cepstral coefficients.

$$mel(f) = 2595 * log10(1 + f/700) \tag{2.1}$$

TABLE 2.6: The frequency band of the Mel filter

Filter	Passband edges (Hz)
Filter 1	[133 267]
Filter 2	[200 333]
Filter 3	[267 400]
...	...
Filter 10	[733 867]
...	...
Filter 20	[1510 1733]
...	...
Filter 40	[5973 6854]

2.4.2.3 Teager Energy Operator (TEO) based features

The presence of additional harmonics other than the F_0-series can indicate the stressful or emotional state of a speaker, and provide a source of characteristic features for the detection and classification of speech under stressful or emotional situations. This speech modeling initiated by Teager [93–95]. Teager noticed that energy plays an important function in hearing and recognition of speech.

The speech signal could be regarded as an effect of amplitude and frequency modulation of separate oscillatory waves and modeled as a combination of several amplitude and frequency-modulated (AM-FM) oscillatory components. A nonlinear model of speech, which represents a discrete-time speech signal $s[n]$ as a sum of M components ($s[n] = \sum_{i=1}^{M} x_i[n]$). By assuming each component of speech can be modeled as an AM-FM sine wave in the discrete-time domain, Kaiser [96] estimate of the speech instantaneous energy, known as the Teager energy operator (TEO) expressed as:

$$\Psi(x[n]) = (a[n])^2 sin(\omega_i^2[n]) \tag{2.2}$$

where $a[n]$ is the instantaneous amplitude of an AM-FM signal and $omega[n]$ is the instantaneous frequency values.

2.4.2.4 Identity vector (i-vector)

I-vector is an efficient feature extraction technique. I-vector extracts the features to identify the linguistic information contained in the speech. The features are represented

to low-dimensional vectors that reflect the speech content. Originally, i-vector was introduced for the speaker recognition task. However, it has recently been demonstrated that i-vector have been successful in various fields of speech processing applications such as speaker diarization [97], accent recognition [98], speaker's gender [99, 100], emotion [101–106], and stress recognition [107].

I-vector representation is a data-driven approach for speech feature extraction that provides a general model for speech processing. A number of speech frames (segment) transform into i-vector space using the i-vector extractor, as shown in Fig. 2.5.

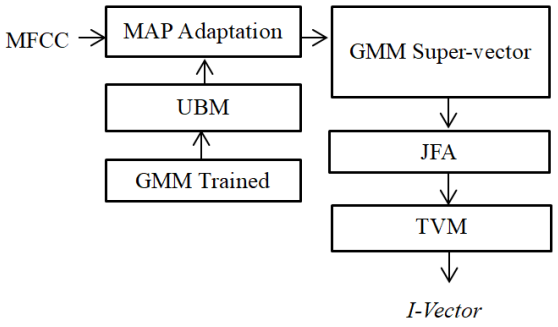


FIGURE 2.5: The basic block diagram of i-vector extractor

For each speech segment, the first-order derivatives of MFCC were used to extract the i-vector feature. The universal background model (UBM) is a large Gaussian mixture model (GMM) trained to represent the gender-independent distribution of the features. The UBM parameters are estimated using the Expectation-Maximization (EM) algorithm to maximize the likelihood of the training data so that the background is universal in a wide-scale database. The Maximum a Posteriori (MAP) algorithm was used to model the speech information in super-vector space by adapting the UBM mean parameters.

A GMM super-vector s is decomposed by JFA into four components: speaker-independent, speaker-dependent, channel-dependent, and residual [108]. Each component is represented in a low-dimensional set of factors, which operate along the principal dimensions (i.e. eigen-dimensions) of the corresponding component. JFA represents the speaker and the channel factor separately, while the i-vector represents them in a single low-dimensional total variability model (TVM). An i- vector model uses a set of low-dimensional total variability factors w to represents each speech sample, known as the i-vector feature. Each factor controls an Eigen-dimension of the total variability matrix T, expressed as follows:

$$s = m + Tw \tag{2.3}$$

where m is the gender-independent mean super-vector (from UBM).

2.4.3 Existing classification methods for emotional speech

In recent years, a great deal of research has been conducted to recognize human emotion using speech information. Many researchers explored several classification methods. The most often used classifier in the stress and emotion recognition task includes the support vector machine (SVM) classifier, the Gaussian mixture model (GMM), the hidden Markov model (HMM), and the neural networks (NN).

2.4.3.1 Support vector machine (SVM)

A support vector machine (SVM) is a supervised machine learning model that is typically used for solving both regression and classification problems [109], such as speech emotion recognition [102–104, 106].

SVM has been widely used to solve binary classification problems. In a binary classification problem, an SVM constructs a hyperplane in a multidimensional vector space, which is then used to separate vectors that belong to two different classes, as shown in Figure 2.6a. A good separation is achieved by the hyperplane that has the largest distance to the nearest training vectors of each class (Figure 2.6b) and the non-optimal hyperplane is shown in Figure 2.6c.

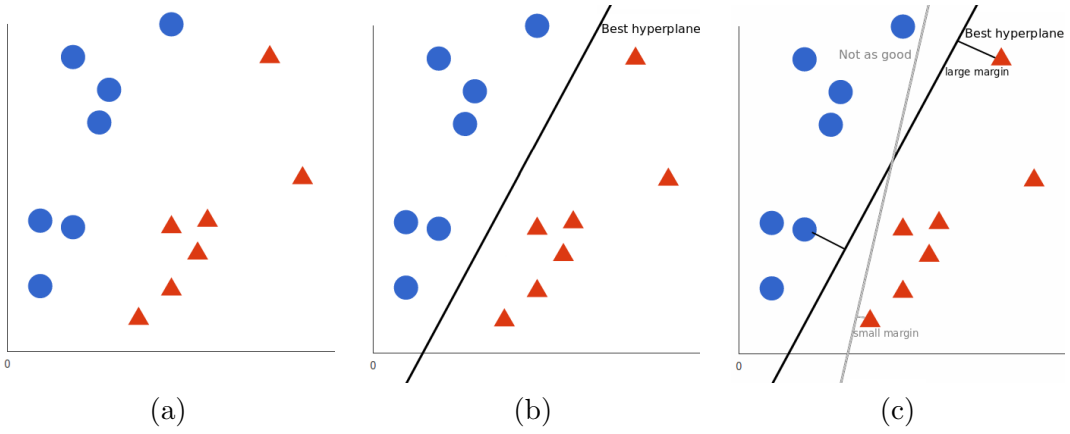


FIGURE 2.6: The SVM's hyperplane: (a) data distribution (blue and red vectors), (b) best hyperplane (black line), (c) not as good hyperplane (grey line).

The two-class SVM method can be expended to a multi-class problem [110]. It is usually performed by reducing the single multi-class problem into multiple binary classification problems, known as the "One Against One (OAO)" strategy. In the OAO, one trains $K(K-1)/2$ binary classifiers for a K multiclass problem; each receives the samples of a pair of classes from the original training set and must learn to distinguish these two classes. At prediction time, a voting scheme is applied: all $K(K-1)/2$ classifiers are applied to an unseen sample and the class that got the highest number of "+1" predictions gets predicted by the combined classifier [111].

2.4.3.2 Gaussian mixture model (GMM)

The Gaussian mixture model (GMM) is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. GMM models the speech as a weighted sum of multivariate Gaussian probability density functions [97, 112, 113]. In this decade, GMM is widely used as a feature modeling and classification algorithm in the speech emotion recognition [38, 78, 79, 102, 114–116], since it can smoothly approximate a wide variety of density distributions.

A Gaussian mixture model is parameterized by two types of values, the mixture component weights and the component means and variances/covariances [117]. For a Gaussian mixture model with K components, the k^{th} component has a mean of μ_k and variance of σ_k, see Figure 2.7. The mixture component weights are defined as ϕ_k for component C_k with the constraint that $\sum_{i=1}^{K} \phi_k = 1$ so that the total probability distribution normalizes to 1.

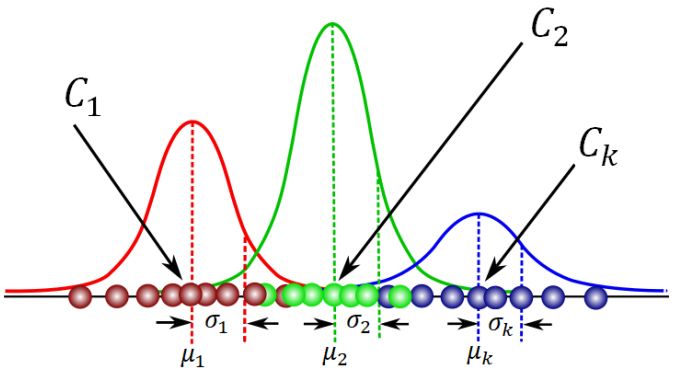


FIGURE 2.7: The illustration of GMM parameters.

For learning the model, GMM uses the expectation-maximization (EM) technique to estimate the mixture model's parameters. the first step (E-step) consists of calculating the expectation of the component assignments C_k for each data point $x_i \in X$ given the model parameters ϕ_k, μ_k, and σ_k. The second step (M-step) consists of maximizing the expectations calculated in the E-step with respect to the model parameters. This step consists of updating the values ϕ_k, μ_k, and σ_k. The entire iterative process repeats until the algorithm converges, giving a maximum likelihood estimate.

2.4.3.3 Hidden Markov model (HMM)

The hidden Markov model (HMM) is a statistical modeling technique that uses the Markov process in modeling the system. The Markov process assumption is simply that the current state depends on the previous state [118]. The Markov chain represents the temporal structure (or pattern) of a feature parameter representing a given class.

HMM have a long experience in speech recognition. The underlying idea is that the statistics of voice are not stationary. Instead of that, voice is modeled as a concatenation of states, each of which models different sounds or sound combinations, and has its own statistical properties. There are two main advantages of HMM's in front of global statistics for emotion recognition [119] first, the structure of HMM's may be useful to catch the temporal behavior of speech; second, HMM technology has been long time studied for speech recognition purposes, being available well-established procedures for optimizing the recognition framework.

The structure of the HMM generally adopted for speech recognition is a left-to-right structure, since phonemes in speech follow strictly the left to right sequence [120]. To illustrate the HMM structure, Figure 2.7 shows the state transition diagram of an HMM. Each of the two hidden states x_i maps to one of the three observable outcomes y_k with some probability b_{ik}. The state transition probabilities a_{ji} are the probabilities of moving from one hidden state to another.

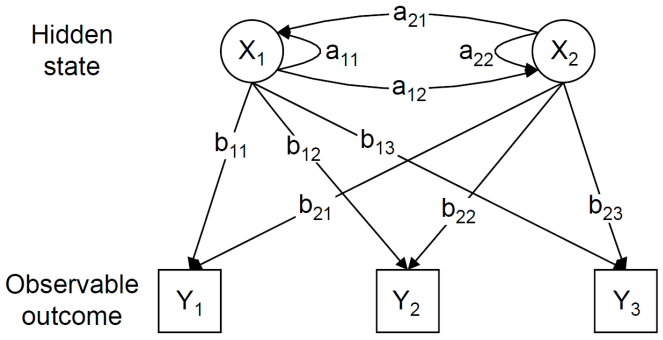


FIGURE 2.8: The illustration of an HMM's state transition diagram.

2.4.3.4 Neural Networks (NN)

Neural networks (NNs) is a nonlinear statistics data modeling tool that is used to model a complex relationship between the input and output data [121] to discover a new pattern. It consists of a group of interconnected artificial neurons, where their connection could be adapted through the change of the weight values during the learning process. NN can be used in two paradigms: supervised learning and unsupervised learning.

NN have been widely used in various pattern recognition problems [122]. The strength of neural networks to discriminate between patterns of different classes has been exploited in a number of speech emotion recognition studies [27, 123–125] showing high levels of success.

In a neural network, there are three essential layers: input Layer, hidden layer, and output layer, as shown in Figure 2.9. The input layer receives the input information of various forms. The hidden layers perform various types of mathematical computation

on the input data and recognize the patterns. The output layer presents the result of rigorous computations performed by the hidden layer.

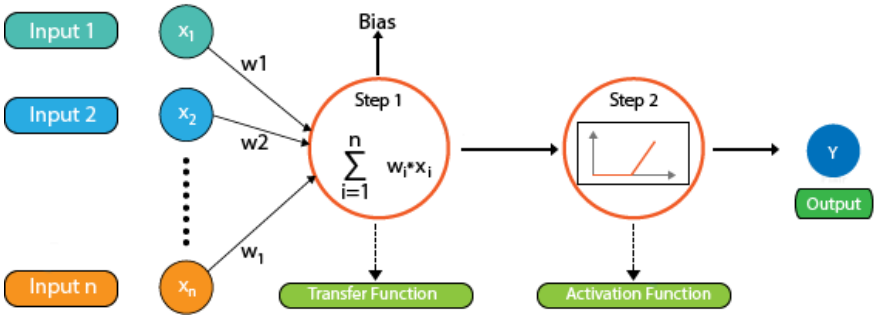


FIGURE 2.9: The illustration of an Neural network.

As shown in Figure 2.9, NN has multiple parameters, such as weights w, biases b, learning rate, batch size etc, that affect the performance of the model. The transfer function can be a simple summation which is a sum of inner dot products of the input to the weights of the connection.
Based on the output of the transfer function, the activation functions compute the appropriate result from the node. The activation function is a non-linear function that able to find an error gradient. Some of the popular activation functions used in Neural Networks are Sigmoid, RELU, Softmax, tanh etc.

2.5 Applications of emotion based technology

Emotion detection has become one of the most important aspects to consider in any project related to Affective Computing. Due to the almost endless applications of this discipline, the development of emotion detection technologies has brought up as a quite profitable opportunity in the corporate sector. This section discusses the possible or existing application of emotion-based technology.

2.5.1 Speech emotion recognition for human-machine interaction

The speech signal communicates linguistic information between speakers as well as paralinguistic information about the speaker's emotions, personalities, attitudes, feelings, levels of stress, and current mental states. Different from human-to-human communication that able to understand the speaker's emotions, human-machine communication suffers from significant inefficiencies because machines cannot understand user feelings or generate emotional responses. Obviously, words are not enough to correctly interpret the mood and intention of a speaker. Thus, the introduction of human social skills to human-machine communication is of paramount importance.

A technology that refers to the communication and interaction between a human and a machine, called human-machine interaction (HMI), has gained increasing attention as they allow humans to control machines through natural and intuitive behaviors. From the human perspective, HMI will be more lifelike and attractive if machines are able to recognize human feelings and respond accordingly. It is expected able to improve the reliability of communication. Therefore, in recent years, HMI technology has started to devote more attention to user attitudes and emotions to increase the acceptability of speech technology for human users.

2.5.1.1 Smart home technology

In today's world, people are often working harder and longer to achieve a more comfortable living. Nevertheless, the pressure and tension due to the increased workload and other challenges have to lead to higher levels of health and mental issues. The employees who experienced high levels of stress have lower engagement are less productive, and often absent to work [126]. Another study previously published by the American Psychological Association also concluded that employees with longer working hours are more frequently linked to family conflicts and stress-related health problems [127]. It indicates that the majority of people are suffering from stress-related health problems and conflicts. As such, it is essential to ensure that the home is conducive for them to distress from the related stresses of their day. When they are going home, the machines will enable a smart home system that reacts to their moods and adjust the lighting or music accordingly and controlled by them using speech. The smart home system automatically detects symptoms of depression, anxiety, bipolar disorder, and allowing a response to such conditions. Thus, machines would no longer be limited to explicit commands and could interact with people in a manner more similar to how we interact with each other.

2.5.1.2 Robot technology

One of the main aims of human-robot interaction is to improve the robot's abilities to interact with humans. In order to achieve an interaction similar to that among humans, robots should be able to communicate in an intuitive and natural way and appropriately interpret human effects during social interactions. Similarly to how humans are able to recognize emotions in other humans, machines are capable of extracting information from the various ways humans convey emotions, including facial expression, speech, gesture, or text and using this information for improved human-computer interaction. This can be described as affective computing [128], an interdisciplinary field that expands into otherwise unrelated fields like psychology and cognitive science and involves the research and development of systems that can recognize and interpret human effects. To leverage

these emotional capabilities by embedding them in humanoid robots is the foundation of the concept of effective Robots, which has the objective of making robots capable of sensing the user's current mood and personality traits and adapt their behavior in the most appropriate manner.

2.5.1.3 Call centers

Over the last decade, enterprises have started to use human-operated call-centers to provide improved services to their customers. The call-center agents answer customer calls and provide information on different aspects of the services provided by the enterprise that they represent. The evaluation criteria of such call-centers depend on the ability of the agents to satisfy their customer needs in the telephone conversation. Therefore, all call-centers have supervisors who monitor the calls and identify if any agent was not able to satisfy a customer. Since the number of calls received in a typical call-center is very high [129], it is not cost-effective to monitor all calls. So the supervisors monitor a subset of calls and identify if any of them had extreme emotional characteristics (such as happy or angry moods). However, the cost involved in human-monitoring of these calls is extremely high. Therefore, automatic monitoring of these calls for recognizing the emotional features is a very important problem from a business perspective.

2.5.2 Speech emotion recognition for industry and society

Speech is the fundamental form of human communication, and much (if not all) human speech is the product of a speaker's emotional state. However, the existing speech processing systems have lacked the effective processing of that emotion. The development of emotional speech technology has the potential to provide significant benefits to the national and international industry and society in general.
In law enforcement and military, national security can benefit from forensic applications of emotion detection (new types of lie detectors, emotional speech analysis of suspects, terrorists, kidnappers, hostages). Public safety, border control, and internet security can benefit from improved automatic speech and speaker recognition systems.
In safety management system, the development of emotional speech technology will open possibilities of new applications such as automatic assessment of the mental state of people working under high-risk and high-stress conditions that require an optimal mental and emotional state (e.g. heavy machinery operators, people working with dangerous chemicals, poisons and radioactive materials, construction workers, pilots, car, and bus drivers surgeons).
In the medical field, mental health and medicine will benefit from the development of automatic systems providing quantitative measures supporting the diagnosis of emotional disorders, such as depression, Alzheimer's, and autism, is currently researching

a diagnostic system for early detection of depression. Also, natural-sounding synthetic speech capable of emotional expression will improve speech aids for mute people and new automatic training systems can be designed to teach autistic children how to recognize emotion in speech.

Chapter 3

Proposed Framework for Stress and Emotions Speech Recognition and Modeling

The Sympathetic Nervous System (SNS) provokes the stress response in humans, carrying psychological, physiological and behavioral symptoms. From a physiological point of view, the increase of SNS activity changes the hormonal levels of the body and provokes various reactions on the body. Therefore, in order to identify stress and emotion, scientists use physiological parameters as observation material and analyze its activity changes. A non-intrusive method, such as speech analysis, becomes popular because it offers easiness in measurement and completely contactless to the user.

The stress measurement method using speech, or known as stress speech recognition (SSR), learns the change of speech patterns in recognizing stress [11]. SSR is a classification system that uses labeled speech utterances for training. A set of relevant stress speech data is required in this training phase. SSRs are robust in precise conditions, but typically, their performance degrades in an imprecise environment. Therefore, more large stress speech data are required to make the SSR model able to adapt to the real condition. Unfortunately, the natural stress speech datasets are hard to be collected.

To address this issue, many researchers explore another approach that no require labeled data for training, known as cluster analysis. The primary difference between classification and clustering is that classification is used in a supervised learning technique where predefined labels are assigned to instances by analyzing their properties, while clustering is used in unsupervised learning where similar data points are grouped, based on their property's similarities. in identify stress and emotion. There are many clustering algorithms that have successfully categorized the stress speech data [35–38]. Most of them used a similarity algorithm to group the data points by computing the distance.

However, it was found that these algorithms become inefficient to be applied in high-dimensional data due to its computation time and memory usage [130], known as the curse of dimensionality. Lately, by a self-learning manner to optimize the clustering objective, a deep clustering algorithms comes to address the curse of dimensionality problem [131]. A deep clustering applies the deep neural network (DNN)-based autoencoder to transform the data from the original space to a lower-dimensional space (embedding space) compactly [132, 133]. By learning in-depth and simultaneously minimize the error, an deep clustering able to present an excellent feature for representing the stress and emotion characteristics.

In some cases, emotion (especially stress) may change when triggered by an event during the speaking [41]. In this fashion, we argue that the prior emotional state should also be monitored so that the emotion of the speaker can be recognized more accurately. By this approach, we can take advantage to deal with larger sets of contextual information [134]. Several studies have successfully modeled the emotion based on its state transition [41, 134–137]. Generally, for predictive modeling or probabilistic forecasting [138], the state transition model is the most used since its convenience for modeling the temporal context in time-series (continuous) data [137, 139, 140].

To this end, we propose an end-to-end framework for recognizing and modeling the stress and emotion in speech. Since classification and clustering have their own superiority, the proposed framework analysis the stress and emotion in the approach of classification, clustering, and state transition modeling, as shown in Figure 3.1.

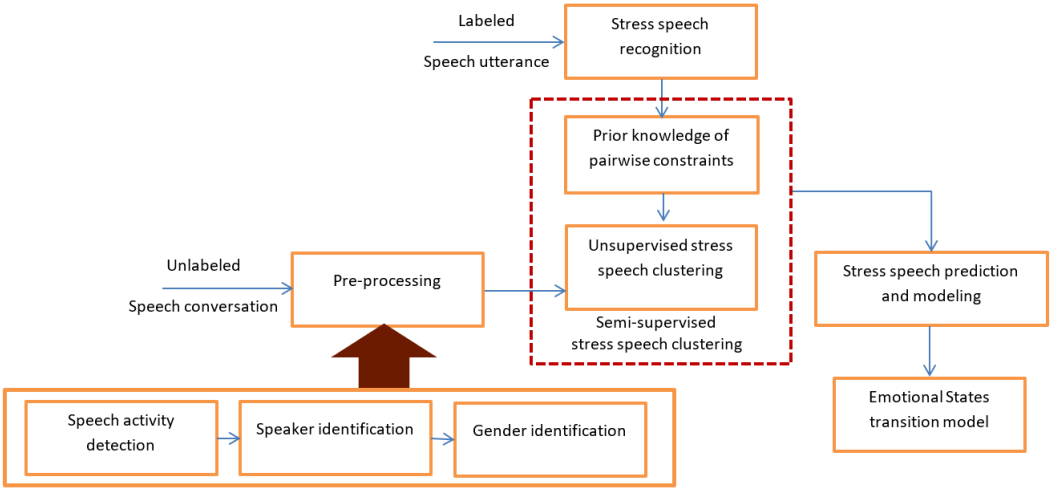


FIGURE 3.1: The proposed framework of stress and emotions speech recognition and modeling

As shown in Figure 3.1, the proposed framework three main parts, (1) a stress and emotion speech recognition (SSR) system, (2) the stress and emotion speech clustering (SSC) system, and (3) a stress and emotion speech modeling (SSM) system. Firstly, we recognize stress and emotion as a classification system that uses labeled stress speech

data. Since the SSR system may degrade its performance due to a real environment (imprecise training condition), in further, we perform the unsupervised SSC where uses unlabeled data for training. Then, to ensure whether the output class corresponds to informational classes, we inject the prior knowledge into the unsupervised SSC, termed as a semi-supervised SSC. We explicitly take the prior knowledge from SSR that known has accurate stress information due to trained by labeled data. Furthermore, for accommodating the change of emotion state during the speaking, we take larger sets of contextual information by capturing the prior emotional states, named as modeling approach (SSM).

3.1 Dataset

The Speech Under Simulated and Actual Stress (SUSAS) database [29, 30] comprises a wide variety of simulated and actual stresses and emotions. The SUSAS database was developed by the linguistic data consortium (LDC). The speech was collected from 32 speakers (13 male and 19 female) within the work environment (pilot cabin) and during a rollercoaster ride. The speakers were reading a fixed list of 35 words presented in random order. The speech corpora represented a narrow-band speech sampled by a 16 A/D converter with a sampling frequency of 8 kHz.

There are two set types of data in the SUSAS database: labeled short utterance data and unlabeled conversation speech data. The labeled short utterance data consist of 1323 female and 1377 male utterances. Each utterance has a stress and emotion label (high stress, low stress, neutral, soft, and angry), as shown in Table 3.1. Moreover, SUSAS provided six unlabeled conversations with three conversation types (single speaker, two speakers of the same gender, and two speakers of different genders). In total, all conversations contained around 50 female and 566 male utterances that presented time series (sequences) data, as shown in Table 3.2.

Table 3.1: The labeled SUSAS dataset used in the experiments

Label	Number of utterances	
	Male	Female
High stress	301	301
Low stress	337	301
Neutral	319	301
Angry	210	210
Soft	210	210

TABLE 3.2: The unlabeled conversation of the SUSAS dataset used in the experiments

Data	Data distribution		Gender	
	Spk_1	Spk_2	Spk_1	Spk_2
1	98	-	Male	-
2	52	50	Male	Female
3	94	-	Male	-
4	118	-	Male	-
5	56	51	Male	Male
6	40	57	Male	Male

3.2 Stress and emotions speech classification

In preliminary work, we use a low stress and non-stress data in the training phase. The binary classification of stress and non-stress is performed using the power of speech and related frequencies as the features. The ensemble method of support vector machine (SVM) and neural network (NN) is used to classify the stress speech. This model is robust for binary classification of stress, but it not quite powerfull for classifying five-classes of stress and emotions. Therefore, in-depth learning-based feature and a deep neural network (DNN) are applied to classify five-classes of stress and emotions.

3.3 Stress and emotions speech clustering

Since a clustering method refers to an unsupervised setting that uses unlabeled conversation data, we perform a pre-processing phase before the clustering process. The pre-processing phase consists of speech activity detection (SAD), speaker identification, and gender identification. The SAD process aims to separate the speech and non-speech segment/frame. Since the conversations are spoken by two-speakers, an speaker identification process is required. Males and females express stress in different ways. Therefore, we perform gender identification.

We then apply an unsupervised deep clustering algorithm to categorize the stress speech data, we named deep time-delay autoencoder embedded clustering (DTEC). DTEC deeply learn the stress speech segment using DDNN-based autoencoder and simultaneously optimizing the clustering assignment by joint supervision of discriminative loss, reconstruction loss, and clustering loss.

Since DTEC has not confirmed yet the compatibility between the output class and informational classes, we incorporate the prior knowledge of pairwise constraints to DTEC. We named semi-supervised deep time-delay embedded clustering (SDTEC). The semi-supervised constraint loss and the unsupervised loss are used simultaneously to supervise the feature representation and the clustering assignment.

3.4 Stress and emotions speech prediction and modeling

Despite its compactness to represent the emotional features, most of the deep clustering algorithms have not considered the prior state yet. To this end, we present a new framework to predict and model the stress and emotions, named the deep time-delay Markov network (DTMN). DTMN analyzes in-depth the stress and emotion speech feature by considering the prior emotional states. Structurally, DTMN contains the Markov modeling that handled by hidden Markov model (HMM) and the neural network architecture of time-delay neural network (TDNN).

Chapter 4

Stress and Emotions Speech Recognition

As discussed in the previous chapter, stress can affect the body, thoughts and feelings, and our behavior. Therefore, emotional awareness is essential for human life by recognizing common stress symptoms so that one can manage their stress. Stress that's left unchecked can trigger many health problems.
Several physiological parameters have been explored to recognize stress and emotion [71]. Using manifold modalities might capture the stress accurately and result in optimal outcomes of recognition. However, several inherent advantages (more readily and economically) make speech-based emotion recognition more exciting to be used in recognizing the stress and emotion, known as the stress speech recognition (SSR) system.

4.1 Binary stress speech classification

Stress speech recognition (SSR) or more generally called speech emotion recognition (SER) is a machine learning-based system that able to recognize and classify the stress and emotion speech data. There are three primary keys for a successful SER system, namely, (1) relevant emotional speech database, (2) extracting effective and efficient features, and (3) designing reliable classifiers using machine learning algorithms.
A set number of labeled dataset is used to train the machine learning for recognizing the stress speech patterns. Therefore, in machine learning, the dataset is the most crucial aspect, especially for the results of classification [141]. However, natural stress and emotion speech data, such as a sample of SUSAS database [29, 30], is hard to be collected massively. Hence, the challenge to improve the predicting capability of SSR system with limited availability of materials data is clearly highlighted.
Many researchers have proposed important speech features which contain emotion information, such as energy, pitch, formant frequency, and Mel-frequency cepstral coefficients

(MFCC) [142]. Loui et.al. [143] presented that the power of the human voice to communicate emotion is well documented in verbal speech as well as in non-verbal vocal sounds, and the human voice is thought to convey emotional valence, arousal, and intensity via its modification of spectral and temporal signals. Arousal is a measure of perceived energy level, ranging from low (calming) to high (exciting). Orthogonally, valence is the polarity of perceived emotions and ranges from negative (sad) to positive (happy). Therefore, in each speech segment, we extract speech power and its related frequency as effective and efficient stress features.

The ability of many classification algorithms has been explored for recognizing stress and emotion in speech. However, the confusion matrices of different methods varied a lot. It indicates that different system architecture has also different capabilities in modeling emotion. It also means that a single classifier hard to performs stably well on all emotion categories, which might be possibly due to bias and variance of the error. To this end, we propose to use an additional algorithm for improving the stability and accuracy of machine learning by combining several features or classifiers, known as ensemble algorithms [144]. The improving strategy of the ensemble algorithms is to reduce the bias and variance of the error [145].

More than two decades, SVM and NN were frequently mentioned as a robust classifier that works well on high bias and variance caused by small training data [146]. Motivated by the improved research results of ensemble learning and the robustness of SVM and NN, we propose an ensemble method for SSR via average aggregating results from a combination of SVM and NN classifiers.

4.1.1 Ensemble SVM-NN

The 10-sets of stress and neutral utterances D are used. We then divide D into 10-groups PD by combination formula. Thus, each group has nine data sets. Each data group is trained using a Support Vector Machine (SVM). After training, the output of SVMs are used as the input of NN. The ensemble SVM-NN framework is shown in Figure 4.1.

We select average technique as an aggregation strategy because of its simplicity. Average is the simplest method for combining several SVMs. Let f_m (m=1,2,...M) be a decision function of the m^{th} NN in the SVM-NN ensemble and C_j (j=1,2,...C) denote a label of the j^{th} class. Then, let N_j=$\{m|f_m(x) = C_j\}$ i.e. the number of NNs whose decisions are known to the j^th class. Then, the final decision of the SVM-NN ensemble f_{SVM-NN} for a given test vector x due to the majority voting is determined by

$$f_{SVM-NN} = \frac{1}{M}\sum_{m=1}^{M} f_m(x) \tag{4.1}$$

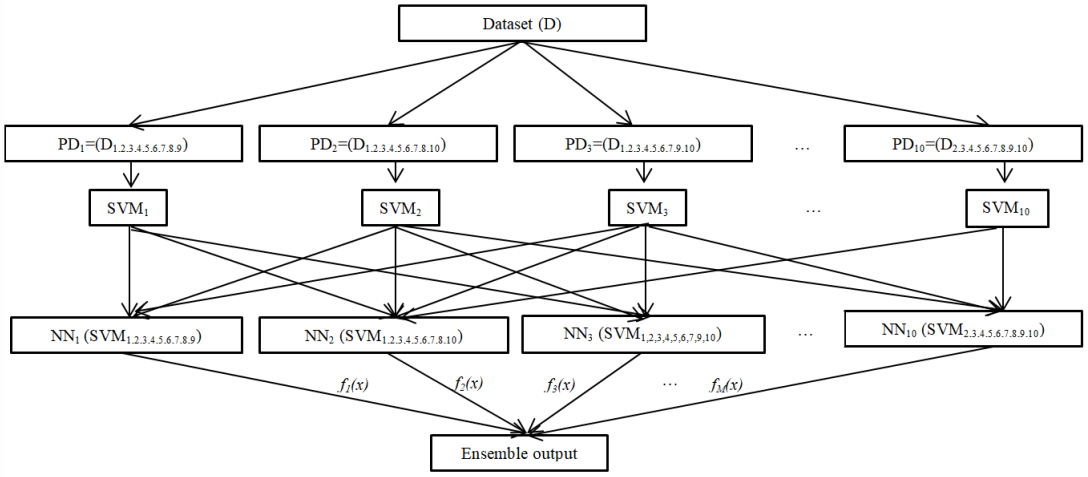


FIGURE 4.1: The ensemble SVM-NN framework

4.1.2 Experimental setup

In this work, we used sample of SUSAS database with 10 words (break, change, degree, destin, east, eight, eighty, fifty, fix, and freeze). Each word is classified into neutral (non-stress) or stress.

4.1.3 Result and discussion

We evaluate the effectiveness of the ensemble SVM-NN in the classification task of 10-samples data from SUSAS using a k-fold validation method, where $k = 10$. The comparison classification result of the ensemble SVM-NN compared to existing methods is shown in Table 4.1.

TABLE 4.1: The comparison classification result

Classifier	Accuracy (%)	
	Male	Female
Single SVM	69.41	72.00
Single NN	90.57	88.84
Ensemble SVM-NN	96.24	95.56

4.1.4 Conclusion

In this chapter, we proposed a stress speech recognition (SSR) system using simple features and a low dataset. The speech power and its related frequency are used as stress features. The 10-female and 10-male utterances of SUSAS dataset are used in training and evaluation. To address this limitation, we used an ensemble method of SVM and NN in improving the classification results. We evaluate the effectiveness of

the ensemble SVM-NN in the classification task of 10-samples data from SUSAS using a k-fold validation method, where $k = 10$. Based on the experimental results, ensemble SVM-NN outperformed both single SVM and NN by presenting the accuracy of 95.9%. We found that, generally, the differences between neutral and stress are the power will decrease and the frequency will increase.

4.2 Stress and emotions speech classification

As discussed in Chapter 4.1, we have been explored a simple approach in the binary classification of the SSR system. A simple feature in binary classification may adequate. However, for multi-class classification with a large dataset, a simple feature could not handle a high bias or variance and may degrade the system performance.
In this decade, an effective and robust speech feature extraction technique that represents the speech content in a low-dimensional vector, called i-vector, has been successfully used in recognizing a speaker's emotion [101–103] and stress [107].
On the other hand, a deep neural network (DNN) has shown its robustness in many speech-based applications. Since DNN can directly optimize a discriminative between classes, it offers a potential promising to represent a robust and compact stress and emotion feature. Thus, we propose to use an acoustic model of DNN-trained to discriminative the classes by mapping the speech variables into a fixed-dimensional embedding vector, called x-vector.
In the back-end processing, the i-vector system usually applies linear discriminant analysis (LDA) to normalize the vector length. LDA is a statistical method to classify the speech pattern by minimizing the inter- and intra-class covariances and finding the best solution for linear problems. Since the stress speech affects a non-linear vocal tract [116], LDA is not reliable. To address this issue, we proposed a general model of LDA, called generalized discriminant analysis (GDA), which applies a non-linear discriminant analysis to map the data into a high-dimensional space using kernel functions.
We propose an end-to-end stress speech recognition (SSR) system using x-vector (DNN embedding feature) followed by generalized discriminant analysis (GDA) for vector length normalization and joint probabilistic linear discriminant analysis (J-PLDA) for scoring.

4.2.1 X-vector system

X-vector considers as the discriminative baseline system, since it is comparable with i-vector systems for text-independent speaker recognition, especially for short utterances [147]. The conceptual baselines of i-vector and x-vector is shown in Figure 4.2.
X-vector is a DNN-based model that maps variable length segments of speech to an embedding space. We built the x-vector system using the framework of [148, 149]. The

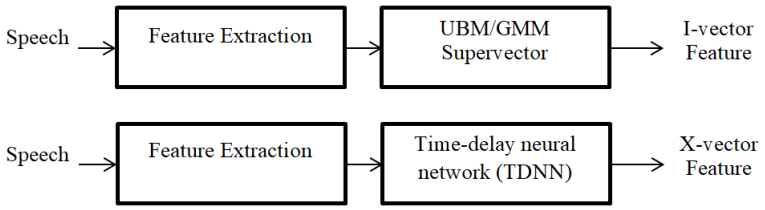


FIGURE 4.2: The baseline of i-vector and x-vector concept

x-vector was composed of the time delay DNN (TDNN) [148, 150] which computed the speech embedding from a variable length segment of the acoustic features. The input consisted of a stack of frames in a short-term temporal context that was then handled by the TDNN architecture, as shown in Figure 4.3.

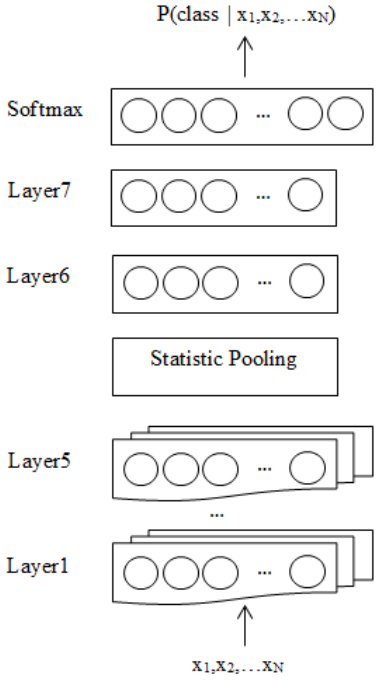


FIGURE 4.3: The x-vector architecture

The TDNN architecture consisted of nine layers that captured the speech context at each level as shown in Table 4.2: five layers for frame level representation; a statistical pooling layer for computing the mean and standard deviation; two segment layers for extracting affine components; and the softmax layer.

We assumed that each input segment had F frames. The first five layers operated at the frame level with a small temporal context in the current frame f by utilizing the rectified linear unit (ReLU) activation and batch normalization. In Table 1, it can be seen that input Layer1 gave the total temporal context of 5 frames $(f-2, f-1, f, f+1, f+2)$. Layer2 was part of Layer1: frames $f-2, f, f+2$, which gave it a total context of 9 frames. Layer3 was the spliced output of Layer2, at frames $f-3, f, f+3$, so that it

TABLE 4.2: The time delay DNN layers configuration

Layer index	Layer context	Number of context	Dimensional output
Layer1	$f-2, f-1, f, f+1, f+2$	5	512
Layer2	$f-2, f, f+2$	9	512
Layer3	$f-3, f, f+3$	15	512
Layer4	f	15	512
Layer5	f	15	1500
Stat Pooling	$[0, F]$	F	3000
Layer6	-	F	3000
Layer7	-	F	512
Softmax	-	F	512xN

where: f is frame. F is all frames. N is number of training data

had a total context of 15 frames. Layer4 and Layer5 are also operated at the frame level but did not provide additional temporal context. The total TDNN network worked on $f-8$ to $f+8$ frames. Therefore, the dimensional vector became tripled (i.e., from 512 to 1536). The statistical pooling layer reached the mean and standard deviation at all frames F. In Layer5, there were 1500-dimensional vectors which doubled after the statistical pooling output. Layer6 and Layer7 processed pooling and normalized into 512-dimensional vectors.

To reduce the sensitivity of the speech length, the DNN was trained to capture a certain duration of the speech section during the test phase. However, memory limitations forced a trade-off between the mini-batch size and the maximum training example length. Hence, we set the speech duration to a maximum of 3 seconds (300 sample frames) and the mini-batch size to 32. The stochastic gradient descent (SGD) was used to train the network for several epochs.

During the training phase, the parameters of the DNN were optimized using the softmax loss. The parameters were defined via a linear transformation with weight and bias vectors that followed by the softmax function and the multiclass cross-entropy loss . The x-vector was extracted from the affine component of $Layer6$, because the softmax layer and $Layer7$ are not needed after the training phase.

In the back-end, GDA maps the feature vector x in space X to the vector $\phi(x)$ in space S. Then, *interclass* S_b and *withinclass* S_w scatter are updated as assuming observations in vector mean $\bar{\phi_c}$ for x in class c that is the centered of S.

$$S_b = \frac{1}{C}\sum_{c=1}^{C} n_c \bar{\phi_c}\bar{\phi_c}^T \tag{4.2}$$

$$S_w = \frac{1}{C}\sum_{c=1}^{C}\sum_{k \epsilon c} \phi(x_k)\phi(x_k)^T \tag{4.3}$$

The traditional LDA kernel function on X-Vector $w1$ and $w2$ as follows:

$$k(w_1, w_2) = \langle w_1, w_2 \rangle \tag{4.4}$$

We use this baseline discriminant analysis:

$$k(w_1, w_2) = \frac{\langle w_1, w_2 \rangle}{\|w_1\| \|w_2\|} \tag{4.5}$$

Furthermore, in this work, we determined Gaussianized cosine as the kernel of the GDA. The Gaussianized cosine kernel based on the following algorithm:

1. Initially, calculates the mean m and the standard deviation v from the training data.
2. All data (training, testing, enrollment) are modeled Gaussianized (every X-Vector w). Then the modified vector becomes $w = \frac{w-m}{v}$
3. The Gaussianized X-Vector normalized in length.
4. The projection matrix P is trained after the training data is calculated.
5. All data is projected in the new feature space for each X-Vector w. The transformation becomes $w = P^T w$
6. The new X-Vector is normalized again in length.
7. This data is calculated cosine kernel according to equation 4.

4.2.2 Experimental setup

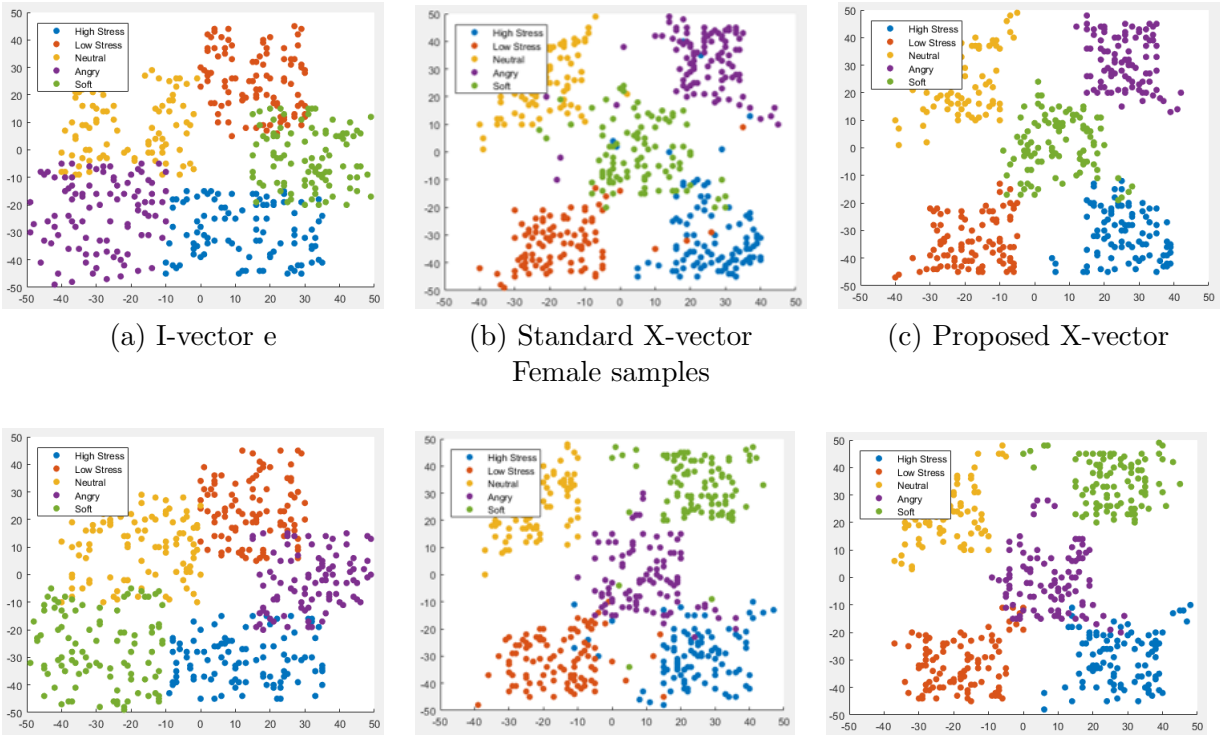
The total data consisted of 1,377 male and 1,323 female utterances. Each utterance was labeled as one of five stress classes: high stress, low stress, neutral, angry, and soft. In the experiments, we divided data 50:50 into training and testing data generated randomly on all of classes.
16-feature vectors were extracted from each speech frame. UBM is a 2,048-component full-covariance GMM that was used to extract the 600-dimensional i-vectors. The nine layers of the TDNN were applied to obtain 512-dimensional x-vectors. At the end of the GDA process, the dimensions of the i-vector and x-vector were reduced to 300-dimensional vectors.

4.2.3 Result and discussion

We visualized the class distribution for each subset using t-Distributed Stochastic Neighbor Embedding (t-SNE). t-SNE builds a set of embedded points in a low space dimension

that have relative similarities with its high dimensions [151]. The embedded points represent the probability distribution of neighbors around each original point. The t-SNE algorithm models the original point as a Gaussian distribution [152].

We visualized 100 random utterances from each class of the i-vector and x-vector for female and male data separately, as illustrated in Figure 4.4.



(a) I-vector e (b) Standard X-vector Female samples (c) Proposed X-vector

(d) I-vector (e) Standard X-vector Male samples (f) Proposed X-vector

FIGURE 4.4: t-SNE visualization of the stress class distribution. (a), (b), and (c) denote for female samples, while (d), (e), and (f) present male samples.

Since x-vector trained gradually (at the frame level in the first five layers and continued in the segment level), x-vector can learn the differences between classes without learning the characteristics of each speaker. Thus, Figure 4.4 shows that x-vector is able to represent the classes as more separate or distinct. On the other hand, i-vector learns the characteristics of each speaker first and then adds specific information (e.g., about stress/emotion); therefore, the classes have a more distributed spread.

We evaluate the effectiveness of x-vector system in the classification task of stress and emotions from the SUSAS dataset and compare it with a simple approach in Chapter 4.1 and i-vector system. The classification result of them is shown in Table 4.3.

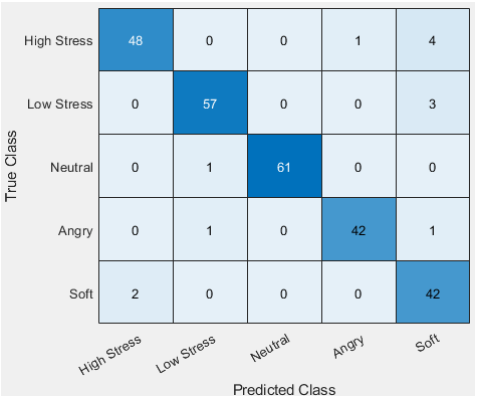
Furthermore, to evaluate the misrecognition tendency, we use a confusion matrix and compare the x-vector system to the i-vector system. In Figure 4.5, each row of the confusion matrix represents the instances of a predicted class while each column represents

TABLE 4.3: The comparison classification result

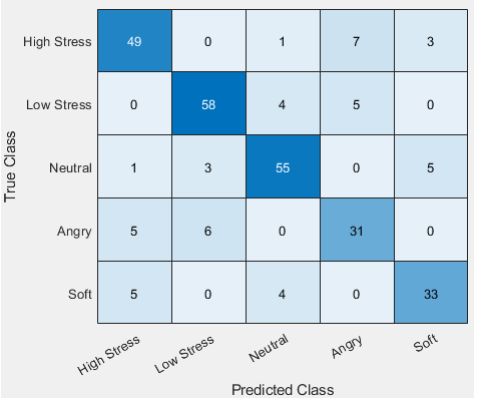
Method	Error rate (%)	
	Male	Female
Ensemble SVM-NN	50.50	48.50
i-vector system	18.12	16.29
Standard x-vector system	8.84	8.97
Proposed x-vector system	5.08	5.31

the instances of a true class.

(a) I-vector system for female data



(b) X-vector system for female data



(c) I-vector system for male data

(d) X-vector system for male data

FIGURE 4.5: The confusion matrix of the proposed and baseline system

Figure 4.5 shows that the proposed system outperformed the baseline system in classifying stress speech. Our system can identify the "neutral" class very accurately, but there are several misrecognitions in other classes. For data from female speakers, the misrecognition occurs between the "stress" and "soft" classes. For data from male speakers, the misrecognition occurs between the "stress" and the "angry" classes. We hypothesize

that the female speakers tended to express their stress in soft (e.g., sadness) while the male speakers tended to express their stress as anger.

4.2.4 Conclusion

A stress speech recognition (SSR) system using the DNN embedding framework concept, called x-vector, has been explored. We found that the proposed x-vector system outperformed i-vector and standard x-vector in extracting stress features in the speech by improving the EER. Based on confusion matrix evaluation, the most misrecognition is between the "stress" and "soft" classes for data from female speakers and between the "stress" and "anger" classes for data from male speakers.

Chapter 5

Stress and Emotions Speech Clustering (Pre-processing)

As mentioned in Chapter 3, the second part of the proposed system is stress and emotion speech clustering. This approach aims to address the SSR system limitation that may degrade due to the real environment. Furthermore, since the clustering method uses unlabeled data, we perform a pre-processing phase to ensure the clustering process presents a good performance. The pre-processing phase consists of speech activity detection (SAD) system, speaker identification, and gender identification.

5.1 Speech activity detection

Speech activity detection (SAD) is an essential part of the speech-based applications [153]. SAD plays a critical role in separating between speech and non-speech segments, such as noise, music background or silence. Typically, SAD is performed as the first task to filter the presence of non-speech segments. Non-speech segments considerably affect the performance of the main system due to carry useless information. In other words, a robust SAD system is an important modal to obtain an accurate detection result of the main system.

Many feature extraction techniques that can reflect the speech features. For instance, energy-based features [154, 155] and Mel-frequency cepstral coefficients (MFCCs) [98, 156–158] are the technique that is frequently used in representing the presence of speech. Both techniques show their robustness in clean conditions, but the performance degrades for noisy signals. In this decade, a modern method, known as the learning-based technique, has been explored in the term of extracting the speech features [98]. The GMM-based feature and its variance have been successfully used in projecting the presence of speech [156, 158, 159]. Furthermore, an effective and sophisticated technique, known

as i-vector, also has shown their ability to present a compact speech feature for a SAD system [156, 160, 161].

A linear method, such as linear discriminant analysis (LDA), has successfully discriminated against the speech and non-speech [157, 162]. More than 30-years, LDA has been commonly used as a standard back-end procedure in a wide range of speech-related tasks. By assuming each class is a Gaussian distribution and all classes share the same covariance matrix, LDA shows its effectiveness in stationary noise conditions. The machine learning model, such as the hidden Markov model (HMM) [158] and support vector machine (SVM) [156, 160], has proven more accurate in non-stationary noise conditions but involve a complex procedure.

Nowadays, deep neural networks (DNNs) have been used and achieves extremely high predictive accuracy in hardly overall machine learning applications, included as a feature compensation and denoising technique [163] in emotion recognition. In the denoising task, DNN learns the entire temporal context of input in-depth and learn its projection by mapping the noisy feature (i-vector) to a denoised space [164]. Besides in denoising tasks, [155, 161, 165] explicitly uses DNN as a channel compensation for a SAD system and proven effective to address a non-stationary noise.

As discussed above, LDA is the most popular and preferable model due to its simpleness and efficiency in recognizing a pattern but susceptible to non-stationary noise. On the other hand, DNN shows its effectiveness as a denoise technique but involves more complex procedures. Many studies have explored various approaches to develop a SAD system. Most of them focus on how to recognize the presence of the speech in high noise conditions and just little works notice the emotional condition of the speaker also. Whereas, the presence of emotion (especially stress condition) affects the performance of speech algorithms [166]. For a standard speech-related system, such as speech recognition [167, 168] or speaker recognition [169, 170], a robust SAD system in the noisy conditions has adequate. However, for the emotion-related system [26, 28, 132], a more powerful SAD system is required because the emotional condition affects the production of speech characteristics.

Finally, we propose a SAD system which not only strong in noisy environments, but also able to compensate the presence of the speech that might be altered because of emotional conditions. The proposed SAD system consists of the i-vector feature extractor and a novel channel compensation method, named as embedded discriminant analysis or EDA. EDA is a channel compensation method that as simple and efficient as LDA but also has an ability to transforms the feature to denoise space like the DNN. EDA distinguishes speech and non-speech effectively by mapping each frame of i-vector feature to a more discriminative space using DNN and modeling its transformation in a projection matrix like the LDA model. We explicitly use a time-delay neural network (TDNN) in the EDA's structure to handle the variations of temporal dependencies caused by the presence of

emotional stress [171]. In the training phase, a large amount of short speech data from the SUSAS dataset are used to create the speech/non-speech model. In the testing phase, the cosine similarity algorithm is used to compute the deviation between the speech/non-speech model and the representation of the audio target, and also for deciding the decision threshold. Based on this threshold, the speech and non-speech boundaries are decided. The effectiveness of the proposed SAD system is evaluated in terms of the stress speech clustering task [132].

5.1.1 Proposed SAD system

The proposed SAD system consists of training and testing phase, as shown in Figure 5.1. In the training phase, the speech and non-speech features are extracted using the MFCC technique. Then, the i-vectors extractor transforms each frame of the speech features to a single low-dimensional i-vector space. By using the i-vector feature of speech and non-speech, EDA is trained to produce the speech and non-speech model. In the testing phase, the same procedure is conducted to the audio target. We perform the cosine similarity algorithm to compute the resemblance score between the audio target and both of speech/non-speech models. Finally, the score based-error evaluation metric is applied for deciding the decision threshold.

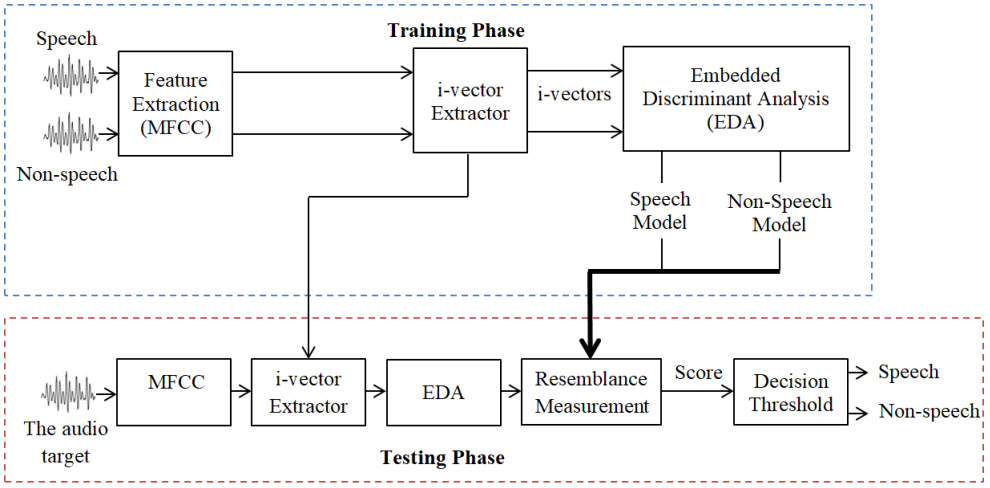


FIGURE 5.1: An end-to-end SAD system in training and testing phase

5.1.1.1 Embedded discriminant analysis

Different to LDA that assumes all classes share the same covariance matrix and find its linear transformation using Gaussian distribution, while EDA initially maps the features to an embedding space, then finds the corresponding transformation using trained neural network.

TABLE 5.1: The EDA network structure

Layer	Input context	Dimensions	Function
Input	-	600	-
Hidden-1	[t-8,t,t+8]	600	ReLU
Hidden-2	[t-5,t,t+1]	600	ReLU+Batch-Norm
Embedding	{0}	400	
Loss layer	Jointly supervision of losses		

We use a one-dimensional convolutional network (known as TDNN) in the EDA structure to address the temporal dynamics dependencies [157, 162]. Structurally, the proposed EDA consists of two hidden layers, the embedding layer, and the loss layer, as shown in Table 5.1.

We propose to use a one-dimensional convolutional network (known as TDNN) in the EDA structure to address the temporal dynamics dependencies caused by emotional conditions of the speaker's [172]. TDNN learns dynamically temporal dependency by generating larger networks from sub-components at across time steps [173]. We applied the sub-sampling (locally-connected) technique on both hidden layers of EDA to make it more efficient. Structurally, the proposed EDA consists of two hidden layers, the embedding layer, and the loss layer, as shown in Table 5.1. EDA capture a total temporal context on [-13,9] [173] that processed by 2 layers. The hidden-1 layer splice together frames $t-8$ through $t+8$ and the hidden-2 layer splices together frames $t-5$ through $t+1$. We applied a rectified linear unit (ReLU) as activation functions on all hidden layers and incorporated a batch normalization in the hidden-2 layer to stabilize the training procedure. During training, the parameters of EDA are optimized under supervision of softmax loss and center loss.

The softmax loss function [174] is defined as follows:

$$\mathcal{L}_S = -\sum_{i=1}^{G} log \frac{e^{W_{z_i}^T y_i + b_{z_i}}}{\sum_{j=1}^{Z} e^{W_j^T y_i + b_j}} \tag{5.1}$$

where $y_i \in \mathbb{R}^d$ denotes i^{th} embedding feature, belonging to $z_{i^{th}}$ class, and Z denotes the number of softmax output (number of classes). W_j is j^{th} column of the weights matrix W and b is bias term. d and G are feature dimensions and the total number of training samples (i-vector), respectively.

The center loss function [175], defined as follows:

$$\mathcal{L}_C = \frac{1}{2} \sum_{i=1}^{G} \parallel y_i - c_{z_i} \parallel^2 \tag{5.2}$$

where c_{z_i} denotes the $z_{i^{th}}$ class center of embedding feature. Equation 5.2 shows that

the intra-class variations are effectively characterized. The c_{z_i} is updated based on mini-batch and the center compute by averaging the embedding feature of the corresponding class iteratively. A scalar α is used to control the learning rate of the center where α is [0 to 1] restrictively.
After training, the transformed features are extracted from the affine component of the embedding layer, hereinafter referred to embedding feature that is formally trained under supervision of softmax loss and center loss, as follows:

$$\mathcal{L} = \mathcal{L}_S + \lambda \mathcal{L}_C \tag{5.3}$$

where λ is a weight for the balance of two-loss functions. If λ is small, the loss is supervised by softmax loss. Otherwise, the supervision is inclined to the center loss.
EDA transforms the i-vector feature from an original space y into embedding space u. We define the transformation model of EDA in a projection function similar to the LDA model, as follows:

$$u = \theta(y) \tag{5.4}$$

where θ denotes the transformation projection matrix of EDA.

5.1.1.2 Resemblance measurement

In the training phase, the SAD system is trained to generate speech and non-speech models. We stored one model vectors per trained segments. Then, those models (speech ϕ_{sp} and non-speech ϕ_{nsp}) are presented in d-dimensional vectors, defined as follows:

$$\begin{aligned} \phi_{sp} &= (\mathcal{N}_1(u_{sp}), \mathcal{N}_2(u_{sp}), ..., \mathcal{N}_d(u_{sp})) \\ \phi_{nsp} &= (\mathcal{N}_1(u_{nsp}), \mathcal{N}_2(u_{nsp}), ..., \mathcal{N}_d(u_{nsp})) \end{aligned} \tag{5.5}$$

where u_{sp} and u_{nsp} are speech and non-speech vectors, respectively. We then compute the deviation between the embedding feature (the output of EDA) of audio target ϕ with both models (speech ϕ_{sp} and non-speech ϕ_{nsp}) using cosine resemblance algorithm [159], formulated as follows:

$$\begin{aligned} \mathcal{S}_{sp}(\phi) &= cos(\phi, \phi_{sp}) - cos(\phi, \phi_{nsp}) \\ &= \frac{\phi^T}{\|\phi\|}\left(\frac{\phi_{sp}}{\|\phi_{sp}\|} - \frac{\phi_{nsp}}{\|\phi_{nsp}\|}\right) \end{aligned} \tag{5.6}$$

where $\| \, . \, \|$ denotes the euclidean distance.

5.1.2 Experimental setup

In this work, we use labeled speech data from 1377 males and 1323 females and more than 1500 non-human speech data (helicopter cockpit noise) from the SUSAS database

TABLE 5.2: he performance comparison result of the proposed SAD system (% Error rate). The performance is presented for different system and speech duration.

System	10-sec	30-sec	60-sec
SSC without SAD system	41.584	44.660	49.675
SSC with Baseline SAD system	32.673	33.981	36.526
SSC with Proposed SAD system	29.703	29.773	29.870

Note: SSC is stress speech clustering system, include ASR system.

that consisting of high noise, low noise and silence. For evaluation, we use the unlabeled conversation speech data as the audio target. The annotations-based ground truth is used to evaluate the SAD system.

We extract the MFCC feature of the speech at 10ms frame rate with 25ms window size. The 13-dimensional MFCC is used as the input of i-vector extractor to produce an frame-level i-vector feature. The UBM super-vector contains 2048 Gaussian mixtures is applied to produce a 600-dimensional i-vector.

We set the EDA's weight balancing parameter for softmax loss and center loss $\lambda = 10^{-2}$ and the controller parameter for learning rate of center $\alpha = 10^{-1}$.

5.1.3 Result and discussion

The effectiveness of the proposed SAD system (EDA-based SAD system) is evaluated in the task of classification of the SUSAS dataset that presented for the SSC system. We perform EDA and LDA as channel compensation method in the proposed SAD system and the baseline SAD system, respectively.

In this experiment, EDA and LDA reduce the vector dimension from 600 (i-vector) to 400. The performance comparison result of the proposed SAD system in terms of EER is shown in Table 5.2 and the example of data segmentation is presented in Fig. 5.2. Generally, all systems present increased error for long speech duration. The proposed SAD system outperforms baseline systems and relatively stable in all speech duration. This indicates that by using EDA, the proposed SAD system able to capture speech information in a short and long temporal context.

Fig. 5.2 shows the proposed SAD system decision results compared with the baseline SAD system, with a correspondence confidence threshold score. Fig. 5.2(a) shows the original speech signal that has 3 types speech/non-speech conditions (silent: 0 to 0.25 seconds, speech: 0.26 to 0.95 seconds, high noise: 1.1 to 1.2 seconds, and low noise: 1.25 to 1.5 seconds). In Fig. 5.2(b) and (d), the SAD score is close to 0 for the silent condition, the negative score for noise conditions, and the positive score for speech conditions. Since EDA able to handle the variations of temporal dependencies, the proposed SAD system presents more sensitive SAD scores (Fig. 5.2(d)). Since LDA uses Gaussian approach to present the speech, some speech frames (Fig. 5.2(c) time: 0.25 to 0.35) are misrecognized as the non-speech.

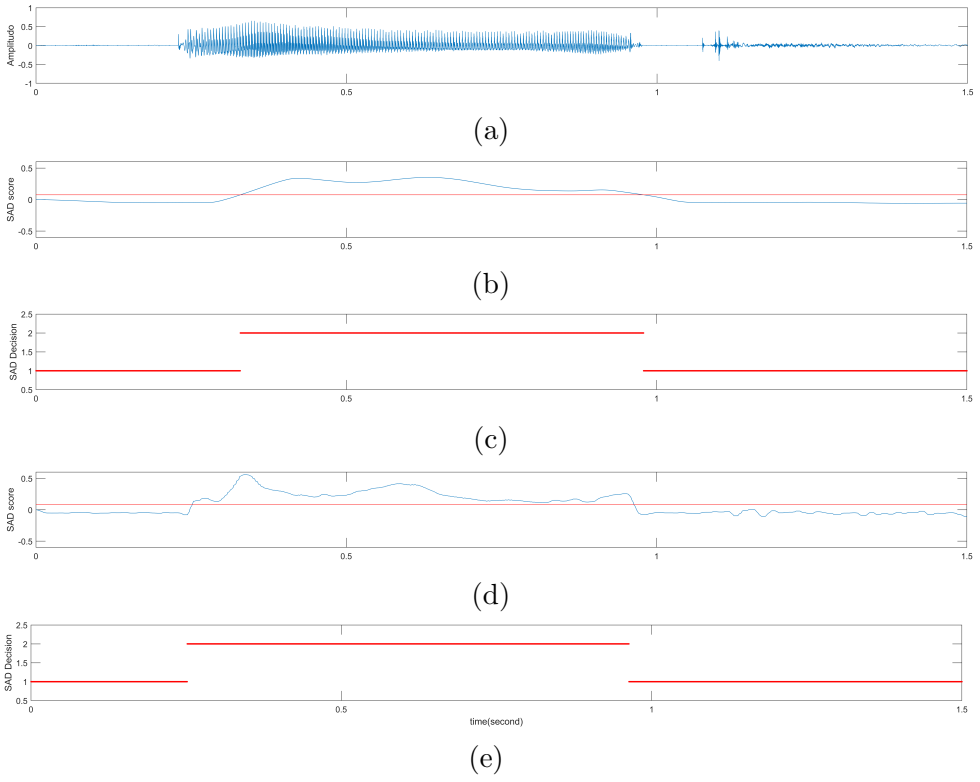


FIGURE 5.2: The example of data segmentation in the time domain, and the corresponding SAD decision. (a) demonstrates the original signal in 1.5 seconds with frame-scale $fs = 8000$. (b) and (d) denote the SAD score for the baseline and proposed SAD system, where the blue indicates the SAD score and the red line is a correspondence threshold. (c) and (e) present the SAD decision for the baseline and proposed SAD system, respectively.

Unlike the LDA-based SAD system, the proposed SAD system shows its capability in capturing the nonlinear relationship between features without requiring the prior assumptions on the input. While LDA probably less adequate due to speech under stress is not always distributed in Gaussian and it may have different covariance. Since EDA notices emotional information on the speech, the proposed SAD system obtains a more accurate in the SAD score that followed by an accurate decision threshold, as shown in Fig. 5.2. It impacts on the ability to distinguish the different non-speech types, such as silent, high noise, and low noise.

5.1.4 Conclusion

In this paper, a compact speech activity detection (SAD) system using i-vector and proposed embedded discriminant analysis (EDA) has been presented. The propose SAD system was not only strong in noisy environments, but also able to compensate the presence of emotional conditions. In the training phase, the speech and non-speech features were extracted using Mel Frequency Cepstral Coefficients (MFCC) technique that was

then transformed to frame-level feature by the i-vector extractor. The proposed EDA transformed the i-vector features of speech and non-speech into denoise space that is supervised by softmax loss and center loss. In order to compensate emotional conditions, EDA was trained using labeled short speech data of SUSAS database to produce a projection function and was used for generating the speech/non-speech models. In the testing phase, the cosine similarity algorithm was used to computes the deviation between the speech/non-speech models and the audio target. The effectiveness of the proposed SAD system was evaluated in terms of equal error rate (EER) and compared to the baseline SAD system as a pre-processing part of the stress speech clustering (SSC) application. Based on the experiment, since EDA able to captured speech information in dynamics temporal context, the proposed SAD system presented a stable EER in short and long speech durations. The proposed SAD system also presented more sensitive SAD scores in different non-speech conditions so that resulting in an accurate decision threshold.

5.2 Speaker verification

Every utterance is not produced in a similar form in all respects [24], even by the same speaker. The voice characteristics may change due to a deviation of the articulator movements caused by various factors, such as stress condition. Moreover, stress has diverse characteristics and different pattern for each individual. Therefore, we apply a speaker verification system in the pre-processing phase.

Speaker verification is the process of accepting or rejecting the identity claimed by a speaker. Most of the applications in which voice is used to confirm the identity of a speaker are classified as speaker verification. In speaker verification, an utterance that spoken by an unknown speaker is compared with a model from the speaker whose identity is being known. If the match is good enough, that is, above a threshold, the identity claim is accepted. Since there are only two choices, acceptance or rejection, speaker verification performance approaches a constant independent of the size of the samples. However, the emotional conditions (especially stress) make the pattern of the fundamental characteristics of speakers changed.

Psycho-physical studies have shown that people sensation of the frequency contents of sounds for speech signals follow a nonlinear scale. In other words, each tone with an actual frequency measured in Hz, is measured in a subjective pitch scale called the "Mel" scale. The robustness of pitch and Mel Frequency Cepstral Coefficients (MFCCs) features in recognizing the speaker in emotional condition has been explored by many works such as [176–180].

The SUSAS database is recorded from the conversation between a pilot and a co-pilot of the Apache helicopter. It means that the conversations are spoken by two-speakers.

Thus, we design a speaker verification system that assigns to separate the time-series utterances and grouping them correspond to the speaker by applying the similarity algorithm. Similarity algorithm is simple and effective matching method that has been widely used in many speech recognition task [181–183]. We investigate three similarity algorithms: Euclidean distance, Mahalanobis distance, and Manhattan distance.

5.2.1 The proposed method of speaker verification

Figure 5.3 shows the proposed method of speaker verification. Each conversation contains the time-series utterances (words). The goal of this work is to group the utterances based on the speaker (speaker-1 or speaker-2). To extract the speaker feature vectors, we use the frequency fundamentals pitch and the MFCC technique. Furthermore, we perform the same technique to the second utterance. We decide that the first speech is spoken by speaker-1. We then measure the similarity between the first and the second feature vectors using the feature-matching algorithms. Euclidean, Mahalanobis, and Manhattan distance algorithm are explored and evaluated their performance in distinguishing the speaker. The standard deviation is used as a threshold. Specifically, if the distance between the two feature vectors is less or equal to the standard deviation, the second utterance is spoken by speaker-1, otherwise, it is spoken by speaker-2.

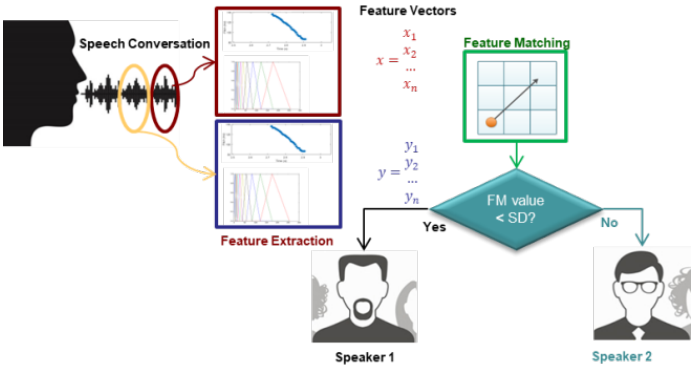


FIGURE 5.3: The method of speaker verification.

5.2.2 Experimental setup

In the preliminary experiment, the standard deviation was compute for each distance algorithm, as follows:

- Euclidean Distance = 0.0383,
- Mahalanobis Distance = 0.0254,
- Manhattan Distance = 0.0341.

10-feature vectors of pitch and 13-feature vectors of MFCC are used as a speaker feature vectors.

5.2.3 Result and discussion

In the feature extraction process, each speech is extracted its pitch and MFCC features. The feature vector sample from the feature extraction process is shown in Figure 5.4.

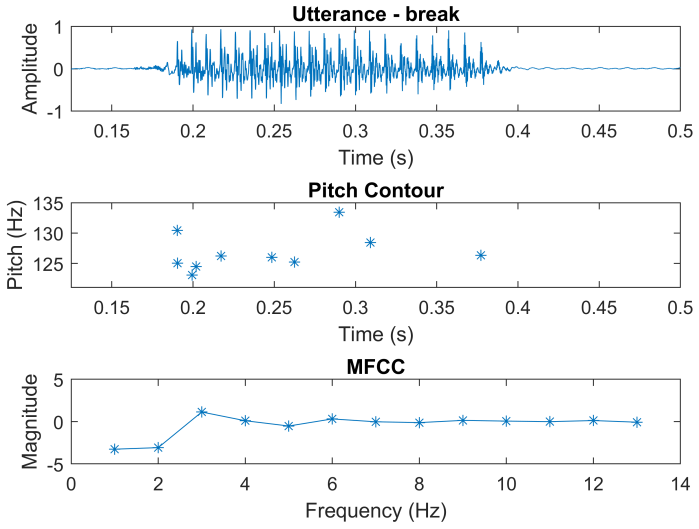


FIGURE 5.4: The feature vector of the utterance "break". Pitch feature is 10-feature vectors and 13-feature vectors for MFCC.

After the thresholding process performed, we evaluate the effectiveness of the speaker verification method. We compare the effectiveness based on the features extraction technique and the distance algorithm used. The comparison performance of feature combination and distance algorithm in the task of speaker verification is shown in Figure 5.5.

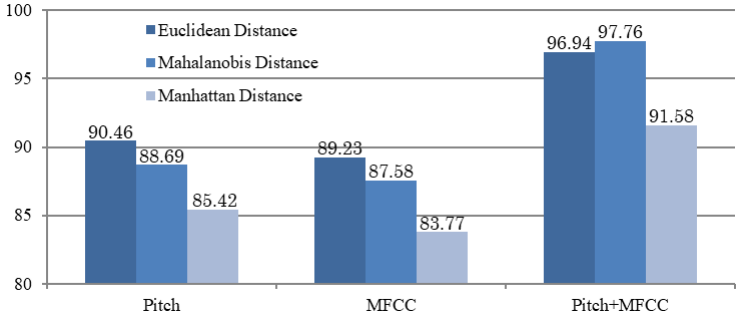


FIGURE 5.5: The comparison performance of feature combination and distance algorithm in the task of speaker verification.

5.2.4 Conclusion

The experiment involved three conversations between two speakers. The system accuracy rate was calculated from the number of words grouped on each speaker. We used pitch and MFCC feature extraction techniques in representing the speaker. Three distance algorithms were explored in the task of computing the distance of two feature vectors. The experimental result shows that Euclidean Distance is better for single feature extracted and Mahalanobis Distance is better for multi-features.

5.3 Gender identification

Naturally, males and females have different speech characteristics caused by their physiological, acoustical, and perceptual differences [184]. Therefore, many speech-based systems, such as speaker verification [185], speech recognition [186], and emotion recognition [187–190], apply a gender identification system in their pre-processing phase to address this gender-dependency phenomenon.

Gender identification is a system that determines the sex of the speaker through speech signals analysis. There are many feature extraction technique can be used to identify the gender in speech. For instance, the Mel-Frequency Cepstral Coefficients (MFCC) [191–193] is commonly used for extracting gender-related features [194] in a particular speech segment. By assuming that the speaker is emotionally neutral, MFCC is able to identify gender effectively. However, it should be noted that voice characteristics may change due to a deviation of the articulator movements caused by a stressful condition [195]. Moreover, males and females respond to stress with different expressions [196]. Hence, stress condition is an important characteristic that should be considered in gender identification system.

On the other hand, a robust feature extraction model (known as i-vector) has been successfully applied to represent gender's features [99] [100]. I-vector is a state-of-the-art feature extraction model in speech-based applications due to its robustness for modeling the intra/inter-domain variabilities. By offering an effective way of representing speaker-specific models, i-vector is also promising for recognizing the speaker's emotional state. By following a Linear Discriminant Analysis (LDA) as a compensation method [98, 100], i-vector was able to discriminate the speaker's gender with high accuracy. The use of machine learning techniques such as Support Vector Machine (SVM) and Artificial Neural Network (ANN) to model a speaker's gender has been explored [191–193]. However, since stress makes gender information distributed non-normally in a long temporal context [197], these approaches become ineffective for recognizing the gender of a stressed speaker.

To address this, the combination between i-vector and a special type of Neural Network architecture, known as the Long Short Term Memory (LSTM), is proposed to identify the speaker gender. LSTM applies a loop connection to its network so that the network able to learn gender information dependency and process entire sequences of data [198]. Thus, the gender identification system has an ability to remember information over arbitrary time intervals to retrieve the gender information in a long temporal context.

5.3.1 Gender identification

As shown in Figure 5.6, the gender identification consisted of three main parts: acoustic feature extraction, i-vector feature extraction, and LSTM modeling. The acoustic features are extracted using Mel-frequency Cepstral Coefficients (MFCC) technique on each segment. The GMM-UBM framework is applied to transform the acoustic features into super-vector space. The joint factor analysis (JFA) decompose a super-vector into a low-dimensional set of components. Then, all components are modeled into a single low-rank factor by the total variability model (TVM) to obtain the i-vector feature. In the back-end phase, the LSTM is used to dynamically model the speaker's gender by considering long-term context dependencies.

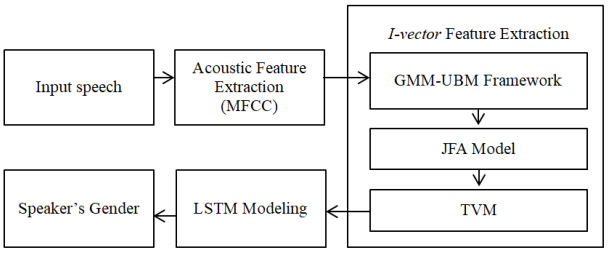


FIGURE 5.6: The gender identification method
GMM-UBM: Gaussian Mixture Model-Universal Background Model

We use the MFCC technique to extract acoustic features. Then, to extract the i-vector feature on each speech segment, the first-order derivatives of MFCC are used. The UBM is a large GMM trained to represent the gender-independent distribution of the features. The Maximum a Posteriori (MAP) algorithm is used to model the gender information in super-vector space by adapting the UBM mean parameters. A super-vector s is decomposed by JFA into four components: gender-independent, gender-dependent, channel-dependent, and residual. TVM is used to represent four components into a low-dimensional total variability factor w, known as the i-vector feature. The total variability matrix T, expressed as follows:

$$s = m + Tw \tag{5.7}$$

where m is the gender-independent mean super-vector (from UBM).

Furthermore, a time-series of i-vector features is used as the input of LSTM to model the gender. We design an LSTM network that has four main parts: a sequence input layer, an LSTM layer, a classification layer, and an output layer, as shown in Figure 5.7.

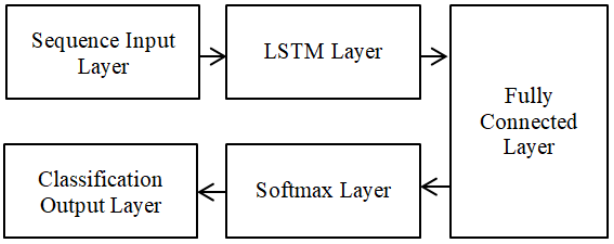


FIGURE 5.7: The architecture of LSTM network

The last layer of the LSTM network is a softmax function that is used to compute probabilities for each class. The output of the Softmax function represents the distribution probability $P(z_j)$ of the possible output class $z_{j,\{j=1...k\}}$, where k is the number of class ($k = 2$, female and male), formulated as follows:

$$P(z_j) = \frac{e^{z_j}}{\sum_{k=1} e^{z_k}} \tag{5.8}$$

5.3.2 Experimental setup

To evaluate the effectiveness of the proposed method, we used speech data from SUSAS. The labeled short utterances data consist of 294 female and 306 male utterances. We used six unlabeled conversations with three conversation types (single speaker, two speakers of the same gender, and two speakers of different genders). Total, all conversations contained 50 female and 566 male utterances that presented time-series (sequences) data with different durations.

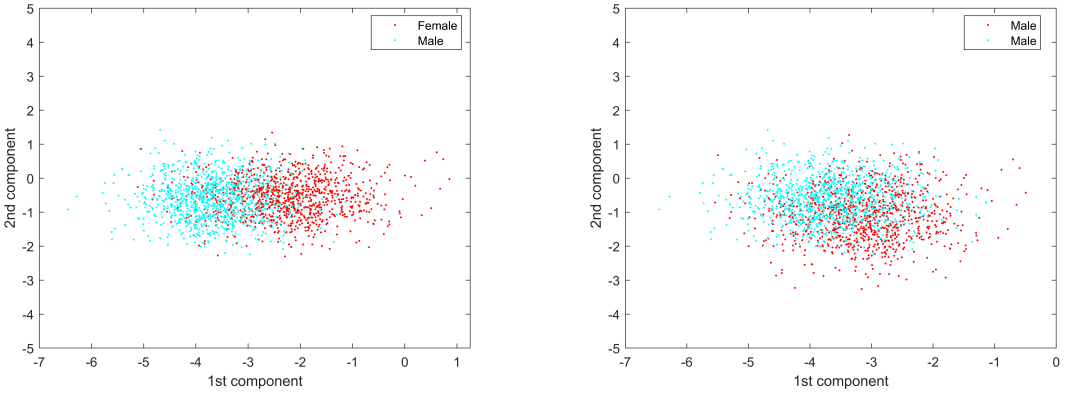
The acoustic features of the time-series input are extracted using the MFCC technique. The sequences of input feature is represented in matrices $M(i, j)$ with $i_{\{i=1,2,...,12\}}$ rows (one row for each feature) and $j_{\{j=1,2,...,t\}}$ columns, where t is number of time-steps (one column for each time-step). The matrix elements denote the Mel variable value of the speech for i feature in j time-step.

I-vector features were extracted from the 12-dimensional MFCC coefficients. We used the gender-dependent UBMs that contains 512 Gaussian mixtures. To model the gender-dependent joint factor analysis (JFA), we trained the same amount of data as for the UBM training. 300 gender factors and 100 channel factors were used as the JFA configuration. These factors were computed from the same data that was used in UBM training. A 400-dimensional matrix T from SUSAS data was computed to obtain a variability matrix of i-vectors. Thus, a 400-dimensional i-vector was generated by adapting the UBM with the MAP algorithm.

A bidirectional LSTM network was used to handle a 400-dimensional input of the i-vectors feature for modeling the speaker's gender. The LSTM network consisted of 100 hidden units, a 2-dimensional fully connected layer, a softmax layer, and a classification layer. For the training phase, the LSTM network's learning rate was set to "0.01". To prevent the gradient explosion problem during training, we used a gradient clipping method by setting the gradient threshold to "1". Furthermore, the stochastic gradient descent with momentum (SGDM) solver was used to optimize the network parameters.

5.3.3 Result and discussion

We visualized the gender vectors of speech samples using a Gaussian distribution model. The vectors were generated from two-bivariate Gaussian mixture distributions that were then modeled into a fit Gaussian mixture. The Gaussian mixture was presented in two-components. Each component was extracted into 1000-vectors that modeled into fitted Gaussian mixture contours, as shown in Figure 5.8.



(a) 2^{nd} conversation (b) 5^{th} conversation

FIGURE 5.8: The fitted Gaussian mixture contours. (a) and (b) denote the visualization of sample data from 2^{nd} and 5^{th} conversation of SUSAS, respectively.

We evaluated the effectiveness of gender identification in identifying the speaker's gender in four experiments that involved six conversations. Each conversation had different durations and different conversation types: single speaker (1^{st}, 3^{rd}, and 4^{th}), two speakers of the same gender (5^{th} and 6^{th}), and two speakers of different genders (2^{nd}).
Table 5.3 shows that the softmax function presented high confidence for all conversation data, meaning that the gender identification method was able to recognize the speaker's gender in all conversation types (single speaker, speakers with the same gender, and speakers with different genders).
We then evaluated the gender identification method's in term of error rate, as shown in Table 5.4. The error rate was computed "1" minus the dividing between the number of correct classifications of the speaker's genders (Spk_1 and Spk_2) and the total number of

TABLE 5.3: The speaker's gender decision based on softmax probability

Data	System Decision		Actual		Softmax decision probability
	Spk_1	Spk_2	Spk_1	Spk_2	
1	Male	-	Male	-	0.9954
2	Male	Female	Male	Female	0.9819
3	Male	-	Male	-	0.9851
4	Male	-	Male	-	0.9862
5	Male	Male	Male	Male	0.9934
6	Male	Male	Male	Male	0.9877

test data Y_{test}, where the number of the correct classifications is the number of utterances where the predicted class Y_{pred} and the actual class Y_{act} were the same.

TABLE 5.4: The error rate of the gender identification method

Data	Data distribution		Y_{test}	$Y_{pred} = Y_{act}$	Error rate (%)
	Spk_1	Spk_2			
1	98	-	98	97	1.03
2	52	50	102	100	1.97
3	94	-	94	93	1.07
4	118	-	118	117	0.85
5	56	51	107	104	2.81
6	40	57	97	95	2.07
Overall			616	606	1.63

We also evaluated the effectiveness of the gender identification by comparing it to the baseline methods. Table 5.5 shows that the i-vector/LSTM error rate reached 1.63%, outperforming the baseline methods. I-vector-based methods showed their robustness in identifying the speaker's gender by overcoming the acoustic feature-based method deficiency. Since the inputs were sequences, the RNN and LSTM were more effective than the LDA due to their capability to memorize the information over arbitrary time. The LSTM capability of learning the long-term dependencies was evidenced by outperforming the accuracy of the RNN.

5.3.4 Conclusion

We proposed a new approach to gender estimation on speech conversation using long short-term memory (LSTM) based on MFCC and I-vector feature extraction. We evaluated the proposed method using six conversations from the SUSAS database spoken by two speakers. The experiment results showed that the proposed method had the error rate at 1.63%.

TABLE 5.5: The comparison performance of gender identification methods in terms of error rate

Method	Error rate (%)
MFCC/ANN	20.56
MFCC/RNN	16.83
MFCC/LSTM	14.58
i-vector/LDA	7.84
i-vector/RNN	3.05
i-vector/LSTM	1.63

Chapter 6

Stress and Emotions Speech Clustering

As described in Chapter 3, a large and relevant dataset are required to build an SSR system that robust in real condition. To address this issue, we use a clustering approach to group stress speech data in a self-learning manner.

Clustering methods refer to unsupervised settings. It means that no labels are given in the learning process, which involves inferring the patterns within datasets without reference to known outcomes or labels. By defining an effective objective in a self-learning manner, clustering methods have been successfully used in many pattern recognition systems, included to categorize stress speech data [35–37]. Typically, clustering methods use a distance or similarity algorithm to group the data points. The effectiveness of distance or similarity algorithms has been studied by [36, 199–202]. However, similarity algorithms generally deteriorate the performance in high-dimensional data. This problem is known as the curse of dimensionality [203].

To this end, the deep clustering methods use DNN-based autoencoder to overcome the curse of dimensionality problem by transforming the original features to a lower-dimensional feature representation (embedding feature). The deep clustering algorithm learns feature representation and clustering assignments simultaneously by supervision of its objective loss functions [204].

In this chapter, we discuss the deep clustering of stress and emotions in two approaches: unsupervised (Chapter 6.1) and semi-supervised (Chapter 6.2).

6.1 Unsupervised stress and emotions speech clustering

Lately, a deep clustering algorithm had shown their superior performance by applying a DNN-based autoencoder that simultaneously learning the clustering assignment [131] deeply. The autoencoder maps the non-linear input parameters by transforming a small

region of each temporal context window [205, 206] into the feature space. However, it should be noted that stress speech information is non-linearly distributed at long and short temporal context [207, 208]. Therefore, we propose to use another DNN type in building the autoencoder that has the ability to learns a temporal context-dependency, know as the time-delay neural network (TDNN). TDNN able to create more large networks from sub-components across time steps.

The deep clustering algorithm works by optimizing their clustering objective iteratively with a self-training target distribution [209]. To toward their objective, a deep clustering algorithm typically applies a network loss [204] and clustering loss [210], and enhance their performance by adding the perspective of the reconstruction loss function [204, 211]. Obviously, this strategy is strengthening the represented feature predictions by pushing the inter-cluster compactness. However, it causes the effect of the inter-cluster similarity ignored [212]. Thus, we propose to add a discriminative loss function to increase the distance of inter-cluster centroid and reducing the intra-cluster variations.

We introduce an unsupervised deep clustering algorithm to categorize the stress and emotions. We named the deep time-delay autoencoder embedded clustering (DTEC). DTEC learn and transform the speech feature from original space to embedding space using TDNN-based autoencoder and simultaneously optimizing the clustering objective by jointly supervision of discriminative loss, reconstruction loss, and clustering loss.

6.1.1 Deep time-delay embedded clustering

DTEC consists of two main phases: (1) TDNN-based autoencoder and (2) clustering objective function. TDNN-based autoencoder transforms a high-dimensional speech segment into a lower-dimensional feature in the embedding space. The clustering objective function optimizes the joint supervision of discriminative loss, reconstruction loss, and clustering loss. The architecture of the DTEC shows in Figure 6.1.

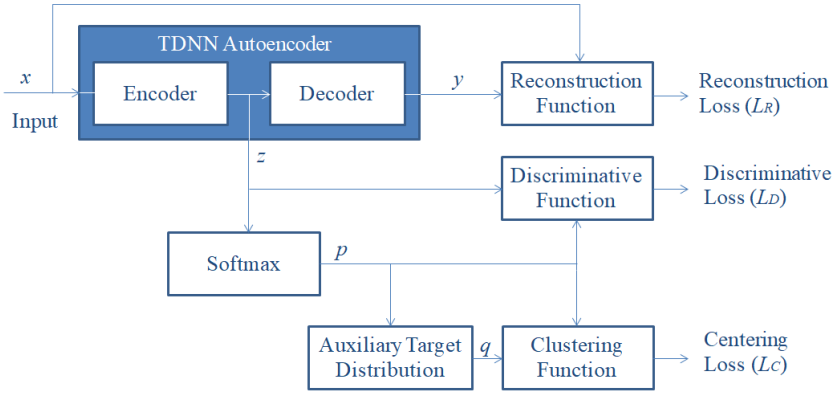


FIGURE 6.1: The architecture of the deep time-delay embedded clustering (DTEC)

Figure 6.1 shows that the TDNN-based autoencoder reconstructs a given input $x = \{x_{i=1...,n}\}$ into its output $y = \{y_{i=1...,n}\}$. The autoencoder composes of an encoder $z = f_W(x)$ and a decoder $y = g_{W'}(z)$. Reconstruction loss L_R compute the difference between the reconstructed feature y and the original input x using mean squared error (MSE), thereby the represented features z are represented properly. To reduce the intra-cluster distance and increase the inter-cluster distance, a discriminative loss L_D function is performed by optimizing the distances between the represented features z and the softmax predicted distribution p. The clustering loss L_C is taken by minimizing the cross-entropy loss between the softmax predicted distribution p and the auxiliary target distribution q.

6.1.1.1 TDNN-based autoencoder

The autoencoder composed of an encoder that adopts the sub-sampled TDNN structure of [173] and a decoder that constructed of the under-sampled reverse TDNN structure of [213]. We use three hidden layers in encoder/decoder pairs which cover the total input temporal context on [t-13,t+9] [173], as shown in Table 6.1.

Table 6.1: The TDNN-based autoencoder network structure

Function	Hidden Layer	Context
Encoder	Layer-1	[-4,+4]
	Layer-2	{-3,+3}
	Layer-3	{-6,+2}
Feature space z	Embedding	{0}
Decoder	Layer-$\hat{3}$	{-6,+2}
	Layer-$\hat{2}$	{-3,+3}
	Layer-$\hat{1}$	[-4,+4]

The rectified linear unit (ReLU) activation function is used in all encoder/decoder pairs [173] with a 256 batch size [214]. Each hidden layer and the embedding layer has dimension of 4000 and 400, respectively [215]. In table 6.1 (encoder), it can be seen that the layer-1 covers from [t-13] to [t+9] with context [t-4,t+4]. Thus, there are 3 activation units which sub-sampled [t-13,t-5], [t-6,t+2], and [t+1,t+9]. The layer-2 covers from [t-9] to [t+5] with context [t-3,t+3] so that has 3 activation units which sub-samples [t-9,t-3], [t-5,t+1], and [t-1,t+5]. The layer-3 covers from [t-6] to [t+2] that has an activation unit that sub-samples [t-6,t+2]. In table 1 (decoder), it can be seen that the decoder structure is a mirror of the encoder. The decoder structure uses a simple reverse TDNN architecture [213]. There is one unit in the layer$\hat{3}$ that covers from [t-6] to [t+2] so that 8-time steps trajectory is produced by the network. The layer$\hat{2}$ cover from [t-9] to [t+5] with context [t-3,t+3] so that has 3 units under-sampled [t-9,t-3], [t-5,t+1], and

[t-1,t+5]. In this layer, each unit produce 6-times trajectory. The layer$\hat{1}$ with context [t-4,t+4] so that has 3 units under-sampled [t-13,t-5], [t-6,t+2], and [t+1,t+9] which each unit produces 8-times trajectory.

6.1.1.2 DTEC's objective function

DTEC optimize the clustering assignments by joint supervision of discriminative loss L_D, reconstruction loss L_R, and clustering loss L_C, defined as:

$$L = L_D + \alpha L_R + \beta L_C \tag{6.1}$$

where α and β are reconstruction and clustering weight parameters.
The discriminative loss is used to optimize the inter/intra-cluster distance [212, 216]. The two feature points should be different if they belong to different clusters and should be similar if belong to the same cluster. Therefore, we formulate the discriminative loss in two-margin variables: distance margin λ_d and variance margin λ_v, that defines the inter-cluster push-force and intra-cluster pull-force, as follows:

$$L_D = \frac{1}{k(k-1)} \sum_{j=1}^{k} max(0,\ \lambda_d - \parallel \mu_j - \mu_m \parallel_2^2) + \mathcal{L} \tag{6.2}$$

where $m \in j, m \neq j$, k is number of cluster, μ_j is mean of p_i in cluster j $\{\mu_j, j = 1, ..., k\}$, and $\mathcal{L}$ define as,

$$\mathcal{L} = \frac{1}{k} \sum_{j=1}^{k} \frac{1}{N_j} \sum_{i=1}^{N_j} max(0,\ \parallel \mu_{ij} - z_{ij} \parallel_2^2 - \lambda_v) \tag{6.3}$$

where N_j is number of points in cluster j, z_i is a represented feature of x_i. We optimize the λ_d and λ_v with assume that are no longer repulsed and can proceed anywhere in the feature space.
We optimize the reconstruction loss using back-propagation through all of the networks. The reconstruction loss function is designed to regulate the loss features in a low-dimensional space z to ensure that the important information of the input x is well represented in term of the reconstruction features y. The reconstruction loss was computed by MSE, as follows:

$$L_R = \frac{1}{n} \sum_{i=1}^{n} \parallel x_i - g_{W'}(z_i) \parallel_2^2 \tag{6.4}$$

where $z_i = f_W(x_i)$, n is number of represented features, f_W and g'_W are encoder and decoder mappings function.
We obtain the cluster prediction probability for each data point from the softmax layer. Since the softmax loss function designed for classification problem [217], the clustering loss is computed from minimizing the cross-entropy loss between the softmax prediction

probability distribution p and the auxiliary target distribution q. The optimization process encourages the softmax layer output to focus on a high probability data. Thus, we define the clustering loss as follows:

$$L_C = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} p_{ij} log(q_{ij}) \tag{6.5}$$

where p_{ij} and q_{ij} are the softmax prediction probability and the auxiliary target distribution of z_i, respectively, that belongs to cluster $j\{j = 1, ..., k\}$, and q_{ij} defined as:

$$q_{ij} = \frac{p_{ij}^2 / \sum_i p_{ij}}{\sum_j (p_{ij}^2 / \sum_i p_{ij})} \tag{6.6}$$

6.1.1.3 DTEC's parameter optimization

We optimize Eq. 6.1 to evaluate the effectiveness of DTEC's objective function. The network contains a TDNN-based autoencoder and the softmax layer on the back-end of the encoder. The cluster assignments are represented as probabilities that were generated by the softmax layer. The encoder transforms the speech input x_i into the represented feature z_i. The decoder reconstructs the represented feature z_i into the reconstructed feature y_i and the softmax produces the class prediction p_i of the x_i. Then, the discriminative loss, reconstruction loss, and clustering loss are computed based on the class prediction p_i, the represented feature z_i, and the reconstructed feature y_i. The total loss was obtained by summing the weighted loss which serves as the reference of the parameters update. Iteratively, the parameters are updated using stochastic gradient descent (SGD) and backpropagation until the network convergence is reached.

6.1.2 Experimental setup

In the training and testing, we use six conversation data (conditioned) of SUSAS that have different duration. The acoustic feature is extracted using Mel-frequency cepstral coefficients (MFCCs) into a 13-dimensional vector with 25ms normalized frame length for each sliding window. On each frame, we append a 200-dimensional i-vector to the MFCC input.
The sensitiveness of the network parameters directly controls the behavior of the DTEC algorithm and has a significant impact on model performance. In the experiments, we set $\lambda_d = 0.1$, $\lambda_v = 0.05$, $\alpha = 0.97$ and $\beta = 0.98$

6.1.3 Result and discussion

we evaluate the effectiveness of DTEC in the clustering assignment of SUSAS data and compare it to several existing clustering methods, such as the standard k-means, the deep embedded clustering (DEC) [210], and the improve deep embedded clustering

(IDEC) [211]. We present the effectiveness in terms of clustering error rate (CER) [218] and normalized mutual information (NMI) [219], as presented in Table 6.2.

TABLE 6.2: The comparison of clustering performance in terms of CER and NMI

Algorithm	Number of cluster	CER (%)	NMI (%)
Standard K-means	4	54.49	45.72
DEC [210]	5	32.82	65.38
IDEC [211]	5	30.45	67.32
Method-1	5	30.43	67.95
Method-2	5	28.39	70.94

Method-1 denotes the DTEC algorithm without discriminative loss L_D. Method-2 denotes the DTEC algorithm with all losses.

Most of the algorithms are categorize into 5 clusters, except k-means is 4 clusters. The deep clustering algorithms outperform the standard k-means with a high gap. It proves that the feature space has great potential to represent the stress speech features. The reconstruction loss L_R shows their superior by reflecting the performance gaps between DEC and IDEC. Since minimizing the KL divergence and minimizing the cross-entropy are equivalent [220], our method-1 performance relatively similar to IDEC. The discriminative loss L_D shows their advantage by improving the performance of method-1 to method-2.

For further analysis, we identify the points that belong to the incorrect cluster using the cluster identification rate, as shown in Table 3. Two clusters have a low identification rate: clusters 2 and 4. Based on our observations, these two clusters are relatively similar emotional characteristics. We suspect they are "low stress" and "soft" [221].

TABLE 6.3: The DTEC identification rate

Cluster	Error rate (%)
1	21.02
2	39.07
3	22.45
4	37.21
5	23.97

6.1.4 Conclusion

In this chapter, we have presented an unsupervised deep clustering algorithm to categorize five classes of stress and emotions, named as deep time-delay embedded clustering (DTEC). DTEC transforms the feature from original space to embedding space and

simultaneously learns the clustering assignment. We designed the autoencoder based on the sub-sampled a time-delay neural network (TDNN) as the encoder and the under-sampled reverse TDNN as the decoder. An encoder/decoder pairs consisted of 3 hidden layers that were covered the temporal input context on [t-13,t+9]. The DTEC clustering objective was jointly supervised by the discriminative loss, reconstruction loss, and clustering loss. The discriminative loss was used to optimize the distance between the represented feature and the softmax predicted distribution. The reconstruction loss was used to minimize the differences between the original feature and the reconstructed feature. The clustering loss was optimized by minimizing the cross-entropy loss between the softmax predicted distributions and auxiliary target distributions. The effectiveness of DTEC was evaluated in clustering task of the SUSAS dataset. Based on the experiment, DTEC outperformed the baseline system in terms of CER and NMI of 2.9% and 5.3%, respectively. The deep clustering algorithms, such as, DEC, IDEC , and DTEC, categorized the stress speech data into five clusters. There were two clusters that have a low identification rate: clusters 2 and 4. These two clusters are relatively similar emotional characteristics. We suspect they are "low stress" and "soft".

6.2 Semi-supervised stress and emotions speech clustering

As discussed in Chapter 6.1, DTEC able to categorize the stress and emotions effectively. Despite its effectiveness, DTEC was not able to confirm yet that the output class corresponds to informational classes because of no measured outcome variable and information about the relationship between the observed clusters.

In many machine learning applications, prior information (annotations) is used to improve learning abilities to give a significant impact on the clustering task [222]. Some works defined the prior information (pairwise constraints) present a relationship between a pair of instances [223]. Typically, the pairwise constraint is divided into two types, the must-link and cannot-link constraints [224]. Must-link constraints are used to associate that two instances are known in the same cluster, while cannot-link constraints specify that two instances belong to different clusters. These constraints can be leveraged as a guide to finding the corresponding clusters [226].

Since DTEC did not use prior information yet to lead to a clustering procedure that is able to enhance the clustering process, we propose a semi-supervised framework of DTEC. We named it semi-supervised deep time-delay embedded clustering (SDTEC). SDTEC improves the effectiveness of DTEC by incorporating semi-supervised information. Specifically, the prior information of pairwise constraints is attached to SDTEC's learning process so that the distance of inter-clusters centroid is farther and the intra-cluster variations are closer. Thus, the state-of-the-art DTEC clustering performance significantly improved by using this prior information of pairwise constraints.

6.2.1 Semi-supervised deep embedded clustering

The SDTEC consisted of the nonlinear transformation via TDNN-based autoencoder and the stress speech recognition (SSR) model-based pairwise constraints, as shown in Figure 6.2.

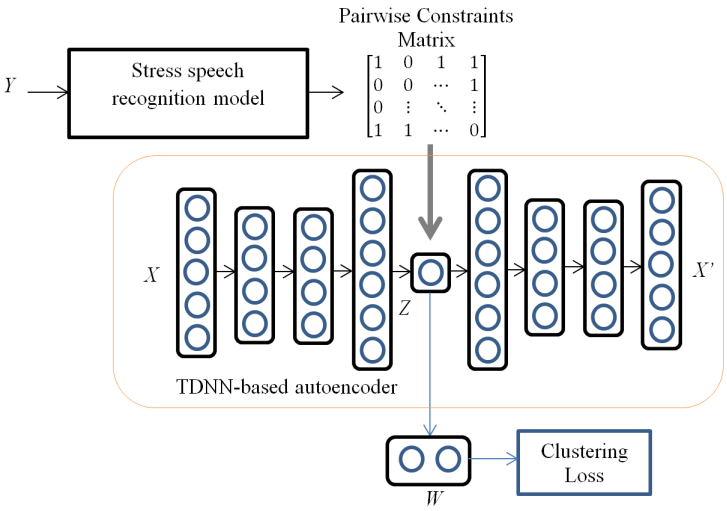


FIGURE 6.2: The architecture of the semi-supervised deep time-delay embedded clustering (SDTEC)

The TDNN-based autoencoder transforms a high-dimensional input data X into a low-dimensional embedding space Z. On the other side, we generated a pairwise constraints matrix from the SSR model, then incorporate it into the embedding feature Z and directly learns the feature representation. The soft assignment (probability of assigning) of each data point W is used to computed the clustering objective functions.

6.2.1.1 Nonlinear transformation

We use the DNN-based autoencoder structure (Chapter 6.1.1.1) to transform the data with a nonlinear mapping $f_\theta : X \to Z$, where θ is the model parameters representation, for representing a low-dimensional stress speech feature in embedding space.

6.2.1.2 SSR model-based pairwise constraint

The classification model able to confirm that the output class corresponds to informational classes because the measured outcome variable and information about the relationship between variables are provided. Therefore, we explicitly take the prior information of pairwise constraints from the SSR model (4.2).
As presented in Chapter 4.2, during the training process of SSR, the softmax loss was used to optimize the network parameters. The parameters were defined via a linear transformation with weight and bias vectors that are followed by the softmax function and the multiclass cross-entropy loss. Furthermore, after the training phase, the softmax layer

represents the probability of a sample belongs to the class labels $P(class|y_1, y_2, ...y_N)$. The whole softmax layer outputs are the distribution of possible clusters given a sample. For initial instances, the pair of must-link and cannot-link constraints of the softmax output distributions are defined by $M = \{(y_i, y_j)$: y_i and y_j belong to the same cluster$\}$ and $C = \{(y_i, y_j)$: y_i and y_j belong to the different clusters$\}$, respectively. Each cluster is presented by a center $\mu_{j,j=1,...,K}$ where K is the number of clusters. Then, a matrix that present the prior information of pairwise constraints (must-link M and cannot-link C) is incorporated into embedding feature Z. The pairwise constraints matrix is defined as follows:

$$A_{y_i,y_j} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nn} \end{pmatrix} \tag{6.7}$$

where $y_{ik} = 1$ if y_i and y_k are assigned to the same cluster, $y_{ik} = -1$ if y_i and y_k are assigned to the different clusters, and $y_{ik} = 0$ if other entities.

6.2.1.3 SDTEC's objective function

We consider incorporating the prior information of pairwise constraints in DTEC's objective function for leading the clustering assignment and feature representation. By providing pairwise constraints, it can specify whether a pair of data examples should be associated in the same class (must-link constraints) or should not be associated (cannot-link constraints). Thus, the same label points become closer and the different label points are away from each other. We define the objective of SDTEC as follows:

$$L = L_u + \gamma L_s \tag{6.8}$$

where L_u is unsupervised loss (Eq. 6.1) and L_s is semi-supervised constraint loss. γ is the hyper parameter that is used to balance both functions. If $\gamma = 0$, SDTEC would be dropped to DTEC. L_s indicates the conformity between the embedding feature $\{z_i\}_{i=1}^n$ with the pairwise constraints matrix A_{y_i,y_j}, defined as follows:

$$L_s = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{n} y_{ik} \parallel z_i - z_k \parallel^2 \tag{6.9}$$

where n is the number of data points.

6.2.2 Experimental setup

In the initial state, the pairs of data points were selected to generate A_{y_i,y_j} matrix based SSR's softmax output distribution and ground truth distribution. If selected data point

relating to the same label with ground truth, it would obtain a must-link constraint. Otherwise, it is a cannot-link constraint. We specify the convergence threshold is 0.1% with 0.01 learning rate. The number of ground truth is used to set the number of clusters K. In this experiment, we set K to 5.

6.2.3 Result and discussion

We evaluate the effectiveness of SDTEC in the clustering task of the SUSAS dataset and compare it with several clustering methods. In this experiment, we set the SDTEC's hyperparameter γ to 10^{-4}, and the number of constraints n_{pc} to $1 \times n$. While the hyperparameters of the unsupervised loss are set the same to DTEC (Chapter 6.1.2). Each method was run independently 10 times, and the average of them is presented in Table 6.4.

TABLE 6.4: The clustering performance comparison

Method	Clustering Error Rate (% CER)
K-means (Km)	55.79
K-means+pairwise constraint (Km+PC)	58.95
K-means+autoencoder (Km+AE)	36.17
K-means+autoencoder+pairwise constraint (Km+AE+PC)	34.55
Self-organizing tree algorithm (SOTA) [226]	31.28
Deep embedded clustering (DEC) [210]	30.14
DTEC (Chapter 6.1)	29.93
SDTEC	24.22

The evaluation result shows that the embedding space is better to represent the feature than the original space, as shown in the performance gap between Km and Km+AE. It indicates that DNN offers a robust feature representation in favor of the clustering assignments. The DNN-based techniques (SOTA, DEC, DTEC, and SDTEC) outperform the k-means and its variants because the DNN was able to represent complicated patterns. SOTA achieves a higher error than other DNN-based techniques since all data were forced to spread on a two-dimensional space so that some important information are ignored. It shows that the updated feature representation based on clustering assignments learning is better feature representations for clustering. As shown in Km+PC and Km+AE+PC compared with Km and Km+AE, the clustering performance is improved by incorporating the pairwise constraints. It indicate that the prior information is an important factor to enhance the clustering performance. SDTEC is a state-of-the-art deep clustering method that outperforms the baseline methods and decrease the error rate of DTEC by 19%. Generally, prior information usage can improve significantly the clustering performance.

The advantage of the SDTEC is its ability to fix the points that belong to incorrect cluster by using the prior information of pairwise constraints. However, we also needed to identify the error rate on each cluster. Each cluster error rate identification for the DTEC and SDTEC are reported in Table 6.5.

TABLE 6.5: The SDTEC and DTEC identification error rate (%)

Cluster	DTEC	SDTEC
1	26.08	23.91
2	34.37	23.43
3	25.51	22.44
4	36.43	23.51
5	27.27	24.79

SDTEC demonstrated a lower error rate than DTEC in all clusters. In DTEC, there are two clusters that have lower errors than others, while in SDTEC, each cluster indicated has an almost similar error rate. It proves that by incorporating the prior information of pairwise constraints, the points that are spread in the sidelines of clusters area can be pull-forced belong to the correct cluster and the incorrect ones can be guided to find the correct cluster.

6.2.4 Conclusion

In this paper, a new stress speech clustering method was proposed, called semi-supervised deep time-delay embedded clustering (SDTEC). The proposed SDTEC able to improve the effectiveness of DTEC by incorporating semi-supervised information that used to guide the clustering procedure. SDTEC consisted of the DNN-based autoencoder and SSR model-based pairwise constraints. The autoencoder was used to transform the data with nonlinear mapping for representing more informative stress speech features in embedding space. Then, the SSR model-based pairwise constraints matrix is incorporated into embedding space and directly learn of feature representation. The semi-supervised constraint loss and the unsupervised loss were used simultaneously to supervise the feature representation and the clustering assignment. The effectiveness of proposed SDTEC was evaluated by comparing it with state-of-the-art clustering methods such as K-means and its variants, SOTA, DEC, and DTEC in terms of clustering error rate (CER). Based on experiment results, the proposed SDTEC gave its best performance by outperforming all baseline methods and able to reduce the clustering error rate of DTEC by 19%. Compared to DTEC, SDTEC has the ability to fix the points that belong to the incorrect cluster by using the prior information of pairwise constraints. SDTEC demonstrated a lower error rate than DTEC in all clusters. Different from DTEC, SDTEC presented

almost similar error rates in all clusters. It proves that by incorporating the prior information, the floated points can be pull-forced belong to the correct cluster and the incorrect ones can be guided to find the correct cluster.

On the other hand, some works showed that the change rate of prosodic features between the target and the prior utterance able to deal with larger sets of contextual information. Hence, to enhance the clustering performance, we interest to leverage the emotional transition modeling approach as a future work.

Chapter 7

Stress and Emotions Speech Prediction and Modeling

To recognize the stress and emotion, most of the existing methods only observe and analyze the speech pattern from the present-time features. However, in real condition, an emotion (especially for stress) could change suddenly because triggered by an event during speaking. Therefore, we argue that the prior emotional state should also be observed so that the emotion of the speaker could be recognized more accurately.

Recognizing the emotion using its state transition has been studied by [41, 134–137] and successfully presented a high accuracy. Markov model has been widely used in various fields of prediction and forecasting [138]. It is caused Markov model offers convenience in modeling the temporal context of time-series (continuous) data [137, 139]. Despite its effectiveness, the Markov model susceptible miss the critical long-term effects [140] due to there are locally dependencies and distance between consecutive hidden states. Thus, we argue that long-term temporal dynamics should be observed more deeply.

The most popular technique today that has a supremacy ability in deeply learning a complex pattern is a deep neural network (DNN). To process a wider temporal context, DNN learns an affine transform for the entire temporal context since the initial layer [173]. In contrast, a time-delay neural network (TDNN) creates more large networks to learns the dependencies inter-contexts.

Thus, we propose to develop a system that able to predict and model the stress and emotions by analyzing its speech features and the prior emotional state. We named this method as the deep time-delay Markov network (DTMN). Structurally, the proposed DTMN contains a Markov model represented by the hidden Markov model (HMM) and a neural network denoted as the time-delay neural network (TDNN). We evaluated the effectiveness of the proposed DTMN by comparing it with several state transition models in the prediction task of the SUSAS dataset. By considering the prior emotional state, the proposed DTMN able to predict the present emotional state accurately. Based on

the prediction result, we then model the tendency of male and female emotional state transitions using a Markov chain.

7.1 Proposed system

The DTMN structurally consists of a Markov model that is denoted by the hidden Markov model (HMM) and a neural network that represented by the time-delay neural network (TDNN), as shown in Figure 7.1. The framework is performed in three phases: the training phase, the prediction phase, and modeling phase.

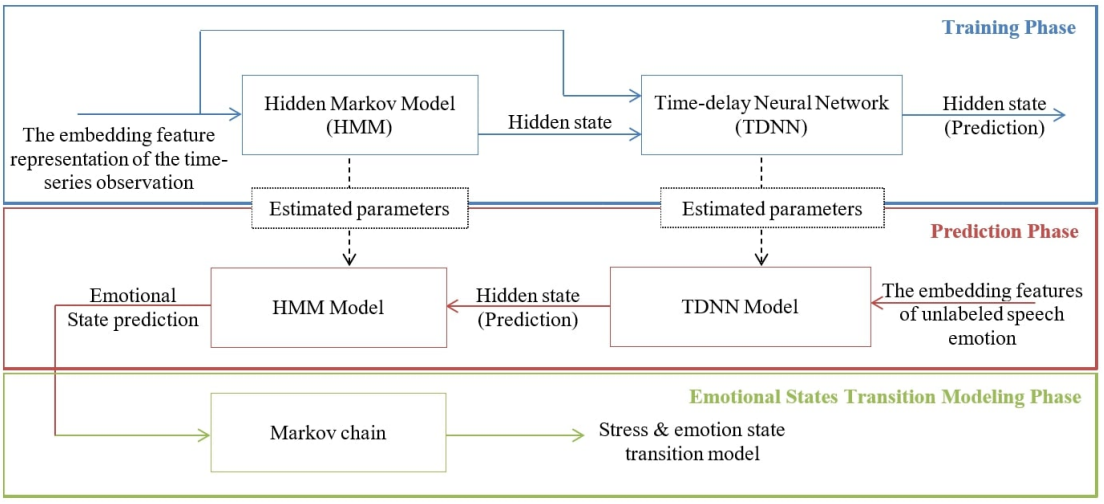


FIGURE 7.1: The proposed framework of deep time-delay Markov network (DTMN). The colored blue indicates the training phase, the color red denotes the prediction phase, and the colored green is the modeling phase.

In the training phase, we train the HMM using the time-series observation to produces the transition probabilities and the hidden states at each time-step. By using the hidden state obtained from HMM and the present speech features, TDNN is trained to predict the present hidden states. After the training phase, the estimated parameters of HMM and TDNN are obtained.

In the prediction phase, the DTMN is used to predict the emotional state of the unlabeled observations. we conduct an opposite procedure with the training phase. First, the TDNN model predicts the present hidden states using the present speech features as input. Then, the HMM model predicts the emotional state of the unlabeled observations using the predicted hidden states.

In the modeling phase, we model the transition pattern of emotion using the Markov chain with the emotional state prediction as input. The Markov chain models five emotional states; high stress, low stress, neutral, soft, and angry.

7.1.1 Deep time-delay Markov model

7.1.1.1 Hidden markov model

Hidden Markov model (HMM) is a Markov chain whose internal state cannot be observed directly but only through some probabilistic function. In other words, the internal state of the model alone determines the probability distribution of the observed variables. This unobservable state is known as the hidden state. The advantage of the hidden states does not need to emphasize about discretization and normalization issues so that we can deal with an arbitrary observation. In addition, the random noise in the observation could be handled by the hidden states. Therefore, the proposed DTMN uses the representation of the hidden states for connecting between observations.

$$\begin{aligned} A &= [a_{i,j}] = P(q_t = i | q_{t-1} = j) \\ E &= [e_{i,j}] = P(y_t = i | q_t = j) \end{aligned} \tag{7.1}$$

where $i, j = \{1...N\}$. Each a_{ij} representing the probability of transition from state i to state j and each e_{ij} expresses the probability of y_t being generated from a state j.

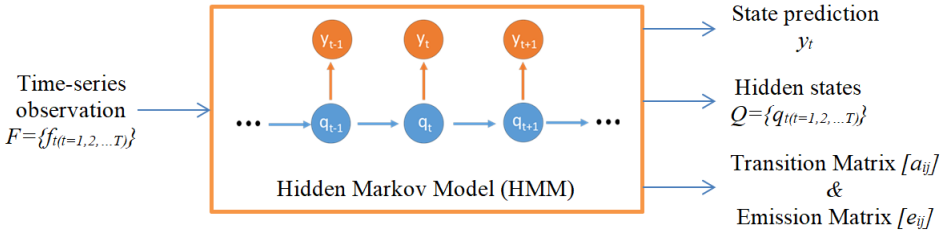


FIGURE 7.2: The hidden Markov model (HMM) training phase

7.1.1.2 Time-delay neural network

We use a fixed-dimension size of convolution networks (known as the time-delay neural network or TDNN) to predict the present hidden states. TDNN is a multilayer artificial neural network architecture that uses modular and incremental design to create more extensive networks from sub-components. It makes TDNN effective in learning the temporal dynamics of the signal even for short-term feature representation [173]. Unlike a standard DNN, in processing a wider temporal context, the first layer of TDNN learns the context in a narrow temporal and continues to the deeper layer. Distinctively, TDNN receives input not only from the hidden state representation at the below layer but also from the activation pattern of the unit output and its context.

In this paper, TDNN is used to model the relation between the hidden states and the observations by applying the relation of the hidden state and the labels (Eq. 7.1).

Concretely, TDNN predicts the present hidden state q_t by taking as input the prior hidden states $q_{t-1...N}$ and the present features f_t. The structure of TDNN is shown in Figure 7.3 and each layer function is summarized in Table 7.1.

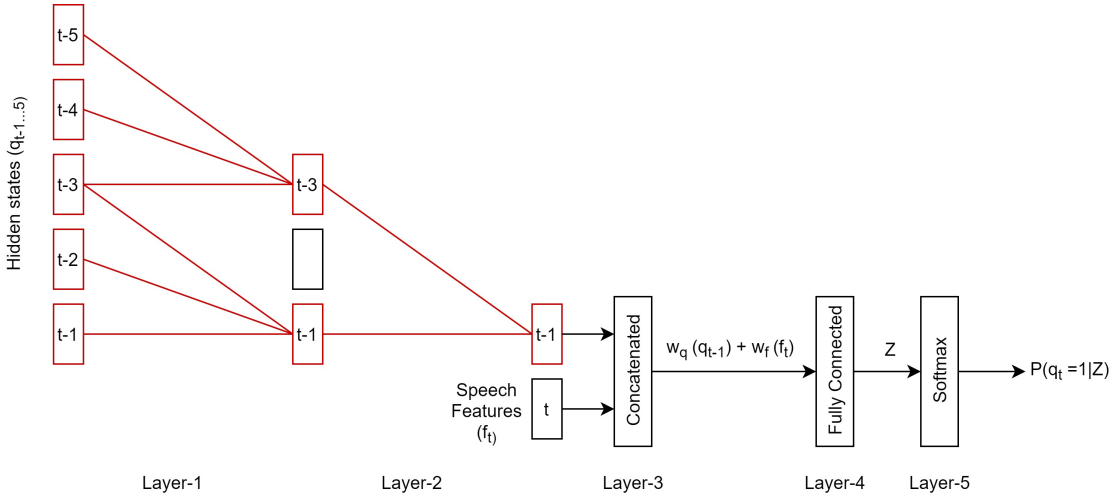


FIGURE 7.3: The structure of the TDNN.

TABLE 7.1: The TDNN layer temporal context structure

Layer	Feature context	Function
Layer-1	$[q_{t-5}, q_{t-1}]$	Without sub-sampled
Layer-2	$\{q_{t-3}, q_{t-1}\}$	Sub-sampled
Layer-3	$\{q_{t-1}, f_t\}$	Concatenated
Layer-4	$\{0\}$	Fully connected
Layer-5	$\{0\}$	Softmax

As shown in Figure 7.3 and Table 7.1, we designed a TDNN with five layers. Layer-1 holds full temporal contexts of prior hidden state from q_{t-5} to q_{t-1} that splices together frames $[0, -2]$. In the Layer-2, we apply the sub-sampling technique (locally connected) [215] so that only two temporal contexts (q_{t-3} and q_{t-1}) are held. Then, we concatenate the present speech features f_t and q_{t-1} feature from the second layer in the Layer-3. A fully connected and softmax layer are performed in the Layer-4 and Layer-5 of TDNN, respectively. A softmax function is used to define the probability by taking a C-dimensional vector Z (from Layer-4) as input and outputs C-dimensional vector τ (real values between 0 and 1). The normalized exponential of the softmax function is expressed as follows:

$$\tau = P(q_t = i|Z) = \frac{e^{Z_c}}{\sum_{d=1}^{C} e^{Z_d}} for d = 1...C \tag{7.2}$$

where $Z = w_q^i, \alpha(q_{t-1}) + w_f^i, \beta(f_t) + b$. w_q and w_f are the coefficients to be estimated. α and β are the functions that is used to transform q_{t-1} and f_t into feature vectors.

We perform binary approach to $\alpha(q_{t-1})$ by assuming that the coordinate of $q_{t-1^{th}} = 1$ and others are zero. The denominator $\sum_{d=1}^{C} e^{z_d}$ is a regularizer that aims to ensure $\sum_{c=1}^{C} \tau = 1$.

7.1.2 Training phase

In the training phase, DTMN is trained to obtain estimated parameters of HMM and TDNN. We perform the training phase in two steps. As shown in Figure 7.2, the first step is to estimate the hidden state q_t based on the labels y_t using the Baum-Welch algorithm, and at the same time, the transition matrix A and emission matrix E are estimated.

After q_t are estimated, the second step is to estimate the parameter of TDNN. We use the structure of the TDNN (Figure 7.3) in the task of supervised prediction. TDNN is trained to predict the hidden state q_t on each time step. Iteratively, we estimate the TDNN's parameters (w_q, w_f, and β) by minimizing the log-likelihood using stochastic gradient descent (SGD).

7.1.3 Prediction phase

After the training phase, we obtain the estimated parameters of HMM (A and E) and TDNN's parameters (w_q, w_f, and β). These estimated parameters are use to built the model of DTMN.

In the prediction phase, we perform an opposite procedure with the training phase. The DTMN model is used to predict the label y_t of the unlabeled observations using present feature f_t and prior hidden state q_{t-1}. By Eq. 7.2, we use f_1 to predict q_t, then q_1 and f_2 are used to predict q_2. Next, to predict q_3, we used (q_2, f_3). This procedure continues until $Q = \{q_{t,(t=1,2,...,T)}\}$ are reached. Since each q_t is random variable and $P(q_t|f)$ 1-by-1 from $t = 1$ to $t = N$, the probability distribution of the labels y_t that gives the prediction for the label, as follows:

$$\begin{aligned} P(y_t = i|f) &= \sum_j P(y_t = i|q_t = j).P(q_t = j|f) \\ &= \sum_j e_{i,j} P(q_t = j|f) \end{aligned} \tag{7.3}$$

7.1.4 Emotional states transition modeling phase

A study [230] defined emotions as discrete patterns of systemic activity. Emotions are categorized clearly and consistently across multiple levels of analysis, such as subjective experiences, physiological activity, and neural activation patterns. It supports that

emotions are discrete systems that are organized in a distributed fashion across the brain.

A discrete system is characterized by a set of states and transitions between the states. To describe formally a discrete event simulation, many works use a stochastic process algebra [231, 232]. In a discrete system, it is able to describe the passing of time and probabilistic choice between a limited number of processes, called discrete stochastic process. Here the universal quantifier is limited to feasible sequences of states to sequences that occur with positive probability. In other words, it is defined as a discrete stochastic process with a finite number of states.

Since emotions are discrete system activity [230], we apply the finite Markov chain to model the states transition of emotion. A finite set of states is high stress, low stress, neutral, soft, and angry. The emotional state updates its state depending on its current features and the prior states as input.

In this emotional states transition modeling phase, a Markov matrix is an $n \times n$ square matrix P such that each element of P is non-negative, and each row of P sums to one. Each row of P can be regarded as a probability mass function over n possible outcomes. let S be a finite set with n elements $x_1, \ldots, x_n$, where the set S is called the state space and $x_1, \ldots, x_n$ are the state values. A Markov chain X_t on S is a sequence of random variables on S that have the Markov property. This means that, for any time-step t and any state $y \in S$,

$$\mathbb{P}\{X_{t+1} = y | X_t\} = \mathbb{P}\{X_{t+1} = y | X_t, X_{t-1,\ldots}\} \tag{7.4}$$

In other words, knowing the current state is enough to know probabilities for future states. In particular, the dynamics of a Markov chain are fully determined by the set of values

$$P(x, y) := \mathbb{P}\{X_{t+1} = y | X_t = x\} \tag{7.5}$$

where $(x, y) \in S$. By construction, $P(x, y)$ is the probability of going from x to y in one unit of time (one step) and $P(x, .)$ is the conditional distribution of X_{t+1} given $X_t = x$, we view P as a stochastic matrix where

$$P_{ij} = P(x_i, x_j) \tag{7.6}$$

7.2 Experimental setup

The experiments are conducted in single personal computer with specification: Intel(R) Core (TM) i7-7700K CPU @ 4.2GHz, 16GB installed memory RAM, and 64-bit Operating system, x64-based processor. For software package, we used Matlab R2017b with several toolboxes, such as deep learning, digital signal processing (DSP) system, econometrics, audio, and signal processing.

We used two labeled conversations data of SUSAS dataset for estimating the two sets of parameters (HMM and TDNN). While for evaluation, we used the six unlabeled conversations that have various lengths of duration. We conditioned the speech input using their activity 5.1, speakers 5.2, and gender 5.3. Then, each speech is represented into a low-dimensional embedding space using SDTEC algorithm 6.2.
In the HMM model, we set the number of hidden states is 80 [140], and the matrix of state transition and the initial state distribution are initialized randomly between 0 and 1. Gaussian distributions is used to determine the emission probabilities.
In the TDNN model, we performed the batch normalization with a 256 batch size to stabilize the training procedure [140]. The rectified linear unit (ReLU) activation function is used on each hidden layer that has a dimension of 4000.
The effectiveness of the proposed DTMN is evaluated to predict the stress and emotions state from the speech data of SUSAS. We then compare it with five state-of-the-art state transition models, as follows:

KNN	:	run KNN with all parameter settings and architecture same with [41]
BN	:	run Bayesian network with all parameter settings and architecture same with [135]
HMM	:	run HMM method with the same settings and architecture in [227]
LSTM	:	run LSTM network with all parameter settings and architecture same with [134]
DMNN	:	run DMNN with same setting and architecture in [140]

7.3 Result and discussion

We demonstrate the effectiveness of the proposed DTMN to predict the present state of stress and emotion, then model their state transition. The proposed DTMN is assigned to predict the state of stress and emotion from the speech data from the SUSAS dataset. The performance of DTMN is evaluated by comparing it with the baseline systems in terms of prediction error rate (PER). Furthermore, we model the state transition of stress and emotions based on the speech label from the prediction result.

7.3.1 Prediction accuracy

The effectiveness of the DTMN is evaluated in predicting the emotional state of the time-series observation. In this experiment, we set the input and the parameters of DTMN as mentioned in Section 7.2. We run each system independently 10-times, and on average of evaluation result is summarized in Table 7.2.
Table 7.2 shows that BN presents a lower error than KNN. It is caused KNN should provide a proper scaling among variable time-steps, while BN depicts the relationships

TABLE 7.2: The evaluation result of the DTMN and the baseline systems in predicting the emotional state

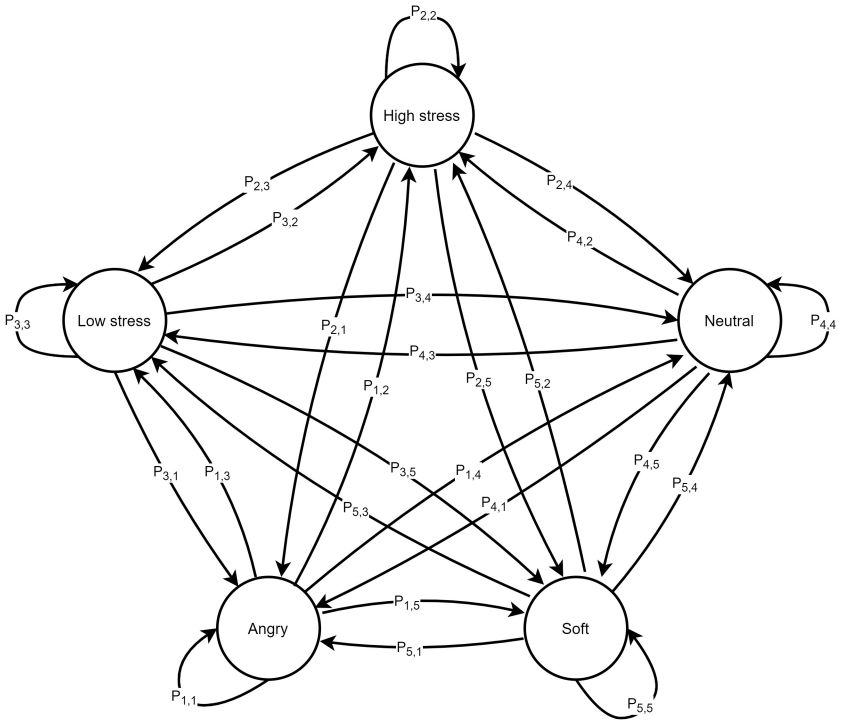
Method	Prediction error rate (% PER)
KNN [41]	48.27
BN [135]	41.63
HMM [227]	28.82
LSTM [134]	24.19
DMNN [140]	10.61
Proposed DTMN	8.55

between variables on each time-step in the manner of conditional independencies. However, on the other hand, BN cannot represent the nonlinear functions of state variables. Hence, BN has a higher error rate than HMM. The performance gap between LSTM and HMM shows that in-depth learning of the hidden state is more effective than statistical machine learning. Although the LSTM has learned the long-term temporal context dependencies, many emotional states are hard to determine or even unobservable. The combination between HMM and DNN (such as DMNN and the DTMN) presents a better ability in solving the LSTM's limitations by demonstrating a lower error rate. By considering the activation patterns over time, the proposed DTMN significantly outperforms the DMNN in predicting the emotional state. The DTMN is a sophisticated emotional state modeling by achieving the prediction error rate (PER) averagely 8.55%.

7.3.2 Emotional states transition modeling

In Section 7.3.1, the proposed DTMN demonstrates an effective result in predicting the stress and emotion by its state transition. It indicates that the proposed DTMN is able to predict accurately the present state based on the prior states. In further, we use a finite Markov chain to model the pattern of emotion transitions. Since males and females express emotion in different ways [228], we present the state transition of males and females in the different diagrams.

In Figure 7.4 shows the emotional state transition model. Table (a) and (b) denote the state transition probability for males and females. $P_{i,j}$ indicates the transition probability from state i to states j. For instance, $P_{1,5}$ is the state transition probability from the state "angry" to state "soft" with the probability "0.02" for males and "0.26" for females. Each Table shows that the sums of each row are one. As an example, the first row of Table (a) represents the transition probability from the state "Angry" to the other states (Angry, high stress, low stress, neutral, and soft) that the sum is one. It indicates that the transition matrix is a stochastic process, i.e $\sum_j P(i,j) = 1$. From Table (a) and (b), it is clear that the highest probabilities of each row and column are diagonal. It indicates that typically emotions are not change in a short time. The



Male

Transition Probability ($P_{i,j}$)		Present state (j)				
		Angry	High stress	Low stress	Neutral	Soft
Prior state (i)	Angry	0.58	0.12	0.19	0.09	0.02
	High stress	0.26	0.59	0.12	0.02	0.01
	Low stress	0.19	0.11	0.58	0.11	0.01
	Neutral	0.04	0.03	0.11	0.78	0.03
	Soft	0.16	0.03	0.02	0.32	0.46

Female

Transition Probability ($P_{i,j}$)		Present state (j)				
		Angry	High stress	Low stress	Neutral	Soft
Prior state (i)	Angry	0.53	0.04	0.11	0.05	0.26
	High stress	0.02	0.61	0.1	0.05	0.22
	Low stress	0.02	0.08	0.62	0.05	0.23
	Neutral	0.04	0.06	0.2	0.65	0.05
	Soft	0.06	0.05	0.05	0.14	0.7

FIGURE 7.4: The state transition model of stress and emotions. Male and female present a similar emotional states transition model. Table (a) and (b) show the transition probability from state i to state j for male and female, respectively.

current emotional state will retain if there are not any typical effective stimuli. On the other side, the highest sum of each column is "neutral" for males and "soft" for females. It proves that females are more emotional than males. Another surprise is females have a probability more easily to soft, while males are more easily to angry after

stressful conditions. It indicates that gender responds to emotional stress in different reacts, both psychologically and biologically, depending on their background experience, behavioral, and physiological domains.

7.4 Conclusion

We proposed a new framework for predicting and modeling the stress and emotions, named as the deep time-delay Markov network (DTMN). DTMN was able to predict the state of stress and emotions by considering its state transition. Structurally, the proposed DTMN consisted of a hidden Markov model (HMM) and the time-delay neural network or TDNN. HMM was used to predict the hidden states at each time-step, while the neural network is applied to learns in-depth the hidden representation of HMM. TDNN predicts the present hidden state using as input the prior hidden states and the features of the present time. We explicitly used a compact feature representation of stress and emotion (embedding features) of SDTEC as the input of DTMN. The effectiveness of the proposed DTMN was evaluated by comparing it with some state transition models such as KNN, LSTM, the Bayesian network, HMM, and DMNN in the task of predicting the emotional state from the time-series data of the SUSAS dataset.

Based on the evaluation result, the proposed DTMN outperformed the baseline state transition systems by achieving the prediction error rate (PER) of 8.55%. In further analysis, we conducted a comprehensive ablation experiment to investigate whether the estimated parameters of HMM and TDNN are related to model performance. Particularly, we investigated a different number of hidden states in the HMM and the various temporal contexts in the TDNN parameters to the prediction result. The experiment result showed the lowest error rate was achieved for the number of hidden states by 80 and the temporal context of TDNN is $[t-1, t-5]$. Furthermore, we performed the Markov chain for modeling the state transition of stress and emotions. The observation result showed that females have a trend longer to stress and become sad after a stressful period. While for males, they tend to be easier to change stress to angry.

Chapter 8

Final Discussion and Conclusions

Emotion plays a vital role in human life. The ability to recognize and make sense of the emotions, known as emotional awareness, makes people care with others and their emotional health. Emotionally healthy people still feel stress, anger, and sadness, but they know how to manage them. Stress is a normal reaction due to changes in environmental conditions that increase the activity of the human physiological system. Speech is one of the physiological parameters that reflect the symptoms of stress and emotion.

We Conducted emotional awareness into two steps. The first step is emotion recognition. It means that the system is able to recognize an time-series emotion class by its characteristics. In order to manage emotions, the second step is to model the state transition of emotions and recognize its patterns. Thus, this thesis presented a system that is able to recognize stress and emotions. The recognition step is consisted of three approaches, i.e., classification, clustering, and prediction. While emotional management step is conducted by modeling the states transition of stress and emotions. The thesis focused on natural speech i.e., speech with naturally expressed (not acted) stress and emotions from Speech under Stress and Actual Stress (SUSAS) database.

This thesis begins with the hypothesize that stress and emotions have a linkage caused by stimuli from the environment. Firstly, we developed the stress and emotions speech recognition (SSR) system that trained using labeled data. The result shows that most misrecognition is between "stress" and "angry" for males, and "stress" and "soft" for females. Thus, we suspected the female tended to express their stress in soft (e.g., sadness) while the male tended to express their stress in anger. Secondly, we developed the stress and emotions speech clustering (SSC) system that trained using unlabeled data. We performed in unsupervised and semi-supervised approach. The result showed that two-cluster present a low identification rate due to both clusters has relatively similar emotional characteristics. We suspect that they are "low stress" and "soft." Thirdly, we developed the stress and emotions speech modeling (SSM) system. We

predicted the present emotional state by analyzing the speech features and the prior emotional state. We then model emotional states transition using finite Markov chain. The result showed that males and females generally present a similar emotional transition representation. However, there are some fundamental differences between males and females. Females have a tendency longer in stress than males but more easily change for other emotions. After a stressful period, females tend to become sad, while males are easier to grow angry.

The non-intrusive measurement method, such as speech, is not as well as the non-invasive methods, such as EEG, in recognizing stress and emotions. However, based on the experiment result, the proposed system presented a low error rate. In other words, this method is quite promising to be used in real life. Therefore, in the future, we interest to implement a smartphone application-based proposed system as an early detection system of emotion.

Bibliography

[1] Mental Health Foundation, "Fundamental Facts About Mental Health", Mental Health Foundation: London, 2016.

[2] D.H. Kluemper, T. DeGroot, S. Choi, "Emotion Management Ability", Journal of Management, 39(4), pp. 878-905, 2013.

[3] S. Sinha, D. Sinha, "Emotional Intelligence and Effective Communication", Management Communication: Trends & StrategiesChapter: Emotional Intelligence and Effective Communication, McGraw Hill, 2007.

[4] O. Serrat, "Understanding and Developing Emotional Intelligence", Knowledge Solutions. Springer, Singapore, 2017.

[5] A.S. Cowen, D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients", PNAS, 114(38), pp. E7900–7909, 2017.

[6] J.K. Kiecolt-Glaser, L. McGuire, T.F. Robles, R. Glaser, "Emotions, morbidity, and mortality: new perspectives from psychoneuroimmunology", Annual review of psychology, 53, pp. 83-107, 2002.

[7] E.M. Sternberg, G.P. Chrousos, R.L. Wilder, P.W. Gold, "The stress response and the regulation of inflammatory disease", Annals of Internal Medicine, 117, pp. 854-866, 1992.

[8] N. Fiedler, R. Laumbach, K. Kelly-McNeil, P. Lioy, Z.H. Fan, et al, "Health effects of a mixture of indoor air volatile organics, their ozone oxidation products, and stress", Environmental health perspectives, 113, pp. 1542, 2005.

[9] J. Du, J. Huang, Y. An, W. Xu, "The Relationship between stress and negative emotion: The Mediating role of rumination", Clinical Research and Trials, 4(1), pp. 5 pages, 2018.

[10] M. Wang and K.J. Saudino, "Emotion regulation and stress", Journal of Adult Development, 18(2), pp. 95-103, 2011

[11] H. Yaribeygi, Y. Panahi, H. Sahraei, T.P. Johnston and A. Sahebkar, "The Impact of Stress on Body Function: A Review", EXCLI Journal, 16, pp. 1057-1072, 2017

[12] R. Smith, R.D. Lane, "Unconscious emotion: A cognitive neuroscientific perspective", Neurosci. Biobehav. Rev., 69, pp. 216–238, 2016.

[13] A. Kumar, P. Rinwa, G. Kaur, L. Machawal, "Stress: Neurobiology, consequences and management", J. Pharm Bioallied Sci, 5(2), pp. 91–97, 2013.

[14] S.M. Smith, W.W. Vale, "The role of the hypothalamic-pituitary-adrenal axis in neuroendocrine responses to stress", Dialogues Clin Neurosci, 8(4), pp. 383–395, 2006.

[15] A.F. T. Arnsten, "Stress signaling pathways that impair prefrontal cortex structure and function", Nat Rev Neurosci, 10(6), pp. 410–422, 2009.

[16] H. Kim, E. Cheon, D. Bai, Y. H. Lee, B. Koo, "Stress and Heart Rate Variability: A Meta-Analysis and Review of the Literature", Psychiatry Investig, 15(3), pp. 235–245, 2018.

[17] Y. Liu, S. Du, "Psychological stress level detection based on electrodermal activity", Behav Brain Res, 341, pp. 50-53, 2018.

[18] M.J. Tipton, A. Harper, J.F.R. Paton, J.T. Costello, "The human ventilatory response to stress: rate or depth?", J. Physiol, 595(17), pp. 5729–5752, 2017.

[19] M. Pedrotti, M.A. Mirzaei, A. Tedescho, J. Chardonnet, F. Merienne, S. Benedetto, T. Baccino, "Automatic Stress Classification With Pupil Diameter Analysis", International Journal of Human-Computer Interaction, 30(3), pp. 220-236, 2014.

[20] G. Giannakakis, M. Pediaditis, D. Manousos, E. Kazantzaki, F. Chiarugi, "Stress and anxiety detection using facial cues from videos", Biomedical Signal Processing and Control, 31, pp. 89–101, 2016.

[21] B.H. Prasetio, H. Tamura, K. Tanno, "Support Vector Slant Binary Tree Architecture for Facial Stress Recognition Based on Gabor and HOG Feature", International Workshop on Big Data and Information Security (IWBIS), Jakarta, Indonesia, pp. 63–68, 2018.

[22] B.H. Prasetio, H. Tamura, K. Tanno, "The Facial Stress Recognition Based on Multi-histogram Features and Convolutional Neural Network", IEEE International Conference on Systems, Man, and Cybernetics (SMC), Miyazaki, Japan, pp. 881–887, 2018.

[23] M. Gavrilescu, N. Vizireanu, "Predicting Depression, Anxiety, and Stress Levels from Videos Using the Facial Action Coding System", Sensors(Basel), 19(17), pp. 1–32, 2019.

[24] J.H.L. Hansen, S. Patil, "Speech Under Stress: Analysis, Modeling and Recognition", Speaker Classification I. Lecture Notes in Computer Science. Müller, C., Eds.; Springer: Berlin, Germany, 4343, pp. 108–137, 2007.

[25] L.D. Vignolo, S.R.M. Prasanna, S. Dandapat, L. Rufiner, D.H. Milone, "Feature optimisation for stress recognition in speech", Pattern Recognition Letters, 84, pp. 1–7, 2016.

[26] B.H. Prasetio, H. Tamura, K. Tanno, "Ensemble Support Vector Machine and Neural Network Method for Speech Stress Recognition", International Workshop on Big Data and Information Security (IWBIS), Jakarta, Indonesia, pp. 57–62, 2018.

[27] K. Tomba, J. Dumoulin, E. Mugellini, O. A. Khaled, S. Hawila, "Stress Detection Through Speech Analysis", International Joint Conference on e-Business and Telecommunications (ICETE), Porto, Portugal, pp. 394–398, 2018.

[28] B.H. Prasetio, H. Tamura, K. Tanno, "Generalized Discriminant Methods for Improved X-Vector Back-end Based Speech Stress Recognition", IEEJ Transactions on Electronics, Information and Systems, 139(11), pp. 1341–1347, 2019.

[29] J.H.L. Hansen Composer, SUSAS LDC99S78. Web Download, Sound Recording. Linguistic Data Consortium: Philadelphia, PA, USA, 1999.

[30] J.H.L. Hansen, Composer. SUSAS Transcript LDC99T33, Sound Recording. Linguistic Data Consortium: Philadelphia, PA, USA, 1999.

[31] E. Cowie, R. Cowie, M. Schröder, "A new emotion database: considerations, sources and scope", The ISCA ITRW on Speech and Emotion, pp. 39-44, 2000.

[32] D.J. France, R.G. Shiavi, S. Silverman, M. Silverman, M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk", IEEE Transactions on Biomedical Engineering, 47(7), pp. 829-837, 2000.

[33] K. Fischer, "Annotating emotional language data", Tech. Rep. 236, Univ. of Hamburg, 1999.

[34] M. Joels, T. Z. Baram, "The neuro-symphony of stress", Nat. Rev. Neurosci, 10, pp. 459–466, 2009.

[35] D. Moungsri, T. Koriyama, T. Kobayashi, "HMM-based Thai speech synthesis using unsupervised stress context labeling", Signal and Information Processing Association Annual Summit and Conference (APSIPA), Siem Reap, Cambodia, pp. 1–4, 2014.

[36] D. Moungsri, T. Koriyama, T. Kobayashi, "Unsupervised Stress Information Labeling Using Gaussian Process Latent Variable Model for Statistical Speech Synthesis", INTERSPEECH, San Francisco, pp. 1517–1521, 2016.

[37] M. R. Morales, R. Levitan, "Mitigating Confounding Factors in Depression Detection Using an Unsupervised Clustering Approach", Computing and Mental Health Workshop (CHI), San Jose, CA, USA, pp. 1–4, 2016.

[38] C.K. Yogesh, M. Hariharan, R. Yuvaraj, N. Ruselita, A.H. Adom, Y. Sazali, P. Kemal, "Bispectral features and mean shift clustering for stress and emotion recognition from natural speech", Computers & Electrical Engineering, 62, pp. 676–691, 2017.

[39] C. Huang, B. Song, L. Zhao, "Emotional speech feature normalization and recognition based on speaker-sensitive feature clustering", International Journal of Speech Technology, 19(4), pp. 805–816, 2016.

[40] N. Hajarolasvadi, H. Demirel, "3D CNN-Based Speech Emotion Recognition Using K-Means Clustering and Spectrograms", Entropy, 21, pp. 1–17, 2019.

[41] T. Pao, J. Yeh, Y. Tsai, "Recognition and analysis of emotion transition in mandarin speech signal", IEEE International Conference on Systems, Man, and Cybernetics (SMC), Istanbul, Turkey, pp. 3326–3332, 2010.

[42] P. Jaak, "Affective neuroscience: the foundations of human and animal emotions", ([Reprint] ed.). Oxford [u.a.]: Oxford Univ. Press. p. 9, 2005.

[43] C. Michel, "What is emotion?", Behavioural Processes, 60(2), pp. 69-83, 2002.

[44] O.M. Wolkowitz, E.S. Epel, V.I. Reus, S.H. Mellon, "Depression gets old fast: do stress and depression accelerate cell aging?", Depresion & Anxiety, 27(4), pp. 327–338, 2010.

[45] T. Radek, Z, Martin, K. Martin, "Emotional Creativity and Real-Life Involvement in Different Types of Creative Leisure Activities", Creativity Research Journal, 28(3), pp. 348–356, 2016.

[46] C.E. Osgood, G.J. Suci GJ, P.H.Tannenbaum, "The Measurement of Meaning", Urbana, Illinois, USA: University of Illinois Press, 1957.

[47] R. Plutchik, "Nature of emotions", American Scientist, 89(4), pp. 344-350, 2002.

[48] P. Ekman, D.T. Cordaro, "What is Meant by Calling Emotions Basic", Emotion Review. 3(4), pp. 364–370, 2011.

[49] D.T. Cordaro, D. Keltner, S. Tshering, D. Wangchuk, L.M. Flynn, "The voice conveys emotion in ten globalized cultures and one remote village in Bhutan", Emotion. 16(1), pp. 117–128, 2016.

[50] D.T. Cordaro, R. Sun, D. Keltner, S. Kamble, N. Huddar, G. McNeil, "Universals and cultural variations in 22 emotional expressions across five cultures", Emotion. 18(1), pp. 75–93, 2018.

[51] D. Keltner, K. Oatley, J.M. Jenkins, Jennifer, Understanding emotions, Hoboken. NJ: John Wiley & Sons, Inc, 2019.

[52] D.L. Schacter, Psychology Ed. 2. 41 Madison Avenue New York, NY 10010: Worth Publishers, 2011.

[53] J.A. Russell, L.F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant", Journal of Personality and Social Psychology. 76(5), pp. 805–19, 1999.

[54] J.A. Russell, "Core affect and the psychological construction of emotion", Psychological Review. 110(1), pp. 145–72, 2003.

[55] R.S. Lazarus, S. Folkman, "Coping and adaptation", The handbook of behavioral medicine, pp. 282-325, 1984.

[56] G. Fink, "Stress: Concepts, Definition and History," in Reference Module in Neuroscience and Biobehavioral Psychology, Melbourne, VIC, Australia, Elsevier, pp. 1-9, 2017.

[57] S. Glazer and C. Liu, "Work, Stress, Coping, and Stress Management," Industrial and Organizational Psychology, 4, pp. 1-45, 2017.

[58] N. Schneiderman, G. Ironson, S.D. Siegel, "STRESS AND HEALTH: Psychological, Behavioral, and Biological Determinants", Annu Rev Clin Psychol, 2005(1), pp. 607–628, 2005.

[59] S. Kaushal, "Contribution of Non Verbal Language in Communication: A Study of Non-Verbal Communication", Asian Journal of Advance Basic Science,2(1), pp. 15-21, 2013.

[60] A.O. Shonubi and A.A. Akintaro, "The Impact Of Effective Communication On Organizational Performance", The International Journal of Social Sciences and Humanities Invention, 3(3), pp. 1904-1914, 2016.

[61] Q. Kang, "Paralanguage", Canadian Social Science, vol. 9, no. 6, pp. 222-226, 2013.

[62] S. Paulmann, D. Furnes, A.M. Bokenes and P.J. Cozzolino, "How Psychological Stress Affects Emotional Prosody", Plos one, 11(11), pp. 1-21 pages, 2016.

[63] A.K. Tiwari, "Non-Verbal Communication-an Essence of Interpersonal Relationship at Workplace", SMS Varanasi Management Insight, 11(2), pp. 109-114, 2015.

[64] H. Hopp, A.S. Troy, I.B. Mauss, "The unconscious pursuit of emotion regulation: Implications for psychological health", Cogn Emot, 25(3), pp. 532-545, 2011.

[65] N.R. Prakash and J. Kaur, "A Study on Physiological Parameters Used To Monitor Stress in Experimentally Induced Stimuli," International Journal of Computer Science and Information Technologies, vol. 6, no. 6, pp. 5244-5246, 2015.

[66] R. Gordan, J.K. Gwathmey, X. Lai-Hua, "Autonomic and endocrine control of cardiovascular function", World J. Cardiol, 7, 204–214, 2015.

[67] J. Wijsman, B. Grundlehner, J. Penders and H.J. Hermens, "Trapezius Muscle EMG as Predictor of Mental Stress", The Wireless Health, San Diego, USA, Oct, 5-7, 2010.

[68] Z. Zhang, "Mechanics of human voice production and control", J. Acoust. Soc. Am, 140, pp. 2614–2635, 2016.

[69] O. Simantiraki, G. Giannakakis, A. Pampouchidou and M. Tsiknakis, "Stress Detection from Speech Using Spectral Slope Measurements", International Symposium on Pervasive Computing Paradigms for Mental Health, Barcelona, 2016.

[70] H. Gao, A. Yuce and J.P. Thiran, "Detecting emotional stress from facial expressions for driving safety", IEEE International Conference on Image Processing (ICIP), Paris, France, 2014.

[71] Alberdi, A., Aztiria, A., Basarab, "A. Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review", Journal of Biomedical Informatics, 59, pp. 49–75 2016.

[72] N. Hjortskov, D. Rissén, A.K. Blangsted, N. Fallentin, U. Lundberg, K. Søgaard, "The effect of mental stress on heart rate variability and blood pressure during computer work", Eur. J. Appl. Physiol, 92(1–2),pp. 84–89, 2004.

[73] S.-h. Seo, J.-t. Lee, "Stress and EEG", in: M. Crisan (Ed.), "Convergence and Hybrid Information Technologies", pp. 426, 2010.

[74] A. Kaklauskas, E.K. Zavadskas, M. Seniut, G. Dzemyda, V. Stankevic, C. Simkevicius, T. Stankevic, R. Paliskiene, A. Matuliauskaite, S. Kildiene, L. Bartkiene, S. Ivanikovas, V. Gribniak, "Web-based biometric computer mouse advisory system to analyze a user's emotions and work productivity", Eng. Appl. Artif. Intell, 24(6), pp. 928–945, 2011.

[75] Y. Daviaux, E. Bonhomme, H. Ivers, E. de Sevin, J. Micoulaud-Franchi, S. Bioulac, C. M. Morin, P. Philip, E. Altena, "Event-Related Electrodermal Response to Stress: Results From a Realistic Driving Simulator Scenario", Human Factors, 62(1), pp. 138–151, 2020.

[76] C.Z. Wei, "Stress emotion recognition based on RSP and EMG signals", Adv. Mater. Res, 709, pp. 827–831, 2013.

[77] W. Liao, W. Zhang, Z. Zhu, Q. Ji, "A real-time human stress monitoring system using dynamic bayesian network", in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2005.

[78] K. Palanisamy, M. Murugappan, S. Yaacob, "Multiple physiological signal-based human stress identification using non-linear classifiers", Electron. Electr. Eng, 19(7), pp. 80–85, 2013.

[79] H. Lu, D. Frauendorfer, M. Rabbi, M.S. Mast, G.T. Chittaranjan, A.T. Campbell, D. Gatica-Perez, T. Choudhury, "StressSense: detecting stress in unconstrained acoustic environments using smartphones", The 2012 ACM Conference on Ubiquitous Computing – UbiComp '12, ACM Press, New York, USA, pp. 351, 2012.

[80] W.J. Ray, H.W. Cole, "EEG alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes", Science (New York, N.Y.) 228 (4700), 750–752, 1985.

[81] A. de Santos Sierra, C. Sanchez Avila, G. Bailador del Pozo, J. Guerra Casanova, "Stress detection by means of stress physiological template", World Congress on Nature and Biologically Inspired Computing, pp. 131–136, 2011.

[82] J. Hernandez, P. Paredes, A. Roseway, M. Czerwinski, "Under pressure: sensing stress of computer users", The 32nd Annual ACM Conference on Human Factors in Computing Systems (CHI), ACM Press, New York, New York, USA, pp. 51–60, 2014.

[83] J.A. Healey, R.W. Picard, "Detecting stress during real-world driving tasks using physiological sensors", IEEE Trans. Intell. Transport. Syst, 6(2), pp. 156–166, 2005.

[84] H. Zhang, Y. Zhu, J. Maniyeri, C. Guan, "Detection of variations in cognitive workload using multi-modality physiological sensors and a large margin unbiased regression machine", Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE Engineering in Medicine and Biology Society, Annual Conference, pp. 2985–2988, 2014.

[85] P. Ren, A. Barreto, J. Huang, Y. Gao, F.R. Ortega, M. Adjouadi, "Off-line and online stress detection through processing of the pupil diameter signal", Ann. Biomed. Eng, 42(1), pp. 162–176, 2014.

[86] K. Sakamoto, S. Aoyama, S. Asahara, "Relationship between emotional state and pupil diameter variability under various types of workload stress", in: B.-T. Karsh (Ed.), Ergonomics and Health Aspects of Work with Computers, Lecture Notes in Computer Science, vol. 5624, Springer, Berlin, Heidelberg, pp. 177–185, 2009.

[87] R.O. Duda, P.E. Hart, D.G. Stork, "Pattern classification", (2nd edition). Wiley, 2001.

[88] E. Douglas-Cowie, N. Campbell, R. Cowie, P. Roach, "Emotional speech: Towards a new generation of databases", Speech Communication, 40(1-2): pp. 33-60, 2003.

[89] L.R. Rabiner, R.W. Schafer, "Digital Processing of Speech Signals", Englewood Cliffs, NJ: Prentice-Hall, Inc., 1978.

[90] R.C. Snell, F. Milinazzo, "Formant location from LPC analysis data", IEEE trans. on Speech and Audio Processing, 1(2), 1993.

[91] R. Serizel, V. Bisot, S. Essid and G. Richard, "Acoustic Features for Environmental Sound Analysis", in Computational Analysis of Sound Scenes and Events, Springer, pp. 71-101, 2017.

[92] L.R. Rabiner, R.W. Schafer, "Theory and Applications of Digital Speech Processing", Upper Saddle River, NJ: Pearson, 2010.

[93] H.M. Teager, "Some observations on oral air flow during phonation", Acoustics, Speech and Signal Processing, IEEE Transactions on, 28(5), pp. 599-601, 1980.

[94] H.M. Teager, S.M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract", Speech production and speech modelling, 1990.

[95] H.M. Teager, S.M. Teager, "A phenomenological model for vowel production in the vocal tract", Speech Science: Recent Advances, 1982.

[96] J. Kaiser, "Some observations on vocal tract operation from a fluid flow point of view", Vocal Fold Physiology: Biomechanics, Acoustics, and Phonatory Control, pp. 358-386, 1983.

[97] A.W. Zewoudie, J. Luque and J. Hernandi, "The use of long-term features for GMM-and i-vector-based speaker diarization systems", EURASIP Journal on Audio, Speech, and Music Processing, 14, 2018.

[98] P. Verma and P.K. Das, "i-Vectors in speech processing applications: a survey", International Journal of Speech Technology, 18(4), pp. 529-546, 2015.

[99] J. Grzybowska and M. Ziolko, "I-vectors in gender recognition from telephone speech", The Twenty-First National Conference on Applications of Mathematics in Biology and Medicine, pp. 57–62, 2015.

[100] M. Wang, Y. Chen, Z. Tang and E. Zhang, "I-vector based speaker gender recognition", IEEE Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 2015.

[101] T. Zhang and J. Wu, "Speech emotion recognition with i-vector feature and RNN model", in IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP), Chengdu, China, 2015.

[102] J. Gomes and M. El-Sharkawy, "Implementation of i-vector Algorithm in Speech Emotion Recognition by using Two Different Classifiers: Gaussian Mixture Model and Support Vector Machine", International Journal of Advanced Research in Computer Science and Software Engineering, 6(9), pp. 8-16, 2016.

[103] L. Mackova, A. Cizmar and J. Juhar, "Emotion recognition in i-vector space", in International Conference Radioelektronika (RADIOELEKTRONIKA), Kosice, Slovakia, 2016.

[104] R. Xia and Y. Liu, "Using i-Vector Space Model for Emotion Recognition", INTERSPEECH, Portland, OR, USA, 2012.

[105] J. Gomes and M. EL-Sharkawy, "i-Vector Algorithm with Gaussian Mixture Model for Efficient Speech Emotion Recognition", International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2015.

[106] R. Xia and Y. Liu, "DBN-ivector Framework for Acoustic Emotion Recognition", INTERSPEECH, San Francisco, USA, 2016.

[107] C. Zhang, G. Liu, C. Yu and J. H. L. Hansen, "I-vector Based Physical Task Stress Detection with Different Fusion Strategies", INTERSPEECH, Dresden, Germany, 2015.

[108] P. kenny, G. Boulianne, P. Quellet, P. Dumouchel, "Joint Factor Analysis Versus Eigenchannels in Speaker Recognition", IEEE Transactions on Audio Speech and Language Processing, 15(4), pp. 1435-1447, 2007.

[109] M. Awad, R. Khanna, "Support Vector Machines for Classification", In: Efficient Learning Machines. Apress, Berkeley, CA, 2015.

[110] Y. Chavhan, M.L. Dhore, P. Yesaware, "Speech Emotion Recognition Using Support Vector Machine", International Journal of Computer Applications, 1(20), 2010.

[111] C.M. Bishop, "Pattern Recognition and Machine Learning", Springer-Verlag New York, 2006.

[112] D. Ververidis, C. Kotropoulos, "Emotional speech classification using Gaussian mixture models", IEEE International Symposium on Circuits and Systems(ISCAS), 3, pp. 2871-2874. 2005.

[113] T.F. Quatieri, "Discrete-time speech signal processing: Principles and Practice", Prentice Hall, 2002.

[114] X. Cheng, Q. Duan, "Speech Emotion Recognition Using Gaussian Mixture Model", International Conference on Computer Application and System Modeling, Atlantis Press, 2012.

[115] I.J. Tashev, Z. Wang, K. Godin, "Speech Emotion Recognition based on Gaussian Mixture Models and Deep Neural Networks", Information Theory and Applications Workshop (ITA), San Diego, CA, USA, 2017.

[116] S.R. Bandela, T.K. Kumar, "Emotion Recognition of Stressed Speech Using Teager Energy and Linear Prediction Features", International Conference on Advanced Learning Technologies (ICALT), 1, Mumbai, India, pp. 422-425, 2018

[117] A. Bhavan, M. Sharma, M. Piplani, P. Chauhan, Hitkul, R.R. Shah, "Deep Learning Approaches for Speech Emotion Recognition", In book: Deep Learning-Based Approaches for Sentiment Analysis. By: Basant Agarwal, Springer Nature Singapore Pte Ltd, 2020.

[118] M. Awad, R. Khanna, "Hidden Markov Model", In: Efficient Learning Machines. Apress, Berkeley, CA, 2015.

[119] A. Nogueiras, A. Moreno, A. Bonafonte, J.B. Marino, "Speech Emotion Recognition Using Hidden Markov Models", Eurospeech, Scandinavia, 2001.

[120] T.L. Nwe, S.W. Foo, L.C. De Silva, "Speech emotion recognition using hidden Markov models", Speech Communication, 41(4), pp. 603-623, 2003.

[121] G.G. Daniel, "Artificial Neural Network", In: Runehov A.L.C., Oviedo L. (eds) Encyclopedia of Sciences and Religions. Springer, Dordrecht, 2013.

[122] C.H. Chen, "Neural Networks in Pattern Recognition and Their Applications", World Scientific Publishing Co Pte Ltd, 1991.

[123] P. Mishra, A. Rawat, "Emotion Recognition through Speech Using Neural Network", International Journal of Advanced Research in Computer Science and Software Engineering, 5(5), pp. 422-428, 2015.

[124] H.K. Palo, M.N. Mohanty, "Comparative Analysis of Neural Networks for Speech Emotion Recognition", International Journal of Engineering and Technology, 7(4), pp. 422-428, 2018.

[125] A.B. Ingale, D.S. Chaudhari, "Speech Emotion Recognition", International Journal of Soft Computing and Engineering, 2(1), pp. 235-238, 2012.

[126] T.D. Street, S.J. Lacey, K. Somoray, "Employee Stress, Reduced Productivity, and Interest in a Workplace Health Program: A Case Study from the Australian Mining Industry, Int J Environ Res Public Health,6(1): 94, 2019.

[127] American Psychological Association, "Toward reducing work stress", Monitor on Psychology, 39(2), 9, 2008.

[128] S.S. Guillen, L.L. Iacono, C. Meder, "Affective Robots: Evaluation of Automatic Emotion Recognition Approaches on a Humanoid Robot towards Emotionally Intelligent Machines", International Journal of Mechanical and Mechatronics Engineering, 12(6), pp. 584-592, 2018 .

[129] R. Fernandez, "A Computational Model for the Automatic Recognition of Affect in Speech", PhD Thesis, MIT Media Arts and Sciences, 2004.

[130] J. Han, M. Kamber, J. Pei, "Advanced Cluster Analysis", Data Mining (Third Edition), The Morgan Kaufmann Series in Data Management Systems, pp. 497–541, 2012.

[131] J. Deng, Z. Zhang, F. Eyben, B. Schuller, "Autoencoder-based Unsupervised Domain Adaptation for Speech Emotion Recognition", IEEE Signal Processing Letters, 21(9), pp. 1068–1072, 2014.

[132] B.H. Prasetio, H. Tamura, K. Tanno, "A Deep Time-delay Embedded Algorithm for Unsupervised Stress Speech Clustering", IEEE International Conference on Systems, Man, and Cybernetics (SMC), Bari, Italy, pp. 1193–1198, 2019.

[133] B.H. Prasetio, H. Tamura, K. Tanno, "Semi-Supervised Deep Time-Delay Embedded Clustering for Stress Speech Analysis", Electronics, 8(11), pp. 1–13, 2019.

[134] R. Zhang, A. Atsushi, S. Kobashikawa, Y. Aono, "Interaction and Transition Model for Speech Emotion Recognition in Dialogue", INTERSPEECH, Stockholm, Sweden, 2017.

[135] H. Xiang, P. Jiang, S. Xiao, F. Ren, S. Kuroiwa, "A Model of Mental State Transition Network", IEEJ Transactions on Electronics, Information and Systems, 127(3), pp. 434–442, 2007.

[136] P. Xiaolan, X. Lun, L. Xin, W. Zhiliang, "Emotional state transition model based on stimulus and personality characteristics", IEEE China Communications, 10(6), pp. 146–155, 2013.

[137] M.A. Thornton,D.I. Tamir, "Mental models accurately predict emotion transitions", Natl Acad Sci, 114(23), pp. 5982–5987, 2017.

[138] M. Awiszus, B. Rosenhahn, "Markov Chain Neural Networks", IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 2018.

[139] F.S. Al-Anzi, D.M. AbuZeina, "A Survey of Markov Chain Models in Linguistics Applications", International Conference on Advanced Information Technologies and Applications (ICAITA), Dubai, UAE, 2016.

[140] M. Yang, W. Tu, W. Yin, Z. Lu, "Deep Markov Neural Network for Sequential Data Classification", The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2, Beijing, China, 2015.

[141] R. Kurczab, S. Smusz and A. J. Bojarski, "The influence of negative training set size on machine learning-based virtual screening," Journal of Cheminformatics, vol. 6, no. 32, 2014.

[142] S. Vaishnav, S. Mitra, "Speech Emotion Recognition: A Review", International Research Journal of Engineering and Technology (IRJET), 3(4), 2016.

[143] P. Loui, J.P. Bachorik, H.C. Li, G. Schlaug, "Effects of voice on emotional arousal," Front Psychol, vol. 4:675, 2013

[144] T.T. Anagnostopoulos, C. Skourlas, "Ensemble Majority Voting Classifier for Speech Emotion Recognition and Prediction", Journal of Systems and Information Technology, 16(3), 2014.

[145] D. Opitz, R. Maclin, "Popular Ensemble Methods: An Empirical Study," Journal of Artificial Intelligence Research, 11, pp. 169-198, 1999.

[146] J.C. Burges Christopher, "A tutorial on support vector machines for pattern recognition", Data Min. Knowl. Discov., 2(2), pp. 121-167, 1998.

[147] L. Xu, R.K. Das, E. Yılmaz, J. Yang and H. Li, "Generative x-vectors for text-independent speaker verification", IEEE Spoken Language Technology (SLT), Athens, Greece, 2018.

[148] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey and S. Khudanpur, "Spoken Language Recognition using X-vectors", The Speaker and Language Recognition Workshop, Les Sables d'Olonne, France, 2018.

[149] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudapur, "X-Vector: Robust DNN Embeddings for Speaker Recognition", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Alberta, Canada, 2018.

[150] D. Snyder, D. Garcia-Romero, D. Povey and S. Khudanpur, "Deep Neural Network Embeddings for Text-Independent Speaker Verification", INTERSPEECH, Stockholm, Sweden, 2017.

[151] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE", Journal of Machine Learning Research, 9, pp. 2579-2605, 2008.

[152] S. Arora, W. Hu and P. K. Kothari, "An Analysis of the t-SNE Algorithm for Data Visualization", Conference on Learning Theory (COLT), Stockholm, Sweden, 2018.

[153] T. Drugman, Y. Stylianou, Y. Kida, and M. Akamine, "Voice Activity Detection: Merging Source and Filter-based Information", IEEE Signal Processing Letters, 23(2), 252-256, 2016.

[154] K. Sang-Kyun, K. Sang-Ick, P. Young-Jin, L. Sanghyuk, and L. Sangmin, "Power Spectral Deviation-Based Voice Activity Detection Incorporating Teager Energy for Speech Enhancement", Symmetry, 8(58), 8 pages, 2016.

[155] P. Sertsi, S. Boonkla, V. Chunwijitra, N. Kurpukdee, and C. Wutiwiwatchai, "Robust Voice Activity Detection Based on LSTM Recurrent Neural Networks and

Modulation Spectrum", Proceedings of APSIPA Annual Summit and Conference, Kuala Lumpur, Malaysia, 2017.

[156] E. Khoury, and M. Garland, "I-Vectors for Speech Activity Detection", The Speaker and Language Recognition Workshop (Odyssey), Bilbao, Spain, 2016.

[157] E. Rentzeperis, C. Boukis, and A. Pnevmatikakis, "Combining Finite State Machines and LDA for Voice Activity Detection", Artificial Intelligence and Innovations (AIAI): from Theory to Applications, IFIP The International Federation for Information Processing, 247, Springer, Boston, MA, 2007.

[158] Y. Liang, X. Liu, M. Zhou, Y. Lou, and B. Shan, "A Robust Voice Activity Detector Based on Weibull and Gaussian Mixture Distribution", International Conference on Signal Processing Systems (ICSPS), Dalian, China, 2010.

[159] O. Ghahabi, W. Zhou, and V. Fisher, "A Robust Voice Activity Detection for Real-time Automatic Speech Recognition", The Conference on Electronic Speech Signal Processing (ESSV), Baden-Württemberg, Germany, 2018.

[160] Z. Huang, Y. Cheng, K. Li, V. Hautamaki, and C. Lee, "A Blind Segmentation Approach to Acoustic Event Detection Based on I-Vector", INTERSPEECH, Lyon, France, 2013.

[161] H. Yamamoto, K. Okabe, and T. Koshinaka, "Robust i-vector extraction tightly coupled with voice activity detection using deep neural networks", Asia-Pacific Signal and Information Processing Association (APSIPA), Annual Summit and Conference, Kuala Lumpur, Malaysia, 2017.

[162] J. Padrell, D. Macho and C. Nadeu, "Robust speech activity detection using LDA applied to FF parameters", The IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, PA, USA, 2005.

[163] R. Chakraborty, A. Panda, M. Pandharipande, S. Joshi, and S.K. Kopparapu, "Front-end Feature Compensation and Denoising for Noise Robust Speech Emotion Recognition", INTERSPEECH, Graz, Austria, 2019.

[164] R. Xia and Y. Liu, "Using Denoising Autoencoder for Emotion Recognition", INTERSPEECH, Lyon, France, 2013.

[165] S. Dwijayanti, K. Yamamori, M. Miyoshi, "Enhancement of speech dynamics for voice activity detection using DNN", EURASIP Journal on Audio, Speech, and Music Processing, 2018(10), 1-15, 2018.

[166] S.E. Bou-Ghazale and J.H.L. Hansen, "A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress", IEEE Trans. on Speech and Audio Processing, 8(4), 429-442, 2000.

[167] M.F. Alghifari, T.S. Gunawan, M.A. binti Wan Nordin, S.A.A. Qadri, M. Kartiwi, and Z.Janin, "On the use of voice activity detection in speech emotion recognition", Bulletin of Electrical Engineering and Informatics, 8(4), 1324-1332, 2019.

[168] D.Sztahó and K.Vicsi, "Speech activity detection and automatic prosodic processing unit segmentation for emotion recognition", Intelligent Decision Technologies, 8(4), 315-324, 2014.

[169] J. Ling, S. Sun, J. Zhu, and X. Liu, "Speaker Recognition with VAD", Pacific-Asia Conference on Web Mining and Web-based Application, Wuhan, China, 2009.

[170] M. Mak and H. Yu, "A study of voice activity detection techniques for NIST speaker recognition evaluations", Computer Speech & Language 28(1), 295–313, 2014.

[171] O.V. Verkholyak, H. Kaya, and A.A. Karpov, "Modeling short-term and long-term dependencies of the speech signal for paralinguistic emotion classification", Tr. SPIIRAN, 18(1), 30-56, 2019.

[172] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series", Arbib, Michael A. (ed.). The handbook of brain theory and neural networks (Second ed.), The MIT press, 255-258, 1998.

[173] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts", in INTERSPEECH, Dresden, Germany, 2015.

[174] S. Wang, Z. Huang, Y. Qian and K. Yu, "Deep Discriminant Analysis for i-vector Based Robust Speaker Recognition", International symposium on Chinese Spoken Language Processing (ISCSLP), Taipei, 2018.

[175] Y. Wen, K. Zhang, Z. Li and Y. Qiao, "A Discriminative Feature Learning Approach for Deep Face Recognition", European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science, 9911, Springer, Cham, 2016.

[176] S. Xu, Y. Liu, X. Liu, "Speaker Recognition and Speech Emotion Recognition Based on GMM", International Conference on Electric and Electronics (EEIC), Hong Kong, China, 2013.

[177] P. Staroniewicz, "Considering basic emotional state information in speaker verification", International Conference on Biometrics and Forensics (IWBF), Limassol, Cyprus, 2016.

[178] M.S. Likitha, S.R.R. Gupta, K. Hashita, A.U. Raju, "Speech based human emotion recognition using MFCC", International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 2017.

[179] G. Disken, Z. Tufekci, L. Saribulut, U. Cevik, "A Review on Feature Extraction for Speaker Recognition under Degraded Conditions", IETE Technical Review, 34(3), 2017.

[180] A. Mansour, Z. Lachiri, "Speaker Recognition in Emotional Context", International Journal of Computer Science, Communication & Information Technology (CSCIT), 2(1), 1–4, 2015.

[181] A. S. Thakur, N. Sahayam, "Speech Recognition Using Euclidean Distance", International Journal of Emerging Technology and Advanced Engineering, 3(3), pp. 587-590, 2013.

[182] G. Aradilla, J. Vepa, H. Bourlard, "Using Posterior-Based Features in Template Matching for Speech Recognition", INTERSPEECH, Pittsburgh, Pennsylvania, 2006.

[183] M.D. Malkauthekar, "Analysis of euclidean distance and Manhattan Distance measure in face recognition", International Conference on Computational Intelligence and Information Technology, Mumbai, India, 2013.

[184] R.A. Lester-Smith, H.S. Brad, "The effects of physiological adjustments on the perceptual and acoustical characteristics of vibrato as a model of vocal tremor", The Journal of the Acoustical Society of America, 140(5), pp. 3827-3833, 2016.

[185] A. Kanervisto, V. Vestman, Md. Sahidullah, V. Hautamaki, T. Kinnunen, "Effects of gender information in text-independent and text-dependent speaker verification", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 2017.

[186] T. Jayasankar, K. Vinothkumar, A. Vijayaselvi, "Automatic Gender Identification in Speech Recognition by Genetic Algorithm", Applied Mathematics & Information Sciences, 11(3), pp. 907-913, 2017.

[187] T. Jayasankar, K. Vinothkumar and A. Vijayaselvi, "Gender-dependent emotion recognition based on HMMs and SPHMMs", International Journal of Speech Technology, 16(2), pp. 133–141, 2013.

[188] F. A. Shaqra, R. Duwairi, M. Al-Ayyoub, "Recognizing Emotion from Speech Based on Age and Gender Using Hierarchical Models", Procedia Computer Science, 151, pp. 37-44, 2019.

[189] I. Bisio, A. Delfino, F. Lavagetto, M. Marchese, A. Sciarrone, "Gender-Driven Emotion Recognition Through Speech Signals For Ambient Intelligence Applications", IEEE Transactions on Emerging Topics in Computing, 1(2), pp. 224-257, 2013.

[190] L. Zhang, L. Wang, J. Dang, L. Guo, Q. yu, "Gender-Aware CNN-BLSTM for Speech Emotion Recognition", International Conference on Artificial Neural Networks (ICANN), Rhodes, Greece, 2018.

[191] G.S. Archana, M. Malleswari, "Gender identification and performance analysis of speech signals", Global Conference on Communication Technologies (GCCT), Thuckalay, India, 2015.

[192] E. Ramdinmawii, V. K. Mittal, "Gender identification from speech signal by examining the speech production characteristics", International Conference on Signal Processing and Communication (ICSC), Noida, India, 2016.

[193] M. Gupta, S.S. Bharti, S. Agarwal, "Support vector machine based gender identification using voiced speech frames", International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, India, 2016.

[194] S.I . Levitan, T. Mishra, S. Bangalore, "Automatic identification of gender from speech", Speech Prosody, 2016.

[195] K.W. Godin, J.H.L. Hansen, "Physical task stress and speaker variability in voice quality", EURASIP Journal on Audio, Speech, and Music Processing, 2015(29), 2015.

[196] J. Marten, "Culture, Gender and The Recognition of The Basic Emotions", Psychologia, 2005(48), pp. 306-316, 2005.

[197] E. Coutinho, B. Schuller, "Shared acoustic codes underlie emotional communication in music and speech—Evidence from deep transfer learning", PLoS ONE, 12(6), 2017.

[198] H.H. Son, "Toward a proposed framework for mood recognition using LSTM Recurrent Neuron Network", Procedia Computer Science, 109, pp. 1028-1034, 2017.

[199] H. Kamper, K. Livescu, S. Goldwater, "An embedded segmental K-means model for unsupervised segmentation and clustering of speech", IEEE Automatic Speech

Recognition and Understanding Workshop (ASRU), Okinawa, Japan, pp. 719–726, 2017.

[200] D. Xu, Y. Tian, "A Comprehensive Survey of Clustering Algorithms", Ann. Data Sci., 2, pp. 165–193, 2015.

[201] K. Wong, "A Short Survey on Data Clustering Algorithms", International Conference on Soft Computing and Machine Intelligence (ISCMI), Hong Kong, pp. 64–68, 2015.

[202] A.S. Shirkhorshidi, S. Aghabozorgi, T.Y. Wah, "A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data", PLoS ONE, 10, pp. 1–20, 2015.

[203] C. Bouveyron, S. Girard, C. Schimid, "High-Dimensional Data Clustering", Elsevier Comput. Stat. Data Anal., (52), pp. 502–519, 2007.

[204] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, J. Long, "A Survey of Clustering With Deep Learning: From the Perspective of Network Architecture", IEEE Access, 6, pp. 39501–39512, 2018.

[205] M. Ravanelli, M. Omolongo, "Automatic context window composition for distant speech recognition," Speech Communication, 101, pp. 34-44, 2018.

[206] A.L. Maas, Q.V. Le, T.M. O'Neil, O. Vinyals, P. Nguyen, A.Y. Ng, "Recurrent Neural Networks for Noise Reduction in Robust ASR," INTERSPEECH, Portland, USA, 2012.

[207] C. Do, Y. Stylianou, "Improved Automatic Speech Recognition using Subband Temporal Envelope Features and Time-delay Neural Network Denoising Autoencoder", INTERSPEECH, Stockholm, Sweden, 2017.

[208] N. Cummins, J. Epps, E. Ambikairajah, "Spectro-temporal analysis of speech affected by depression and psychomotor retardation," IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada, 2013.

[209] S. Choe, S. Chung, Y. Ji, H. Kang, "Orthonormal Embedding-based Deep Clustering for Single-channel Speech Separation," arXiv:1901.04690 [eess.AS], 2019.

[210] J. Xie, R. Girshick,, A. Farhadi, "Unsupervised deep embedding for clustering analysis," The International Conference on Machine Learning (ICML), New York, USA, 2016.

[211] X. Guo, L. Gao, X. Liu, J. Yin, "Improved Deep Embedded Clustering with Local Structure Preservation," The International Joint Conference on Artificial Intelligence (IJCAI), Melbourne, Australia, 2017.

[212] S. Tripathi, A. Ramesh, A. Kumar, C. Singh, P. Yenigalla, "Learning Discriminative features using Center Loss and Reconstruction as Regularizer for Speech Emotion Recognition," arXiv:1906.08873 [cs.SD], 2019.

[213] P. Y. Simard, Y. LeCun, "Reverse TDNN: An Architecture For Trajectory Generation," presented at the Advances in Neural Information Processing Systems (NIPS), Colorado, USA, 1991.

[214] X. Feng, Y. Zhang, J. Glass, "Speech Feature Denoising and Dereverberation via Deep Autoencoders for Noisy Reverberant Speech Recognition," IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Florence, Italy, 2014.

[215] V. Peddinti, G. Chen, D. Povey, S. Khudanpur, "Reverberation robust acoustic modeling using i-vectors with time delay neural networks," INTERSPEECH, Dresden, Germany, 2015.

[216] N. Le and J. Odobez, "Robust and Discriminative Speaker Embedding via Intra-Class Distance Variance Regularization," INTERSPEECH, Hyderabad, India, 2018.

[217] B. B. Meier, I. Elezi, M. Amirian, O. Durr, T. Stadelmann, "Learning Neural Models for End-to-End Clustering," The Workshop on Artificial Neural Networks in Pattern Recognition (ANNPR), Siena, Italy, 2018.

[218] R. K. Kumar, L. Birla, K. S. Rao, "A robust unsupervised pattern discovery and clustering of speech signals," Pattern Recognition Letters, 116, pp. 254-261, 2018.

[219] H. Wang, T. Lee, C.C. Leung, B. Ma, H. Li, "Acoustic Segment Modeling with Spectral Clustering Methods," IEEE/ACM Trans. on Audio, Speech, and Language Processing, 23(2), pp. 264-277, 2015.

[220] M. Sbert, M. Chen, J. Poch, A. Bardera, "Some Order Preserving Inequalities for Cross Entropy and Kullback–Leibler Divergence," Entropy, 20(12), pp. 1-10, 2018.

[221] A.M. Shahsavarani, E.A.M. Abadi, M.H. Kalkhoran, "Stress: Facts and Theories through Literature Review," International Journal of Medical Reviews, 2(2), pp.230-241, 2015.

[222] O. Chapelle, B. Scholkopf, A. Sien, "Semi-Supervised Learning", The MIT Press: London, UK, 2007.

[223] I. Davidson, S. Basu, "A Survey of Clustering with Instance Level Constraints", ACM Trans. Knowl. Discov. Data, 1, 2007.

[224] K. Wagstaff, C. Cardie, "Clustering with Instance-level Constraints", International Conference on Machine Learning (ICML), Stanford, CA, USA, pp. 1103–1110, 2000.

[225] G. Xu, Y. Zong, Z. Yang, "Constraint-based Clustering Algorithm", Applied Data Mining, CRC Press: Boca Raton, FL, USA, pp. 89–92, 2013.

[226] S.L. Suarez Gomez, J.D. Santos Rodriguez, F.J. Iglesias Rodriguez, F. De Cos Juez, "Analysis of the Temporal Structure Evolution of Physical Systems with the Self-Organising Tree Algorithm (SOTA): Application for Validating Neural Network Systems on Adaptive Optics Data before On-Sky Implementation", Entropy, 19, 103, 2017.

[227] V.V. Nanavare, S.K. Jagtap, "Recognition of Human Emotion from Speech Processing", Procedia Computer Science, 49, pp. 24–32 2015.

[228] A. Lausen, S. Annekathrin, "Gender Differences in the Recognition of Vocal Emotions, Frontiers in psychology, 9(882), pp. 1–22, 2018.

[229] R. Meyes, M. Lu, C. Waubert de Puiseau, T. Meisen, "Ablation Studies in Artificial Neural Networks. arXiv:abs/1901.08644, 2019.

[230] Nummenmaa, L., Saarimäki, H. Emotions as discrete patterns of systemic activity. *Neuroscience Letters.* **693**, 3–8 (2019).

[231] Harrison, P.G., Strulo, B. Stochastic Process Algebra for Discrete Event Simulation. *In: Baccelli F., Jean-Marie A., Mitrani I. (eds) Quantitative Methods in Parallel Systems. Esprit Basic Research Series.* Springer, Berlin, Heidelberg, (2019).

[232] Zhai, J., Yang, Q., Su, F., Xiao, J., Wang, Q., Li, M. Stochastic Process Algebra Based Software Process Simulation Modeling. In Trustworthy Software Development Processes, International Conference on Software Process (ICSP). Vancouver, Canada (2009).