

DEVELOPMENT OF SPEECH RECOGNITION SYSTEM
IN SOCIAL SIGNAL PROCESSING APPLICATION

NGEW CHI NEE

This Report Is Submitted In Partial Fulfilment of Requirements for the Bachelor
Degree of Electronic Engineering (Wireless Communication)

Faculty of Electronics and Computer Engineering
Universiti Teknikal Malaysia Melaka

JUNE 2015



UNIVERSITI TEKNIKAL MALAYSIA MELAKA
FAKULTI KEJURUTERAAN ELEKTRONIK DAN KEJURUTERAAN KOMPUTER

BORANG PENGESAHAN STATUS LAPORAN
PROJEK SARJANA MUDA II

DEVELOPMENT OF SPEECH RECOGNITION SYSTEM

Tajuk Projek : IN SOCIAL SIGNAL PROCESSING APPLICATION

Sesi Pengajian :

1	4	/	1	5
---	---	---	---	---

Saya NGEW CHI NEE

(HURUF BESAR)

menyku membenarkan Laporan Projek Sarjana Muda ini disimpan di Perpustakaan dengan syarat-syarat kegunaan seperti berikut:

1. Laporan adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan membuat salinan laporan ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. Sila tandakan () :

SULIT*

*(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972).

TERHAD**

***(Mengandungi maklumat terhad yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)

TIDAK TERHAD

Disahkan oleh:



(LANDATANGAN PENULIS)


DR. ABDUL MAJID BIN DARSONG
(GOLD DAK) PENYERAH KEMAH
PUSAT KEJURUTERAAN ELEKTRONIK DAN KEJURUTERAAN KOMPUTER
Universiti Teknikal Malaysia Melaka (UTeM)
Hang Tuah Jaya
76100 Durian Tunggal, Melaka


Tarikh: 5/6/2015

Tarikh: 5/6/2015

"I hereby declare that this report is the result of my own work except for quotes as cited in the references"

Signature : 
Name : NGEW CHI NEE
Date : 5th JUNE 2015

“I hereby declare that I have read this report and in my opinion this report is sufficient in terms of scope and quality for the award of Bachelor of Electronic Engineering (Wireless Communication) With Honours”

Signature : 
Supervisor's Name : DR. ABD MAJID BIN DARSONO
Date : 5th JUNE 2015

ACKNOWLEDGEMENT

First of all, I would like to convey my gratefulness to my supervisor, Dr. Abdul Majid bin Darsono, for his guidance, motivation and invaluable supervision throughout my research work.

Here also, I would like to thank Dr. Masrullizam bin Mat Ibrahim and En. Ahamed Fayeez bin Tuani Ibrahim, who are the lecturers in Faculty of Electronic Engineering and Computer Engineering in sharing their knowledge and recommendation.

Lastly, I wish to express my deepest gratitude to my parents and my family for their continuous encouragement. I would like to thank my friends for their support and encouragement.

ABSTRACT

Social Signal Processing (SSP) has been widely used in robotic and computer as one of the Artificial Intelligent (AI) in contribute to human machine interaction. One of the examples of SSP is to recognize human emotions. In this research, a system which capable to recognize different states of emotion in speech is successfully developed using Support Vector Machine (SVM) technique. The first two main objectives of this research are to develop a speech emotion recognition system and graphical user interface (GUI) using MATLAB software. Besides that, performance of the system also has been studied based on the percentage accuracy. Linguistic Data Consortium (LDC) is used as the database. The features contained in the LDC voice samples are extracted and used to develop the dataset. The methods used to extract the features include energy, pitch, formant, Mel-Frequency Cepstrum Coefficient (MFCC) features. Statistic such as mean value is calculated. A training model is introduced to train the classifier in the system and a testing model is used to analyse the system. 'Happiness', 'Anger', 'Sadness' and 'Neutral' are examined. Gender dependent and independent test are studied to analyse the gender impact on the performance of emotion recognition. The result shows that the gender independent test has higher accuracy than the gender dependent test. Besides that, male has better emotion performance than the female. In general, the proposed system has higher performance than the existing system.

ABSTRAK

Pemrosesan Isyarat Sosial (SSP) telah digunakan secara meluas dalam bidang robotik dan komputer. Ia digunakan sebagai salah satu jenis Kepintaran Buatan (AI) yang menyumbang kepada penyelidikan interaksi antara mesin dengan manusia seperti pengecaman emosi melalui suara. Dalam kajian ini, satu sistem yang dapat mengecam emosi melalui suara telah dibangunkan dengan menggunakan pengklasifikasi Sokongan Mesin Vektor (SVM). Objektif kajian ini adalah untuk membangunkan satu sistem yang dapat mengecam emosi melalui suara dan membangunkan GUI (Antara Muka Pengguna Grafik) dengan menggunakan perisian MATLAB. Selain itu, prestasi sistem untuk pengecaman emosi juga perlu dikaji. *Linguistic Data Consortium* (LDC) telah digunakan sebagai pangkalan data emosi. Ciri-ciri seperti tenaga, nada, formant, *Mel-Frequency Cepstral Coefficient* (MFCC) yang dipetik daripada suara telah digunakan. sebagai set data ucapan beremosi dan seterusnya statistik kasar telah dikirakan. Sistem ini dibangunkan melalui dua peringkat, iaitu model latihan dan model uji. Emosi yang dikaji dalam kajian ini ialah gembira, marah, sedih dan neutral. Prestasi sistem ini telah dianalisa berdasarkan model jantina dan model bebas jantina. Keputusan menunjukkan bahawa prestasi model bebas jantina adalah lebih tinggi daripada model berdasarkan jantina. Selain itu, lelaki mempunyai pengecaman emosi yang lebih baik daripada wanita. Secara amnya, sistem ini mencapai tahap prestasi yang lebih baik daripada sistem yang telah sedia ada.

CONTENTS

CHAPTER	TITLE	PAGE
	PROJECT TITLE	i
	REPORT STATUS APPROVAL FORM	ii
	DECLARATION	iii
	SUPERVISOR APPROVAL	iv
	ACKNOWLEDGEMENT	v
	ABSTRACT	vi
	ABSTRAK	vii
	TABLE OF CONTENTS	viii
	LIST OF FIGURES	xii
	LIST OF TABLES	xiv
	LIST OF ABBREVIATION	xv
I	INTRODUCTION	1
	1.1 Background Study	2

1.1.1	Speech Recognition System	2
1.1.2	Social Signal Processing	3
1.2	Problem Statement	4
1.3	Objectives	5
1.4	Motivation	5
1.5	Scope of Research	6
1.6	Thesis Outlines	7
II	LITERATURE REVIEW	8
2.1	Speech Signal	9
2.2	Emotions	10
2.3	Speech Emotion Recognition System	11
2.3.1	Emotional Speech Databases	13
2.3.2	Features Extraction	15
2.3.2.1	Vocal Tract Spectrum Features	16
2.3.2.2	Prosodic Features	19
2.3.2.3	Non-linear Features	21
2.3.3	Features Classification	22
2.3.3.1	Hidden Markov Model (HMM)	22
2.3.3.2	Gaussian Mixture Model (GMM)	23
2.3.3.3	Support Vector Machine (SVM)	24

III	METHODOLOGY	27
3.1	Emotional Speech Database Set Up	28
3.2	Description of the Methodology	29
3.3	Feature Extraction Method	31
3.3.1	Pitch	31
3.3.2	Energy	32
3.3.3	Formant	32
3.3.4	Mel-Frequency Cepstral Coefficient	32
3.4	Feature Selection	32
3.5	Feature Classification	33
3.6	Software Requirement	34
IV	RESULTS AND DISCUSSION	35
4.1	Accuracy of Multiclass Approach	36
4.2	Classification Analysis	37
4.2.1	Different Database Testing	38
4.2.2	Feature Extraction	39
4.2.3	Feature Selection	43
4.2.4	Cross-Validation Test	44
4.2.5	Gender Dependent Test	45
4.3	Graphical User Interface (GUI)	46

V	CONCLUSION	49
	5.1 Conclusion	50
	5.2 Potential of Commercialization	51
	5.3 Recommendation and Research Advancement	51
	REFERENCES	52
	APPENDIX A	55

LIST OF FIGURES

NO	TITLE	PAGE
1.1	Basic block diagram of speech recognition	2
1.2	Behaviour cues and social signal	4
2.1	Speech signal for the word ‘one’	9
2.2	The model of emotions in 2D with a valence and an arousal axis	10
2.3	Basic block diagram of speech emotion recognition system	11
2.4	Categories of speech features	15
2.5	Block diagram of MFCCs	17
2.6	Block diagram of estimating formant frequency using LPC	18
2.7	HMM topology used in emotion recognition	22
2.8	SVM with hyper plane	24
2.9	Example of multi-level SVM classification for four different classes	26
3.1	Flow chart of the proposed speech emotion recognition system	30
3.2	Block diagram of proposed feature extraction	31
3.3	Flow chart of proposed SVM classification	33

4.1	Feature extraction of a female utterance	40
4.2	Feature extraction of a male utterance	41
4.3	MFCC of 'happy' utterance	42
4.4	Feature selection using mutual information	43
4.5	GUI layout	46
4.6	Simulation part of GUI	47
4.7	Result of GUI Layout	48

LIST OF TABLES

NO	TITLE	PAGE
2.1	Summary of existing database for emotion recognition	14
2.2	Analysis of vocal tract spectrum features	16
2.3	Acoustic characteristics of different types of emotion categories	20
4.1	Comparison of Multiclass SVM Accuracy (%)	36
4.2	Performance accuracy (%) for two different databases	38
4.3	Accuracy of SER System for different cross validation number	44
4.4	Accuracy of Gender Dependent Test of Proposed SER System	45

LIST OF ABBREVIATIONS

AI	-	Artificial Intelligent
DCT	-	Discrete Cosine Transform
GMM	-	Gaussian Mixture Model
FFT	-	Fast Fourier Transform
HMM	-	Hidden Markov Model
MFCC	-	Mel Frequency Cepstral Coefficients
LPC	-	Linear Predictive Coding
LPCC	-	Linear Prediction Cepstral Coefficients
OAA	-	One Against All
SER	-	Speech Emotion Recognition
SSP	-	Social Signal Processing
SVM	-	Support Vector Machine
TEO	-	Teager Energy Operator
ZCR	-	Zero-Crossing Rate

CHAPTER 1

INTRODUCTION

This chapter provides an introduction of the research entitled “Development of Speech Recognition System in Social Signal Processing Applications”. The chapter is organized in six sections. Background study, objectives, motivation of the project, problem statements, scope of the research and outline of this report is presented.

1.1 Background Study

1.1.1 Speech Recognition System

Speech recognition is a technology that recognizes the spoken languages and converts to a text or to instruction by digitizing the sound and matching its pattern against the stored patterns [1]. Speech recognition has been widely applied especially in voice user interfaces such as in car system, telephony, education, robotic control and etc. For example, speech recognition system has been implemented into telephony especially in smart phones as additional function such as voice dialling, speech-to-text processing, and others control. Few of the most well-known speech recognition system are Google Voice Search Engine, LG's Voice Mate, Apple's Series and Samsung's S Voice. Speech recognition system with high accuracy in recognizing the languages has become an important research challenge in recent years of research and development.

Figure 1.1 shows the basic block diagram of speech recognition system. Basically, the speech will input to the system through a microphone and undergoes speech feature extraction to interpret the speech. Then, the classifier will recognize the input pattern and classify it as the output. Extracting the features from the voice signal of a word is a very important process because it is the first step in the system and the accuracy of the system is depending on the data that extracted from the signal voice. Therefore, good features must be extracted from the speech in order to achieve higher accuracy of the system performance.

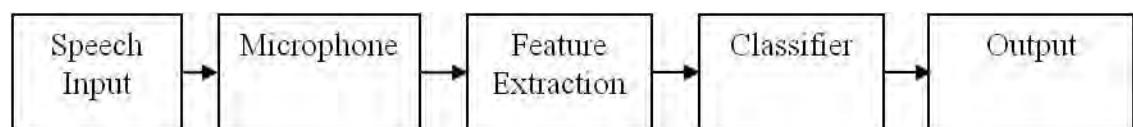


Figure 1.1: Basic block diagram of a speech recognition system.

The speech input can be in the form of speaker-independent or speaker-dependent speech. User is not required to train the system for the speaker-independent speech recognition, for example, the system is developed to operate for any speaker. However, a particular speaker is essential to read text sections for the training in the speech recognition system.

1.1.2 Social Signal Processing

Social Signal Processing (SSP) [2] is the combination of research in human-science and computer science. The key ideas of SSP are to modelling the social interaction and providing the abilities for computers to understand and synthesize the non-verbal behaviour cues. Figure 1.2 shows the categories of social signals. In this research, voice quality of vocal behaviour in social signal is studied. This part is discussed in detail in Chapter 2 for the literature review of feature extraction.

Recently SSP has been widely used in robotic and computer as one of the Artificial Intelligent (AI) in contribute to human machine interaction. The latest 2014 robot that equipped with social signal processing is the “Pepper” introduced by Japan billionaire Masayoshi. Pepper is a robot that programmed to “feel” the human emotions through recognizing human expression and human voice tones. Besides that, with the implementation of cloud AI, “Pepper” able to share the human feelings among others “Peppers” and develop their own approaches accordingly. This is not the first time that social signal processing is applied into robotic; in 1999 the Sony Corporation has developed AIBO entertainment robot dog, Honda with the ASHIMO robot and etc.

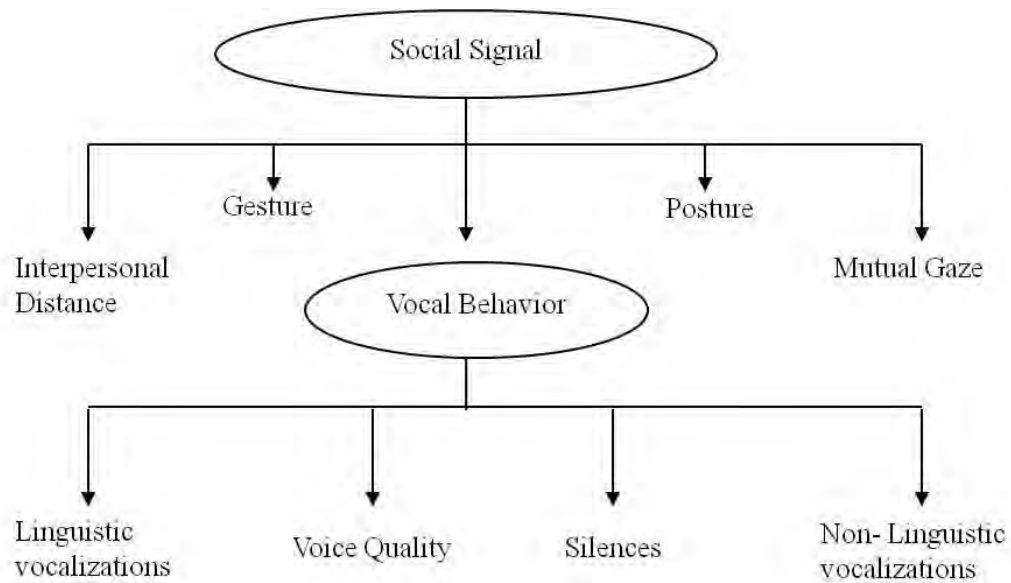


Figure 1.2: Behaviour cues and social signal

1.2 Problem Statement

Speech emotional database is critically important in SER as it is used as the input to the system. However, due to the existence of different cultural, different environment and different speaking styles, the speech features are directly affected and the expression of emotion is also differed.

Speech features includes spectral features and prosodic features, for example, formants, pitch and energy are discovered useful in SER. However, one single type of speech feature used in the research is insufficient to classify the emotions correctly. This is because different emotions may exhibit similar properties.

1.3 Objectives

There are three objectives to achieve in this project which are:

- a) To develop a reliable speech emotion recognition (SER) system using MATLAB software.
- b) To develop a graphical user interface (GUI) for the SER system proposed.
- c) To analyse the accuracy of the developed system in recognizing emotion occurrence in particular speech signal.

1.4 Motivation

Speech is the primary approach to communicate with each other and the voice capability allows people to interact and exchange ideas. Hence, the spoken language becomes one of the main attributes of humanity. Emotions occur naturally in daily life and every emotion has their own characteristics. SER becomes wide area of interest for researchers in human-machine interaction recently. SER has various applications in the fields of security, medicine, learning, and entertainment. This research will open up new possibilities for development of service industry and healthcare application. In service industry, such as Maxis, Celcom, and Telekom can used the SER system for monitoring the performances of their operators on duty in attending the caller manner. Besides that, in medicine field, it can be used as an analysis of emotional state of human being for psychology purpose.

1.5 Scopes of Research

The scope of this research is to investigate the relationship between the techniques of feature extraction used for speech features such as spectral features and prosodic features with the accuracy of the SER system.

In this research, a number of speakers are randomly selected from a group of non-professional and first time trained volunteers are requested to record emotional speeches in English language to collect the speech utterance. The speakers are requested to act in four different basic states of emotions such as anger, happy, sad and fear. The methods that used to extract the features include energy, pitch, formant and Mel-Frequency Cepstral Coefficient (MFCC). Only audible segments are considered in analysing the features. Then, analysis SER techniques such as mean, maximum, minimum and standard deviation are carried out. As for the classifier, Support Vector Machine (SVM) is employed. This project is mainly software based and thus the system is implemented using MATLAB software only.

1.6 Thesis Organization

In this research, a speech emotion recognition system, which capable to recognize emotions in speech signal is developed. This thesis is structured in five chapters. Chapter 1 explained the background study, objectives, motivation, problem statement, and scope of this project.

Chapter 2 gives an overview of the literature review related to the speech emotion recognition. The chapter describes different aspects; among them are speech signal properties, emotion models, SER structure system which included the block diagram of SER system, types of database, then the feature extraction methods and lastly is the classification schemes used in emotion recognition.

Chapter 3 is devoted to the methodology of developing the SER system in this research. Flow charts for feature extraction and classification techniques that designed in this research are explained in this chapter. Furthermore, possible prosodic correlates of emotions in speech that can be extracted such as pitch and energy and feature vectors are described in details.

Chapter 4 gives the detailed analysis of the experimental results and comparison of the results obtained with the existing projects. Lastly, chapter 5 gives the achievement of this project, discusses the project's market potential and the future improvements.

CHAPTER 2

LITERATURE REVIEW

This chapter is organized in four sections. Speech is the most fundamental communication tools in human life and has its own properties. In the beginning of this chapter, the basic properties of the speech are discussed. A precise idea of what emotions are is essential in order to recognize the emotions. Hence, a short brief of two-dimensional emotion model is covered in this chapter. After that, a general structure of the speech emotion recognition (SER) system is presented and every stage in designing the SER system such as speech emotional database, feature extraction and feature classification are discussed.

2.1 Speech signal

Speech is the primary approach to communicate with each other and the voice capability allows people to interact and exchange ideas. Speech recognition has developed as a recent research area in Human-Machine Interaction. Speech signal acts as the input to the speech recognition. According to Nyquist theorem, a continuous signal can be sampled without loss of information if only the components of frequency contained in the signal less than one half of the sampling rate. For example, human ears only used the frequency which is less than 8000 Hz for perception of speech, so it only requires 16 kHz of sampling rate for most speech processing purposes.

A good knowledge of basic properties of the speech signal is essential in the development of speech recognition system. One of the speech properties is that it will slowly varying over a short period of time and remain stationary over a long period of time. As the speaker usually takes some time to speak the word when recording starts, the beginning of the speech signal recording is corresponding to silence or background noise. These properties of the speech signal are showed in Figure 2.1.

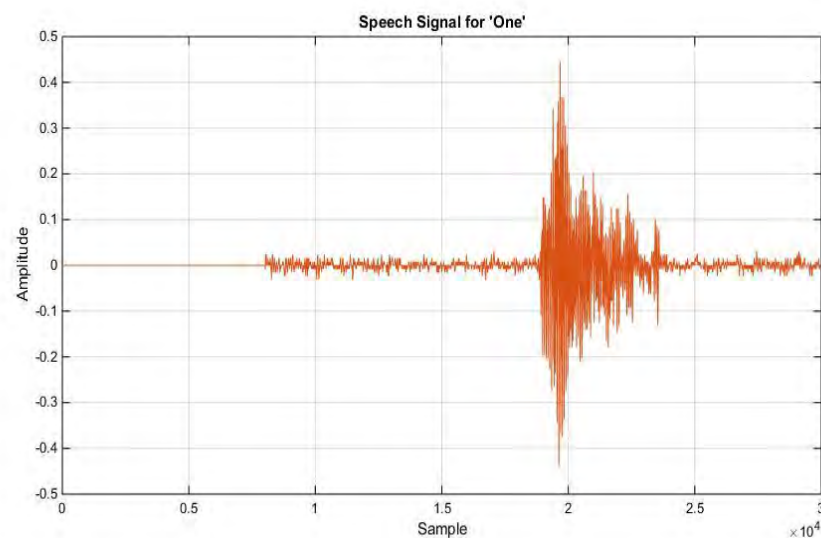


Figure 2.1: Speech signal for the word 'one'

2.2 Emotions

'Palette theory' [3] states that any emotion can be decomposed into its primary emotions as the way that any colour is formed from primary colour. The basic emotions are happy, fear, sad, anger, disgust, surprise and neutral. Figure 2.2 shows the model of emotions in two-dimensional with valence axis, which is from negative to positive, and arousal axis, which is usually from high to low. For example, from the Figure 2.2, happy is an affective state with moderate arousal and high valence level whereas sad is characterized by low arousal and negative valence.

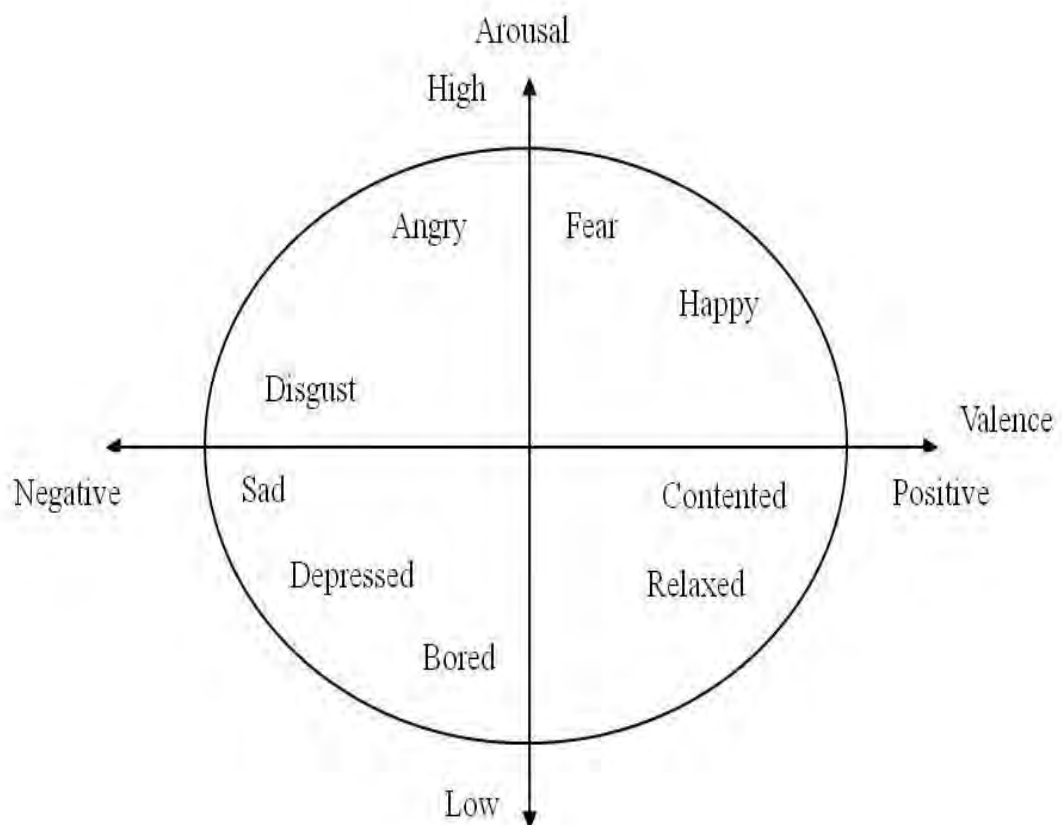


Figure 2.2: The model of emotions in 2D with a valence and an arousal axis.