Lang Resources & Evaluation (2019) 53:17–55
https://doi.org/10.1007/s10579-018-9428-9

ORIGINAL PAPER

Developing a Thai emotional speech corpus from Lakorn (EMOLA)

Sawit Kasuriya[1] · Thanaruk Theeramunkong[2,4] · Chai Wutiwiwatchai[3] · Piyawat Sukhummek[1]

Published online: 28 November 2018
© Springer Nature B.V. 2018

Abstract Advances in emotional speech recognition and synthesis essentially rely on the availability of annotated emotional speech corpora. As a low resource language, the Thai language critically lacks corpora of emotional speech, although a few corpora have been constructed for speech recognition and synthesis. This paper presents the design of a Thai emotional speech corpus (namely EMOLA), its construction and annotation process, and its analysis. In the corpus design, four basic types with twelve subtypes of emotions are defined with consideration of the Pleasure-Arousal-Dominance emotional state model. To construct the corpus, a series of Thai dramas (1397 min) were selected and its video clips of approximately 868 min were annotated. As a result, 8987 transcriptions (of conversation turns)

✉ Thanaruk Theeramunkong
thanaruk@siit.tu.ac.th

Sawit Kasuriya
sawitk@nectec.or.th

Chai Wutiwiwatchai
chai@nectec.or.th

Piyawat Sukhummek
jonechov@hotmail.com

[1] School of Information, Computer and Communication Technologies, Sirindhorn International Institute of Technology, Thammasat University, 99 Phaholyothin Road, Khlong Luang, Pathumthani 12120, Thailand

[2] School of Information, Computer and Communication Technologies, Sirindhorn International Institute of Technology, Thammasat University, 99 Moo 18 Paholyothin Road, Klong Luang, Rangsit, Pathumthani 12121, Thailand

[3] National Electronics and Computer Technology Center (NECTEC), 112 Phahonyothin Road, Khlong Nueng, Khlong Luang District, Pathumthani 12120, Thailand

[4] Academy of Science, Royal Society of Thailand, Sanam Sueapa, Khet Dusit, Bangkok 10300, Thailand

were derived in total, with each transcription tagged as one basic type and a few subtypes. Finally, an analysis was conducted to describe the characteristics of this corpus in three sets of statistics: collection-level, annotator-oriented and actor-oriented statistics.

Keywords Simulated emotional speech corpus · Thai emotional speech corpus · Pleasure-Arousal-Dominance emotional state model · Collection-level · Annotator-oriented · Actor-oriented statistics

1 Introduction

While researchers in speech technology have presented various efficient techniques and applications that can recognize and/or synthesize the human voice, there are still only a few proposals on how to cope with feelings in the human voice. It is a challenging task to characterize and to recognize emotional states from human speech due to difficulties in defining emotions which are caused by both cultural and individual differences in perception and the expression of emotion. In order to advance in the recognition of emotion in a particular language, it is necessary to collect and construct an emotional speech corpus of that language with annotation of moods in a systematic and consistent way. In the past, speech corpora for recognizing emotion were constructed in several languages, such as English, Spanish, Chinese, and Japanese and an intensive survey on emotional speech corpora construction was presented by Ververidis and Kotropoulos (2006). Collection of natural emotional speech is difficult since we cannot know where or when speeches expressing a particular emotion will occur. Therefore, most researchers have constructed emotion corpora by asking a professional actor to perform a mood state when speaking (simulated emotional speech) or by creating a situation to elicit a response from a person expressing the target emotion in his/her speech (elicited emotional speech) (Busso et al. 2008).

Currently, to the best of our knowledge, there is no extant Thai emotional speech corpus. In this paper, to represent emotion in speech, two emotional state models are applied: (1) a numerical state model, namely Pleasure-Arousal-Dominance (PAD) and (2) a categorical state model, including four basic emotional types with twelve subtypes. For the sake of the ease of emotional speech collection with clear emotional state, we decided to construct a corpus with simulated emotional speech where speech is uttered by professional actors and actresses. Here, the three steps of corpus construction are transcription (subtitling), metadata preparation (formatting) and emotion annotation (labeling). In this work, the selected series of Thai dramas contain approximately 1520 min of video clips. These are segmented into 8987 turns of conversation, each turn is transcribed and enriched with metadata for facilitating the tagging process, and finally tagged by using the two emotional state models. To characterize our corpus, we have extracted a number of statistics based on speakers, annotators and emotion tags.

In the rest of our study, Sect. 2 provides the background on human emotion theories as well as a literature review on emotional speech corpus construction. Our corpus design and construction are described in Sect. 3. The corpus design includes two main concerns: the tag format based on Document Type Definition (DTD) in the form of Extensible Markup Language (XML) and the tagging guidelines for corpus annotators. As for the corpus construction, the three-step process is depicted in the order of subtitling, formatting and labeling. Section 4 presents a number of statistics extracted from our constructed corpus, namely EMOLA (i.e. Thai emotional speech corpus from Lakorn) in several aspects, mainly for analyzing the tagging results and the annotators' style of tagging. Finally, a conclusion and further works are summarized in Sect. 5.

2 Literature review

This section presents the background on human emotion theories as well as a literature review on emotional speech corpus construction. As for the former, definitions of psychological emotional states and definition issues are addressed, followed by how emotional states are expressed in speech signals. For the latter, a number of previous works related to construction of emotional speech corpora are described.

2.1 Theories of human emotion

While research on human emotion has increased significantly over the past two decades in several fields, including psychology, neuroscience, endocrinology, medicine, history, sociology, and computer science, numerous theories have been developed to explain the origin, neurobiology, experience, and function of emotions, such as primary emotions (Plutchik 1980, 1984), basic emotions (Stein and Oatley 1992), normal emotions (Kaiser and Scherer 1998), and emotional responses (Scherer 1986).

Among these, several studies have explored relations between emotions and speech signals with various emotion types, such as anger, happiness, sadness, fear, neutral (Scherer and Tannenbaum 1986). While human emotion is of two modes: speaker mode versus listener mode, human emotion can be interpreted by these two categories: vocal expression and perception of emotion. In this work, we focus on perception since it is difficult to establish the emotional intention of speech, compared with directly interpreting the vocal emotion that we perceive, thus our task is to create a system that recognizes emotions. Thus, the most challenging part of our task is how to cope with different interpretations of individuals on emotions when they perceive speech. The definition of emotion and the classification of emotion are described in the next section.

2.1.1 Definition and classification of emotion

So far there have been several approaches proposed by psychologists to clarify human emotion. These approaches are based on different theories and their definitions of emotion are inconsistent, uncertain, and arguable. In this subsection, some interesting theories are briefly reviewed. As an example of an early modern theory, Plutchik proposed the so-called Plutchik's wheel of emotions to illustrate different emotions in a complete and comprehensive way (Plutchik 1980). The wheel is composed of eight primary bipolar emotions: joy versus sadness; anger versus fear; trust versus disgust; and surprise versus anticipation. Additionally, each emotion is expressed by a color and it is possible to express emotion intensity by color intensity and to mix emotions like mixing colors to form different emotions. As parts of Plutchik's ten postulates, it is possible to analyze and interpret emotion by using basic emotions, emotion combination, emotion opposites, emotion similarity, and emotion intensity. So far there have been two fundamental approaches in research on emotion classification: (1) emotions as discrete and fundamentally different constructs and (2) emotions as points characterized by a dimensional basis in a coordinate system (Cowie and Cornelius 2003).

(1) Emotions as discrete categories

A naive and straightforward way to explain emotions is to use emotional categories, each of which is expressed by an emotional word, referring to a human's state of emotion. In this discrete emotion theory, an innate set of basic emotions that are cross-culturally recognizable is defined. These basic emotions can be distinguished by an individual's facial expression and biological processes (Colombetti 2009). However, two issues are (1) how to define the set of basic emotions, and (2) how a listener perceives an emotion. In research on the first issue, Paul Ekman and his colleagues (Ekman 1992) conducted an intensive cross-cultural study on basic emotions and reached the conclusion that the six basic emotions are anger, disgust, fear, happiness, sadness, and surprise. The work reported that each emotion has particular characteristics attached, allowing them to be expressed in varying degrees, and that they act as discrete categories rather than an individual emotional state.

In an independent work (Bann and Bryson 2012), Bann and Bryson have proposed a theory that people convey their understanding of emotions through the language they use that surrounds emotion keywords. They suggest that the more distinct language is used to express a certain emotion, the more distinct the perception of that emotion is, and thus more basic. By experiments, Bann and Bryson's most semantically distinct emotion set is coincidentally the same as the basic emotion set proposed in Ekman et al. (1972).

As for the second issue, how a listener perceives an emotion, it is common that different listeners may have different opinions and then provide different emotional words to describe a certain emotional state. Matching between emotional states and emotional words is somehow subjective according to individual perception due to language usage, experience, circumstance, and the personality of the listener.

(2) Emotions in a dimensional model

As the second approach, an alternative is to design a set of primitive properties, each of which refers to a dimension in a systematic space. Several researchers prefer this approach for both theoretical and practical reasons. As the pioneer of modern psychology, Wilhelm Max Wundt proposed in 1897 that emotions can be described by three dimensions: “pleasurable versus unpleasant”, “arousing versus subduing” and “strain versus relaxation” (Wundt 1897). Around half a century after this first proposal, in 1954, Harold Schlosberg named three dimensions of emotion: “pleasantness–unpleasantness”, “attention–rejection” and “level of activation” in 1954 (Schlosberg 1954).

A so-called Positive Activation-Negative Activation (PANA) model, originally created by Watson and Tellegan in 1985, suggests that a positive effect and a negative effect are two separate systems (Watson and Tellegan 1985). In the PANA model, the vertical axis represents a low to high positive effect and the horizontal axis represents a low to high negative effect. The dimensions of valence and arousal lie at a 45-degree rotation over these axes. Recently, the circumplex emotion model developed by Posner and his colleagues has suggested that emotions are distributed in a two-dimensional circular space, containing arousal, and valence dimensions (Posner et al. 2005). This two-dimensional model of emotion attempts to characterize human emotions by incorporating valence and arousal and intensity dimensions.

Another model developed by Cowie and Cornelius suggested two dimensions, namely, activation and evaluation spaces (Cowie and Cornelius 2003). A three-dimensional model, the PAD emotional state model, was developed by Albert Mehrabian and James A. Russell, to describe and measure emotional states in the dimensions of Pleasure, Arousal and Dominance (Mehrabian and Russell 1974; Russell and Mehrabian 1977; Mehrabian 1996). Originally, some researchers used sixteen scales for a pleasure dimension, nine scales for an arousal dimension, and nine scales for a dominance dimension (Mehrabian 1995). They strongly believe these three dimensions characterize all emotions and any other concepts of emotional states. More specifically, Zhang et al. (2008) found anger to have PAD values of [− 0.90, + 0.79, + 0.95] while happiness tended to be expressed by [+ 0.68, + 0.68, + 0.43] on average. In Havlena and Holbrook (1986), anger is [− 0.85, + 0.23, − 0.32] and happiness is [+ 0.93, − 0.12, + 0.45].

2.1.2 Emotional states in speech

Compared with non-verbal communication, such as facial expression and gesture, speaking is a straightforward way to exchange information and emotions. To study emotions in speech is one of the important aspects towards understanding people’s emotions. Even speech conveys emotions in two parts: content semantics and acoustic properties, so understanding emotion via acoustic properties extracted from speech signals is important. In many cases, speech with the same content may be interpreted differently in emotional terms when it is uttered with different acoustic patterns. Some psychologists have summarized emotional states relating to acoustic properties rather than content, for example, Scherer (1995). It is well known that amplitude, energy of speech signal or pitch variation can express differences between anger and happiness while pitch contours show differences between states of sadness and pleasantness.

Cowie and Cornelius (2003) have studied relations between speech and emotions and have proposed a method to describe emotions. This study reported that 84% of clips are given a neutral label and only a very few clips present some other emotions (Cowie and Cornelius 2003). In general, most utterances produced by speakers are emotionally neutral. It was reported that acoustic stress, physical parameters, and speaker attitude affect emotions in speech. Some works state that acoustic stress (i.e., pitch movement) and other physiological/acoustic properties in many situations can be used as a clue to indicate emotions in speech (Johnstone and Scherer 1999). Moreover, there has been much evidence in the literature that a speaker's attitude also affects his/her emotional states in speech (Schubiger 1958; Crystal 1975, 1976; O'Connor and Arnold 1973).

2.2 Previous works on emotional speech corpus construction

As material for research on emotional speech, we need to construct an emotional speech corpus with tagging information. The constructed speech corpus can be used for emotional speech analysis and for evaluating and comparing the performance of emotional speech recognition systems. So far there have been many emotional speech corpora constructed under different characteristics such as corpus language, number of speakers (subjects), corpus purpose, number of emotion states, and type of collected speech. Ververidis and Kotropoulos (2006) have listed sixty-four emotional speech corpora and provided significant details of each corpus. We additionally enumerate thirty-seven emotional speech corpora that have been recently reported during 1995–2016 as shown in Table 1. Among these corpora, 34 of them are monolingual corpora while the remaining three corpora include two, two and four in various languages i.e., 10 English, 8 Japanese, 5 Chinese, 4 Italian, 4 French, 2 German, 1 Basque, 1 Dutch, 1 Greek, 1 Indian, 1 Indonesian, 1 Persian, 1 Polish, 1 Slovenian, and 1 Spanish. The corpora were constructed under different circumstances to express particular emotions and each different mother tongue may not express emotions in the same way or same style due to various cultural differences and the individual nature of the languages. In these works, the researchers strongly believe that emotion is related to vocabulary since some words can express emotions. Furthermore some language use a descriptive emotion word (verb) for both action and feeling expressions, such as, "love" (Kövecses 2003).

As shown in the fifth column of Table 1, the corpora were constructed with different numbers of speakers (subjects), ranging from one speaker to a few hundred speakers. The corpus with the most speakers was constructed by Cole in 2005 with 780 children participating in the project (Cole 2005; Ververidis and Kotropoulos 2006). Also, they vary according to the type of speaker, which includes children versus adults, male versus female, native versus non-native, actor versus non-actor, and general purpose versus specific purpose, but some of them are of mixed types. For example, there are four corpora which collected emotions from children (Ververidis and Kotropoulos 2006; Dadkhah et al. 2008). Besides speaker characteristics, the corpora also vary according to the number of annotators who participated in labeling each utterance with emotion states. Most of the corpora

Table 1 A list of available emotional speech corpora (EN, English; JA, Japanese; FR, French; SL, Slovenian; ES, Spanish; IT, Italian; PL, Polish; EU, Basque; ZH, Chinese; NL, Dutch; FA, Persian; EL, Greek; HI, Hindi; ID, Indonesian)

No	Year	Source (or name)	Lang	Subject	Aim	Emotion		Type
						Category	Dimension	
1	1995	Greasley et al. (1995), Reading/Leeds Emotion in Speech Project	EN	Interviews on radio/TV programs	R	AG, DG, FR, HP, SD	–	NL
2	1998	Iida et al. (1998)	JA	1 ALS patient	S	AG, JY, SD	–	SM
3	2000	Douglas-Cowie et al. (2000, 2003, 2005), The Belfast Database	EN	100 speakers (TV and studio recordings) 125 speakers (updated)	R	16 emotions	ACT, EVA	NL, EL
4	2002	Hozjan et al. (2002), Nwe et al. (2003), IESSDB	FR, EN, SL, ES	Two professional actors for each language	A, S	AG, DG, FR, JY, NT, SD, SP	–	SM
5	2003	Campbell (2003), JST/CREST ESP corpus	JA	Daily recording of volunteers' natural spoken interactions	R	–	ACT, EVA	EL
6	2004	Zovato et al. (2004), Multi-style Emotional Speech Database	IT	One female and two males (professional Italian speakers)	S	AG, HP, NT, SD	–	SM
7	2005	Cichosz and Slot (2005, 2007)	PL	8 actors (4F, 4M)	R	AG, BD, FR, HP, NT, SD	–	SM
8	2005	Burkhardt et al. (2005), EMODB	DE	10 actors (5F, 5M)	R	AG, BD, DG, FR, JY, NT, SD	–	SM
9	2005	Yamagishi et al. (2005), ATR Japanese Speech Database	JA	Professional narrators (1F, 1M), annotated by 9 males	S	JY, NT, PL, RO, SD	–	SM
10	2005	Abrilian et al. (2005), EmoTV	FR	48 subjects in TV interviews	R	AG, DP, DB, DG, EX, FR, IR, Joy, NT, PA, SD, SN, SP, WR	INT, VAL	NL

Table 1 continued

No	Year	Source (or name)	Lang	Subject	Aim	Emotion		Type
						Category	Dimension	
11	2006	Wu et al. (2006), MASC	ZH	68 speakers (23 F, 45M)	R	AG, EC, PN, NT, SD	–	EL
12	2006	Zhang et al. (2006)	ZH	8 speakers (4F, 4M)	U	AG, FR, JY, NT, SD	–	SM
13	2006	Saratxaga et al. (2006)	EU	Two professional actors (1F, 1M)	S	AG, DG, FR, HP, NT, SD, SP	–	EL
14	2006	Martin et al. (2006), The eNTERFACE'05 Audio-Visual Emotion Database	EN	42 Subjects (14 nationalities)	R	AG, DG, FR, HP, SD, SP	–	NL, EL
15	2006	Laskowski and Burger (2006), ISL Meeting Corpus	EN	18 meetings (avg. 5 participants per meeting)	A	–	VAL(+,0,-)	SM
16	2007	Wang et al. (2007), Li (2015), An Expressive speech Corpus of Standard Chinese (ESCSC)	ZH	6 professional actors (3F, 3M) + 14 non-professional actors	A	4 functional moods of Chinese: declarative, interrogative, exclamatory, and imperative	–	SM
17	2007	Douglas-Cowie et al. (2007), HUMAINE Database	EN, FR	(Note 1)	R	N/A	(Note 2)	NL, EL
18	2008	Haq et al. (2008), SAVEE	EN	4 actors	R	AG, DG, FR, HP, NT, SD, SP	–	SM, NL
19	2008	Busso et al. (2008), IEMOCAP	EN	10 actors (5F, 5M)	R	AG, DG, EC, FR, FS, HP, NT, SD, SP, OT	–	SM, NL
20	2008	Grimm et al. (2008), VAM database	DE	104 speakers (VAM-Video, VAM-Audio, VAM-Faces)	R	AG, DG, FR, HP, NT, SD, SP (for VAM-Faces only)	ACT (calm-excited), VAL (+, −), DOM (weak-strong)	NL

Table 1 continued

No	Year	Source (or name)	Lang	Subject	Aim	Emotion		Type
						Category	Dimension	
21	2008	(Trong et al. 2008)	NL	28 participants	A	AG, AM, BD, DG, EC, FR, FS, HP, MD, SP, RL, WD	ARO, VAL	SM
22	2008	Dadkhah et al. (2008)	FA	56 Iranian children (26B, 28G, 10–11 year old)	A	AG, FR, HP, SD	–	IN
23	2008	Fu et al. (2008), Beihang University Mandarin Emotion Speech Database	ZH	7 actors (4M, 3F)	R	AG, DG, JY, SD, SP	–	SM
24	2008	Kostoulas et al. (2008), A Real-World Emotional Speech Corpus for Modern Greek	EL	43 (20F, 23M), 12–56 years old	R	AG, CF, DL, HA, PS, NT	–	EL
25	2008	Arimoto et al. (2008)	JA	10 University students (5F, 5M)	R	–	ADG	EL
26	2008	Mori et al. (2008, 2011), UU Database	JA	14 paid persons	A	–	PLE, ARO, DOM, CRE, INT, POS	SM
27	2009	Moriyama et al. (2009), KEIO-ESD	JA	a male speaker (Game)	S	47 emotions	–	SM
28	2009	Fersini et al. (2009)	IT	5–10 speakers	R	AG, FR, JY, NT, SD	–	SM
29	2009	Koolagudi et al. (2009), IITKGP-SESC	HI	10 professional artists from All India radio (5F, 5M)	A	AG, CP, DG, FR, HP, NT, SC, SP	–	SM
30	2011	Arimoto et al. (2011), Spontaneous Emotional Speech Database	JA	13 university students (4F, 9M)	R	FR, SP, SD, DG, AG, EC, JY, AC, NT, OT	–	NL

Table 1 continued

No	Year	Source (or name)	Lang	Subject	Aim	Emotion		Type
						Category	Dimension	
31	2012	Fersini et al. (2012), Italian Emotional DB-Real Emotions	IT	30 trials (7 courts)	R	AG, NT, SD	–	NL
32	2012	Sneddon et al. (2012), The Belfast Natural Induced Emotion Database	EN	3 sets: 570 clips (44F, 70M), 650 clips (45F, 37M), 180 clips (30F, 30M)	R	Set 1: AM, DG, FR, FS, SP Set 2: AM, AG, DG, SP, FR, SD Set 3: AM, DG, FR	–	EL
33	2013	Ringeval et al. (2013), RECOLA	FR	46 participants (27F, 19M)	A	–	ARO, VAL (±)	SM
34	2014	Costantini et al. (2014), EMOVO	IT	6 professional actors (3F, 3M)	R	AG, DG, FR, JY, NT, SD, SP	–	SM
35	2014	Bao et al. (2014), CASIA	ZH	219 speakers from films (2 h)	R	–	–	SM
36	2014	Lubis (2014), Lubis et al. (2015), IDESC	EN, ID	English: 12 speakers (8F, 4M) Indonesia: 18 speakers (6F, 12M)	R	AG, CT, HP, NT,SD	ARO, VAL	NL

Table 1 continued

No	Year	Source (or name)	Lang	Subject	Aim	Emotion		Type
						Category	Dimension	
37	2016	Lubis et al. (2016), AV Emotion Corpora in Japanese	JA	14 native speakers	R	24 non-prototypical emotional states	ARO, VAL	SM

Note (1): Naturalistic data: Belfast (125), EmoTV (48), Castaway Reality Television Database (300 min) - Induced data: Sensitive Artificial Listener (4), Activity Data/Spaghetti Data (60), Belfast Driving Simulator Data (30), EmoTABOO, Green Persuasive Dataset (8), DRIVAWORK (24)

Note (2): Castaway Reality Television Database: POS-NEG Activity, Data/Spaghetti Data: ACT, DRIVAWORK: STR

Aim: R, Recognition; A, Analysisl; S, Synthesis; U, Unknown

Lang: EN, English; JA, Japanese; FR, French; SL, Slovenian; ES, Spanish; IT, Italian; PL, Polish; ZH, Chinese; NL, Dutch; FA, Persian; EL, Greek; HI, Hindi; ID, Indonesian; EU, Basque; DE, German

Type: EL, Elicited; SM, Simulated; NL, Natural; IN, Self-reported interview

Categories: AC, Acceptance; AM, Amusement; AG, Anger; BD, Boredom; CP, Compassion; CF, Confusion; CT, Contentment; DL, Delighted; DP, Despair; DG, Disgust; DB, Doubt; EX, Exaltation; EC, Excitement; FR, Fear; FS, Frustration; IR, Irritation; JY, Joy; HA, Hot anger; HP, Happiness; MD, Malicious delight; NT, Neutral; PA, Pain; PN, Panic; PS, Pleased; PL, Polite; RL, Relief; RO, Rough; SD, Sadness; SC, Sarcastic; SN, Serenity; SP, Surprise; WD, Wonderment; WR, Worry; and OT, Others

Dimension: ACT, Activation; ADG, Anger degree; ARO, Arousal; CRE, Credibility; DOM, Dominance; INR, Interest; INT, Intensity; PLE, Pleasantness; VAL, Valence; POS, Positive; NEG, negative; STR, Stress states

Gender: M, Males; F, Females; B, Boys; G, Girls required at least three annotators for tagging and the majority rule is usually applied i.e., this label is assigned when it obtains more than half of the votes.

In terms of computing aim or purpose (the sixth column in Table 1), an emotion corpus may be designed for either emotion recognition, emotion synthesis, or analysis (such as psychological study and/or market research) but some of them may have an unspecified purpose. From the viewpoint of emotion definition, while there are various emotional states in human beings, researchers designed emotional states (categories) in their corpus depending on their application tasks and the objectives of the corpus. Among 38 corpora listed in Table 1, thirty-three corpora were tagged with categorical emotions, twelve of which include emotions expressed by a dimensional model and seven of them have both categorical emotion and dimensional emotion as specified in the seventh and eighth columns in Table 1. The number of categorical emotions (the seventh column) varies from a single emotion (it exists or does not exist) to over 10 emotion states, depending on the definitions of emotions that the researchers are interested in for their application. The most popular set of emotion states are {'anger', 'happiness', 'neutral', and 'sadness'}, {'anger', 'fear', 'happiness', 'sadness', and 'surprise'}, and {'anger', 'fear', 'happiness', 'sadness', 'surprise', and 'disgust'}. While the four most common emotions are 'anger', 'happiness', 'neutral', and 'sadness', some popular complementary emotions are 'disgust', 'fear' and 'surprise'. However, some works did not include 'neutral' as a human's emotion state since it was treated as 'no emotion'. One major factor to develop an emotional speech corpus for emotion recognition is the need to define emotional states clearly in annotation. Matching between emotional states and emotional words is somehow subject to individual perceptions due to language usage, experience, circumstance, and the personality of the listener. Another issue in tagging emotion states in speech is that sometimes speech may include more than one categorical emotion. With regard to dimensional emotion (the eighth column), some corpora were developed to keep a number of numeric values or polarity, in place of categories, as in emotion dimensions, such as levels of Activation, Evaluation, Intensity, Valence, Positive–Negative, Stress, Arousal, Pleasantness, Dominance, Credibility, or Interest (Mori et al. 2011).

The last characteristic (the ninth column) indicates the circumstances in which emotional speech is acquired. Three typical ways to acquire emotional speech are (1) natural method, (2) elicited method, and (3) simulated or acted method. In the first type, natural speech is collected from spontaneous conversation in unrestricted circumstances when emotions are expressed naturally. Even though genuine emotion is the most important type of speech for research, it is hard to collect due to its sparsity in nature, i.e., we do not know when and where people will speak with genuine emotion. Most utterances naturally produced by speakers are emotionally neutral and speech with emotions rarely occur. Due to this difficulty, some researchers recorded videos and audios of people's reactions to customer service or some specific scenes, where people usually come to give information or complain, for the purpose of obtaining speech with emotions.

With regard to the second type, i.e. elicited speech, another way to obtain speech close to natural speech is to create a situation that will evoke emotions (Ververidis and Kotropoulos 2006). Researchers create a situation to make participants express the target emotion. After the participants were stimulated to target an emotion, then they may express that emotion naturally. To make this method succeed, researchers must design a task or a game with scenarios to involve the participants. They usually make some difficult conditions for the participants to undergo in a particular task or game, for example, a difficult spelling test (Bachorowski 1999) or interacting with a malfunctioning system in a Wizard-of-Oz scenario (Batliner et al. 2003).

In the last type, a simulated method is widely used in place of natural or elicitation methods since it is the easiest way to collect emotion speech data (Moriyama et al. 2009; Li 2015). The elicitation process requires one to ask professional actors to express or simulate emotional speech. Speech collected by this method is known as acted speech or simulated speech since speakers have been instructed to produce the target emotion by using self-induced simulation.

3 EMOLA: a Thai emotional speech corpus from Lakorn

This section presents two types of designs: a corpus design and a corpus construction design. In the former, the source of materials and emotion types for our corpus are explained. In the latter, the corpus construction process and supporting tools are described in detail. The output from these designs is a Thai emotional speech corpus, namely EMOLA, using a Thai TV drama series, so-called “Lakorn” (drama series in Thai).

3.1 Corpus design

As a convenient way to collect data for our corpus, we use spoken speech with simulated emotions where professional actors and actresses express or simulate emotional speech according to performance scripts, normally including daily emotions close to a real situation. The most common emotions observed in the Thai drama series, are anger, happiness, sadness, fear, joy, jealousy, envy and rage. In this work, the selected Thai TV drama series includes around 10–20 leading roles, acted by popular professional actors and actresses, producing several dialogues (conversation turns) in a scene where emotions are expressed.

Even though the actors and actresses sometimes overact, they are skillful in expressing their emotions. By using this series, we were able to use several forms of data such as video, audio, or even drama scripts online. However, in reality, we found that the online drama scripts do not match the speech in the actual performance, causing us to prepare the scripts or subtitles ourselves. In the corpus design, we segment the drama series into a number of short video/audio clips, each of which corresponds to a conversation turn, attached to the script is a subtitle and its timestamp (start and ending times).

As for our final outcome, each conversation turn is annotated with two levels of categorical emotions and one three-dimensional emotional scale. In assigning categorical emotions to the conversation turns, the first level is for a basic emotion, selecting one of the four emotions, i.e., ‘anger’, ‘happiness’, ‘neutral’, or ‘sadness’, while the second level includes optional emotions, choosing in order of relevance, a few from twelve emotional labels: ‘anger’, ‘doubt’, ‘excitement’, ‘fear’, ‘fun’, ‘jealousy’, ‘happiness’, ‘hate’, ‘rage’, ‘sadness’, ‘satisfaction’, and ‘surprise’. We chose one of the most popular models, the Pleasure-Arousal-Dominance (PAD) emotional model, for describing dimensional emotion (feeling) in this work. In PAD, each emotion is described by three measurements, each of which is generally scaled from − 1.0 to 1.0 and they are theoretically independent from each other. The PAD is somehow subjective and varies according to culture, language usage, circumstances, annotators, and the settings for the experiment.

3.2 Corpus construction design and tools

This section presents the design of our corpus construction process, which is composed of three sub processes: data enrichment (transcribing, time-stamping, actor ID & environment tagging), data preparation (video segmentation & XML formatter), and emotion annotation as shown in Fig. 1. In the design of our emotional speech corpus, a Thai TV drama series video (Lakorn) will be transcribed and each conversation turn in the video will be marked with its start/end timestamps, corresponding actor ID(s) and environments, and kept as a subtitle file. After that, the timestamps are used to segment the video into a set of video clips. These video clips, together with their corresponding transcriptions, actor IDs and noise types will be formatted and kept as XML-based transcription/metadata files. Finally, the transcription/metadata files together with the video clips are presented to an annotator for emotion labeling, and the result will be kept as XML-based corpus metadata. The corpus metadata and video clips constitute our emotion-tagged speech corpus.

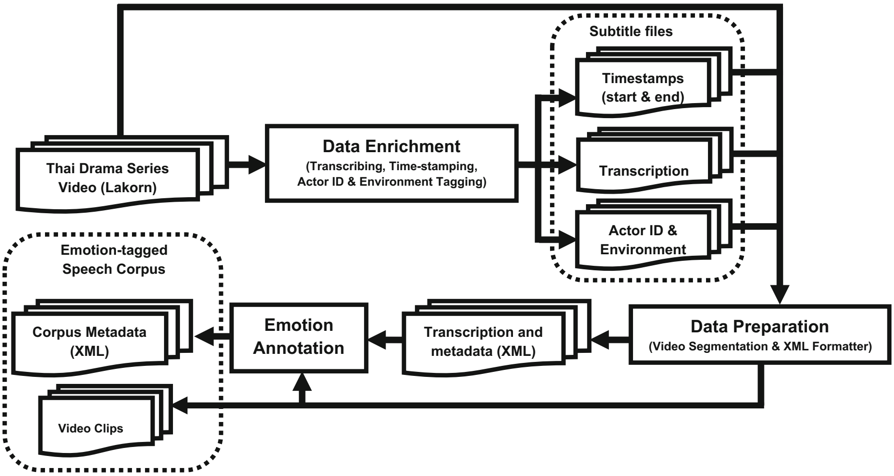

Subtitle files
Timestamps (start & end)
Transcription
Actor ID & Environment
Thai Drama Series Video (Lakorn)
Data Enrichment
(Transcribing, Time-stamping, Actor ID & Environment Tagging)
Emotion-tagged Speech Corpus
Corpus Metadata (XML)
Video Clips
Emotion Annotation
Transcription and metadata (XML)
Data Preparation
(Video Segmentation & XML Formatter)

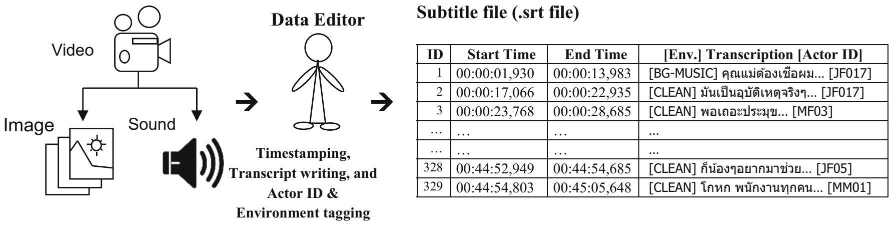
Fig. 1 Three steps in EMOLA corpus construction

3.2.1 Data enrichment: transcribing, time-stamping, actor ID & environment tagging

We collected emotional speech from TV drama series as our raw material, which consisted of video files extracted from four DVD discs. These video files are arranged into a number of episodes in order. As no subtitle or transcription is provided, each video file is manually transcribed into texts with their timestamps. By setting conversation turns (approximately one speaker's utterance) as a unit of transcription, it is possible to focus on the real emotions of a particular scene and usually we were able to assign one main emotion for each utterance. Besides the transcription and timestamps of the speech, we also tag the ID of the speaker who makes the speech, as well as the environment type of the speech. The three environment types are none (clean speech), melody (background music) and noise (environmental sound).

Figure 2 displays the process of data enrichment where a data editor analyzes a video (audio and visual components) and then provides start/end timestamps, transcription (subtitle), actor ID and environment tagging in the format of a subtitle file (.srt). In practice, the original video files in the VOB format (.vob) are translated to the AVI-formatted files (.avi) in order to be compatible with a subtitling-support tool, namely *Subtitle Edit*. As free software, the Subtitle Edit tool enables a data editor to associate transcriptions, timestamps, actor ID and environment tag into a video file by outputting a subtitle file as a supplement file. As shown in Fig. 3, we can play a part of the video in the top-right corner of the program window, and then the program will output the speech (or conversation) waveform corresponding to the part of the video playing in the bottom-right corner of the window. In the figure, we blurred the faces of speakers (actors) due to copyright reasons, but during the process of emotion labelling, the annotator watched the original video without any blurring.

At this point, the data editor can determine the period of the waveform that he would like to provide a transcription for and replay the waveform several times to confirm a suitable waveform period for that transcription. In the format of `[Environment] (Transcription [Actor ID]) +`, the input of the data editor is inserted in a text box, located in the middle-left part of the window. Here, `[Environment]` is used to specify one of three environment types: `[CLEAN]`, `[BG-MUSIC]` and `[EX-MUSIC]`, standing for clean speech, speech with background


Video
Image
Sound
Data Editor
Timestamping, Transcript writing, and Actor ID & Environment tagging
Subtitle file (.srt file)

ID	Start Time	End Time	[Env.] Transcription [Actor ID]
1	00:00:01,930	00:00:13,983	[BG-MUSIC] คุณแม่ต้องเชื่อผม... [JF017]
2	00:00:17,066	00:00:22,935	[CLEAN] มันเป็นอุบัติเหตุจริงๆ... [JF017]
3	00:00:23,768	00:00:28,685	[CLEAN] พอเถอะประมุข... [MF03]
…	…	…	...
…	…	…	...
328	00:44:52,949	00:44:54,685	[CLEAN] ก็น้องๆอยากมาช่วย... [JF05]
329	00:44:54,803	00:45:05,648	[CLEAN] โกหก พนักงานทุกคน... [MM01]

Fig. 2 Data Enrichment: Timestamping, Transcript writing, and Actor ID & Environment tagging

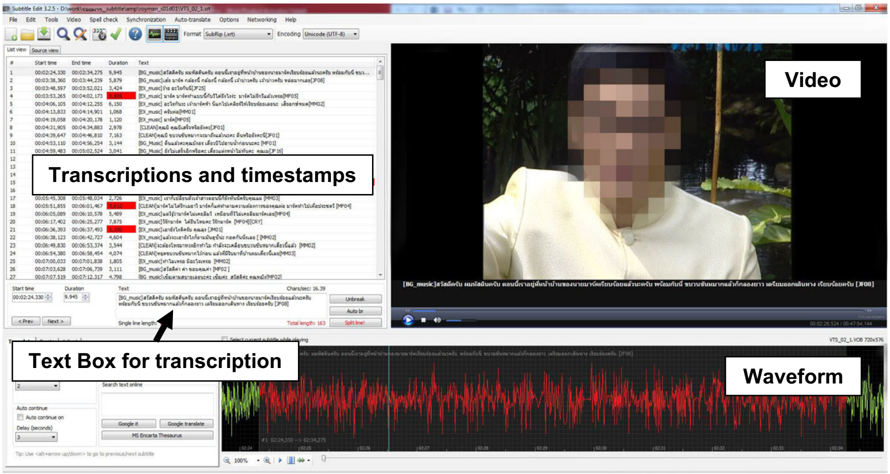


Fig. 3 A sample screen using Subtitle Edit

music and speech with exciting music/sound, respectively. While `Transcription` is the transcription usually presented in the Thai language, `[Actor ID]` is the bracketed identifier of the actor who utters that speech, corresponding to the `Transcription`. In the situations that multiple actors speak simultaneously, multiple pairs of `Transcription` and `[Actor ID]` can be inserted. The list of transcriptions and timestamps tagged for each time interval (waveform period) are shown in the top-left corner of the window.

3.2.2 Data preparation: video segmentation and XML formatter

To prepare the video data for emotion annotation, video segmentation and XML formatting are performed as described below.

(1) Video Segmentation

After obtaining the subtitle files (`.srt`), the video file (`.avi`) is segmented into a number of video clips (`.avi`) according to the time-stamping in the subtitle files using the video and audio converter (tool) named `ffmpeg`. Since the `ffmpeg` can be used as a script, it is very convenient to divide into several video clips by executing a series of `ffmpeg` commands but with different parameters. With `ffmpeg` commands, the original video will be divided into a set of video clips, indexed by three digits at the end of the output filename. These indices are used for mapping the clips with their transcriptions.

(2) XML Formatter

To keep transcription and metadata in a standard format, the XML is applied and the resulting file has the XML extension (`.xml`). We have designed the original XML DTD to be simple and to match our purpose. The top-left part of Fig. 4 shows the Document Type Definition (DTD) (`.dtd`) designed for describing transcriptions (subtitles) and metadata in the form of a filename of the video source as well as a set of video-clip filenames with their corresponding subtitles, start timestamps, duration intervals, actor ID and speaking environment. On the other hand, the right hand side of Fig. 4 illustrates an example of the data description based on the XML DTD on the left. For clarity, the translation of each transcription in the XML-tagged data is given in the bottom-left hand side of Fig. 4.

3.2.3 Emotion annotation

The last process (the third process in Fig. 1), emotion annotation, is performed to manually tag emotion for each video clip (utterance). As part of our corpus design, we decided to keep both the emotional category and the three-dimensional emotion

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE dataset [
<!ELEMENT dataset (sourcefile, data+)>
<!ATTLIST dataset name CDATA #REQUIRED>
<!ELEMENT sourcefile (#PCDATA)>
<!ELEMENT data (id, videoclip, subtitle, starttime,
                interval, actorid, environment)>
<!ELEMENT videoclip (#PCDATA)>
<!ELEMENT subtitle (#PCDATA)>
<!ELEMENT starttime (#PCDATA)>
<!ELEMENT interval (#PCDATA)>
<!ELEMENT actorid (#PCDATA)>
<!ELEMENT environment (#PCDATA)>
]>
```

1. คุณแม่ต้องเชื่อผมนะครับ ผมไม่ได้ฆ่าไพเค้า ผมรักไพจริงๆ
   ผมไม่มีวันที่จะฆ่าเค้าได้หรอกครับ
   (Mom, believe me. I did not kill Phai. I really loved Phai. I would never be able to kill her.)

2. มันเป็นอุบัติเหตุจริงๆนะครับคุณแม่ คุณแม่ได้โปรดเห็นใจผมเถอะนะครับ
   (It was really an accident, mom. Mom, please sympathize with me.)

3. พอเถอะประมุข แม่ขอแล้วหยุดทุกอย่างไว้ตรงนี้แค่นี้เถอะ
   (Stop explaining, Pramook. I (mom) beg you to stop everything at this point.)

…
…

328. ก็น้องๆอยากมาช่วยงานคุณนี่คะ
   (As your brothers/sisters would like to help your work.)

329. โกหก พนักงานทุกคนเนียรู้จักนิสัยผมดี ถ้าผมไม่ได้สั่งหรือขอให้ช่วย
   พวกเค้าจะไม่ยุ่งเด็ดขาด
   (You lie ! All the staff here know my character well. If I do not order or beg for help, they will never get involved.)

```
<dataset name="rm_s0103d02.xml">
<sourcefile> rm_s0103d02.avi </sourcefile>

<data>
<id> 1 </id>
<videoclip> rm_s0103d02_001.avi </videoclip>
<subtitle> คุณแม่ต้องเชื่อผมนะครับ ผมไม่ได้ฆ่าไพเค้า ผมรักไพจริงๆ
ผมไม่มีวันที่จะฆ่าเค้าได้หรอกครับ </subtitle>
<starttime> 00:00:01,930 </starttime>
<interval> 12.05 </interval>
<actorid> JF017 </actorid>
<environment> BG-MUSIC </environment>
</data>

<data>
<id> 2 </id>
<videoclip> rm_s0103d02_002.avi </videoclip>
<subtitle> มันเป็นอุบัติเหตุจริงๆนะครับคุณแม่
คุณแม่ได้โปรดเห็นใจผมเถอะนะครับ </subtitle>
<starttime> 00:00:17,066 </starttime>
<interval> 5.87 </interval>
<actorid> JF017 </actorid>
<environment> CLEAN </environment>
</data>
…
…
<data>
<id> 328 </id>
<videoclip> rm_s0103d02_328.avi </videoclip>
<subtitle> ก็น้องๆอยากมาช่วยงานคุณนี่คะ </subtitle>
<starttime> 00:44:52,949 </starttime>
<interval> 1.74 </interval>
<actorid> JF05 </actorid>
<environment> CLEAN </environment>
</data>

<data>
<id> 329 </id>
<videoclip> rm_s0103d02_329.avi </videoclip>
<subtitle> โกหก พนักงานทุกคนเนียรู้จักนิสัยผมดี
ถ้าผมไม่ได้สั่งหรือขอให้ช่วย พวกเค้าจะไม่ยุ่งเด็ดขาด </subtitle>
<starttime> 00:45:05,648 </starttime>
<interval> 10.85 </interval>
<actorid> MM01 </actorid>
<environment> CLEAN </environment>
</data>

</dataset>
```

Fig. 4 Document Type Definition (DTD) for transcription (subtitle) and metadata (top left), an example of XML-tagged data (right), and the translation of each transcription in the XML-tagged data (bottom left)

of each utterance, for two purposes: emotion recognition and analysis of the relations between category and dimension. For consistency, we had held a training course and several meetings for discussion among the six emotion annotators, who were trained to annotate all utterances in the whole corpus in the same manner. The annotators were requested to use the self-developed software namely "EmoAnnotator" to label emotions for each transcription. By using EmoAnnotator, an annotator can watch the video clip with both audio and visual functions, and its corresponding transcription for emotion labeling. The guidelines (instructions) for the annotators are as follow.

I. Provision of Annotator Information: The annotator fills in the details of the annotator with annotator ID, name, and age.

II. Error Checking: The annotator has to check for any errors before giving an emotion label to a transcription. If any unacceptable errors are found in the data, no label is assigned to the transcription. The following three types of unacceptable errors are considered.

    (1) Incorrect Segmentation: Ideally one video clip includes one (or more) complete utterance(s) from only one speaker with only one emotion under only one environment (clean, background music, noise). However, sometimes two types of incorrect segmentation are found.

        a. Over-segmentation: a video clip that includes multiple utterances from multiple speakers with more than one emotion.
        b. Under-segmentation: a video clip that includes incomplete (unfinished) utterances.

    (2) Multiple Speakers or Two speakers: Sometimes it is impossible to make a video clip which includes utterance(s) from only one speaker since two or more speakers are speaking at the same time.

    (3) Wrong Transcription: For some reason, the transcription (subtitle) does not match with the video clip it corresponds to. Sometimes the transcription is completely different from the video clips and therefore we judge it to be an error. However, in some cases, only some words are missing or are added in the transcription and then it is not judged to be a wrong subtitle but just an incomplete subtitle. Such incomplete subtitles are not considered as errors since the transcription is partially useful.

    If the annotators mark one of the above errors, they have no need to perform any of the following steps.

III. Annotation of Primary Emotion: According to our design, the four possible primary emotions are anger, happiness, sadness, and neutral. These primary emotions are easy to distinguish. Therefore, all annotators were trained to give the most suitable primary emotion to a video clip.

IV. Annotation of Secondary Emotion: In some situations, if the video clip does not match perfectly with any of the primary emotions, the annotators are instructed to select any (one or more) of twelve secondary emotions, which are anger, confusion, disgust, excitement, fear, happiness, jealousy, pleasure, rage, sadness, satisfaction, and surprise. Moreover, in the case that none of these twelve emotions matches with the emotion in the video clip, the annotators are allowed to manually add a new emotion that is appropriate into the text box provided.

V. Determination of Three-Dimensional Emotion: Besides the primary and secondary emotional categories, annotators are told to provide the values of Pleasure-Arousal-Dominance (PAD) space, which are the three-dimensional emotions illustrated in Sect. 2.1.1. Any of the three PAD dimensions is rated in the scale of − 1.0 to 1.0, with an interval scale of 0.25. Thus, the nine possible values for each dimension are − 1.00, − 0.75, − 0.50, − 0.25, 0.00, + 0.25, + 0.50, + 0.75, and + 1.00. Here, positive feeling (+ value) refers to pleasure, non-arousal, and submissive. On the other hand, negative feeling (− value) means displeasure, arousal, and dominance.

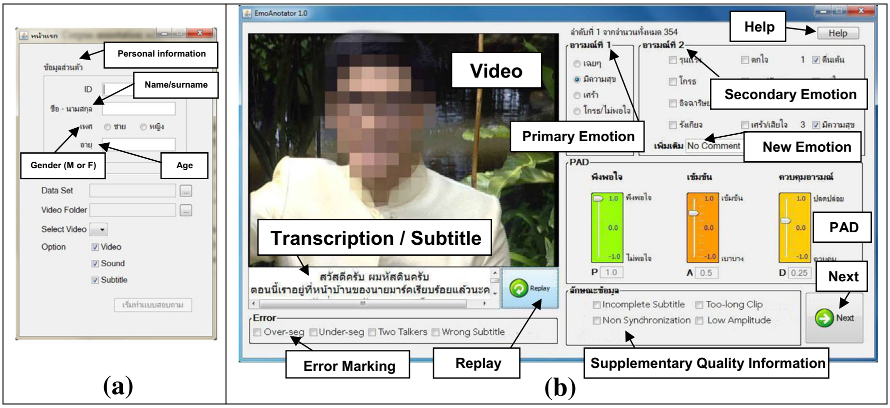
VI. Provision of Supplementary Quality Information: Moreover, the annotators are asked to provide more information particularly related to the audio/visual quality of the video clip, including incomplete subtitles, non-synchronization, too-long clips, and low amplitude. Even if these quality problems are not very serious, it is useful to keep such information.

Similar to the format for the transcription and metadata, we also designed the Document Type Definition (DTD) (`.dtd`), as shown in Fig. 5, for describing the personal information of the annotator (including his/her identification, name, gender, and age), the input filename of the transcription and metadata, the annotation settings (that is the usage of video, sound, and transcription during annotation), the annotation result (primary emotion, secondary emotion and three-dimensional emotion), and the supplementary quality information (incomplete subtitle, non-synchronization, too-long a clip, and low amplitude).

As an example, the EmoAnnotator tool starts from requesting an annotator to input his/her information as shown in Fig. 6a and then the annotation window is displayed as shown in Fig. 6b. At the window for inputting annotator information, the annotator has to input his/her identification, name/surname, gender, and age, as well as the name of the data set and video-clip folder/filename he/she would like to annotate. He/she can select to show or hide the video, audio (sound) and subtitle during annotation for further exploration of the effects of annotation environment on tagging quality. However, in our corpus construction, the annotators are advised to turn on all of these three options. In the window for emotion annotation, there are eleven components which are video display, transcription (subtitle) display, replay button, error-marking checkboxes, help button, primary-emotion radio button, secondary-emotion ranked checkboxes, new-emotion text boxes, PAD sliders, checkboxes for supplementary quality information and next-video-clip button.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE emotionannotation [
   <!ELEMENT emotionannotation  (annotatorprofile, setting, annotateddata+)>
   <!ELEMENT annotatorprofile   (id, name, gender, age)>
   <!ELEMENT settings           (datasetname, videoflag, soundflag, transflag)>
   <!ELEMENT annotateddata      (videoclip, subtitle, error?, label1?, label2-1?, label2-2?, label2-3?,
                                 label2-4?, label2-5?, label2-6?, label2-7?, label2-8?, label2-9?,
                                 label2-10?, label2-11?, label2-12?, label3?, pleasure?, arousal?,
                                 dominance?, incsubtitle?, nonsync?, longclip?, lowamp?)>
   <!ELEMENT error              (overseg, underseg, multispeaker, wrongtrans)>
   <!ELEMENT id          (#PCDATA)>                <!ELEMENT name          (#PCDATA)>
   <!ELEMENT gender      (#PCDATA)>                <!ELEMENT age           (#PCDATA)>
   <!ELEMENT datasetname (#PCDATA)>                <!ELEMENT videoflag     (#PCDATA)>
   <!ELEMENT soundflag   (#PCDATA)>                <!ELEMENT transflag     (#PCDATA)>
   <!ELEMENT videoclip   (#PCDATA)>                <!ELEMENT subtitle      (#PCDATA)>
   <!ELEMENT overseg     (#PCDATA)>                <!ELEMENT underseg      (#PCDATA)>
   <!ELEMENT multispeaker (#PCDATA)>               <!ELEMENT wrongtrans    (#PCDATA)>
   <!ELEMENT label1      (#PCDATA)>                <!ELEMENT label2-1      (#PCDATA)>
   <!ELEMENT label2-2    (#PCDATA)>                <!ELEMENT label2-3      (#PCDATA)>
   <!ELEMENT label2-4    (#PCDATA)>                <!ELEMENT label2-5      (#PCDATA)>
   <!ELEMENT label2-6    (#PCDATA)>                <!ELEMENT label2-7      (#PCDATA)>
   <!ELEMENT label2-8    (#PCDATA)>                <!ELEMENT label2-9      (#PCDATA)>
   <!ELEMENT label2-10   (#PCDATA)>                <!ELEMENT label2-11     (#PCDATA)>
   <!ELEMENT label2-12   (#PCDATA)>                <!ELEMENT label3        (#PCDATA)>
   <!ELEMENT pleasure    (#PCDATA)>                <!ELEMENT arousal       (#PCDATA)>
   <!ELEMENT dominance   (#PCDATA)>                <!ELEMENT incsubtitle   (#PCDATA)>
   <!ELEMENT nonsync     (#PCDATA)>                <!ELEMENT longclip      (#PCDATA)>
   <!ELEMENT lowamp      (#PCDATA)>
]>
```

Fig. 5 Document Type Definition (DTD) for describing personal information of the annotator (identification, name, gender, and age), the input filename of transcription and metadata (dataset name), the annotation settings (the usage of video, sound, and transcription during annotation), the primary emotion (Label 1), secondary emotion (Label 2), optional emotion (Label 3), three-dimensional emotion (Pleasure, Arousal, and Dominance), and supplementary quality information (incomplete subtitle, non-synchronization, too-long a clip and low amplitude)


หน้าแรก
Personal information
ข้อมูลส่วนตัว
ID
Name/surname
ชื่อ - นามสกุล
เพศ ○ ชาย ○ หญิง
อายุ
Gender (M or F)
Age
Data Set
Video Folder
Select Video
Option ☑ Video ☑ Sound ☑ Subtitle
เริ่มทำแบบสอบถาม
(a)
EmoAnotator 1.0
ลำดับที่ 1 จากจำนวนทั้งหมด 354
Help
Video
อารมณ์ที่ 1
○ เฉยๆ
● มีความสุข
○ เศร้า
○ โกรธ/ไม่พอใจ
Primary Emotion
อารมณ์ที่ 2
□ รุนแรง □ ตกใจ 1 ☑ ตื่นเต้น
□ โกรธ
□ อิจฉาริษยา
□ รังเกียจ □ เศร้า/เสียใจ 3 ☑ มีความสุข
Secondary Emotion
เพิ่มเติม No Comment
New Emotion
PAD
พึงพอใจ 1.0 0.0 -1.0 พึงพอใจ ไม่พอใจ P 1.0
เข้มข้น 1.0 0.0 -1.0 เข้มข้น เบาบาง A 0.5
ควบคุมอารมณ์ 1.0 0.0 -1.0 ปลดปล่อย ควบคุม D 0.25
Transcription / Subtitle
Replay
Error
□ Over-seg □ Under-seg □ Two Talkers □ Wrong Subtitle
Error Marking
ลักษณะข้อมูล
□ Incomplete Subtitle □ Too-long Clip
□ Non Synchronization □ Low Amplitude
Supplementary Quality Information
Next
(b)

สวัสดีครับ ผมหัสดินครับ
ตอนนี้เราอยู่ที่หน้าบ้านของนายมาร์คเรียบร้อยแล้วนะค

Fig. 6 **a** Annotator information input window, and **b** emotion annotation window

By watching a video clip and reading the transcription or replaying the video clip with the replay button on the left side of the window, an annotator can mark one or more error types that may occur in the bottom left of the window. If any error is marked for the video clip, the annotator cannot perform any further annotation. He can use the ‘Help’ button on the top right hand corner to access the annotation instructions or guidelines. Otherwise, he/she can tag one primary emotion, as well as optionally a few secondary ranked emotions or a new emotion, for the video clip

Table 2 Major statistics of the material for the EMOLA corpus

1.	Total length of video (the whole series) (A)	1397 min (23 h 17 min)
2.	Length of video that have subtitles (B)	868 min (14 h 28 min)
3.	Ratio of subtitled video to the whole video	62.13 percent
4.	Number of male actors	20 actors
5.	Number of female actors	31 actresses
6.	Number of transcriptions (conversation turns)	8987 transcriptions
7.	Number of single speaker's transcriptions (scenes)	8779 transcriptions
8.	Number of multiple speakers' transcriptions (scenes)	208 transcriptions (453 speaker-scenes)
9.	Number of speaker scenes	9232 speaker-scenes
10.	Number of speaker scenes by male actors	3963 speaker-scenes
11.	Number of speaker scenes by female actors	5269 speaker-scenes
12.	Average transcription length	5.80 s
13.	Minimum transcription length	1.18 s
14.	Maximum transcription length	72.40 s
15.	Standard derivation of transcription length	3.80 s
16.	Environments (numbers of transcriptions)	Clean (3434) Background music (3715) and noise (1838)

Table 3 Two most assigned labels

Annotator ID.	Assigned Label #1	Assigned Label #2
AnnF01 (39)	Happiness (37.79%)	Anger (36.21%)
AnnF02 (37)	Anger (33.25%)	Happiness (22.80%)
AnnF03 (29)	Error (36.16%)	Anger (24.47%)
AnnF04 (28)	Neutral (49.13%)	Error (34.45%)
AnnM01 (26)	Error (27.83%)	Sadness (25.71%)
AnnM02 (23)	Neutral (37.16%)	Anger (21.76%)

on the top right part of the window and provide three values for P, A and D dimensions in the middle right of the windows. The sliders will be green if an annotator feels positive (+ 1.00) and red if he has a negative feeling (− 1.00). If any additional or supplementary quality on the speech is available, the annotator can check one or more options in the bottom right of the windows. The 'Next' button should be pressed to watch the next video clip.

During labeling, annotators can take a break by terminating the program and later resuming their work by selecting the data set, the video folder and the video file as shown in Fig. 6a. The annotation results are stored in the form of XML as shown in Fig. 5. The annotators work independently on the process; they are not allowed to take any advice from the others. Because of this, all labels are subjective and based on an individual's judgment.

4 Statistics of annotated data and evaluation

A Thai emotional speech corpus can be constructed using the process described in the previous section. To characterize the constructed corpus, we have provided three levels of statistics: collection-level, annotator-oriented, and actor-emotion-oriented statistics.

4.1 Collection-level statistics related to transcription/metadata

A popular Thai TV drama series was selected for construction of an emotion annotated corpus, namely EMOLA. Table 2 illustrates a number of the major statistics of the material for corpus construction, related to the transcription and metadata. This corpus material included a video with a length of approximately 23 h and 17 min of which only 14 h and 28 min of the video are subtitled since the others include silence, music, background noise, or non-speech portions. Therefore, sixty-two percent of the video was transcribed. Of a total of 8987 transcriptions, 208 of them (2.31%) are transcriptions that have multiple actors speaking simultaneously while most transcriptions (97.69% or 8779 transcriptions) contain a single actor's speech. The 208 transcriptions with multiple actors speaking occupy 453 speaker scenes. In this study, a transcription with *n* speaking actors will be counted as *n* speaker scenes. The corpus has a total of 9232 speaker scenes (5269 speaker scenes by 20 actors and 3963 speaker scenes by 31 actresses). The average, minimum and

Table 4 The number of emotional labels and their associated labels in the second level

ID.	Cat.1/Cat.2		AG	CF	DG	EC	FR	HP	JL	JY	RG	SD	ST	SP	Total
AnnF01	AG	w.	1007.10	737.15	246.25	152.78	355.97	7.03	79.62	10.48	130.87	148.33	81.42	297.00	3254
		w/o w.	2402.00	1994.00	685.00	501.00	1028.00	17.00	230.00	29.00	386.00	435.00	188.00	915.00	8810
	HP	w.	18.92	581.90	13.73	232.93	149.82	585.73	14.43	226.63	0.58	39.42	1419.60	112.30	3396
		w/o w.	43.00	1374.00	37.00	676.00	356.00	1466.00	36.00	675.00	2.00	92.00	2809.00	314.00	7880
	SD	w.	51.85	304.95	10.03	43.65	338.13	4.08	4.32	1.00	4.62	665.53	154.48	56.35	1639
		w/o w.	130.00	784.00	30.00	133.00	848.00	11.00	14.00	3.00	12.00	1375.00	321.00	176.00	3837
	NT	w/o w.													463
	ER	w/o w.													235
AnnF02	AG	w.	1684.67	380.08	109.42	65.50	88.50	2.33	181.83	15.33	79.33	206.67	9.83	164.50	2988
		w/o w.	2033.00	517.00	200.00	132.00	144.00	5.00	298.00	23.00	174.00	337.00	13.00	295.00	4223
	HP	w.	21.17	77.17	4.17	104.58	0.83	1101.25	15.83	183.42	0.33	4.67	525.75	9.83	2049
		w/o w.	25.00	101.00	7.00	214.00	2.00	1414.00	25.00	353.00	1.00	9.00	777.00	20.00	2965
	SD	w.	29.33	69.00	4.50	14.33	15.83	4.83	2.67	0.33	0.33	1398.33	13.17	36.33	1589
		w/o w.	43.00	120.00	11.00	31.00	33.00	9.00	4.00	1.00	1.00	1483.00	23.00	72.00	1848
	NT	w/o w.													520
	ER	w/o w.													1841
AnnF03	AG	w.	1523.02	17.07	22.00	8.78	40.73	3.17	9.83	4.78	403.53	57.07	5.50	103.52	2199
		w/o w.	2088.00	37.00	57.00	22.00	102.00	6.00	19.00	12.00	855.00	134.00	13.00	199.00	3544
	HP	w.	2.00	7.75	0.00	90.40	1.50	432.07	0.00	63.73	1.98	4.83	622.82	15.92	1243
		w/o w.	5.00	14.00	0.00	227.00	4.00	875.00	0.00	172.00	7.00	10.00	1072.00	30.00	2416
	SD	w.	24.83	19.82	1.50	6.08	78.32	0.50	1.00	0.00	88.73	1095.98	20.00	70.23	1407
		w/o w.	58.00	48.00	5.00	14.00	196.00	1.00	2.00	0.00	204.00	1364.00	37.00	163.00	2092
	NT	w/o w.													888
	ER	w/o w.													3250

Table 4 continued

ID.	Cat.1/Cat.2		AG	CF	DG	EC	FR	HP	JL	JY	RG	SD	ST	SP	Total
AnnF04	AG	w.	376.78	22.20	102.73	0.00	11.20	5.33	41.87	1.00	230.90	40.45	7.17	22.37	862
		w/o w.	584.00	42.00	187.00	0.00	25.00	8.00	86.00	2.00	433.00	82.00	12.00	41.00	1507
	HP	w.	2.00	0.25	2.75	22.00	0.00	154.83	2.17	44.83	0.00	0.67	133.50	0.00	363
		w/o w.	2.00	1.00	6.00	64.00	0.00	333.00	4.00	123.00	0.00	2.00	298.00	0.00	835
	SD	w.	8.42	5.92	0.75	0.00	6.58	3.33	0.00	0.50	24.17	186.50	5.33	9.50	251
		w/o w.	22.00	17.00	2.00	0.00	17.00	9.00	0.00	1.00	54.00	241.00	12.00	22.00	397
	NT	w/o w.													4415
	ER	w/o w.													3096
AnnM01	AG	w.	556.32	404.85	172.94	27.60	320.84	4.25	54.87	29.08	46.40	337.30	33.83	118.72	2107
		w/o w.	1133.00	889.00	438.00	78.00	856.00	13.00	128.00	65.00	123.00	840.00	81.00	314.00	4958
	HP	w.	19.23	335.80	15.70	101.35	95.47	331.13	18.37	176.25	0.67	83.12	767.63	42.28	1987
		w/o w.	48.00	631.00	40.00	303.00	254.00	852.00	48.00	476.00	2.00	178.00	1501.00	108.00	4441
	SD	w.	103.07	592.65	41.00	7.97	466.83	4.25	13.00	13.42	6.45	940.53	50.92	70.92	2311
		w/o w.	251.00	1302.00	114.00	22.00	1198.00	12.00	36.00	31.00	18.00	1867.00	106.00	191.00	5148
	NT	w/o w.													81
	ER	w/o w.													2501

Table 4 continued

ID.	Cat.1/Cat.2		AG	CF	DG	EC	FR	HP	JL	JY	RG	SD	ST	SP	Total
AnnM02	AG	w.	997.87	83.70	65.25	3.83	8.45	1.00	15.67	1.67	585.70	89.83	1.17	101.87	1956
		w/o w.	1458.00	138.00	106.00	6.00	17.00	1.00	29.00	4.00	1003.00	150.00	3.00	152.00	3067
	HP	w.	0.20	8.33	0.00	30.42	0.50	241.12	2.00	112.08	18.20	10.70	761.78	33.67	1219
		w/o w.	1.00	12.00	0.00	69.00	1.00	506.00	2.00	256.00	30.00	18.00	1064.00	49.00	2008
	SD	w.	19.58	38.50	3.33	0.33	18.25	0.33	2.25	0.00	32.33	791.00	4.83	68.25	979
		w/o w.	43.00	84.00	7.00	1.00	38.00	1.00	4.00	0.00	73.00	906.00	8.00	124.00	1289
	NT	w/o w.													3340
	ER	w/o w.													1493

Note: Cat. 1 is the set of primary emotion labels (AG, Anger; HP, Happiness; SD, Sadness; NT, Neutral; ER, Error); Cat. 2 is the set of secondary emotion labels (AG, Anger; CF, Confusion; DG, Disgust; EC, Excitement; FR, Fear; HP, Happiness; JL, Jealousy; JY, Joyful; RG, Rage; SD, Sadness; ST, Satisfaction; SP, Surprise). Moreover, ‘w.’ refers to “with weighing”, and ‘w/o w.’ refers to “without weighing”

Table 5 Number of transcriptions by label agreement patterns

Emotion label	Total majority	Majority							No majority			Total
		6-0-0-0-0 (1.00)	5-1-0-0-0 (1.00)	4-2-0-0-0 (1.00)	4-1-1-0-0 (1.00)	3-2-1-0-0 (1.00)	3-1-1-1-0 (1.00)	2-1-1-1-1 (1.00)	2-2-2-0-0 (0.33)	2-2-1-1-0 (0.50)	3-3-0-0-0 (0.50)	
Anger	2240	399	618	308	393	414	96	12	44.67	82.00	105.00	2471.67
Happiness	1524	142	403	386	184	359	47	3	41.33	68.50	121.00	1754.83
Sadness	754	5	33	146	133	358	66	13	70.00	112.00	132.50	1068.50
Neutral	1317	93	299	298	189	371	62	5	39.33	76.50	54.50	1487.33
Unlabelled	1957	113	742	346	341	341	66	8	38.67	87.00	122.00	2204.67
Total	7792	752	2095	1484	1240	1843	337	41	234.00	426.00	535.00	8987.00

maximum transcription lengths are 5.80, 1.18, and 72.40 s, respectively, with a standard deviation of 3.80 s. Some transcriptions are from clear speech, some from scenes with background music and the rest are speech in a noise environment.

In general, the interpretation of emotion in speech is subjective in nature. In this work, we recruited six annotators (four females and two males) to annotate emotion in all the transcriptions in this corpus. Later, the results of the annotated data were analyzed and evaluated in order to compare individual perceptions as can be seen in the next section.

4.2 Annotator-oriented statistics of emotion annotation

After finishing the preparation of the transcriptions and metadata, the emotion annotation was performed. It should be noted that all the video clips were individually labeled by six annotators (four females and two males) aged between 23 to 39. Note that 'F' refers to female annotators and 'M' refers to male annotators in the annotator ID. For example, AnnF01 is a female annotator and AnnM01 is a male annotator. This section describes the statistics of the emotion annotations. All the annotators were asked to select one label per transcription (utterance), however, some of them may not have been able to label them due to some errors which occurred in the data. Table 3 shows the details of the number of emotional labels in

Pairwise Comparison		Female actors					Male actors					Average
		ActF01	ActF02	ActF03	ActF04	ActF05	ActM01	ActM02	ActM03	ActM04	ActM06	
AnnF01	AnnF02	0.4865	0.3879	0.5683	0.3154	0.4904	0.4456	0.4393	0.3494	0.4443	0.3105	0.4238
	AnnF03	0.3796	0.3422	0.4852	0.2910	0.3952	0.3428	0.3301	0.2530	0.4230	0.2128	0.3455
	AnnF04	0.1536	0.1407	0.1783	0.1120	0.1739	0.0771	0.0699	0.0525	0.1152	0.0313	0.1105
	AnnM01	0.4050	0.3820	0.3922	0.3241	0.4543	0.3566	0.2897	0.1747	0.3644	0.2033	0.3346
	AnnM02	0.4097	0.2488	0.3758	0.3129	0.3449	0.2775	0.2543	0.1747	0.2819	0.2400	0.2921
AnnF02	AnnF03	0.4698	0.5105	0.5703	0.5074	0.5026	0.5048	0.4617	0.4312	0.5767	0.4056	0.4941
	AnnF04	0.2167	0.2581	0.2229	0.2236	0.2223	0.1755	0.1650	0.1457	0.1577	0.1457	0.1933
	AnnM01	0.4822	0.5026	0.4752	0.4304	0.5332	0.4697	0.4332	0.4027	0.4826	0.4731	0.4685
	AnnM02	0.4501	0.3475	0.4331	0.3741	0.2996	0.3554	0.3304	0.2496	0.2924	0.3583	0.3491
AnnF03	AnnF04	0.2657	0.3281	0.3150	0.2499	0.3400	0.2568	0.2726	0.2466	0.2104	0.2804	0.2766
	AnnM01	0.4832	0.5268	0.4367	0.4023	0.5108	0.4640	0.3919	0.3518	0.4565	0.4421	0.4466
	AnnM02	0.4060	0.3764	0.4497	0.3891	0.3136	0.3577	0.3899	0.3365	0.3032	0.3933	0.3715
AnnF04	AnnM01	0.2204	0.2572	0.1885	0.1740	0.2141	0.1500	0.1421	0.1423	0.1612	0.1694	0.1819
	AnnM02	0.2997	0.3802	0.3491	0.3399	0.3491	0.2714	0.2753	0.3202	0.3086	0.2208	0.3114
AnnM01	AnnM02	0.3778	0.3377	0.3353	0.3388	0.2826	0.2812	0.2218	0.2074	0.2212	0.3317	0.2936

(a)

	AnnF01	AnnF02	AnnF03	AnnF04	AnnM01	AnnM02
AnnF01	1.0000	0.4238	0.3455	0.1105	0.3346	0.2921
AnnF02	0.4238	1.0000	0.4941	0.1933	0.4685	0.3491
AnnF03	0.3455	0.4941	1.0000	0.2766	0.4466	0.3715
AnnF04	0.1105	0.1933	0.2766	1.0000	0.1819	0.3114
AnnM01	0.3346	0.4685	0.4466	0.1819	1.0000	0.2936
AnnM02	0.2921	0.3491	0.3715	0.3114	0.2936	1.0000

During merging, similarity (agreement) between two clusters can be calculated by two alternative approaches: single linkage (s_s) and complete linkage (s_c).

[Single Linkage]

$$s_s(C_i, C_j) = max_{x \in C_i, y \in C_j} s(x, y)$$

[Complete Linkage]

$$s_c(C_i, C_j) = min_{x \in C_i, y \in C_j} s(x, y)$$

(b)

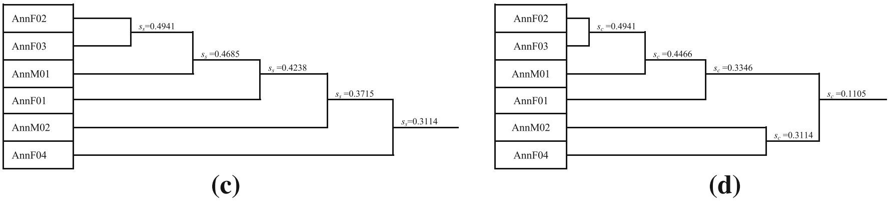


Fig. 7 Annotator clustering based on emotion label agreement (Cohen's kappa): single linkage and complete linkage. **a** Emotion label agreement (Cohen's kappa) with respect to the total number of speaker scenes, by actor. **b** Similarities between annotators based on average emotion label agreement. **c** Clustering by single linkage. **d** Clustering by complete linkage

Table 6 Number of transcriptions by label agreement patterns for three annotators (AnnF02, AnnF03 and AnnM01)

Emotion label	Total majority exists	Majority exists		No majority exists	Total
		3-0-0-0-0 (1.00)	2-1-0-0-0 (1.00)	1-1-1-0-0 (0.33)	
Anger	2332	1111	1221	178.33	2510.33
Happiness	1616	842	774	145.67	1761.67
Sadness	1581	779	802	169	1750
Neutral	284	14	270	123	407
Unlabelled	2398	1293	1105	160	2558
Total	8211	4039	4172	776	8987

two levels of categories assigned by each annotator (Level 1: anger, happiness, sadness, and neutral, as well as Level 2: anger, happiness, sadness, confusion, disgust, excitement, fear, jealousy, joy, rage, satisfaction, and surprise). We observed some variations on the annotated labels among the annotators as follows.

Discussion on the level-1 labels

(1) AnnF01 and AnnF02 assigned 'Happiness' and 'Anger' labels to many utterances.
(2) AnnF01, AnnF02, AnnF03, and AnnM01 rarely assigned a 'neutral' label while AnnF04 and AnnM02 assigned a 'Neutral' label to many utterances.
(3) AnnF03, AnnF04 and AnnM01 assigned an 'Error' label to many utterances.
(4) The two most assigned labels for each annotator are summarized in Table 3.

Discussion on the level-2 labels

(1) The results of the second level seem to match well with the results of the first level. 'Anger' in the first level seems to match well with 'Anger' in the second level. Utterances that are assigned with a 'Happiness' label in the first level usually have 'Happiness' or 'Satisfaction' labels in the second levels while those with 'Sadness' in the first level also receive 'Sadness' in the second level.
(2) AnnF01, AnnF02, and AnnM01 sometimes tag 'Anger' utterances with 'Confusion as their second-level labels. AnnF03, AnnF04, and AnnM02 sometimes tag 'Anger' utterances with 'Rage' as their second-level labels.
(3) AnnF01 usually assigns 'Sadness' utterances with 'Fear' in the second level while AnnM01 often assigns 'Sadness' utterances with 'Confusion in the second level.

From these observations, it can be seen that some annotators have similar tagging concepts while others have different concepts.

In addition to a comparison of the individual annotators in Table 4, we also investigated annotator agreement on annotated labels as shown in Table 5. From this table, we found that 7792 transcriptions have majority emotion labels and 1195 transcriptions have no majority labels. The most popular emotion label was 'Anger' while the least popular label was 'Sadness'. Moreover, there were ten possible scenarios of annotator agreement. Each scenario was assigned a point (1.0) to the majority label. Moreover, in the case that the majority votes were equally shared with two or more labels, the score was divided into the number of labels shared. In Table 4, *a-b-c-d-e* indicates the number of annotators for each emotion, who provided the same labels, in descending order. For example, the column 6-0-0-0-0 of the first row ('Anger') has a value of 399 which means there are 399 transcriptions assigned with 'Anger' by all six annotators. The column 3-2-1-0-0 of the first row (Anger) had a value of 414 which means there were 414 transcriptions assigned with 'Anger' by three annotators, another emotion with two votes, and another emotion with one vote. The column 2-2-2-0-0 of the first row ('Anger') had a value of 44.67. This means 134 transcriptions were assigned with Anger by two annotators and another two emotions, each with two votes. Since the 2-2-2-0-0 had three equal emotions, each emotion obtained 0.33 (= 1/3) points and with 134 transcriptions, 0.33 × 134 is equal to 44.67. This means that the points for the 3-3-0-0-0 scenario were 0.5 (= 1/2) points for each annotated label and 0.33 (= 1/3) points for the 2-2-2-0-0 scenario.

The details of the ten possible scenarios were grouped into five groups according to the largest number of votes which are described as follows.

(1) The first case is idealistic labeling, in which all six annotators share the same opinion (6-0-0-0-0) in labeling an utterance. There are only 752 out of 8987 utterances. In other words, all annotators agreed in giving the same label by approximately 8%. Sadness is the most complex emotional state with 5 utterances, while anger causes the clearest emotional state with 399 utterances.

(2) In the case of 5-1-0-0-0 one annotator did not label the emotion states in the same way as the others. In this case the number of unlabeled data is highest with 742 utterances, but sadness has only 33 utterances. The total number of matched utterances is highest among all scenarios with 2095.

(3) The majority label was voted by four annotators. There are two sub-scenarios: (1) 4-1-1-0-0 where the winning label was voted by four annotators and the other two annotators voted for two other labels, and (2) 4-2-0-0-0 where the winning label was voted by four annotators and the other two annotators voted for one other label. The total number of utterances in this scenario was 2724, which outnumbered all the other majority cases.

(4) For the cases where the majority is three, there are three scenarios: (1) 3-1-1-1-0, (2) 3-2-1-0-0, and (3) 3-3-0-0-0. The first two scenarios (2180 transcriptions) have the majority vote. The last scenario (1070 transcriptions) had two emotions as majority votes and each emotion obtained 0.5 points. Therefore, there were 535 weighted transcriptions for 3-3-0-0-0.

Table 7 Emotion distribution annotated by AnnF02, AnnF03 and AnnM01 for five major actors and actresses

	ActF01	ActF02	ActF03	ActF04	ActF05
AnnF02					
Anger	686 (50.74%)	234 (33.19%)	365 (60.23%)	211 (61.52%)	135 (44.55%)
Happiness	396 (29.29%)	142 (20.14%)	138 (22.77%)	60 (17.49%)	77 (25.41%)
Neutral	49 (3.62%)	42 (5.96%)	25 (4.13%)	17 (4.96%)	13 (4.29%)
Sadness	221 (16.35%)	287 (40.71%)	78 (12.87%)	55 (16.03%)	78 (25.74%)
Sum	1352 (100.00%)	705 (100.00%)	606 (100.00%)	343 (100.00%)	303 (100.00%)
No error	1352 (80.57%)	705 (73.28%)	606 (87.83%)	343 (78.85%)	303 (88.86%)
Errors	326 (19.43%)	257 (26.72%)	84 (12.17%)	92 (21.15%)	38 (11.14%)
Total	1678 (100.00%)	962 (100.00%)	690 (100.00%)	435 (100.00%)	341 (100.00%)
AnnF03					
Anger	534 (52.40%)	171 (29.95%)	303 (58.95%)	140 (49.65%)	103 (42.56%)
Happiness	222 (21.79%)	82 (14.36%)	96 (18.68%)	51 (18.09%)	57 (23.55%)
Neutral	75 (7.36%)	72 (12.61%)	43 (8.37%)	29 (10.28%)	20 (8.26%)
Sadness	188 (18.45%)	246 (43.08%)	72 (14.01%)	62 (21.99%)	62 (25.62%)
Sum	1019 (100.00%)	571 (100.00%)	514 (100.00%)	282 (100.00%)	242 (100.00%)
No error	1019 (60.80%)	571 (59.36%)	514 (74.49%)	282 (64.83%)	242 (70.97%)
Errors	657 (39.20%)	391 (40.64%)	176 (25.51%)	153 (35.17%)	99 (29.03%)
Total	1676 (100.00%)	962 (100.00%)	690 (100.00%)	435 (100.00%)	341 (100.00%)
AnnM01					
Anger	530 (44.92%)	207 (31.46%)	207 (36.13%)	126 (36.21%)	106 (37.32%)
Happiness	319 (27.03%)	140 (21.28%)	171 (29.84%)	89 (25.57%)	81 (28.52%)
Neutral	3 (0.25%)	11 (1.67%)	9 (1.57%)	2 (0.57%)	3 (1.06%)
Sadness	328 (27.80%)	300 (45.59%)	186 (32.46%)	131 (37.64%)	94 (33.10%)
Sum	1180 (100.00%)	658 (100.00%)	573 (100.00%)	348 (100.00%)	284 (100.00%)

Table 7 continued

	ActF01	ActF02	ActF03	ActF04	ActF05
No error	1180 (70.41%)	658 (68.40%)	573 (83.04%)	348 (80.00%)	284 (83.28%)
Errors	496 (29.59%)	304 (31.60%)	117 (16.96%)	87 (20.00%)	57 (16.72%)
Total	1676 (100.00%)	962 (100.00%)	690 (100.00%)	435 (100.00%)	341 (100.00%)
	ActM01	ActM02	ActM03	ActM04	ActM05
AnnF02					
Anger	492 (39.17%)	165 (33.54%)	145 (30.72%)	125 (44.01%)	56 (23.83%)
Happiness	396 (31.53%)	230 (46.75%)	121 (25.64%)	22 (7.75%)	153 (65.11%)
Neutral	106 (8.44%)	38 (7.72%)	68 (14.41%)	15 (5.28%)	13 (5.53%)
Sadness	262 (20.86%)	59 (11.99%)	138 (29.24%)	122 (42.96%)	13 (5.53%)
Sum	1256 (100.00%)	492 (100.00%)	472 (100.00%)	284 (100.00%)	235 (100.00%)
No error	1256 (80.05%)	492 (85.12%)	472 (82.81%)	284 (81.84%)	235 (74.84%)
Errors	313 (19.95%)	86 (14.88%)	98 (17.19%)	63 (18.16%)	79 (25.16%)
Total	1569 (100.00%)	578 (100.00%)	570 (100.00%)	347 (100.00%)	314 (100.00%)
AnnF03					
Anger	367 (36.77%)	110 (26.19%)	101 (24.82%)	99 (39.92%)	21 (11.54%)
Happiness	209 (20.94%)	151 (35.95%)	67 (16.46%)	20 (8.06%)	106 (58.24%)
Neutral	162 (16.23%)	100 (23.81%)	141 (34.64%)	27 (10.89%)	37 (20.33%)
Sadness	260 (26.05%)	59 (14.05%)	98 (24.08%)	102 (41.13%)	18 (9.89%)
Sum	998 (100.00%)	420 (100.00%)	407 (100.00%)	248 (100.00%)	182 (100.00%)
No error	998 (63.73%)	420 (72.66%)	407 (71.40%)	248 (71.68%)	235 (74.84%)
Errors	568 (36.27%)	158 (27.34%)	163 (28.60%)	98 (28.32%)	79 (25.16%)
Total	1566 (100.00%)	578 (100.00%)	570 (100.00%)	346 (100.00%)	314 (100.00%)
AnnM01					

Table 7 continued

	ActM01	ActM02	ActM03	ActM04	ActM05
Anger	339 (30.76%)	93 (20.67%)	84 (19.18%)	81 (30.68%)	36 (18.27%)
Happiness	345 (31.31%)	228 (50.67%)	143 (32.65%)	40 (15.15%)	114 (57.87%)
Neutral	18 (1.63%)	4 (0.89%)	7 (1.60%)	4 (1.52%)	3 (1.52%)
Sadness	400 (36.30%)	125 (27.78%)	204 (46.58%)	139 (52.65%)	44 (22.34%)
Sum	1102 (100.00%)	450 (100.00%)	438 (100.00%)	264 (100.00%)	197 (100.00%)
No error	1102 (70.37%)	450 (77.85%)	438 (76.84%)	264 (76.30%)	197 (62.74%)
Errors	464 (29.63%)	128 (22.15%)	132 (23.16%)	82 (23.70%)	117 (37.26%)
Total	1566 (100.00%)	578 (100.00%)	570 (100.00%)	346 (100.00%)	314 (100.00%)

(5) For the cases where the majority was two, there were three scenarios; (1) 2-1-1-1-1, (2) 2-2-2-0-0, and (3) 2-2-1-1-0. The first scenario (41 transcriptions) had the majority vote. The last two scenarios had no majority emotion. There were 702 transcriptions for 2-2-2-0-0, which are equivalent to 702 × 1/3 = 234 weighted transcriptions, and 852 transcriptions for 2-2-1-1-0, which are equivalent to 852 × 1/3 = 426 weighted transcriptions.

(6) In summary, 7792 out of 8987 transcriptions had majority votes while only 1195 transcriptions had no majority votes for three scenarios, i.e., 3-3-0-0-0, 2-2-2-0-0, and 2-2-1-1-0.

Moreover, we analyzed annotator clustering based on inter-annotator agreement (Cohen's kappa). In Fig. 7, for each actor of ActF01-ActF05 and ActM01-ActMo6, the agreement among annotators was calculated based on Cohen's kappa. Then the average was calculated as the measure of agreement between two annotators as shown in the last column of Fig. 7a. The averages are summarized in Fig. 7b to present the measure of agreement between any pair of annotators. Based on the definition in Fig. 7c, the clustering results of single linkage and complete linkage are shown in Fig. 7d, e, respectively. In conclusion, the three annotators who had the most similar opinions in emotion labeling, are AnnF02, AnnF03, and AnnM01 for both single and complete linkage when the threshold was set to 0.44. We further investigated annotator agreement on labels for only these three annotators (AnnF02, AnnF03 and AnnM01) and the results are shown in Table 6. From this table, it can be seen that 8211 transcriptions had a majority of emotion labels and 776 transcriptions had no majority labels. The most popular agreed emotion label was 'Anger' (2332) while the least popular label was 'Neutral' (284). Two possible scenarios with majorities were 3-0-0 and 2-1-0 (1.0 point for the majority), whereas one scenario with "No majority" was 1-1-1 (1/3 points each).

4.3 Actor-emotion-oriented statistics and analysis of emotion annotation data

Emotion annotation results can be applied to investigate the emotion similarities among actors. In this analysis, the results of the three most similar annotators: AnnF02, AnnF03, and AnnM01 were used. In Table 2, there were 51 actors (20 males and 31 females) in this corpus, so for the sake of simplicity, we selected the TOP-5 dominant (most seen) actresses (ActF01, ActF02, ActF03, ActF04 and ActF05), who were 18, 69, 19, 48, and 22 years old and those of actors (ActM01, ActM02, ActM03, ActM04, and ActM05) aged 24, 26, 62, 53, and 18 respectively, to represent all the actors since the total number of those actors' utterances were 7486, which is almost 83.30% of all utterances in the corpus. The emotion distributions of ten major actors/actresses tagged by AnnF02, AnnF03 and AnnM01 are shown in Table 7. Here, the leading actors were ActF01 and ActM01and their conversations covered 36% (3249 utterances) of the drama. On average, the order of emotion expressions was 'Anger' > 'Happiness' > 'Sadness' > 'Neutral'. According to Table 7, it can be seen that most actors usually expressed 'Anger', except ActF02, ActM03, and ActM04 who expressed more 'Sadness' and ActM02 and

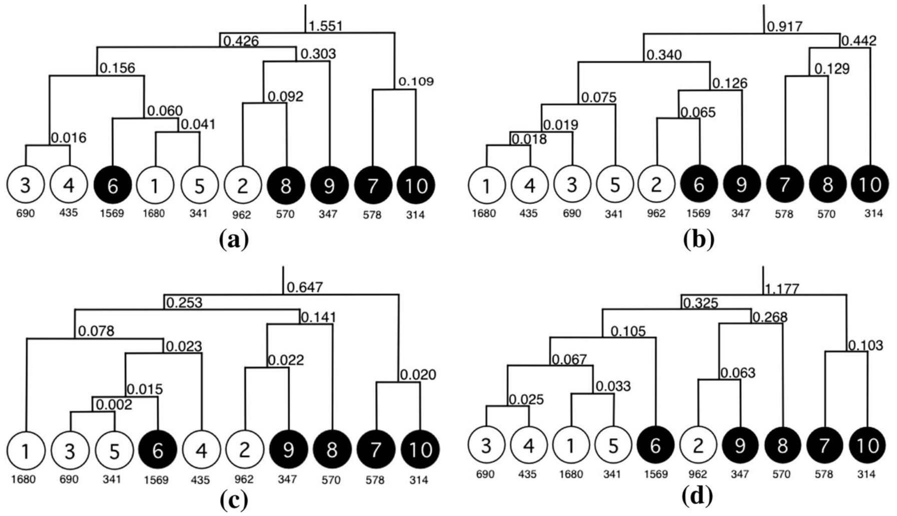


Fig. 8 Complete linkage clustering of actors' emotions labeled by three annotators (**a** AnnF02, **b** AnnF03, and **c** AnnM01) and **d** sum of three annotators

ActM05 who expressed more 'Happiness'. On average, 'Neutral' was the least expressed emotion. Moreover, on average the female actors generated more 'Anger' and 'Sadness' than the male actors. The 'Error' row indicates the number of transcriptions (conversation turns) for which the annotator could not assign an emotion. The 'Sum' and 'No Error' rows were identical in number but different in proportions. The 'Total' row presents the summation of 'Error' and 'No Error', equal to the number of total transcriptions (conversation turns) of each actor. The ActM03's emotions seemed vague, resulting in different emotion labels being assigned by the three annotators. The ActM04 usually expressed 'Anger' and 'Sadness' (more than 80% of total conversion turns). In a nutshell, the male actors generally expressed more 'Happiness' while the female actors expressed more 'Anger' and 'Sadness'.

To analyze the actor characteristics, we introduced Kullback–Leibler (KL) divergence, which is a measure for expressing the difference between two probability distributions (the proportions of two actors' emotions). We calculated $p_j^i(x)$, the probability of an emotion (x) labelled by an annotator (i) assigned to an individual actor (j), by Eq. (1). The sum of the probabilities of all four emotions (anger, happiness, neutral, and sadness) for each actor is 1.

$$p_j^i(x) = \frac{N_j^i(x)}{\sum_{x \in E} N_j^i(x)} \tag{1}$$

For example, $p_{ActF02}^{AnnF01}(AG)$ refers to the probability of the anger (AG) emotion made by ActF02, labelled by AnnF01. The KL divergence of two actors (here A and B) judged by an annotator (i) is defined by the difference between two probability distributions of all the target emotions (x) of those two actors (A and B) assigned by an annotator (i). The formal description is given in Eq. (2).

$$D_{KL}^{i}(A,B) = D_{KL}(p_A^i(x)||p_B^i(x)) = \sum_{x \in E} p_A^i(x) \log_2 \frac{p_A^i(x)}{p_B^i(x)} \tag{2}$$

Since the KL divergence was not symmetric ($D_{KL}^{i}(A, B) \neq D_{KL}^{i}(B, A)$), we applied Jensen-Shannon divergence (JSD), a symmetric KL-divergence-based measure instead, as shown in Eq. (3).

$$D_{JSD}^{i}(A,B) = \frac{1}{2} D_{KL}^{i}(A,B) + \frac{1}{2} D_{KL}^{i}(B,A) \tag{3}$$

To investigate further similarities among the actors, we used JSD to express the distance between an actor pair, and then applied the complete linkage method to the TOP-10 group of actors (5 females, 5 males). The result is shown in Fig. 8 (note that node 1 = ActF01, 2 = ActF02, 3 = ActF03, 4 = ActF04, 5 = ActF05, 6 = ActM01, 7 = ActM02, 8 = ActM03, 9 = ActM04 and 10 = ActM05, and using black circles refer to actors while white circles refer to actresses). At first glance, the female actors seemed to share a similar proportion of emotions while the male actors also expressed a similar proportion of emotions, except for ActM01 (No. 6 in the figure) and ActF02 (No. 2 in the figure). Figure 8a shows that ActF03 (No.1) and ActF04 (No.4) share the most similar emotions according to AnnF02. Moreover, most male actors, ActM02, ActM03, ActM04, and ActM05 (Nos. 7-10), are considered to have similar emotions in the drama. As shown in Fig. 8b, AnnF03 decided that ActF01 (No. 1) and ActF04 (No. 4) were the most similar pair with 0.018 of complete linkage. AnnF02 (No. 2) was the only female actor who had similar characteristics to two of the male actors, ActM01 and ActM04 (Nos. 6 and 9). Figure 8c showed similarities in the results based on AnnM01’s annotations. Although the result was similar to that of the previous annotators (AnnF02 and AnnF03), the decisions show that ActF03 (No. 3), ActF05 (No. 5) and ActM01 (No. 6) are the most similar with 0.002 and 0.015. On average, the three annotators (Fig. 8d), which is most of the female actors (Nos. 1, 3, 4, and 5) expressed their emotions in a similar way to the male actors (Nos. 7-10) even though they had a high JSD divergence. ActM01 (No. 6) is a male actor whose emotion distribution was close to that of the female actors. On the other hand, ActF02 (No. 2) is a female actor whose emotion distribution was similar that of the male actors.

5 Conclusion and future work

This paper presents the construction of a Thai emotional speech corpus, namely EMOLA, using a Thai drama series (Lakorn). The design, construction, and annotation process are discussed. In the corpus design, four basic emotion types of anger, happiness, sadness and neutral were selected and twelve subtypes of emotions: anger, confusion, disgust, excitement, fear, happiness, jealousy, pleasure, rage, sadness, satisfaction, and surprise, were used. In addition to the categories of emotion, there are also emotions based on the Pleasure-Arousal-Dominance (PAD) emotional state model, where each emotion is represented by three values: P, A, and D. An XML DTD was designed to store transcriptions, metadata and emotion tags. The mapping between the categories of emotion and the PAD representatives was analyzed. In the process of corpus construction, we transcribed and gave metadata to 8987 video clips with approximately 868 min (from a video of 1397 min in total), while we also assigned one basic type and a few subtypes to each video clip. This corpus was developed from a Thai drama series in which there were 20 male actors and 31 female actors. However, 208 utterances were produced by multiple actors. All utterances have been annotated by six annotators (AnnF01, AnnF02, AnnF03, AnnM01, AnnM02 and AnnM03). The characteristics of the corpus were investigated in three aspects: the video material, the annotators, and the actors. The relationship between basic emotions (level-1) and subtypes of emotions (level-2) and labeling of agreement patterns among the annotators were analyzed. According to an analysis of the work of the annotators, AnnF02, AnnF03 and AnnM01 were the top-3 annotators who mostly shared the same opinions in emotion labeling. Moreover, we applied Kullback–Leibler divergence and Jensen-Shannon divergence to measure the similarities between two actors' emotions and we also used the complete linkage clustering to group the actors' characteristics. The results for all three annotators show that ActF01 and ActF03 were the most similar in this corpus. This database is expected to be a resource to help us understand the use of certain emotions in the Thai language and this should then be useful for modeling the relation between video material, actors, and annotators. In the future, we plan to use this corpus for emotion recognition in Thai speech. The annotation differences among annotators will be investigated with regard to speech mood recognition. Since emotion is not usually expressed throughout the whole of a speech utterance, the location of moods in the utterance should also be studied.

Acknowledgements This work was partially supported by a SIIT graduate student scholarship, the Center of Excellence in Intelligent Informatics, Speech and Language Technology and Service Innovation (CILS), Thammasat University, the Center of Excellence in Intelligent Informatics and Service Innovation (IISI), SIIT, Thammasat University, and the Thailand Research Fund under the Grant Number RTA6080013.

References

Abrilian, S., Devillers, L., Buisine, S., & Martin, J.-C. (2005). EmoTV1: Annotation of real-life emotions for the specification of multimodal affective interfaces. In *HCI International.*

Arimoto, Y., Ohno, S., & Iida, H. (2008). Automatic emotional degree labeling for speakers' anger utterance during natural Japanese Dialog. In *LREC.*

Arimoto, Y., Ohno, S., & Iida, H. (2011). Assessment of spontaneous emotional speech database toward emotion recognition: Intensity and similarity of perceived emotion from spontaneously expressed emotional speech. *Acoustical Science and Technology, 32*(1), 26–29.

Asghar, D., Moloud, P., & Peymaneh, S. (2008). The pattern of Facial Expression among Iranian Children. In *Proceedings of Measuring Behavior* (pp. 172–173). Maastricht.

Bachorowski, J.-A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science, 8*(2), 53–57.

Bann, E. Y., & Bryson, J. J. (2012). The conceptualisation of emotion qualia: Semantic clustering of emotional tweets. In *Computational models of cognitive processes: Proceedings of the 13th neural computation and psychology workshop* (pp. 249–263). World Scientific.

Bao, W., Li, Y. A., Yang, M., Li, H., Chao, L., & Tao, J. (2014). Building a Chinese Natural Emotional Audio-visual Database. In *12th international conference on signal processing (ICSP)* (pp. 583–587).

Batliner, A., Fischer, K., Huber, R., Spilker, J., & Nöth, E. (2003). How to find trouble in communication. *Speech Communication, 40*(1), 117–143.

Burkhardt, F. A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. *Interspeech, 5,* 1517–1520.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., et al. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*, 335–359.

Campbell, N. (2003). Databases of expressive speech. In *Proceedings of oriental COCOSDA workshop.*

Cichosz, J., & Slot, K. (2005). Low-Dimensional feature space derivation for emotion recognition. In *Ninth European conference on speech communication and technology* (pp. 477–480).

Cichosz, J., & Slot, K. (2007). Emotion recognition in speech signal using emotion-extracting binary decision trees. In *Proceedings of affective computing and intelligent interaction.*

Cole, R. (2005). *The CU kids' speech corpus.* The Center for Spoken Language Research (CSLR). http://cslr.colorado.edu/.

Colombetti, G. (2009). From affect programs to dynamical discrete emotions. *Philosophical Psychology, 22*(4), 407–425.

Costantini, G., Iaderola, I., Paoloni, A., & Todisco, M. (2014). EMOVO Corpus: an Italian emotional speech database. In *LREC* (pp. 3501–3504).

Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are express in speech. *Speech Communication, 40*(1), 5–32.

Crystal, D. (1975). *The English tone of voice*. London: Edward Arnold.

Crystal, D. (1976). *Prosodic systems and intonation in English*. Cambridge: Cambridge University Press.

Dadkhah, A., Pourmohammadi, M., & Shirinbayan, P. (2008). The pattern of Facial Expression among Iranian Children. In: *Measuring behavior 2008.* Psychonomic Soc Inc, 1710 Fortview Rd, Austin, TX 78704, USA.

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication, 40*(1), 33–60.

Douglas-Cowie, E., Cowie, R., & Schroder, M. (2000). A new emotion database: considerations, sources and scope. In *ISCA tutorial and research workshop (ITRW) on speech and emotion.*

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., Mcrorie, M., et al. (2007). The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. In *Affective computing and intelligent interaction* (pp. 488–500).

Douglas-Cowie, E., Devillers, L., Martin, J.-C., Cowie, R., Savvidou, S., Abrilian, S., et al. (2005). Multimodal databases of everyday emotion: Facing up to complexity. In *Ninth European conference on speech communication and technology.*

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*(3/4), 169–200.

Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). *Emotion in the human face: Guide-lines for research and an integration of findings: Guidelines for research and an integration of findings*. Oxford: Pergamon.

Fersini, E., Messina, E., & Archetti, F. (2012). Emotional states in judicial courtrooms: an experimental investigation. *Speech Communication, 54*(1), 11–22.

Fersini, E., Messina, E., Arosio, G., & Archetti, F. (2009). Audio-based emotion recognition in judicial domain: A multilayer support vector machines approach. In *International workshop on machine learning and data mining in pattern recognition* (pp. 594–602). Springer.

Fu, L., Mao, X., & Chen, L. (2008). Speaker independent emotion recognition based on SVM/HMMs fusion system. In *International conference on audio, language and image processing, 2008 (ICALIP2008)* (pp. 61–65). IEEE.

Greasley, P., Setter, J., Waterman, M., Sherrard, C., Roach, P., Arnfield, S., et al. (1995). Representation of prosodic and emotional features in a spoken language database. In *Proceedings of the XIIIth ICPhS.*

Grimm, M., Kroschel, K., & Narayanan, S. (2008). The Vera am Mittag German audio-visual emotional speech database. In *IEEE international conference on multimedia and expo.*

Haq, S., Jackson, P. J., & Edge, J. (2008). Audio-visual feature selection and reduction for emotion classification. In *Proceedings of AVSP* (pp. 185–190).

Havlena, W. J., & Holbrook, M. B. (1986). The varieties of consumption experience: Comparing two typologies of emotion in consumer behavior. *Journal of Consumer Research, 13*(3), 394–404.

Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A., & Nogueiras, A. (2002). Interface databases: Design and collection of a multilingual emotional speech database. In *LREC.*

Iida, A., Campbell, N., Iga, S., Higuchi, F., & Yasumura, M. (1998). Acoustic nature and perceptual testing of corpora of emotional speech. In *ICSLP.*

Johnstone, T., & Scherer, K. R. (1999). The effects of emotions on voice quality. In *Proceedings of the XIVth international congress of phonetic sciences* (pp. 2029–2032). Citeseer.

Kaiser, S., & Scherer, K. R. (1998). Models of 'normal' emotions applied to facial and vocal expression in clinical disorders. In J. Flack, F. William & J. D. Laird (Eds.), *Emotions in psychopathology: Theory and research* (pp. 81–98).

Koolagudi, S. G., Maity, S., Kumar, V. A., Chakrabarti, S., & Rao, K. S. (2009). IITKGP-SESC: Speech database for emotion analysis. In *International conference on contemporary computing* (pp. 485–492). Berlin: Springer.

Kostoulas, T., Ganchev, T., Mporas, I., & Fakotakis, N. (2008). A real-world emotional speech corpus for modern greek. In *LREC.*

Kövecses, Z. (2003). *Metaphor and emotion: Language, culture, and body in human feeling*. Cambridge: Cambridge University Press.

Laskowski, K., & Burger, S. (2006). Annotation and analysis of emotionally relevant behavior in the ISL meeting corpus. In *LREC.*

Li, A. (2015). *Encoding and decoding of emotional speech: A cross-cultural and multimodal study between Chinese and Japanese*. Berlin: Springer.

Lian-hong, C., Dan-dan, C., & Rui, C. (2007). TH-CoSS,a Mandarin Speech Corpus for TTS. *Journal of Chinese Information Processing*, *02*.

Lubis, N. A. (2014). Construction and analysis of Indonesian emotional speech corpus. In *17th oriental chapter of the international committee for the co-ordination and standardization of speech databases and assessment techniques (COCOSDA)* (pp. 1–5).

Lubis, N., Gomez, R., Sakti, S., Nakamura, K., Yoshino, K., Nakamura, S., et al. (2016). Construction of Japanese audio-visual emotion database and its application in emotion recognition. In *LREC.*

Lubis, N., Sakti, S., Neubig, G., Toda, T., & Nakamura, S. (2015). Construction and analysis of social-affective interaction corpus in English and Indonesian. In *Oriental COCOSDA held jointly with 2015 conference on asian spoken language research and evaluation (O-COCOSDA/CASLRE)* (pp. 202–206).

Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006). The eNTERFACE'05 audio-visual emotion database. In *Data engineering workshops, 2006* (p. 8). IEEE.

Mehrabian, A. (1995). Relationships among three general approaches to personality description. *The Journal of Psychology, 129*(5), 565–581.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology, 14*(4), 261–292.

Mehrabian, A., & Russell, J. A. (1974). *Approach to environmental psychology*. Cambridge, MA: MIT Press.

Mori, H., Satake, T., Nakamura, M., & Kasuya, H. (2008). UU database: A spoken dialogue corpus for studies on paralinguistic information in expressive conversation. In *International conference on text, speech and dialogue* (pp. 427–434). Berlin: Springer.

Mori, H., Satake, T., Nakamura, M., & Kasuya, H. (2011). Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics. *Speech Communication, 53*(1), 36–50.

Moriyama, T., Mori, S., & Ozawa, S. (2009). A synthesis method of emotional speech using subspace constraints in prosody. *Journal of Information Processing, 50*(3), 1181–1191.

Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication*, *41*(4), 603–623.

O'Connor, J., & Arnold, G. (1973). *Intonation of colloquial English*. London: Longman.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory, research, and experience* (Vol. 1, pp. 3–31). New York: Academic Press.

Plutchik, R. (1984). Emotions: A general psychoevolutionary theory. In *Approaches to emotion* (pp. 197–219).

Posner, J., Russell, J. A., & Petersona, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dvelopmental and Psychopathology, 17*(3), 715–734.

Ringeval, F., Sonderegger, A., Sauer, J., & Lalanne, D. (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE international conference and workshops on* (pp. 1–8). IEEE.

Russell, J., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality, 11,* 273–294.

Saratxaga, I., Navas, E., Hernaez, I., & Luengo, I. (2006). Designing and recording an emotional speech database for corpus based synthesis in Basque. In *Proceedings of fifth international conference on language resources and evaluation (LREC)* (pp. 2126–2129).

Scherer, K. R. (1986). Vocal affect expression: a review and a model for future research. *Psychological Bulletin, 99*(2), 143.

Scherer, K. R. (1995). Expression of emotion in voice and music. *Journal of Voice*, *9*(3), 235–248.

Scherer, K. R., & Tannenbaum, P. H. (1986). Emotional experiences in everyday life: A survey approach. *Motivation and Emotion, 10*(4), 295–314.

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

Schubiger, M. (1958). *English intonation, its form and function*. Tübingen: M. Niemeyer Verlag.

Sneddon, I., McRorie, M., McKeown, G., & Hanratty, J. (2012). The Belfast induced natural emotion database. *IEEE Transactions on Affective Computing, 3*(1), 32–41.

Stein, N. L., & Oatley, K. (1992). Basic emotions: Theory and measurement. *Cognition and Emotion, 6*(3–4), 161–168.

Trong, K. P., Neerincx, M. A., & Van Leeuwen, D. A. (2008). Measuring spontaneous vocal and facial emotion expressions in real world environments. In *Proceedings of measuring behavior 2008* (pp. 170–171). Maastricht.

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

Wang, X., Li, A., & Tao, J. (2007). An expressive speech corpus of standard Chinese. In *O-COCOSDA2007.* Hanoi, Vietnam.

Watson, D., & Tellegan, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin, 98,* 219–235.

Wu, T., Yang, Y., Wu, Z., & Li, D. (2006). MASC: A speech corpus in Mandarin for emotion analysis and affective speaker recognition. In *2006 IEEE Odyssey-the speaker and language recognition workshop* (pp. 1–5).

Wundt, W. M. (1897). Outlines of psychology. In http://psychclassics.asu.edu/index.htm, *Classics in the history of psychology.* Toronto: York University 2010.

Yamagishi, J., Onishi, K., Masuko, T., & Kobayashi, T. (2005). Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis. *IEICE TRANSACTIONS on Information and Systems, 88*(3), 502–509.

Zhang, S., Ching, P., & Kong, F. (2006). Acoustic analysis of emotional speech in Mandarin Chinese. In *International symposium on chinese spoken language processing* (pp. 57–66).

Zhang, S., Xu, Y., Jia, J., & Cai, L. (2008). Analysis and modeling of affective audio visual speech based on PAD emotion space. In *6th international symposium on Chinese spoken language processing* (pp. 1–4). Kunming, China.

Zovato, E., Sandri, S., Quazza, S., & Badino, L. (2004). Prosodic analysis of a multi-style corpus in the perspective of emotional speech synthesis. In *ICSLP 2004* (Vol. 2, pp. 1453–1457). Prentice Hall.