ASSESSING THE NATURALNESS OF MALAY EMOTIONAL VOICE CORPORA

Mumtaz B. Mustafa[1, 4], Raja N. Ainon[1, 4], Roziati Zainuddin[1,4], Zuraidah M. Don[2,4], Gerry Knowles[3,4]

[1]Faculty of Computer Science and Information Technology, University of Malaya, Malaysia
[2]Faculty of Language and Linguistics, University of Malaya, Malaysia
[3]Lingenium Sdn Bhd, Kuala Lumpur, Malaysia
[4]Computational Speech Group, ICT Research Cluster, University of Malaya
mumshaka4@siswa.um.edu.my,ainon,roziati,zuraida@um.edu.my,g.knowles@lancaster.ac.uk

ABSTRACT

This research reports the development and evaluation of Malay emotional voice corpora through listening evaluation, and how the numbers of emotion choices offered to evaluators affect the result of the evaluation. The voice corpora comprises of three emotions, namely anger, sadness and happiness being expressed by two male and two female actors. The voice corpora were evaluated in two separate listening tests involving a number of Malay native evaluators balanced for gender, age and profession. In the first listening test, evaluators were given twenty five choices of emotions to choose from. For the second test, the number of emotion choices is only five. Each test was conducted separately with different group of evaluators. The results of the two tests are grossly different with the emotion identification rate of the first test lower than the second test.

Index Terms— Malay emotional voice database, listening evaluation, emotion identification rate, forced choice approach

1. INTRODUCTION

Synthesizing emotional speech that is both natural and pleasant as human speech was a challenge in speech synthesis research. Many of the existing speech synthesis system were unable to effectively synthesize emotional speech with high degree of naturalness with proper articulation, prosody and voice quality [1, 2, 3]. New generation speech synthesis systems such as the corpus based and parametric statistical speech synthesis system have the ability to generate synthetic speeches which is very close to human speech in term of naturalness and intelligibility [4]. These new generations speech synthesis system generates synthetic voices by making reference to recorded human speech unit inventory.

In order to generate high quality synthetic voices, the recorded voice database must be of high quality as well. Naturalness and quality of recording is usually determined by conducting listening evaluations, which commonly consist of Emotion Identification Rate (EIR) [5, 6, 7, 8], and Mean Opinion Score (MOS) [9, 10]. EIR test is commonly conducted by giving a limited number of choices to evaluators to choose. Is this approach sufficient enough for screening the quality and naturalness of recorded corpora? Forced choice approach for emotion identification has been criticized as it allows evaluators to have a predictive pattern based on the choices offered. Adding more choices can reduces any emerging pattern that may be identified by evaluators, which can distort the emotion identification. This research investigates how the choices offered can affect the EIR and the naturalness scores of Malay emotional speech corpora.

Malay is one of the western branches of Austronesian languages which are widely spoken among Malay-speaking countries such as Malaysia and Indonesia [11]. There are some common features between Malay language and English language. Firstly, Malay language is a phonetic language and it is written in Roman characters. Secondly, all syllables are pronounced almost equally and it is thus, considered as a non-tonal language. In general, there are 6 main vowels and 26 consonants in standard Malay. Nineteen of the consonants, namely /b, d, f, h, j, k, l, m, n, ŋ ("ng"), p, q, r, s, t, v, w, y, z/ are phonetically close to their English equivalents [11].

Malay is an under-resourced language with few resources by way of recorded speech particularly emotional speech corpora. This research propose the development of Malay emotional voice corpora that are rich in content, phonetically balanced and have a fairly good coverage of the language [12]. Exactly what constitute a phonetically balanced database was not well discussed in the literature. This research, therefore, interpret phonetically balanced database as having the maximum numbers of phone,

syllable and word combinations. The naturalness and quality of the emotional corpora was evaluated in a listening test. This research also investigates how the choices offered can affect the EIR and the naturalness scores of Malay emotional speech corpora.

The remainder of this paper is organized as follows. Part 2 discusses the preparation of data for developing the emotional voice corpora. Part 3 describes the listening evaluation procedures. Evaluation results are presented in part 4. Part 5 discusses the results and part 6 summarises the main conclusions.

2. DEVELOPMENT OF MALAY EMOTIONAL CORPORA

This research creates Malay emotional speech corpora which can be applied for the development of emotional speech synthesis system as well for the evaluation and understanding of the Malay emotional voices. This section describes the procedures for the development of Malay emotional voice corpora designated as MESD.

2.2. Sentence creation

This research propose the creation of 1,000 Malay sentences that have a good mixture of Malay words, syllable and phones to be applied for developing Malay voice database. These sentences were constructed by referring to various written sources such as local Malay newspaper (43%), educational text books (39%), and other general reading materials (18%). The length of the sentences constructed range from 3 words to about 12 words per sentence, and the average sentence length is about 5.5 words.

Table 1. Syllable structure types and tokens

Structure	Types	Tokens	Tokens of the five commonest types
V	5	219	a = 91, u = 43, i = 43, e = 26, o = 16
VC	50	374	an = 115, ar = 20, at = 19, in = 19, as = 18
CV	90	3,400	di = 221, ba =124, me = 124, ke = 122, la = 115
CVC	561	2,606	kan = 189, ber = 79, ter = 58, lah = 55, per = 52

The corpus consists of 5,534 word tokens (2,763 different word types), 12,666 syllables and 39,996 phones. Sentences constructed were to include all possible syllable types, namely vowel (V), vowel-consonant (VC), consonant-vowel (CV) and consonant-vowel-consonant (CVC), so that the voice database for Malay is phonetically balanced. The 1,000 sentences comprises of 5 types of V, 50 types of VC, 90 types of CV and 561 types of CVC. Table 1 gives the frequencies of types and tokens, together with the number of tokens for each of the five most frequent syllable types. The syllables 'di', 'ke', 'kan', 'ber', 'ter', 'an' and 'per' are very common because they are also affixes, while 'lah' is also an enclitic. The majority of the words in Malay are bi-syllabic or tri-syllabic, which are the most common word structures in Malay, together accounting for 97.52% of Malay words [13]. To reflect this, these two word types make up 77% of the words in the database. Figure 1 shows the number of syllables in the 2,763 words used for this research.

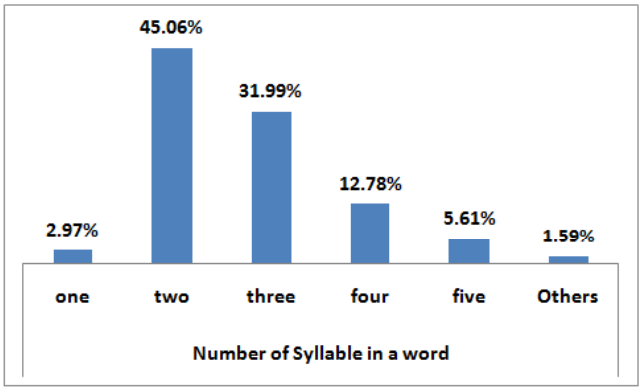

Figure 1: Proportions of numbers of syllables in words in the Malay G2P database extracted from the corpus of 1,000 Malay sentences.

2.3. Choices of emotion

In order to develop a high quality emotional voice database, it is important to understand the types and classification of human emotional speech. Emotion can be described as change in the state of readiness for maintaining or modifying the relationships with the environment [14]. There are various types and ways human can express emotion including crying, laughing, shouting, and also by speech. In the area of emotional speech synthesis, only a handful of emotional speech are being simulated and evaluated. As far as emotional speech synthesis is concerned, the common six emotions that were regularly considered are sadness, happiness, anger, fear, surprise and disgust [3, 7, 8, 14] as they represent the most commonly distinguishable types of emotional speech.

Classification of emotion in emotional speech synthesis enables researchers to focus on groups of emotions that are clearly distinguishable and easily differentiated by users to speech synthesis systems. The choice of emotions is attributed to how human emotions can be grouped and classified. A number of literatures have attempted to classify human emotions into meaningful groups.Table 2 lists some of the classification of basic human emotion by

different theorists. Among the different classifications; anger, sadness, happiness and fear are the most commonly recognized as basic emotions.

Table 2. Major emotional classification by emotion theorist

Theorist	Basic Emotion
James [15]	Fear, grief, love, rage
McDougall [16]	Anger, disgust, elation, fear, subjection, tender-emotion, wonder
Izard [17]	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise
Plutchik [18]	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
Ekman) [19]	Anger, disgust, fear, joy, sadness, surprise
Tomkins [20]	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise
Frijda [21]	Desire, happiness, interest, surprise, wonder, sorrow
Parrot [22]	Love, joy, surprise, anger, sadness, fear

2.4. Recording the MESD voice database

For preparing the emotional voice databases comprises of three types of emotions, namely anger, happiness and sadness, four professional actors, two male (MS 1 and MS2) and two female (FS1 and FS2), have contributed their voices. The same 1,000 sentences were used for recording the emotional voices for anger, happiness and sadness. Using the same set of sentences for recording different types of emotion, enables greater control of contextual factors and allows a better comparison of speech prosody among the three emotions. However, the lack of emotional cue on the sentences makes it difficult for the voice contributors to express the target emotion. During recording, voice contributors were advised to minimize the variation of expression and apply a single domain that they are familiar with. Each recording session focused on one single expression and voice contributors were given pre-training session before the actual recording. Table 3 shows the duration of each recording session, with total recording time about 29.7 hours.

3. EVALUATION OF MESD VOICE DATABASE

The quality and naturalness of the recorded emotional voices of MESD database were evaluated using a series of listening evaluation. The objective of the evaluation is to validate the appropriateness of emotional cue of each recorded emotional utterances. Forced choice approach for emotion identification test has been criticized as it allows evaluators to have a predictive pattern based on the choices offered. Adding more choices can reduce any emerging pattern that may be identified by evaluators that can distort the emotion identification. The listening test of MESD is divided into two sessions, the first session (FET), which offers evaluators a choice of 25 emotions for the EIR test and second session (SET), which only offers 5 choices of emotions. The main difference between FET and SET is the number of choices offered to evaluators. The objective of separating the listening evaluation into FET and SET is to investigate how the choices offered to evaluators affect the result of the emotion identification.

Table 3. The length of each voice database in MESD

Voices	Recording time (hour)		
	Angry	Happy	Sad
MS1	2.17	2.30	2.75
MS2	2.25	2.50	2.80
FS1	2.00	2.50	2.75
FS2	2.25	2.70	2.70

Table 4. List of 25 choices of emotion for the First Evaluation Test (FET) categorized according to Parrot [22]

Primary Emotion	Secondary Emotion	Tertiary emotion (the 25 choices)
Joy	Cheerful	Happiness (Gembira)
		Joy (Riang)
		Cheerful (Ceria)
		Delight (Suka)
	Zest	Excitement (Kegirangan)
	Contentment	Pleasure (Kepuasan)
Surprise	Surprise	Surprise (Hairan)
Anger	Exasperation	Frustration (Kegagalan)
	Rage	Anger (Marah)
		Fury (Geram)
		Hate (Benci)
		Resentment (Dendam)
	Disgust	Disgust (Jijik)
	Envy	Jealous (Cemburu)
Sadness	Suffering	Hurt (Melukai)
	Sadness	Sadness (Sedih)
		Sorrow (Dukacita)
	Shame	Shame (Malu)
		Regret (Kesal)
	Sympathy	Sympathy (Kasihan)
Fear	Horror	Fear (Takut)
		Shock (Terkejut)
	Nervousness	Worry (Risau)
		Distress (Kesengsaraan)
Neutral (not part of Parrot model)		

100 native Malay listeners of diverse gender, age and profession were involved in the listening test. MESD database has 12,000 recorded utterances (1,000 sentences x 4 speakers x 3 emotions), and these utterances was randomly divided into 50 set of folders, each containing 240 voice snippets. Each evaluator only evaluates any one of the fifty sets. Table 4 lists the 25 choices of emotions offered for FET. For SET, the five choices are anger, sadness, happiness, fear and excitement.

We randomly divide 50 evaluators for FET and the remaining 50 for SET. The listening evaluation was conducted in studio environment to reduce external noises. Evaluators were given ample time to carry out the test and they can listen to the voice snippet as many time before answering 3 questions pertaining to the voice snippet as applied in [8] which are:

- Q1: Emotion identification rate (EIR) using forced choice method (Evaluators have choices of 25 types of emotions for FET and 5 choices for SET).
- Q2: Quality assessment score (High, average or low) (QS)
- Q3: Effort score (ES) taken to identify the emotion (1: very easy to 7: very hard)

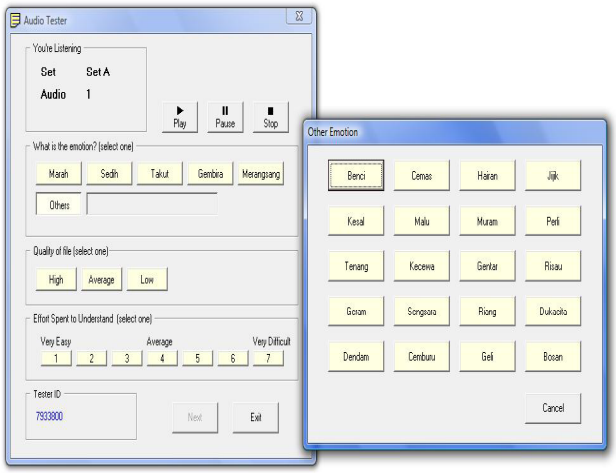

Figure 2: The audio tester for the first listening evaluation (FET)

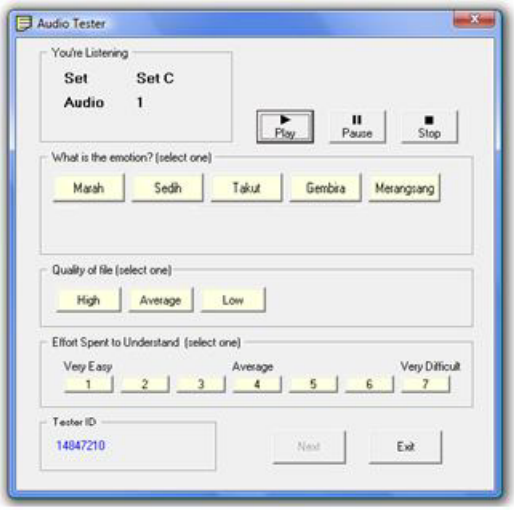

Figure 3: The audio tester for the second listening evaluation (SET)

In facilitating the listening test, an audio evaluation and processing system has been developed to record and analyze the results of the evaluation of each evaluator. An identification number is generated by the system to save separately, the answer given by each evaluator. The audio evaluation system allows evaluators to response to all three questions effectively and also enable fast statistical evaluation. Figure 2 shows the listening evaluation system for FET, which offered 25 choices of emotion and figure 3 shows the listening evaluation system for SET with 5 choices of emotion.

4. RESULTS OF LISTENING EVALUATION

For FET, the emotion of anger had the highest identification rate (76.22%) and sadness the lowest (60.87%). In general, the three investigated emotions were confused with other types of emotions especially emotions that share common matrix as shown in table 4. For example, happiness was confused with similar emotions like joy and excitement. For SET, the three emotions scored a higher identification, with sadness the highest (90.18%) and happiness the lowest (87.47%). Table 5 tabulates the EIR (Q1) for FET and SET. The differences in identification rate of FET and SET clearly follows the number of choices offered. For FET, male sadness and happiness are better identified than that of female, and in SET, male happiness and anger has better identification. We also found that shorter utterances have better recognition than longer ones. Table 6 shows the results of perceptual evaluation of recorded emotional speech classified according to gender and utterance length.

The quality score (Q2) of anger is the highest (4.86) and happiness the lowest (4.62). We found that the number of choices offered has no effect on the quality assessment as both FET and SET have similar quality scores. For the effort score test (Q3), we found that more effort is required by evaluators to identify an emotion when more choices are given. Happiness has the highest effort score for both FET and SET (4.02 and 2.11 respectively). For FET, sadness (3.88) has higher effort score than anger (3.58). Whereas, for SET, anger (1.73) score higher than sadness (1.68). In both evaluations, neutral speech has the lowest effort score compared to emotional speech. We found that the effort score depends on the number of choices offered and the way the speech is expressed by the actors. Table 7 shows the quality score (Q2) and effort score (Q3) for both tests highlighting the best scores.

5. DISCUSSION

The listening evaluation of FET and SET produces difference results for EIR and ES, which may caused by number of factors including number of choices offered,

Table 5. Emotion identification rates (Q1: EIR) by evaluator from the first evaluation (FET) and second evaluation tests (SET). Emotions choice with identification less than 2.00% is grouped as others

	Recorded speech					
Recognized as	Anger		Sadness		Happiness	
	FET	SET	FET	SET	FET	SET
Anger	**76.22**	**89.79**	0.02	0.23	0.64	2.17
Sadness	0.06	0.69	**60.87**	**90.18**	0.07	1.03
Happiness	0.21	4.08	0.11	0.88	**70.29**	**87.47**
Joy	0.19	-	0	-	5.86	-
Surprise	0.44	-	0.07	-	2.37	-
Sorrow	0.04	-	3.69	-	0.02	-
Disgust	4.81	-	0.33	-	1.11	-
Excitement	0.33	5.44	0.03	-	13.84	7.45
Fear	0	-	20.16	8.71	0.08	1.88
Hate	10.32	-	0.19	-	0.21	-
Jealous	3.14	-	0.27	-	0.81	-
Neutral	0.15	-	2.81	-	0.97	-
Worry	0.05	-	6.17	-	0	-
Others	4.04	-	5.28	-	3.73	-

Table 6. Recognition rates of emotions (Q1: EIR) for FET and SET categorized according to gender and utterance length

Categories			Emotional Identification Rate (%)		
			Anger	Sadness	Happiness
Gender	Male	FET	74.59	65.92	74.66
		SET	91.05	88.87	89.73
	Female	FET	77.85	55.82	65.92
		SET	88.53	91.49	85.21
Utterance	Long	FET	71.97	51.98	63.35
		SET	86.36	86.75	81.92
	Short	FET	80.47	69.76	77.23
		SET	93.22	93.61	93.02

Table 7. Quality assessment (Q2: QS) and effort score (Q3: ES) for FET and SET categorized according to gender and utterance length

	Test	Emotion		
		Anger	Sadness	Happiness
Quality assessment (Q2)	FET	4.86	4.78	4.61
	SET	4.86	4.77	4.63
Effort score (Q3)	FET	3.58	3.88	4.02
	SET	1.73	1.68	2.11

biasness when dividing the evaluators for FET/SET, as well as many other random factors such as demographic, ages and gender of evaluators. This research attempts to reduce the effect of evaluator's biasness by ensuring a balanced and diverse group of evaluators for demographic, age and gender.

We found that the quality score (QS) of FET is very similar to SET, which indicates that evaluators from both tests have similar perception about the quality of recording. The similarity of QS for FET and SET indicate non existence of evaluator's biasness of FET and SET.

The dissimilarity of EIR and ES between FET and SET therefore, is attributed to the number of choices offered to evaluators. Variance of SET (162.59) is higher than FET (60.89) and is significantly different as indicated by ANOVAs test ($p < 0.05$). This indicates that number of choices offered does influence the EIR and ES, but not the QS of recorded utterances.

In this research the EIR for male voices is generally better than female voices (two out of three male voices has better EIR than female voices for both FET and SET). Does this mean that male is better in uttering emotional speech for Malay? It is impossible to make that conclusion; however, the result from this evaluation shows that the emotional speech varies among genders and speakers, an important factor to be considered when preparing voice database for developing speech synthesis system. This is because, a poor emotional voice database can impair the

naturalness of the synthetic voices regardless the effectiveness of the speech synthesis system. The EIR and ES of shorter utterances are higher than long utterances for all three emotions for SET and FET. One reason for this is because of dilution of prosody for longer utterances that make it difficult for evaluators to accurately identify the emotional content.

7. CONCLUSIONS

The forced choice approach for emotion identification has been criticized as it allows evaluators to have a predictive pattern based on the choices offered. This research found that the number of choices offered to evaluators can influence the outcome of a listening evaluation. As the sentences applied for recording emotional corpora has no emotional meaning, evaluators relies only on speech factors such as prosody and voice quality to determine the types of emotion. Evaluator's confusion escalates when the choices of emotion offered are similar to each other.

Listening evaluation is an important form of evaluation for research in speech synthesis. However, providing a choice of emotion for EIR limits the usefulness and validity of the test as evaluators are induced to choose only the types of emotion offered. In real life application, users will not be provided with a list of choices to determine the types of emotional speech generated by emotional speech synthesis systems. One possible solution to enhance the credibility of forced choice method of listening evaluation is to merge forced choice and open choice method by allowing evaluators to freely state the types of emotion they heard which are not available in the list of choices. This research has also build diverse and complete emotional voice corpora with high quality of recording, which can be applied for the development of emotional speech synthesis systems for Malay.

8. ACKNOWLEDGMENTS

This work is supported by a research grant from University of Malaya, Malaysia.

9. REFERENCES

[1] K.R. Scherer, "Vocal communication of emotion: A review of research paradigms," Klaus R. Scherer. *Speech communication,* vol. 40, pp. 227-256, 2003.

[2] P. Lieberman, and S.B. Michaels, "Some aspects of fundamental frequency and envelop amplitude as related to the emotional content of speech," *Journal of Acoustical Society of America*, vol. 34, pp. 922-927, 1962.

[3] M. Schröder, "Emotional speech synthesis: A review," *In: Proc Eurospeech,* vol.1, pp. 561-564, 2001.

[4] P. Taylor, *Text-to-Speech Synthesis,* Cambridge University Press, 2009.

[5] J. Yamagishi, K. Onishi, T. Masuka, and T. Kobayashi, "Acoustic modeling of speaking style and emotional expression in HMM-Based speech synthesis," *IEICE Trans. Inf. & Syst*., vol. E88-D, no. 3, pp. 502-509, 2005.

[6] H. Tsuzuki, H. Zen, K. Tokuda, T. Kitamura, M. Bulut, and S. Narayana, "Constructing emotional speech synthesizers with limited speech database," *In: Proc. ICSLP*, vol. 2, pp. 1185-1188, 2004.

[7] F. Burkhardth, and W. Sendlmeie, "Verification of acoustic correlation of emotional speech using formant-synthesis," *In: Proc. ISCA workshop on speech and emotion,* pp. 151-156, 2006.

[8] R.C. Barra, J. Yamagishi, S. King, J. M. Montero, and J. Macias-Guaras, "Analysis of statistical parametric and unit selection speech synthesis system applied to emotional speech," *Speech Communication,* vol. 52(5), pp. 394-404, 2010.

[9] S. King, K. Tokuda, H. Zen, and J. Yamagishi, "Unsupervised adaptation for HMM-based speech synthesis," *In: Proc. Interspeech 2008,* pp. 1869-1872, 2008.

[10] S. Chomphan, and T. Kobayashi, "Tone correctness improvement in speaker dependent HMM-based Thai speech synthesis," *Speech Communication*, 50: 392-404, 2008.

[11] W.K. Loo, H.O. Siew, and Z. Roziati, "Building a unit selection speech synthesiser for Malay language using Festvox and hidden markov model toolkit (HTK)," *Chiang Mai University Journal of Natural Sciences,* 6(1), pp. 149-158, 2007.

[12] E. Navas, I. Hernáez, and I. Luengo, "An Objective and Subjective Study of the Role of Semantics and Prosodic Features in Building Corpora for Emotional TTS," *In IEEE Transactions on audio, speech and language processing,* vol. 14, no. 4, pp. 1117-1127, 2006.

[13] B. S. Teoh, *The Sound System of Malay Revisited*, Kuala Lumpur, Dewan Bahasa dan Pustaka, Malaysia, 1994.

[14] A. Iida, N. Campbell, F. Higuchi, M. Yasumura, "A corpus-based speech synthesis system with emotion," *Speech communication*, vol.40, pp. 161-187, 2003.

[15] W. James, "What is an emotion? Mind," 9, 188-205, 1884.

[16] W. McDougall, *An introduction to social psychology*. Boston: Luce, 1926.

[17] C. Izard, *Human Emotions*. Plenum Press, New York, 1977.

[18] R. Plutchik, *A general psychoevolutionary theory of emotion.* In R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory, research, and experience,* vol. 1. Theories of emotion, pp. 3-33, New York: Academic, 1980.

[19] P. Ekman, W. V. Friesen, and P. Ellsworth, *What emotion categories or dimensions can observers judge from facial behavior?* In P. Ekman (Ed.), *Emotion in the human face,* pp. 39-55, New York: Cambridge University Press, 1982.

[20] S. S. Tomkins, *Affect theory*. In K. R. Scherer & P. Ekman (Eds.), *Approaches to emotion,* pp. 163-195, Hillsdale, NJ: Erlbaum, 1984.

[21] N. Fridja, *The emotion*, Cambridge, U.K. Cambridge University Press, 1986.

[22] W. Parrott, *Emotions in Social Psychology,* Psychology Press, Philadelphia, 2001.