Journal of Engineering and Applied Sciences 2 (8): 1262-1267, 2007
ISSN: 1816-949X

Research on a Methodology to Model Speech Emotion

Y.U. Dong-Mei and Jian-An Fang
College of Information Science and Technology, Donghua Unviversity,
Shanghai 201620, China

Abstract: Research was conducted to develop a methodology to model the emotional content of speech as a linear function of time and speech features. In this study, emotional type is coming from 6 base emotions (anger, disgust, fear, joy, sadness and surprise). But, in the application is not defined as a simple word, is quantified using the dimensions valence and arousal and the value of valance or arousal is expressed as some interval. Results demonstrate that the model is more excellence than others.

Key words: Emotion, speech, model, valance and arousal, interval

INTRODUCTION

Emotions are an essential part of human life; they influence how we think, adapt, learn, behave and how humans communicate with others. Computer models of emotions constitute a fascinating research topic and, although there are some pioneering works in this domain (Abelson, 1963), only recently it has received the attention it deserves. The work of researches like Dyer (1987), Ortony *et al.* (1994), Picard (1997), Cañamero (2001) and Rafael (2006) etc., have enormously contributed to the development of the area. The kind of approaches, motivations and goals that researchers interested in computer and emotions pursue are very different.

Affective computing includes implementing emotions and therefore can aid the development and testing of new and old emotion theories. However, affective computing also includes many other things, such as giving a computer the ability to recognize and express emotions, developing its ability to respond intelligently to human emotion and enabling it to regulate and utilize its emotions (Picard, 1997). Generally, communication between emotions through the model of speech, gesture, music and action etc. especially, there are many motivations in identifying the emotional state of speakers. In human-machine interaction, the machine can be made to produce more appropriate responses if the state of emotion of the person can be accurately identified (Tin *et al.*, 2003). There are many motivations in identifying the emotional state of speakers. In human-machine interaction, the machine can be made to produce more appropriate responses if the state of emotion of the person can be accurately identified. Current research can be put importance to emotion recognition about speech, from the reported findings on features of speech and emotional states (Williams and Stevens, 1981; Frick, 1985) all of these studies depend on measuring speech features representing speech contents including fundamental frequency (F0) contour, continuous acoustic variables and voice quality, respectively. A comparison of these studies reveals the treating emotion as a discrete variable involves ambiguously selecting the number of emotions (Mark *et al.*, 2006). To resolve this ambiguity, Schubert (1999) recommends representing an emotion as a continuous multidimensional variable.

Because speech changes with time, the emotion communicated by the speech can also change with time. Because the emotion can vary throughout the sentence selection, a time-varying method of measuring emotion is more appealing than describing speech with a single emotion. The goal of this study, is to develop a methodology to create valid models of time-varying continuous emotional content for a speaker.

In the following study, an introductory overview of emotion representation, this is followed by a discussion of the background of the speech emotion. Based on this analysis, a new approach to structure emotion has been developed with a focus on the part of provided method. Results of an experiment based on this approach are described which show the result of the new approach.

EMOTION REPRESENTATION

There are 2 main approaches to structure emotion. Emotions can be grouped into theories that focus on how emotions arise and how they are perceived and theoriesfocusing on how observed emotions could be categorized or structured in the field psychology.

Corresponding Author: Y.U. Dong-Mei, College of Information Science and Technology, Donghua Unviversity, Shanghai 201620, China

Discrete emotion theories and the concept of basic emotions: From psychology on emotions, psychologists have suggested a different number of these, ranging from 2-18 categories, but there has been considerable agreement on the following 6: Anger, disgust, fear, joy, sadness and surprise.

Although this category is be accepted by many researchers, examinations of the semantics of basic emotion terms by researcher, Wierzbicka (1992) showed contradictive results. She explained that for some languages certain words describing basic emotions do not exist (such as the word anger for the Ilongot language of the Philippines or the Ifaluk language of Micronesia) and concluded that the basic emotions are just cultural artifacts of the English language.

Dimensional emotion theories and the Circumplex model of affect: Use dimensions rather than discrete categories to describe the structure of emotions about dimensional emotion theories. According to a dimensional view, all emotions are characterized by their valence and arousal. However, arousal and valence are not claimed to be the only dimensions or to be sufficient to differentiate equally between all emotions, but they have proved to be the main dimensions (Russell, 1983).

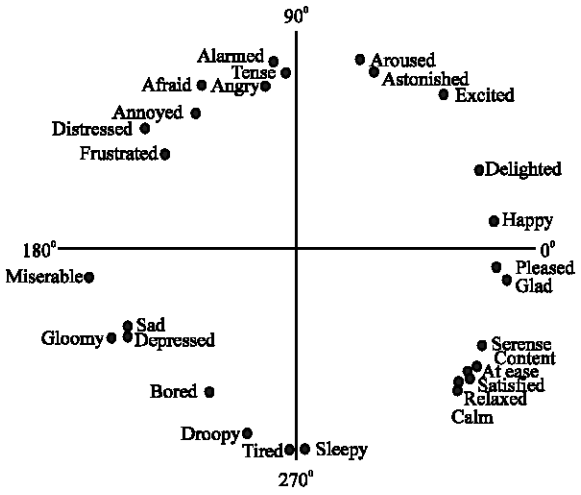

Fig. 1: Circumplex model of affect (Russell, 1983)

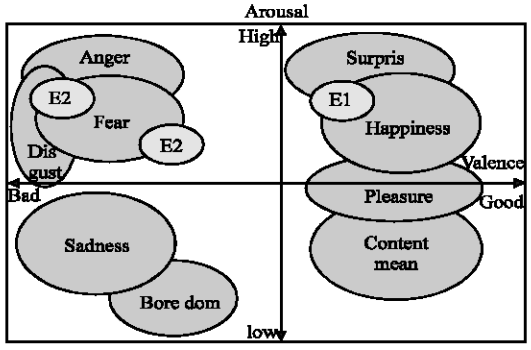

Fig. 2: Mixed emotions in a two-dimensional space (Peter and Antje, 2006)

Mixed emotions of a two-dimensional space: Defining the structure of emotion has been a troublesome task in the past and still is. A researcher (Antje, 2006) has put forward a new method: Mixed emotions.

From Fig. 2, we can see naming emotions with words is problematic. For example, just as the word angry can have very different meanings as well (such as, anger pulus anxiety plus depression when confronted with loss of important data). Adding to this that the category borders are blurry, we suggest to abandon labeling emotions with words. This study will use the content of 2.3 to define emotions.

BACKGROUND

Database of emotion corpus: The study aims to develop a methodology to model the emotion of speech coming from some speakers.

Have been selected several variables, including the genre of speech emotion to model, the number of sentence selections to be appraised, the duration of every speaker reading all sentences, the specific speaker to express emotion while reading every sentence and the sampling rate of the cursor in the two-dimensional space.

Performers need sampling rate of the cursor in two-dimensional when they express some emotion. Actually, the sampling rate should probably equal the time constants of the system (Liung, 1999).

Table 1: Sample sentences of emotion database

Emotion	Sentence
Anger	You are always late.
	Don't come here.
	You are not fair to me.
Disgust	I don=t want to wear this dress.
	I don't like this color.
	I don=t want to go.
Fear	Nobody accompanies me.
	The water is too deep.
	I am going to die.
Joy	Hey, you pass the exam.
	Your baby is so cute.
	I have succeeded. I won.
Sadness	I feed so sad.
	My little dog died.
	Life is meaningless, I want to die.
Surprise	Snake! Snake!
	He has woken up!
	Is it real?

Specifically, design an emotion database and set up for text-independent emotion classification studies (Scherer *et al*., 2001). The database includes short sentences covering the 6 basical emotions, namely anger, disgust, fear, joy, sadness and surprise.

A total of 10 native Mandarin language speakers (5 males and 5 females) are employed to generate 180 utterances. The recording is done in a quiet environment using a mouthpiece microphone.

To ensure that each speaker is able to express the emotion of every sentence, the duration of the session with each performer should be limited (Schubert, 1999). Thus, let performs express all emotion for a lot of sentences, where A (a = 1,2,.....A) is the total number of sentences selections. To be maximally informative, the *A* sentences selections need to differ and vary considerably (Table 1).

Characteristics of speech emotion: There are 2 broad types of information in speech, one is the semantic part, one the other hand, refers to the implicit messages such as the emotional state of the speaker. For speech emotion modeling, the identification of the speech features that represent the emotional state of the speaker is an important step.

Prosodic parameters have been found to represent the majority of emotional content in verbal communication (Murray and Arnott, 1993; Scherer, 2003). Of these, fundamental frequency (pitch), energy and speaking rate are widely observed to be the most significant characteristics (Battiner *et al*., 2003; Lee *et al*., 2004; Donn *et al*., 2006). The studying content of fundamental frequency and energy includes many parameters, look at the following Table 2.

Table 2 is describing in the previous research and the literature relating to automatic emotion detection from speech based on the acoustic correlates.

PROVIDED METHOD

Overview of the system: The study of speech emotion is a difficult problem, but the system is same almost as general pattern recognition process. In the model, the speech signal is sampled at 22.05 KHz and coded with 16 bits PCM. The signal samples are segmented into frames of 16 ms each with 9 ms overlap between consecutive frames. The typical values of fundamental frequency of speakers range from 100-200 Hz. The window size of duration 16 ms, covers approximately two periods of fundamental frequency (Cairns and Hamsen, 1994).

Model structure: Once the speech feature is collected, the model structure is ascertained. Each parameter in the model need defined. Described in detail the output of model using the following expression:

$$\hat{y}_a(t \mid \theta) = f(u_a(t), u_a(t-1), ..., e(t), e(t-1), ...) \quad (1)$$

Where:

$\hat{y}_a(t \mid \theta)$: Emotion output for utterance selection a;
$u_a(t)$: Feature vector for utterance selection a;
e(t) : 2-D white noise process with zero mean;
f() : Function representing the model of output structure;
θ : d -dimensional vector containing all of the parameters needed describing f().

Table 2: Characterisitics of specific emotions

No.	Speech property	Speech feature	Emotion
1	F_0 mean	Increased	Anger, joy, surprise, fear;
2		Normal	Surprise;
3		decreaased	Sadness, disgust, fear;
4	F_0 range	Wider	Anger, joy, surprise,disgust,fear;
5		Narrower	Sadness, disgust, fear;
6	Energy	Increased	Anger, joy;
7		Normal	Disgust, fear;
8		Decreased	Disgust, sadness;
9		-	Surprise;
10	Speaking rate	High	Anger, joy, disgust, fear; 1112
		Normal	Sadness, fear;
13		Low	Surprise;
14	Formants	F_1 mean increased	Anger, sadness, disgust, fear;
15		F_1 mean decreased	Joy;
16		F_2 mean higher	Anger;
17		F_2 mean lower	Anger, sadness, disgust, fear;
18		F_3mean higher	Anger;
19		F_1 bandwidth increased	Joy;
20		F_1 bandwidth decreased	Sadness, disgust, fear;
21		-	Surprise

Equation 2 is describing emotional output when speaker express utterance selection. Generally, while speaker express reading uncertainty utterances, equation can be described.

For this study, only think about one linear model, which is Auto-Regression with extra inputs (ARX) structure. Given m-dimensional input data u(t) and two-dimensional output data, y(t) the ARX model structure can be described using the following expression:

$$\begin{aligned} &y(t)+A_1(\theta)y(t-1)+...+A_{n_a}(\theta)y(t-n_a)\\ &=B_0(\theta)u(t)+...+B_{n_b}(\theta)u(t-n_b)+e(t) \end{aligned} \quad (2)$$

Where

y(t) : Vector consisting of valence and arousal at time t;
u(t) : Vector consisting of the features from Table 2;
$A_k(\theta)$ 2×2 matrix; $B_k(\theta)$ 2 × m Mmatrix;
e(t) : 2-D white noise process with zero mean;
N_a : Maximum number of auto-regressive terms in the model;
N_b : Maximum number of lagged inputs in the model;
θ : d-dimensional vector containing all of the nonzero elements of $A_k(\theta)$ and $B_k(\theta)$.

For Eq. 2, can be defined the following form:

$$\begin{aligned} &y(t)+A_1(\theta)\,y(t-1)+A_2(\theta)\,y(t-2)\\ &=B_0(\theta)\,u(t)+B_1(\theta)\,u(t-1)+B_2(\theta)\,u(t-2)+e(t) \end{aligned} \quad (3)$$

Verification validation: In this study, stipulating artificially the model structure is linear, namely Eq. 3 is linear. Assume the following is tenable:

$$\begin{aligned} &\hat{y}(t\,|\,\theta)=[I+A_1(\theta)q^{-1}+A_2(\theta)q^{-2}]^{-1}\\ &[B_0(\theta)+B_1(\theta)q^{-1}+B_2(\theta)q^{-2}]u(t) \end{aligned} \quad (4)$$

At the same time, the following equation is reasonable:

$$q^{-k}y(t)=y(t-k) \quad (5)$$

Now, simultaneous Eq. 4 and 5, then, y(t) n be expressed as follows:

$$\begin{aligned} y(t)&=[I+A_1(\theta)q^{-1}+A_2(\theta)q^{-2}]^{-1}[B_0(\theta)\\ &+B_1(\theta)q^{-1}+B_2(\theta)q^{-2}]u(t)\\ &+[I+A_1(\theta)q^{-1}+A_2(\theta)q^{-2}]^{-1}e(t) \end{aligned} \quad (6)$$

Which order,

$$\begin{aligned} \alpha=&[I+A_1(\theta)q^{-1}+A_2(\theta)q^{-2}]^{-1}*\\ &[B_0(\theta)+B_1(\theta)q^{-1}+B_2(\theta)q^{-2}] \end{aligned} \quad (7)$$

$$\beta=[I+A_1(\theta)q^{-1}+A_2(\theta)\,q^{-1}]^{-1} \quad (8)$$

Therefore, $y(t)=\alpha u(t)+\beta(t)$, the model is linear, the problem is proved.

RESULTANT EXPERIMENT

Emotional appraisals for eighteen basic utterances were measured using the dimensions valence and arousal. The utterances come from the corpus database (Table 1). The procedure of sampling is processing in a quiet enough environment. The three utterances of expressing every emotion is read continuously by read, but different emotions has the intervals of 9 ms. For every speaker, he can read representation of every emotion according of difference order, namely, don' = t need read the content of Table 1 from A to Z. In the model that above mentioned, the best result in the model using 16 of the 21 speech features and 38 parameters, as shown in the following:

$$A_1(\theta)=\begin{pmatrix}\theta_1 & \theta_2\\ 0 & \theta_3\end{pmatrix}\quad A_2(\theta)=\begin{pmatrix}\theta_4 & \theta_5\\ 0 & \theta_6\end{pmatrix} \quad (9)$$

$$B_0(\theta)=\begin{pmatrix}0&0&0&0&0&0&0&0&\theta_7&\theta_8&0&0&0&0&0&0&0&0&0\\ 0&0&\theta_9&\theta_{10}&\theta_{11}&0&\theta_{12}&0&0&0&\theta_{13}&0&0&\theta_{14}&\theta_{15}&\theta_{16}&\theta_{17}&\theta_{18}\end{pmatrix} \quad (10)$$

$$B_1(\theta)=\begin{pmatrix}0&\theta_{19}&\theta_{20}&\theta_{21}&0&0&\theta_{22}&0&\theta_{23}&0&0&0&0&0&0&0&0&0\\ \theta_{24}&0&\theta_{25}&\theta_{26}&\theta_{27}&0&\theta_{28}&0&0&0&\theta_{29}&\theta_{30}&\theta_{31}&0&0&0&0&\theta_{32}\end{pmatrix} \quad (11)$$

$$B_2(\theta)=\begin{pmatrix}0&0&0&\theta_{33}&0&0&0&0&\theta_{34}&0&0&0&0&0&0&0&0&0\\ \theta_{35}&0&\theta_{36}&0&0&0&0&0&0&\theta_{37}&0&0&0&0&0&\theta_{38}\end{pmatrix} \quad (12)$$

There are several comments to be made about the parameters in Eq. 4-12. Firstly, the parameters in matrices B_0, B_1 and B_2 correspond to the contribution of each input (the columns) to each output (the rows). Secondly, the structure of $A_1(\theta)$ and $A_2(\theta)$ indicates how previous emot-ional appraisals affect the current emotional appraisal.

For the model structure, the output of emotion about sadness is shown in Fig. 3.

For this experience, only display the emotion result of sadness, other emotions' results about average accuracy are summarized in Table 3 average accuracy.

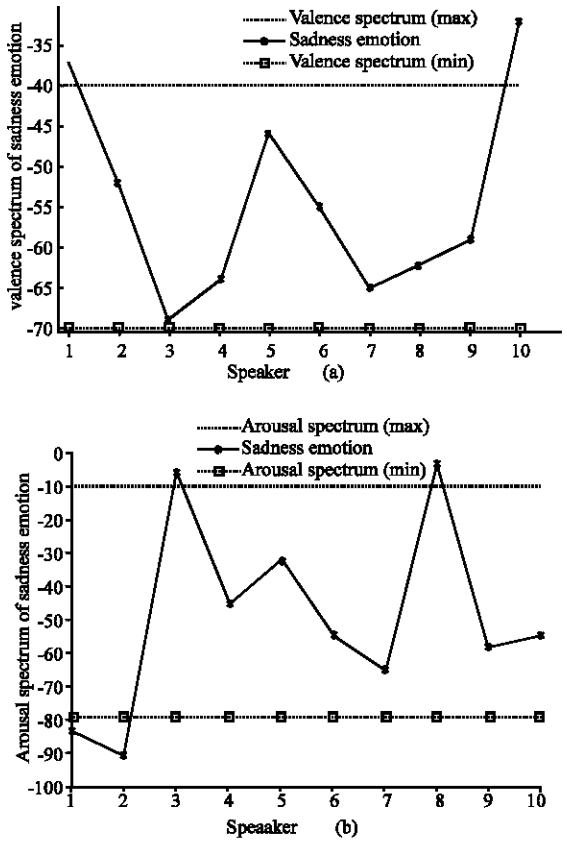

Fig. 3: The result of sadness emotionand = s (a) Valence and (b) Arousal

Table 3: Comparison with other methods

Approach accuracy(%)	Emotions classified	Average
ASSESS method (McGilloway *et al.*,2000)	Afraid, happiness, Neutral, sadness, anger	55
Patter recognition by Dellaert *et al.*(1996)	Happiness, sadness, Anger and fear	79.5
Neural network by Nicholson *et al.*(1999)	Joy, teasing, fear, sadness, Disgust, anger, surprise, neutral	50
Proposed method of this study	Anger, disgust, fear , joy, Sadness, surprise	77.4

Looking at Table 3, these other approaches include statistical patter recognition and neural net work classification, Note that the systems differ in speaker dependency, text dependency, the number and type of emotions classified and the size of the database used. Nevertheless, it provides a crude comparison of the different approaches.

CONCLUSION

In this study, a model of emotional state of utterances is proposed. Although the model is linear, from Table 3, the validity is already verified. At one time, the method of modeling is simpler than other ones, such as McGilloway *et al.* (2000), Dellaert *et al.* (1996) and Nicholson *et al.* (1999).

There are several suggested areas to investigate in future works. Firstly, the emotion corpus is limited, enlarging database is very important; secondly, modeling speech emotion has not been considered semantic, more features may be needed to create valid models for this research; Third, in the models, should use several different techniques to verify and should consider other model structures can get improved results.

REFERENCES

Abelson, R.P., 1963. Computer Simulation of "Hot" Cognition. In S.S. Tomkins and S. Messick (Eds.), Computer simulation of personality. New York: Wiley.

Batliner, A., K. Fischer, R. Huber, J. Spilker and E. Noth, 2003. How to find trouble in communication. Speech Commun., 40: 117-143.

Cairns, D.A. and J.H.L. Hansen, 1994. nonlinear analysis and classification of speech under stressed conditions, Acoust, Soc. Am., 96: 3392-3400.

Cañamero, D., 2001. Emotions and adaptation in autonomous agents: A design perspective. Cyber. Sys., 32: 507-529.

Christian Peter, 2006. Antje Herbon, Emotion representation and physiology assignments in digital systems. Interacting with Computers, 18: 139-170.

Dellaert, F., T. Polzin and A. Waibel, 1996. Recognizing Emotion in Speech. Fourth International Conference on Spoken Language Processing, 3: 1970-1973.

Donn Morrison, Ruili Wang, Liyanage and C. De Silva, 2006. Ensemble methods for spoken emotion recognition in call-centres, Speech Communication.

Dyer, M.G., 1987. Emotions and their computations: three computer models. Coginition and Emotion, 1: 323-347.

Frick, R., 1985.Communicating Emotion : The Role of Prosodic Features. Psycho. Bull., 97: 412-429.

Lee, C.M., S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee and S. Narayannan, 2004. Emotion recognition based on phoneme classes. Proceedings of the International Conference on Spoken Language Processing (ICSLP 2004), Jeju Island, Korea.

Ljung, L., 1999. System Identification: Theory for the user, 2nd Ed. Upper Saddle River, NJ: Prentice-Hall.

Mark, D., A. Korhonen, David and M. Clausi, 2006. Ed Jernigan, Modeling Emotional content of Music Using System Identification, System, Man and Cybernetics-Part B: Cybernetics, 36: 588-599.

McGilloway, S., R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk and S. Stroeve, 2000. Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark, ISCA Workshop on Speech and Emotion, Belfast.

Murray, I. And J. Arnott, 1993. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. Acoustical Soc. Am., 93: 1097-1108.

Nicholson, J., K. Takahashi and R. Nakatsu,1999. Emotion recognition in speech using neural networks, 6th International Conference on Neural Information Processing, ICONIP, 2: 495-501.

Ortony, A., G. Clore and A. Collins, 1994. The cognitive structure of emotions. Cambridge: Cambridge University press.

Picard, R.W., 1997.Affective computing, Cambridge, MA: MIT Press.

Rafael Pérez y Pérez, 2006. Employing emotions to drive plot generation in a computer-based storyteller. Cognitive Systems Research.

Russell, J.A., 1983. Pancultural aspects of the human conceptual organization of emotions. Personality Soc. Psychol., 45: 1281-1288.

Scherer, K.R., 2003. Vocal communication of emotion: A review of research paradigms, Speech Commun., 40: 227-256.

Scherer, K.R., R. Banse and H.G. Wallbott, Emotion inferences from vocal expression correlate across languages and cultures. Cross-Cultural Psychol., 32: 76-92.

Schubert. E., 1999. Measurement and time series analysis of emotion in music. Ph. D. dissertation, School of Music and Music Education, Univ., New South Wales, Sydney, Australia.

Tin Lay Nwe, 2003. Say Wei Foo, Liyanage C. De Silva, Speech emotion recognition using hidden Markov models. Speech Commun., 41: 603-623.

Wierzbicka, A., 1992. Talking about emotions: Semantics, culture and cognition. Cognition and Emotion, 6: 3-4.

Williams, C.E. and K.N. Stevens, 1981. Vocal Correlate of Emotional States. In: Darby, J.K. (Ed.), Speech Evaluation in Psychiatry. Grune and Stratton, Inc., pp: 189-220.