Research On Prosody Conversion of Affective Speech Based on LIBSVM and PAD Three Dimensional Emotion Model

Xiaoyong Lu[1, a], Tao Pan [2, b]

[1]College of Computer Science and Engineering, Northwest Normal University LanZhou 730070, China;

[2] LanZhou Resources&Environment Voc-Tech College, Lanzhou 730021, China.

[a]lu@nwnu.edu.cn, [b]pant@163.com

Keywords: PAD emotion model, five-scale tone model, Library for Support Vector Machines （LIBSVM）support vector regression, generalized regression neural network, Prosody Conversion.

Abstract. This paper proposes a framework for prosody conversion of emotional speech based on LIBSVM support vector regression model and PAD three dimensional emotion model. We design an emotional speech corpus including 11 kinds of emotional utterances. Each utterance is labeled the emotional information with PAD value. A five-scale tone model is employed to model the pitch contour of emotional speech at the syllable level. A LIBSVM SVR-based prosody conversion model is proposed to realize the transformation of pitch contour, duration and pause duration of emotional speech according to the PAD values of emotion and context information of text. Speech is then re-synthesized with the STRAIGHT algorithm by modifying pitch contour, duration and pause duration, and is compared with the results obtained by the generalized regression neural network. Experimental results show that the modified speech achieves 3.8 of average Emotional Mean Opining Score (EMOS).

Introduction

Emotion is an important part of the voice's naturalness and expressive. In the interactive process, it carries a wealth of information. In 1995, Professor Picard MIT of the United States of America proposed the concept of emotional computing, and published the first monograph "Affective computing"[1] in 1997.Speech signals have two channels conveying semantic information: one for linguistics, and the other for extra-linguistics, such as emotion. For harmonic Human-Machine Interaction, it is very important to consider these two channels synchronously[2].

Currently emotional speech synthesis mainly adopts speech synthesis methods based on Hidden Markov Model (HMM) statistical parameter[3] and large-corpus based concatenation[4]. Although the former can use the method of speaker adaptation transform[5][6] to realize emotional speech synthesis, the quality of statistical parameter speech synthesis is difficult accepted by users. Though speech synthesis by large-corpus based concatenation method can achieve high quality speech synthesis, it is very difficult to record different emotional corpus. Therefore, some studies propose a method to realize emotional speech synthesis through using the relationship between the emotional features and the acoustic characteristics of speech combining text analysis and emotional state. The paper [6] and paper [7] introduce different emotional psychological model to high expression speech synthesis. The paper [8] achieves emotional speech conversion by using PAD three dimensional emotion model and paper [9] uses SVR predicting emotional prosody parameter. But the all of these studies lack the modeling of fundamental frequency contour.

In order to convert F0 envelop in emotional speech conversion, the researcher structures 11 kinds of typical emotional text corpus, records relative voice corpus, labeling the PAD of voice corpus with psychological method, builds syllabic F0 model with Five-scale Tone Model[10], and constructs the predicting model of emotional speech prosody parameter with Support Vector Regression algorithms (LIBSVM) [11]. According to the statement of the PAD value and contextual features predict target emotional speech prosody parameters. At the same time, using LIBSVM support vector regression predicts the same output parameters and compares it. Finally, using STRAIGHT[12] algorithm

achieves emotional voice conversion. The results show that after converting voice can better perform the relative emotion by the text.

PAD three dimensional emotion model

The main methods of emotional description include discrete areas of emotion representation and the emotional dimension of the continuous change dimension representation. Category description method cannot indicate the relative relationship and changes among emotions and it is difficult to describe the mixed emotions. Therefore, this paper adopts PAD three dimensional emotion model[13] to describe emotional speech so that can expand the research of emotional speech to quantify emotional computational research.

Corpus construction and PAD evaluation

Text corpus. The content of text corpus not only considers certain length, but also has abundant emotion characteristics. In the project of text corpus construction, the researcher adopts the method that sentences with no emotional embeds into 11 kinds of typical emotional segments. In this way, it will stimulate the need of emotional characteristics easily than isolated sentences. In the experiment, the researcher designs 10 emotional section based on specific situation for each emotion, and each emotional segment embeds in a non-emotional bias statement, then forms 110 different segments. In the choosing of non-emotional bias statements, the researcher takes the length combination method. Five to six statements are longer which have 150 syllables, and four to five statements are shorter which have 50 syllables. In this way, there are 2200 syllables.

Speech corpus. In the process of recording, the researcher chooses a female mandarin recorder who is not a professional actress to record in the recording studio. At first it requires to record neutral speech and then emotional speech. When recording neutral speech, the recorder is required to use a poker-face and unchanged tone and speed to read text. When recording other 10 kinds of emotional speech, the first step is to set a specific scene to stimulate recorder's relative emotion, and then to read text and record. For example, with recording sad emotional voice, by watching sad movie segments and pictures stimulate recorder's sad emotion. After simulating recorder's emotion, the recorder is required to explore natural emotional expression to read 10 emotional segments. The final recorded voice file with WAV form single track after 16k Hz sampling and 16 bit quantization.

PAD evaluation. After recording speech corpus, the researcher adopts emotional scale of Chinese edition[14], which is improved by the psychological center, Chinese Academy of Sciences, to evaluate the PAD score of recording speech corpus. Seeing from table I, the recording emotional voice can express the selected 11 kinds of emotion basically.

Conversion framework of emotional speech prosody based on Support Vector Regression

Conversion framework. This paper proposes emotional voice prosody conversion system based on SVR (see Fig. 1). It comprises two parts of training and conversion.

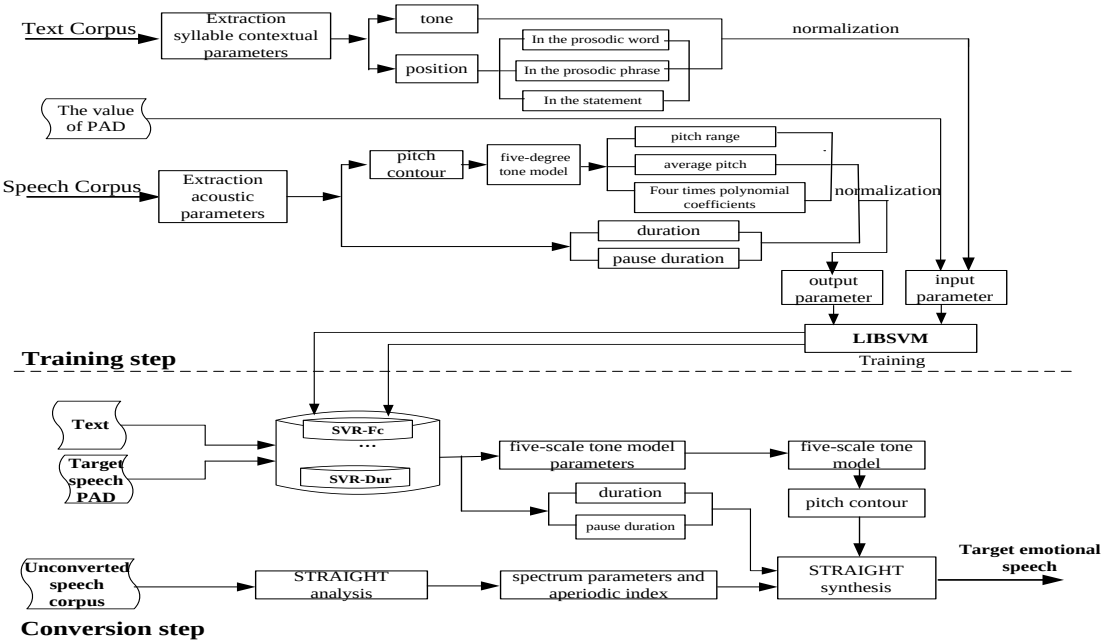

Fig. 1 Framework of emotional speech prosody's conversion

In the process of training, the first step is to extract syllabic context characteristic parameter from text corpus, and then making the PAD score which is from each syllabic context parameter and evaluation is as the condition property of LIBSVM. Meanwhile, F0 contours, duration and pause duration are extracted from speech corpus. F0 contours is modeled on five-scale tone model with five-scale tone model, duration and pause duration being the input parameter. Inputting and outputting parameters in a normalized way, obtaining nine regression models by LIBSVM, leading different input parameter to corresponding output.

Firstly, in the conversion stage, on the basis of test corpus get contextual parameters to be converted speech syllable and PAD value target speech as the input of LIBSVM. Using 9 SVR models which get from the processing of training, respectively predict five-scale tone model parameters, duration and pause duration of target emotional speech's syllable. And using five-scale tone model generates the syllabic pitch envelop of target emotional speech. Meanwhile, STRAIGHT is used to obtain speech spectrum parameters and aperiodic index to switch speech. Finally, by using the generated F0 envelop, predicted duration and pause duration, and getting frequency spectrum parameters and aperiodic key from analyzing STRAIGHT can synthesize target emotional speech.

Model pitch contour with five-scale tone model. Although pitch values and pitch ranges vary with gender and age, the shape of pitch contour is stable for a certain tone within a Mandarin. We use the five-scale tone model to depict the shape of pitch contour for four kinds of tone of Mandarin. The five-scale tone model fits a pitch contour with a 4th order polynomial in log domain by eliminating the influence of the pitch range and average pitch, as shown in Equation (1).

$$F_{0i}(t) = \log^{-1}\left[f_c + f_d \bullet f_{0i}(t) \right] \tag{1}$$

$$f_{0i}(t) = a + bt + ct^2 + dt^3 + et^4 \tag{2}$$

Where $\log^{-1}(\bullet)$ denotes the inverse log operation with base 10, t is the normalized time ranging $[0...1]$. f_c is the median fundamental frequency of speakers in logarithm domain that reflect the pitch level of a speaker. f_d is the range of fundamental frequency in logarithm domain. f_{0i} is a quadratic curve as shown in Equation (2) denoting the shape of pitch contour for four tones of Mandarin dialect, and i is the ith tone ranging $[1...4]$ □ $F_{0i}(t)$ is the generated pitch contour of tone i.

The original F0 value is used to obtain F0 mean value and f0 range. Then, the time of normalization will be worked out with syllable F0 points. Four polynomial coefficient based on syllables is intended to obtain through Equation (2). Lastly, syllable F0 contours based on five-scale tone model is obtained during the normalized time through Equation (1).

Prediction model based on SVR. Recently, more robust semi-parametric methods like SVR have been successfully applied to the prediction of WS and other time series. SVR, an extension of Support Vector Machines (SVM), was proposed by [15]. SVR pursues the best trade-off between the model's Empirical Error and the model complexity. This compromise is achieved by constraining SVR regression function $f(\bullet)$ to the hyperplanes function class, and employing a margin, also called insensitive tube, around the hyperplane. Moreover, $f(\bullet)$ only depends on a reduced set of the training data called the Support Vectors (SV), those which correspond to the active constraints in the optimization problem.

Formally, given a data set of the form $(x_i, y_i) \in R^N \times R$, the SVR dual optimization problem is formulated as:

$$\max imizeW(\alpha_i, \alpha^*) = \frac{1}{2}\sum_{i,j=1}^{m}(\alpha_i - \alpha_i^*)\ (\alpha_j - \alpha_j^*)\left\langle \phi(x_i), \phi(x_j) \right\rangle - \varepsilon\sum_{i=1}^{m}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{m}(\alpha_i^* + \alpha_i^*)$$

$$\text{Subject to} \sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0 \qquad (3)$$

$$0 \le \alpha_i \le C, \forall i = 1, \ldots, m$$

$$0 \le \alpha_i^* \le C, \forall i = 1, \ldots, m$$

where C is the complexity penalization term, and α, α^* correspond to the dual variables for the active constraints[16].

Evenmore, SVR is able to perform a non-linear regression due the kernel trick. Colloquially, it consist in providing SVR with a specific kernel function which maps data from the input space to a high dimensional feature space where a linear regression is performed. Once Eq. (3) is solved and the hyperplane function found, a future value can be predicted employing Eq. (4).

$$f\left(X, \alpha, \alpha^*\right) = \sum_{i=1}^{S}\left(\alpha_i - \alpha_i^*\right)\kappa(X_i, X) + b. \qquad (4)$$

Input and output parameters of SVR model. Author extract PAD value of every sentence, tone information of every syllable, the position of every syllables in the prosodic word, the position of every syllable in the prosodic phrase, the position of the syllables in the sentence as input parameters of LIBSVM model, and extract f0 mean value(fc), f0 range value(fd), parameters of f0 counter based on five-scale tone model, duration, silent duration as output parameters of LIBSVM model.

EXPERIMENTAL RESULTS

Performance evaluation of five-scale tone model. In order to structure LIBSVM emotional speech prosody conversion model efficiently, the researcher use five-scale tone model to convert all emotional syllabic pitch contour and to establish each F0 model at the syllable level before training LIBSVM model. In order to test five-scale tone model's modeling performance to F0 envelop of syllable, this paper calculates root-mean-square error between pitch contour and original pitch contour of five-scale tone model (see Table 1).

$$\sigma = \left(\frac{\sum d_i^2}{n}\right)^{1/2} \qquad (5)$$

In Eq. (5), d_i means the deviation of measured value and mean value, and $i = 1, 2, 3, \ldots n$.

Table 1 the RMSE value of five-scale tone model

Emotional type	Mean RMSE(Hz)
Neutral	3.0766
Relax	2.7608
Surprise	2.7692
Gentlenss	3.7665
Joy	3.7163
Contempt	6.2576
Disgust	4.6085
Fear	6.8551
Sad	5.0393
Anxiety	3.3261
Anger	6.3221

In order to test the five-scale tone model's performance, we calculates the RMSE between model predicted pitch contour and original pitch contour as shown in Table 1. We can see that the maximum RMSE error is 6.9 Hz. Therefore, the five-scale tone model can satisfy the requirement of modeling pitch contour.

Selecting Parameters of SVR. Obtaining optimal model parameter is to achieve optimal regression prediction. The experiments are all based on LIBSVM work-box. The researcher adopts the way of 5 repeated overlapping to obtain separately each model's optimal punishment parameter and optimal nuclear parameter. It will make model reach high-point. The parameter will be showed in Table 2.

Table 2 Optimal parameters of SVR

Model Category	Optimal Parameters		
	c	g	p
Fc	256.0	0.0313	0.0313
Fd	0.5	0.5	0.0156
a	0.5	8.0	0.25
b	2.0	0.125	1.0
c	4.0	1.0	4.0
d	8.0	0.5	2.0
e	4.0	0.5	8.0
dur	1024.0	0.0313	0.0078
sil	1024.0	0.0625	0.0010

Performance evaluation of SVR prediction model. According to the conversion system of the model in this paper, 4/5 experimental corpus are used to LIBSVM model training. After testing more times and cross validation to this model, the researcher gains the optimal parameter of this model. Also the model reaches the optimal state. In this foundation, the researcher regards the rest of 1/5 experimental corpus as testing data, that uses SVR model in the process of training, and then obtains five-scale tone model's parameters of the corresponding emotional statement, duration and pause duration, and also uses five-scale tone model to generate F0 envelop of syllable. In order to prove the performance of SVR, the researcher takes correlation analysis to predicted value and original value of SVR and GRNN models (see Table 3).

Table 3 the mean value of 11 kinds of emotions' correlation coefficient

Emotional type	Mean R	
	SVR	GRNN
Neutral	0.8606	0.7166
Relax	0.8674	0.6625
Surprise	0.5763	0.3958
Gentlenss	0.7129	0.6208
Joy	0.7407	0.5607
Contempt	0.7067	0.3577

Disgust	0.8612	0.5733
Fear	0.5047	0.5004
Sad	0.5375	0.3453
Anxiety	0.8450	0.7355
Anger	0.7561	0.3184

In Table 3, R means correlation coefficient and R=1 means completely correlate. In SVR model, the highest mean correlation value is relax emotion, and the lowest value is fear which is 0.5047, but those all higher than GRNN model's mean correlation value. Therefore, compared with GRNN model, all forecasting features reach a better correlation under SVR model.

Subjective evaluation of the result of the conversion. The researcher explores EMOS to make a subjective evaluation of emotional speech after converting. EMOS evaluation method is mainly focused on the evaluation of the degree of emotional expression, and it uses 5 level rating standards to evaluate the similarity between voice transformation and emotional expression in terms of the original speech. The EMOS evaluation will be divided speech quality into 5 grades, that is, excellent, good, middle, poor and bad and each grade is given value of 5 points, 4 points, 3points, 2 points and 1 point. In the experiment, the researcher chooses 10 students (they are undergraduates, 5 girls and 5 boys) who are never know EMOS. Each student selects randomly 110 sentences to score from 440 experimental conversion sentences. In the evaluation, the first step is to play original emotional speech as a natural voice standard which EMOS score is 5 points, and then EMOS will be used to score the emotional similarity of being tested speech. Finally, after listening scoring results of the evaluation of voice average calculates the final EMOS score, and also calculates its 95% confidence interval, as shown in Figure 2.

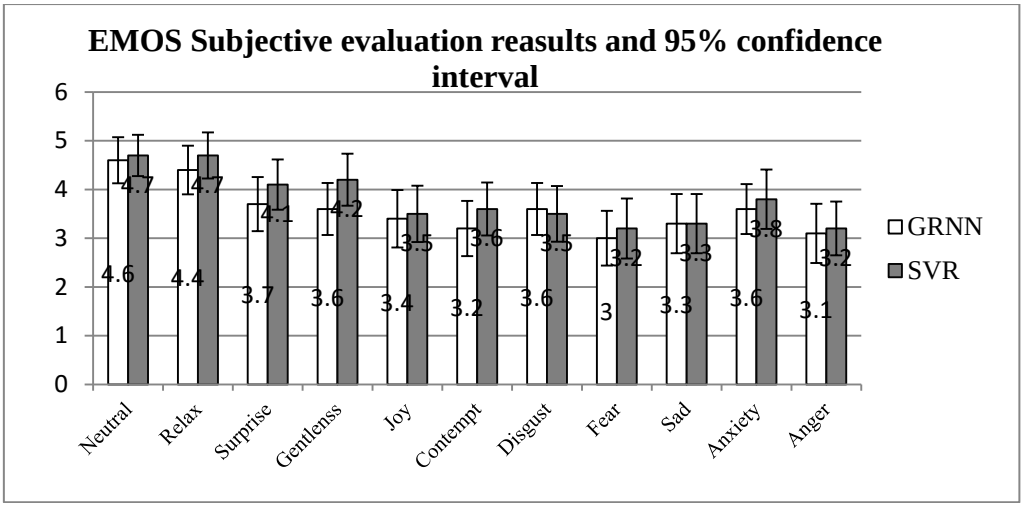

Fig. 2 EMOS Subjective evaluation results and 95% Confidence interval

From Figure 2 similar to the typical emotional manifestation of fear, anger and sadness, it not just is in its prosodic features, but pays more attention to the voice, facial expression and psychological aspects reflect. Using some acoustic prosodic features of speech is not able to more fully reflect its emotional components. Therefore, it is caused lower EMOS value of some typical emotions.

Conclusions

The researcher establishes a prosodic model of one person's different emotional speech conversion with LIBSVM and finally gets related emotional speech by using PAD three dimensional emotion model and five-scale tone model. The results show that the conversion method in this paper is a practical way under the circumstance of less corpus data. However, because emotional voice is not only presented in the change of F0, so it also needs to add other voice parameters to deeply analyze. The further research will apply PAD three dimensional emotional model to statistical parameter voice synthesis, and achieve emotional voice synthesis of HMM.

Acknowledgements

The research leading to these results was partly funded by Science and Technology Fund of Gansu for Young Scholars (Grant No. 1506RJYA126) and Youth Teacher Scientific Capability Promoting Project of Northwest Normal University (No. NWNU-LKQN-13-23).

References

[1] Picard R W, Picard R. Affective computing[M]. Cambridge: MIT press, 1997.

[2] Han Wenjing, Li Haifeng, Ruan Huabin, et al.l.Review on Speech Emotion Recognition [J]. Journal of Software, 2014, 25(1):37-50. (In Chinese).

[3] Heiga Zen, Keiichi Tokuda, and Alan W. Black. Statistical parametric speech synthesis [J].Speech Communication, 51(11):1039-1064,2009. (In English)

[4] Cai Lianhong, Cui Dandan, Cai Rui.TH-CoSS, a Mandarin Speech Corpus for TTS[J]. JOURNAL OF CHINESE INFORMATION PROCESSING, 2007,21 (2): 94-99. (In Chinese)

[5] T. Nose, J. Yamagishi, T. Masuko, T. Kobayashi.A style control technique for HMM-based expressive speech synthesis[J].IEICE Trans. Inf. & Syst., vol.E90-D, no.9, pp.1406-1413, Sept. 2007. (In English)

[6] Xu Jun, Cai Lianhong.Hierarchical prosody analysis and modeling for emotional conversions[J]. Journal of Tsinghua University(Natural Science),2009,49(S1): 1274-1277. (In Chinese)

[7] Hongwu Yang, Helen M. Meng, Lianhong Cai.Modeling the acoustic correlates of expressive elements in text genres for expressive text-to-speech synthesis[C].In INTERSPEECH, 2006,pp.1806-1809. (In English)

[8] Zhiyong Wu,Helen M. Meng,Hongwu Yang,Lianhong Cai.Modeling the Expressivity of Input Text Semantics for Chinese Text-to-Speech Synthesis in a Spoken Dialog System[J].IEEE Transaction on Audio, Speech, and Language Processing, 2009,17(8):1567-1577.(In English)

[9] Cui Dandan. Analysis and Conversion for Affective Speech [D].Beijing:Tsinghua University, 2007. (In Chinese)

[10] Weitong Guo,Hongwu Yang,Dong Pei,Qingqing Liang.Prosody Conversion of Chinese Northwest Mandarin Dialect based on Five Degree Tone Model[J].JDCTA: International Journal of Digital Content Technology and its Applications, Vol. 6, No. 17, pp. 323 ~ 332, 2012. (In English)

[11] Chih-Chung Chang,Chih-Jen Lin. LIBSVM : a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3), 27:1-27:27. (In English)

[12] H. Kawahara, I. Masuda Katsuse, and A. de Cheveigne. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds [J]. Speech Communication, Vol. 27, No. 3-4, pp. 187-207, 1999. (In English)

[13] Mehrabian, A.Correlations of the PAD Emotion Scales with self-reported satisfaction in marriage and work[J]. Genet Soc Gen Psychol Monogr, 1998,124(3):311-34. (In English)

[14] Xiaoming Li, Haotian Zhou.The Reliability and Validity of the Chinese Version of Abbreviated PAD Emotion Scales[J].Affective Computing and Intelligent Interaction,2005,3784(1),513-518. (In English)

[15] H. Drucker, C. Burges, L. Kaufman, A. Smola, V. Vapnik, Support vector regression machines[J].Adv. Neural Inf. Process. Syst. 1997,9 :155-161. (In English)

[16] B. Sch€olkopf, A. Smola,.Learning with Kernels[M]. The MIT Press, 2002. (In English)