Journal of Southeast University (English Edition) Vol. 27, No. 4, pp. 357 – 360 Dec. 2011 ISSN 1003—7985

Emotional speaker recognition based on prosody transformation

Song Peng[1] Zhao Li[1] Zou Cairong[1,2]

([1] Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing 210096, China)
([2] Foshan University, Foshan 528000, China)

Abstract: A novel emotional speaker recognition system (ESRS) is proposed to compensate for emotion variability. First, the emotion recognition is adopted as a pre-processing part to classify the neutral and emotional speech. Then, the recognized emotion speech is adjusted by prosody modification. Different methods including Gaussian normalization, the Gaussian mixture model (GMM) and support vector regression (SVR) are adopted to define the mapping rules of F0s between emotional and neutral speech, and the average linear ratio is used for the duration modification. Finally, the modified emotional speech is employed for the speaker recognition. The experimental results show that the proposed ESRS can significantly improve the performance of emotional speaker recognition, and the identification rate (IR) is higher than that of the traditional recognition system. The emotional speech with F0 and duration modifications is closer to the neutral one.

Key words: emotion recognition; speaker recognition; F0 transformation; duration modification

doi: 10.3969/j.issn.1003 – 7985.2011.04.002

Speaker recognition refers to automatically identifying a speaker's identity based on his or her voice. Great progress has been made in the past several decades, but there are still many problems that need to be resolved, such as environmental noise, channel effects and emotion variability. In this paper, the emotion variability which means different emotional states between training and testing processes are analyzed and compensated for. Several methods have been presented to resolve it. A structured method which aims at making the system be robust to the emotional variations in the speaker's voice in the training phase was proposed in Ref. [1], where the registered speakers are asked to provide all kinds of emotional states. However, when only the neutral speech is provided for training, it is difficult to obtain the emotional features and train the emotional model. A neutral-to-emotional GMM transformation method[2] was presented to train the emotional GMM model from his or her neutral speech, which assumes that if two speakers' neutral speech satisfies similar distributions, so does their emotional speech. However, when a new emotional utterance is added to the dataset, it needs to retrain the emotional model. Tao et al.[3] discussed all kinds of prosody conversion methods which aim at synthesizing emotional speech. Wu et al.[4] investigated rules based feature modification for emotional speaker recognition, and achieved more promising results than traditional speaker recognition systems.

Based on these approaches, a novel emotional speaker recognition system is presented in this paper. The transformation of prosodic parameters such as F0s and durations are investigated, which are adapted to the neutral ones according to the mapping rules. The speaker models are trained from only the neutral utterances of the registered speakers, and the unknown emotional states are recognized by an emotion recognition process. Several evaluation experiments are conducted, and the results show that the proposed ESRS can efficiently improve the IR compared with the baseline speaker recognition system.

1 Framework of Emotional Speaker Recognition

1.1 Baseline GMM-UBM framework

The Gaussian mixture model (GMM) is one of the mainstream methods of speaker recognition. A GMM denoted as λ is composed by M component densities, given as

$$p(\boldsymbol{X}\mid\lambda) = \sum_{i=1}^{M}\omega_i b_i(\boldsymbol{X}) \tag{1}$$

where $\boldsymbol{X}$ is a D-dimensional feature vector; M is the number of GMM components; ω_i is the prior probability of the i-th component; and

$$b_i(\boldsymbol{X}) = \frac{1}{(2\pi)^{D/2}\left|\boldsymbol{\Sigma}_i\right|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_i)'\sum_i^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_i)\right\} \tag{2}$$

For a group of S speakers, each speaker is modeled by a GMM depicted by his or her models $\lambda_1, \lambda_2, \ldots, \lambda_S$. For a given unknown speech sequence, the maximum posterior probability $p(\lambda_i\mid\boldsymbol{X})$ indicates that it belongs to the speaker model λ_i.

In the GMM based speaker recognition, a universal background model (UBM) is first trained using the expectation maximization (EM) method from a large number of utterances of different speakers[5], aiming at improving the adaptation abilities of the acoustic model to a different environment. When a new speaker is enrolled, the model of the new speaker λ_{MAP} is adapted from the UBM λ_{UBM} using the maximum a posteriori (MAP) method. In the recognition mode, the two models are coupled and the recognition method is often referred to as GMM-UBM. For an unknown

Received 2011-06-22.

Biographies: Song Peng (1983—), male, graduate; Zhao Li (corresponding author), male, doctor, professor, zhaoli@seu.edu.cn.

Foundation items: The National Natural Science Foundation of China (No. 60872073, 60975017, 51075068), the Natural Science Foundation of Guangdong Province (No. 10252800001000001), the Natural Science Foundation of Jiangsu Province (No. BK2010546).

Citation: Song Peng, Zhao Li, Zou Cairong. Emotional speaker recognition based on prosody transformation[J]. Journal of Southeast University (English Edition), 2011, 27(4): 357 – 360. [doi: 10.3969/j.issn.1003 – 7985.2011.04.002]

utterance, the recognition result is determined by the log-likelihood ratio $R(\boldsymbol{X})$ defined as

$$R(\boldsymbol{X}) = \log p(\lambda_{\text{MAP}} \mid \boldsymbol{X}) - \log p(\lambda_{\text{UBM}} \mid \boldsymbol{X}) \quad (3)$$

1.2 Emotional speaker recognition system

Fig. 1 gives the flowchart of the proposed ESRS, which is divided into three modules: emotion recognition, prosody transformation and speaker recognition.

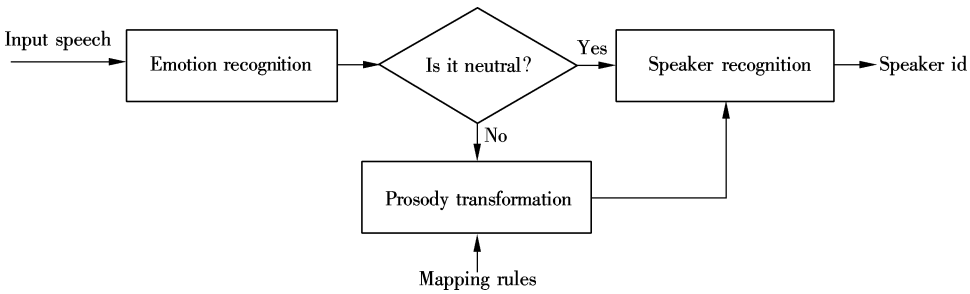

Fig. 1 Flowchart of ESRS

In the proposed ESRS, the first step is emotion recognition. The support vector machine (SVM) has been proved to outperform many other generative classifiers. A novel GMM supervector based SVM method is adopted to do the emotion recognition[6-7]. Then, since the speaker models are trained from the neutral speech, the mappings between emotional and neutral speech are necessary. In the system, the prosodic features of the emotional speech are converted to those of the neutral one according to the mapping rules. Finally, the converted emotional speech is used as the input of the GMM-UBM based speaker recognition module, and the recognition decision is made by the matching score on each speaker's model.

2 Prosody Transformation Algorithms

2.1 F0 transformation

One of the main motivations of using prosody transformation is that the correlations exist between source excitation and spectral envelope. By reducing the F0 differences between the speech of children and adults, the recognition rate of the speech recognition system can be greatly improved[8]. Another reason is that the prosodic features such as F0s and durations vary greatly in different emotional states. For instance, Tao et al.[3] added prosody modifications to synthesize the expressive emotional speech, which resulted in better performance.

2.1.1 Gaussian normalization

The common method of F0 transformation is known as normalizing the means and variances, which is based on the assumption that the F0s of source and target speakers have Gaussian distributions. Denoting the F0s of source and target speakers by f_x and f_y, respectively, the conversion function is given by

$$F(f_x) = \mu_y + \frac{\sigma_y}{\sigma_x}(f_x - \mu_x) \quad (4)$$

where μ_x and μ_y denote the mean, and σ_x and σ_y denote the covariance for f_x and f_y, respectively. This method is often used as a reference method for its simplification and good performance; however, it globally transforms the features ignoring the relationships between source and target speakers[9].

2.1.2 GMM transformation

A GMM transformation method is adopted for the F0 transformation. The aligned F0s of source and target speakers are jointly modeled by a GMM, and the converted F0 is obtained as

$$F(f_x) = E(f_y \mid f_x) = \sum_{i=1}^{M} p_i(f_x)\left[\mu_i^y + \frac{\Sigma_i^{yx}}{\Sigma_i^{xx}}(f_x - \mu_i^x)\right] \quad (5)$$

$$p_i(f_x) = \frac{\omega_i N(f_x, \mu_i^x, \Sigma_i^{xx})}{\sum_{k=1}^{M} \omega_k N(f_x, \mu_k^x, \Sigma_k^{xx})} \quad (6)$$

where $\boldsymbol{\mu}_i = \begin{bmatrix}\mu_i^x \\ \mu_i^y\end{bmatrix}$ and $\boldsymbol{\Sigma}_i = \begin{bmatrix}\Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy}\end{bmatrix}$, and $p_i(f_x)$ is the probability of f_x belonging to the i-th component of the GMM.

2.1.3 SVR transformation

Unlike traditional statistical methods, SVR performs a nonlinear mapping between source and target. Based on the theory of structural risk minimization, it can obtain better prediction using a small training dataset. By introducing the kernel function, it can make a nonlinear function become linear in a higher space and overcome the over-fitting problem[10]. A typical mapping function takes the form

$$F(f_x) = w\varphi(f_x) + b \quad (7)$$

where w and b are regressors; $\varphi(f_x)$ is a nonlinear mapping function to a higher dimensional space. Introducing the Lagrange function and the dual problem, the function (7) can be modified as

$$F(f_x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(f_x, f_x^j) + b \quad (8)$$

where $\alpha_i \geqslant 0$, $\alpha_i^* \geqslant 0$ are Lagrange multipliers; n is the frame number; f_x^j is the j-th frame of F0; and $K(f_x, f_x^j)$ is the kernel function. The radial basis function (RBF) kernel is adopted in this paper, and it is given as

$$K(f_x^i, f_x^j) = \exp\left[-\frac{1}{2}\left(\frac{\| f_x^i - f_x^j \|}{\sigma}\right)^2\right] \quad (9)$$

2.2 Duration modification

An average linear ratio is used for the duration modification, so the change of duration can be written as

$$\hat{d}_n = d_e \frac{\bar{d}_n}{\bar{d}_e} \tag{10}$$

where d_e and $\hat{d}_n$ are the lengths of frames of emotional and converted speech, respectively; $\bar{d}_n$ and $\bar{d}_e$ are the average lengths of neutral and emotional frames, respectively. The leading and trailing silences are excluded from the utterances.

A splicing/adding strategy is adopted. The excessive frames in each utterance are reduced, while the appropriate frames are repeated. The time domain pitch synchronous overlap-add (TD-PSOLA) technique is employed to manipulate the durations of emotional speech to map those of neutral speech.

3 Evaluation Experiments

Several experiments were designed to evaluate the performance of the presented ESRS. The spectral parameters were 13-order Mel-frequency cepstral coefficients (MFCCs) together with their delta coefficients, and the log domain F0s were used for analysis and transformation. The speech was analyzed at a 20 ms frame length with a 10 ms overlap, and the number of components of GMM and UBM was optimized as 64. Three types of strategies were compared: the baseline GMM-UBM method, the ESRS using emotion recognition, and the ESRS with prior known emotion states. The IR was employed to evaluate the whole performance of the ESRS. In addition, the accuracy of the GMM supervector based SVM method was used to assess the performance of the emotion recognition, and, in order to evaluate the quality of the converted emotional speech, an ABX test with 20 experienced listeners was performed to compare different types of prosody transformation algorithms.

3.1 Dataset description

An emotion dataset was established for the experiments, which included four types of emotional styles (happiness, sadness, fear and anger) and a neutral one. 40 sentences with no apparent emotional tendency were provided; 10 (5 male and 5 female) graduates in our laboratory were hired to speak the sentences in five types of speaking styles, which were recorded at an 11.025 kHz sampling rate with 16-bit precision, and the durations were around 2 to 5 s. The utterances with ambiguous emotions were asked to repeat the recording. Finally, 2 000 utterances (1 000 for training and the rest for testing) were collected as the corpus.

3.2 Results and discussion

As can be seen from Tab. 1, the GMM supervector based SVM emotion recognition method can obtain a satisfying result. It is obvious that the neutral speech obtains the highest recognition accuracy, while the fearful one obtains the worst one.

The IR results of the proposed processing algorithms are shown in Tab. 2. The symbols "case 1", "case 2", and "case 3" stand for the proposed ESRS with different F0 transformation methods based on Gaussian normalization, GMM and SVR, respectively; "case 4" corresponds to the baseline GMM-UBM method with prior known emotion. In different strategies, the higher the IR is, the better the performance. It can be found that the IRs of case 1 to case 3 which adopt the prosody transformation algorithms are higher than that of the baseline GMM-UBM method. Among the strategies using emotion recognition and prosody transformation, case 3 achieves the best performance, case 1 the worst. Another phenomenon can be found that different strategies obtain similar results for neutral speech, and case 4 has better performance than case 1 to case 3. The main reason is that the wrong recognized emotion state in emotion recognition may degrade the whole performance of emotional speaker recognition.

Tab. 1 Accuracy of emotion recognition

Emotion type	Accuracy/%
Happiness	81.3
Sadness	87.2
Fear	80.3
Anger	89.5
Neutrality	96.1

Tab. 2 IR results for speaker recognition %

Methods	Happiness	Sadness	Fear	Anger	Neutrality	Average
Baseline	39.2	58.7	45.3	41.2	98.5	56.6
Case 1	50.6	65.1	53.5	43.8	97.4	62.1
Case 2	51.5	69.2	58.6	45.7	98.1	64.6
Case 3	52.3	70.1	60.1	46.2	98.2	65.4
Case 4	53.6	72.2	62.9	48.3	98.5	67.1

In order to evaluate the similarity performance of different prosody transformation methods, an ABX test was performed to judge whether X was closer to A or B. Here X was the target neutral speech, and A and B were converted emotional speech using SVR or other methods. Fig. 2 summarizes the preference results of all kinds of transformation methods. The symbols "SVR", "Gaussian normalization" and "GMM" denote F0 transformation based on SVR, Gaussian normalization and GMM involving duration modification, respectively, and the "duration only" corresponds to the emotion transformation with only duration modification. It is obvious that the converted emotional speech using the SVR method is closest to the target neutral one. Furthermore, it is found that the performance of only duration transformation is not too bad compared with the prosody transformation based on the SVR, so this introduces a question: What is important in emotion transformation?

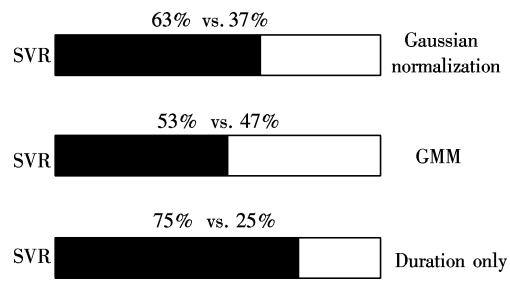

Fig. 2 Results for similarity performance

4 Conclusion

A novel emotional speaker recognition system is proposed in this paper. The emotion recognition and emotional-to-neutral state transformation are employed to enhance the emotional speaker recognition performance. Experimental results show that the presented system using prosody transformation can greatly improve the IR compared with the baseline system, and the SVR based transformation method obtains better performance than the traditional methods. Meanwhile, the performance of the proposed system is more or less influenced by the errors of emotion recognition. Further research will focus on this issue. Maybe adding an error correction module or using other solutions instead of emotion recognition will obtain better results.

References

[1] Scherer K L, Johnstone T, Klasmeyer G, et al. Can automatic speaker verification be improved by training the algorithms on emotional speech? [C]//*International Conference on Spoken Language Processing*. Beijing, China, 2000: 807 - 810.

[2] Shan Z Y, Yang Y C, Ye R Z. Natural-emotion GMM transformation algorithm for emotional speaker recognition [C]//8*th Annual Conference of the International Speech Communication Association*. Antwerp, Belgium, 2007: 782 - 785.

[3] Tao J H, Kang Y G, Li A J. Prosody conversion from neutral speech to emotional speech [J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006, **14** (4): 1145 - 1154.

[4] Wu Z H, Li D D, Yang Y C. Rules based feature modification for affective speaker recognition [C]//*International Conference on Acoustics, Speech, and Signal Processing*. Toulouse, France, 2006: 661 - 664.

[5] Reynolds D A, Quatieri T F, Dunn R B. Speaker verification using adapted Gaussian mixture models [J]. *Digital Signal Processing*, 2000, **10** (1): 19 - 41.

[6] Campbell W M, Sturim D E, Reynolds D A, et al. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation [C]//*International Conference on Acoustics, Speech, and Signal Processing*. Toulouse, France, 2006: 97 - 100.

[7] Hu H, Xu M M, Wu W. GMM supervector based SVM with spectral features for speech emotion recognition [C]//*International Conference on Acoustics, Speech, and Signal Processing*. Honolulu, Hawaii, USA, 2007: 413 - 416.

[8] Sinha R, Ghai S. On the use of pitch normalization for improving children's speech recognition [C]//10*th Annual Conference of the International Speech Communication Association*. Brighton, UK, 2009: 568 - 571.

[9] Wu Z Z, Kinnunen T, Chng E S, et al. Text-independent F0 transformation with non-parallel data for voice conversion [C]//11*th Annual Conference of the International Speech Communication Association*. Makuhari, Japan, 2010: 1732 - 1735.

[10] Basak D, Pal S, Patranabis D C. Support vector regression [J]. *Neural Information Processing—Letters and Reviews*, 2007, **11** (10): 203 - 224.

基于韵律变换的情感说话人识别

宋 鹏[1] 赵 力[1] 邹采荣[1,2]

([1] 东南大学水声信号处理教育部重点实验室,南京 210096)

([2] 佛山科学技术学院,佛山 528000)

摘要:为了解决由情感变化引起的说话人识别性能下降问题,提出了一种新的情感说话人识别系统. 首先,通过引入情感识别作为前端处理模块,对中性语音和情感语音进行分类. 然后,对情感语音进行韵律修正,分别采用高斯归一化、高斯混合模型(GMM)和支持向量回归(SVR)等方法建立情感语音和中性语音的基频映射规则,并根据平均线性变化率对时长进行了修正. 最后,对韵律修正后的情感语音进行识别. 实验结果表明,提出的情感说话人识别系统可以有效地提高情感说话人识别的性能,识别率相比传统方法有了显著的提高. 并且通过基频和时长修正的情感语音更接近于中性语音.

关键词:情感识别;说话人识别;基频转换;时长修正

中图分类号:TN912.3