A speech emotion enhancement method for hearing aid

Shulan Xia*, Jilin Wang

College of Electrical Engineering, Yancheng Institute of Technology, Yancheng 224051, Jiangsu, PR China

Received 1 October 2014, www.cmnt.lv

Abstract

In this paper, emotional perception of the hearing-impaired patients for hearing aid is investigated, and a speech enhancement algorithm is proposed, which is text-independent and requires less and non-parallel training data. In addition, the conversion of prosodic and spectral parameters is also studied. The Eigenvoice Gaussian mixture model (EV-GMM) is used to transform the F0s and spectral parameters, which is built using multiple pre-stored sources emotional and target neutral speech sentences. In the training and testing stages, the duration modification is utilized to improve the performance of EV-GMM training and converted output quality and an adaptive median filter is proposed to smooth the trajectory of the converted speech. Perceptual and objective experiments are presented, simulation results corroborate the effectiveness of the proposed algorithms.

Keywords: hearing aid, emotional speech enhancement, EV-GMM, duration modification

1 Introduction

Emotional communication is very vital for social activities. However, hearing-impaired patients lack this communication for their defective hearings. Although in recent years, the hearing aid research in China is deeper and deeper [1-4], but the little study focused on emotional problems for hearing-impaired patients has yet still been done. The emotion enhancement for hearing aid is an efficient technique to compensate this problem, which refers to transforming the emotional character of the source emotion to enhance the emotional feelings of the hearing-impaired patients.

One strategy for emotion enhancement is based on voice conversion methods. For example, the GMM-based voice transformation algorithm is directly applied to carry out the emotion transformation in [5], it is found the prosody features mainly dominate the emotion state and the transformation of spectral parameters is not sufficient for conveying the required target emotion. In [6], the relevance of the speech components including F0, residual signal and spectral envelope is investigated.

Another alternative method is based on unit selection. In [7], the GMM classification and regression tree (CART) model has been adopted to transform the prosody from neutral speech to target one. In [8], the F0 contours are generated using hidden Markov model (HMM) and the syllable contours are selected from the database using the cost function. In [9], the emotion transformation based on prosodic unit selection is explored and discussed. It is obvious when the corpus is large, these approaches seem to work better than prosodic and spectral voice conversion methods, however, the emotion database is too large, and hard to design and label, etc.

Based on these works, the emotional speech enhancement system, which is completely text-independent and need less non-parallel training data, is proposed. The feature modifications of pitch, duration and spectral envelope are investigated, and an efficient EV-GMM framework is proposed to improve the performance of emotion enhancement. Meanwhile, the duration modification is incorporated in the training and testing phase to improve the enhancement performance, and the efficient adaptive median filter is also adopted to smooth the enhanced speech and reduce the discontinuity problem.

The paper is organized as follows. Section 2 gives a brief introduction of the emotion corpus. Section 3 describes the proposed prosody and spectrum enhancement methods. Then different experimental results are discussed in section 4. Finally, the conclusions are made in section 5.

2 DATASET

A mandarin emotional speech corpus including five types of emotional states (happiness, sadness, anger, fear and surprise) and one neutral state was established for the experiments. Two broadcast professionals including one male named LIN and one female named HUA were hired for the recording. 100 sentences with no apparent emotional tendency were provided for the material. Each sentence in the six simulating speaking styles was uttered resulting total 600 sentences for each speaker. The speech sentences were recorded in a quiet lab environment with 11.025KHz sampling rates and 16 bit precision, each of them has around 3~4ms valid speech.

The corpus for each speaker is divided into two parts in this paper: source set and target set. The first includes four types of pre-stored emotional speaking styles and one testing emotional one, and the latter contains the neutral speech.

3 Prosody and spectrum transformation

It is well known the prosody transformation plays an important role in emotion transformation, so it can be regarded

***Corresponding author** e-mail: xslnj@126.com

as a particular voice transformation focusing on prosodic level. In this section, the transformations of prosodic features including F0 and duration are investigated, and a spectral transformation is also performed to enforce the performance.

3.1 F0 TRANSFORMATION

3.1.1 Baseline F0 transformation

One typical F0 transformation is based on GMM which is previously applied to spectrum transformation [10, 11]. Denoting the F0s of source emotional and target neutral speech by f_x and f_y respectively, f_x and f_y are modelled by a joint GMM, the probability is as follows

$$p(f_x, f_y \mid \lambda) = \sum_{i=1}^{M} \alpha_i N(f_x, f_y, \mu_i, \Sigma_i) . \tag{1}$$

And the converted function can be written as

$$F(f_x) = E(f_y \mid f_x) = \sum_{i=1}^{M} p_i(f_x)[\mu_i^y + \frac{\Sigma_i^{yx}}{\Sigma_i^{xx}}(f_x - \mu_i^x)], \tag{2}$$

$$p_i(f_x) = \frac{\alpha_i N(f_x, \mu_i^x, \Sigma_i^{xx})}{\sum_{k=1}^{M} \alpha_k N(f_x, \mu_k^x, \Sigma_k^{xx})} , \tag{3}$$

where $\boldsymbol{\mu_i} = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix}$ and $\boldsymbol{\Sigma_i} = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix}$ are the mean and covariance matrices of the *i*-th component, and the size of GMM components is *M*. In this method, the parallel utterances of source and target are needed for the model training.

3.1.2 Proposed F0 transformation

The proposed F0 transformation is based on EV-GMM algorithms, which is motivated by the voice transformation based on speaker adaptation [12] and many-to-one voice transformation [13], the EV-GMM has similar form as prior GMM, except for the mean value of source speaking style, which takes the form as

$$\mu_i^x = b_i(0) + B_i \omega , \tag{4}$$

where $b_i(0)$ is the bias value for the *i*-th component, $B_i=[b_i(1),b_i(2),\ldots,b_i(J)]$ is a matrix consisting with *J* basis vectors, and $\omega = [\omega_i(0), \omega_i(2), ..., \omega_i(J)]^T$ is a *J*-dimensional weight vector.

The EV-GMM allows the adaptation of arbitrary input new emotional speech by adjusting the values of the weight vector ω, which is estimated by the maximum likelihood eigendecomposition (MLED) [14] method as follows,

$$\hat{\omega} = \int p(f_x^u, f_y \mid \lambda^{ev}) df_y , \tag{5}$$

where λ^{ev} means the EV-GMM model, and f_x^u is a new source emotional F0 sequence for training. Using the GMM trained with pre-stored source emotional and target neutral speech pairs as the initial model and expectation maximization (EM) algorithm, an adapted EV-GMM for the new source emotional speech can be achieved. The transformation of F0s is directly performed using the adapted F0 EV-GMM model.

3.2 DURATION MODIFICATION

As is well known, different emotions have different speaking rates, which are determined by the number of frames. The simple duration modification based on an average linear ratio is adopted and it takes the form as

$$\hat{D}_n = D_e \frac{\bar{D}_n}{\bar{D}_e} , \tag{6}$$

where D_e and $\hat{D}_n$ are the durations of source emotional and converted speech, $\bar{D}_e$ and $\bar{D}_n$ are the average durations of source emotional and target neutral speech respectively.

Duration as an important emotional feature should be taken into account to separate different emotions. Unfortunately, it is overlooked by traditional voice transformation which mainly bases on spectral and F0 modifications. A new strategy is developed to address this issue in emotional transformation framework, the durations of source emotional utterances are modified to map those of target neutral ones in the training process of EV-GMM, the durations of source emotional speech can be seen as a pre-processing module, and It is depicted in Figure 1.

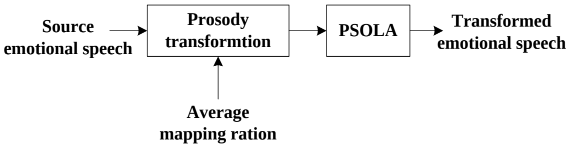


FIGURE 1 A flowchart of duration modification

3.3 SPECTRUM TRANSFORMATION

The spectral transformation works as styles of F0s and the transformation is also straightforward performed using the spectral adapted EV-GMM, which is based on the GMM transformation method as proposed by Kain [6].

3.4 POST-PROCESSING BY RAMF

The emotion transformation is performed on a frame-by-frame basis. One main shortcoming is the discontinuities in the converted speech. A ranked-order based adaptive median filter (RAMF) [15] technique previously applied to image processing has been presented here to remove the unrepresentative points and smooth the converted speech. Assuming *W* is a rectangle filtering window, S_{cur} are the values of current points, $S_{\min}$, $S_{\max}$, S_{med} , are the minimum, maximum and median values of points in the filtering windows. The modified RMAF can be seen a two-level structure: level *A* and level *B*.

$$\text{Level } A: \begin{array}{l} A_1 = S_{med} - S_{\min} \\ A_2 = S_{med} - S_{\max} \end{array} . \tag{7}$$

If $A_1 > 0$ and $A_2 < 0$, then turn to level *B*, otherwise, increase the size of *W* to repeat level *A* until the size of $W > S_{\max}$, then S_{cur} is used as the output.

$$\text{Level } B: \begin{array}{l} B_1 = S_{cur} - S_{\min} \\ B_2 = S_{cur} - S_{\max} \end{array} . \tag{8}$$

If $B_1 > 0$ and $B_2 < 0$, then S_{cur} is used as the output, or S_{med} is adopted as an output.

4 Evaluation experiments

Several objective and subjective experiments were designed to evaluate the performance of the proposed emotion transformation algorithm. On one hand, an objective experiment is conducted to measure the distance of F0 contours between source emotional and target neutral speech. On the other hand, the subjective tests including ABX, mean opinion score (MOS) were conducted by listening test with ten experienced listeners. Five kinds of transformation methods were compared. They are the traditional GMM based transformation method using parallel data, proposed EV-GMM transformation method considering pitch only (PO), pitch and duration (PD), pitch, spectrum and duration (PSD), and pitch, spectrum and duration adding RAMF (PSDR).

In order to train the EV-GMM model, four types of emotions including happiness, sadness, anger and fear were used as pre-stored source speech, and a neutral one was used as target speech, all of them was phonetically balanced utterances and aligned by dynamic time warping (DTW) technique. The first 50 (sentence: 1-50) parallel utterances for each emotional and neutral pairs totally 250 were used for training EV-GMM. The testing emotional transformation is conducted between surprise and neutral, and a traditional GMM based transformation method using surprised and neutral parallel training dataset were used for comparison. Sentences 91-100 of surprise and neutral were used for evaluation, and sentences 51-90 were for the training of tradetional GMM method.

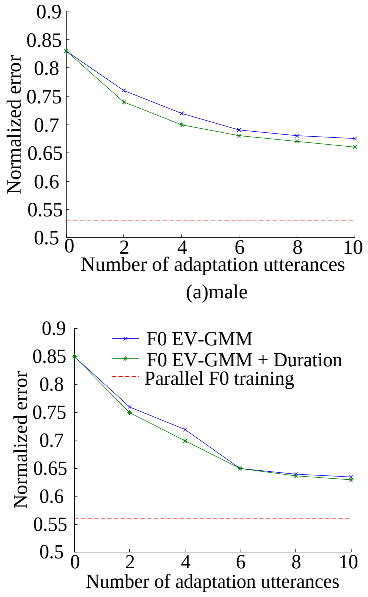
It is noted that F0s were in log-scaled domain, and the 16 order line spectral frequencies (LSFs) were extracted. The sizes of EV-GMMs for F0 and spectrum were optimized as 256 and 512, respectively. the numbers of GMMs for F0 and spectrum were set as 16 and 64, respectively.

4.1 OBJECTIVE EVALUATION

The objective evaluation experiment was performed to assess the performance of prosody transformation, the error measurement in this section is a mean error normalized by the initial F0 distance between source and target speech, which can be written as

$$\xi = \frac{\frac{1}{N}\sum_{i=1}^{N} | y_i - F(x_i) |}{\frac{1}{N}\sum_{i=1}^{N} | y_i - x_i |} , \tag{9}$$

where x_i and y_i are the F0 values for source and target speech respectively, *N* is the number of frames, and *F*( ) refers to the F0 converted function.



(a)male

(b)female

FIGURE 2 Normalized error

Figure 2 depicts the performance of presented prosody transformation method with different number of adaptation data. It can be found that the mean error decreases with more adaptation data, since it correspondences a more accurate F0 modeling. Incorporating the duration modification can greatly enhance the performance of F0 transformation. Moreover, it can be seen when the training data is above four for male and six for female, the errors remain approximately consonant, which means the number of sentences is enough to model the F0 distributions.

4.2 SUBJECTIVE EVALUATIONS

In order to evaluate the similarity between converted and target neutral speech, an ABX test was designed to judge whether X is close to A or B, where X means the converted emotional speech, A and B either the source surprised or target neutral speech.

Table 1 shows the percentages of the converted speech that were closer to the target using the above mentioned methods. It is obvious proposed EV-GMM (with the PCR of 79%) based can achieve satisfactory results compared to traditional GMM using parallel training data (with the PCR of 85%), and the prosody including pitch and duration mainly contribute to the emotion transformation, the spectrum transformation and RAMF module can enhance the transformation performance.

TABLE 1 Percentage of correct responses using ABX test

Methods	**Percentage of correct response (%)**	
	Male	*Female*
Parallel training	85.1	86.3
PO	69.1	65.3
PD	75.5	72.4
PSD	77.2	76.2
PSDR	79.4	77.5

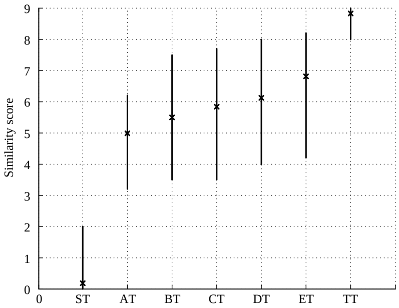


FIGURE 3 Similarity test

A MOS test is also conducted to assess the quality of the converted emotional speech, each converted speech based on GMM and proposed EV-GMM was shown to the listeners, who were asked to rate the quality using 10-point scores with 0 for “totally different” and 9 for “identical”. The pairs of speech include source speech (S), target speech (T), converted speech using PO (A), converted speech using PD (B), converted speech using PSD (C), converted speech using PSDR (D) and converted speech using parallel training data (E) respectively. Different sentences were used to make the pairs, so listener can judge the similarity between different speech pairs.

Figure 3 summarized the result of MOS, the mean score in each case was marked "X", and the vertical solid lines indict the variances of scores. The "ET" using parallel training data outperforms other pairs, and "AT" which adopts F0 only for EV-GMM transformation performs worst, the mean score of "DT" is nearer to that of "ET", which indicts the proposed method can achieve an acceptable performance compared with ideal transformation methods using parallel data.

5 Conclusions

A novel emotional speech enhancement method for hearing aid is proposed in the paper, which is based on EV-GMM and relaxes the constraints of parallel training data. The idea of the algorithm is to transforming the emotional character of the source emotion to enhance the emotional feelings of the hearing-impaired patients. Objective test shows the mean error is small and comparable to that using the baseline GMM based transformation using the parallel corpus. The subjective performances also demonstrate the efficiency of presented method that is indicated by listening tests.

Further work will be focused on subjective evaluations for hearing-impaired persons.

Acknowledgments

The work was supported by the National Natural Science Foundation of China under Grant No. 61301219, No. 61375028, No. 61273266 and No. 61301295, the Natural Science Foundation of Jiangsu Province under Grant No. BK20130241.

References

[1] Ruiyu Liang J X, Jian Zhou, Cairong Zou, Li Zhao 2013 An improved method to enhance high-frequency speech intelligibility in noise *Applied Acoustics* **74**(1) 71-8 (*in Chinese*)

[2] Liang Rui-Yu, Xi Ji, Zhao Li, Zou Cai-rong, Huang Cheng-wei 2012 Experimental study and improvement of frequency lowering algorithm in Chinese digital hearing aids A*cta physica sinica* **61**(13) 134305(1-11) (*in Chinese*)

[3] Wang Q, Zhao L, Qiao J, Zou C 2010 Acoustic feedback cancellation based on weighted adaptive projection subgradient method in hearing aids *Signal Processing* **90**(1) 69-79

[4] Liang R, Zou C, Zhao L, Wang Q, Xi J 2012 Experimental study on enhancement method for high-frequency hearing loss in Chinese digital hearing aids *Acta Acustica* **37**(5) 527-33 (*in Chinese*)

[5] Kawanami H, Iwami Y, Toda T, Saruwatari H, Shikano K 2003 GMM-based Voice Conversion Applied to Emotional Speech Synthesis *in Proc Eurospeech, Geneva Switzerland* 2401-4

[6] Barra R, Montero J M, Macias-Guarasa J, Ferreiros J, Pardo J M 2007 On the limitations of voice conversion techniques in emotion identification tasks *in Proc Interspeech* 2233-6

[7] Tao J H, Kang Y G, Li A J 2006 *IEEE Trans on Audio, Speech, and Language Processing* **14**(4) 1145-54

[8] Wu C H, Hsia C C, Liu T H, Wang J F 2006 *IEEE Trans on Audio Speech and Language Processing* **14**(4) 1109-16

[9] Erro D, Navas E, Herndez I, Saratxaga I 2010 *IEEE Trans on Audio Speech and Language Processing* **18**(5) 974-83

[10] Kain A, Macon M W 1998 Spectral voice conversion for text-to-speech synthesis *in Proc ICASSP Seattle USA* 285-8

[11] Inanoglu Z 2003 Transforming Pitch in a Voice Conversion. Framework *Master's thesis St Edmund's College University of Cambridge* 28-32

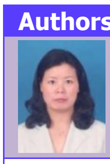
[12] Mouchtaris A, Spiegel J V D, Mueller P 2004 Non-Parallel Training for Voice Conversion by Maximum Likelihood Constrained Adaptation *in Proc ICASSP Montreal Canada* **1** 1-4

[13] Toda T, Ohtani Y, Shikano K 2006 Eigenvoice Conversion Based on Gaussian Mixture Model *in Proc Interspeech Pittsburgh USA* 2446-9

[14] Kuhn R, Junqua J, Nguyen P, Niedzielski N 2000 *IEEE Trans Speech and Audio Processing* **8**(6) 695-707

[15] Hwang H, Haddad R A 1995 *IEEE Transactions on Image Processing* **4**(4) 499-502

Authors

	**Xia Shulan, China** **Current position, grades:** master, the associate professor of Yancheng Institute of Technology, China. **Scientific interest:** signal processing, speech signal processing. **Experience:** an expert in the field of signal processing.
	Wang Jilin, China **Current position, grades:** master, the associate professor of Yancheng Institute of Technology, China. **Scientific interest:** signal processing, speech signal processing. **Experience:** an expert in the field of signal processing.