Data Pre-processing in Emotional Speech Synthesis by Emotion Recognition

Ling CEN, Minghui DONG, Paul CHAN
Institute for Infocomm Research (I2R), A*STAR, 1 Fusionopolis Way, Singapore 138632
E-mail: {lcen, mhdong, ychan}@i2r.a-star.edu.sg Tel: +65-963756291

***Abstract*— Synthesizing emotional speech by means of conversion from neutral speech allows us to generate emotional speech from many existing Text-to-Speech (TTS) systems. How much of the target emotion can be portrayed by the generated speech is largely dependent on the emotion data used to train the mapping function for voice transformation. In this paper, we introduce a method to pre-process the emotion database by detecting the emotions from speech using machine learning methods. A selection criterion is proposed to yield a refined database based on the results from emotion recognition. The experimental results have shown that the proposed data pre-processing method can effectively improve the naturalness of the synthesized speech by better portraying the targeted emotion. The quality of speech synthesized using the smaller database is comparable to that using the whole database. The computational load is reduced due to the reduction of the data used for training the transformation model.**

I. INTRODUCTION

The emotional states of a speaker are implicitly expressed in human speech together with linguistic information. To realize natural communication between human and computers, the expression of emotions is incorporated into synthesized speech so that the computer is able to convey emotions in speech. This has attracted a lot of interest over the recent years [1-8].

Voice conversion technology aims to automatically transform the voice with a source speaking style into that with a specified target speaking style. This has been used to generate emotional speech by means of conversion from neutral speech. This is desirable as it allows us to generate emotional speech from many existing Text-to-Speech (TTS) systems. The system for converting neutral speech to emotional speech can be broken down into two stages, namely, training and transformation stages. In the training stage, the transformation function for voice conversion is formulated based on the information from the source and target voices. This function is then employed in the transformation stage to modify the source voice so that the converted speech can match the characteristics of the target voice and portray the targeted speech emotion.

The transformation function is derived by analyzing the voice samples of source and target speech. How well the emotional characteristics of the target voice may be portrayed by the converted speech is largely dependent on the quality of the training data used for deriving the transformation model. Quality here refers not only to voice quality but also to the degree of emotion expressed in the training speech. Another problem encountered in speech emotion research is that there still lacks a precise method of defining and quantifying each emotional state. Consequently, there is no explicit way to express emotions in speech. Methods of expression of different emotions may be widely varied over different cultures, gender and ages. When the training data is recorded, even though the professional actors are employed to deliver utterances conveying different specific emotions, how much of the intended emotion may be accurately perceived by each listener is yet uncertain.

To address this, a data pre-processing method is proposed. Firstly, an emotion-detection system is build by combining four classification methods, namely, Probabilistic Neural Network (PNN), Support Vector Machines (SVMs), k-Nearest Neighbor (KNN), and Linear Discriminant Analysis (LDA). This is to predict the emotional states of the speech samples in the training database. A selection criterion is used to choose better speech samples based on the accuracy of emotion prediction. A refined and reduced database is then obtained using these selected samples for training purposes. After processing the database, the Gaussian Mixture Model (GMM) linear transformation is utilized to model the mapping function between neutral and emotional speech. With the help of data-preprocessing, we can choose to only use the speech samples that are predicted to better portray the target emotion in training. This may not only improve the naturalness of the generated speech, but also reduce the computational cost incurred at the training stage.

The remaining part of this paper is organized as follows. Section II elaborates on the data pre-processing method. Section III presents the GMM-based voice conversion methods. Section IV follows with the experiment results. Section V ends off with the concluding remarks.

II. PROPOSED DATA PRE-PROCESSING METHOD

The basic idea in our method is to use good samples with distinct emotions for training the transformation model. Manual selection is a time-consuming task. Many factors may affect its accuracy, such as the listeners' cultural background, gender and age. The amount of time having spent in continuously listening may also affect accuracy, as listeners may begin to experience listening fatigue. To address this, we propose a method to automatically select speech samples from the training database to obtain a refined database. Instead of using the whole database, the transformation model between neutral and targeted emotional speech is trained using this database. In this way, not only can we synthesize speech with

more distinct emotions, it will also require less computational load.

The procedural flowchart for data pre-processing is illustrated in Figure 1, where $c_{PNN}, c_{SVM}, c_{KNN}$ and c_{LDA} are the emotion classes estimated from the 4 classifiers. Acoustic features are firstly extracted for speech samples with targeted emotions. These features are then used to predict the emotions of the samples by individually using the PNN, SVM, KNN, and LDA methods. Each of the classifiers achieves one predicted class for each speech sample. Whether or not one sample is selected is based on the degree of unanimity of the 4 classifiers and actual class according to a selection criterion. This will be elaborated in the following subsections.

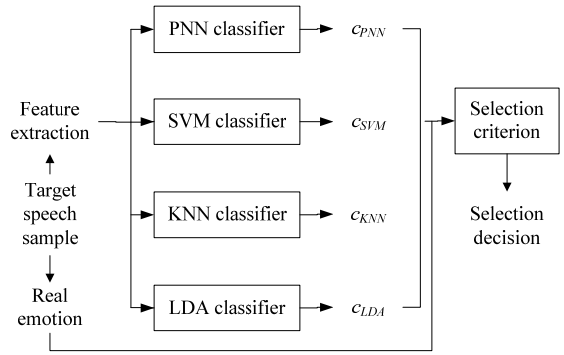

Fig. 1 Structure of data pre-processing system.

A. Acoustic Features

The acoustic features are extracted as described in [9]. Three short time cepstral features, i.e. Perceptual Linear Prediction (PLP) Cepstral Coefficients, Mel-Frequency Cepstral Coefficients (MFCC), and Linear Prediction-based Cepstral Coefficients (LPCC) are derived as acoustic features for emotion detection. Besides these, Delta and Acceleration of these raw features are also used. To reduce the acoustic variation in different utterances, utterance-level normalization is carried out by subtracting the mean and dividing by the standard deviation of the features. To capture longer time characteristics of the speech signal, utterance segmentation is performed. The frames within an utterance are grouped into several segments, each of which consists of 40 frames, with an overlap of 20 frames. The statistics, including median, mean, standard deviation, maximum, minimum, and range of the framed-based features, is calculated for each segment. The feature dimensionality is reduced from 792 to 150 using the Principal Component Analysis (PCA) by mapping the original feature data onto a lower-dimensional space.

B. Classification Methods

To achieve reliable prediction results for emotion recognition, four classifiers based on the PNN, SVM, KNN, and LDA methods, are employed. The PNN is a Bayesian statistical classifier that uses a Parzen estimator to approximate class dependent Probability Density Functions (PDF) [10]. It has been largely used to solve classification problems due to its simple training process, quick convergence, and ease implementation. The SVMs method is developed by Vladimir Vapnik [11] and his colleagues at AT&T Bell Labs in the mid 90's. The traditional techniques for pattern recognition are usually based on the minimization of empirical risk learned from training datasets, while SVMs aim to minimize the structural risk to achieve optimum performance. SVMs have shown better generalization performance than traditional techniques in solving classification problems. The KNN algorithm is a supervised learning algorithm, where a sample is classified based on the majority of *k*-nearest neighbor category. Here, *k* is a positive integer, typically small. If $k = 1$, then the object is simply assigned to the class of its nearest neighbor. LDA is a technique for classifying a set of observations into predefined classes by constructing a set of linear functions of the predictors. When the LDA is used for emotion classifications, a linear discriminant function is constructed for each emotion. Given a new speech sentence, the discriminant functions corresponding to all the emotion classes are evaluated and the observation is assigned to a class if its discriminant function has the highest value.

For every speech sample, an emotion class is estimated by each of the four classifiers. Together with the actual emotional state of this sample, selection decision is made according to a selection criterion described in the following sub-section.

C. Selection Criterion

Although there are some studies considering the ensemble of multiple classifiers in speech emotion recognition [12, 13], the problem to be solved here is different from a pure recognition problem. We are aiming to find good samples with distinct emotion expression. A wrong estimation may be caused by two factors. One, the emotion of that particular utterance is not distinctly expressed in speech. Two, the classifier fails to give a correct prediction due to its low recognition accuracy. To discard the samples whose emotion is wrongly estimated due to the former factor but give more chances to select the samples having wrong estimation due to the latter, a selection criterion is introduced to build up a refined database.

Let the actual emotion of one speech sample be c_r, and the emotion classes estimated by the PNN, SVM, KNN, and LDA be $c_i, i \in \{1,2,3,4\}$, respectively. For each emotion class, the recognition accuracy of each of the 4 classifiers is represented as $a_i, i \in \{1,2,3,4\}$. If the emotion class of one sample can be correctly recognized by all classifiers, this sample is considered to have distinct emotional characteristic and will be selected. Conversely, if all classifiers give the wrong predictions for a sample, it will be discarded as its emotion is not conspicuously expressed. This can be expressed as:

$$\begin{aligned} &\text{if } c_i = c_r \text{ for any } i \in \{1,2,3,4\}, d_s = 1, \\ &\text{if } c_i \neq c_r \text{ for any } i \in \{1,2,3,4\}, d_s = 0, \end{aligned} \quad (1)$$

where $d_s \in \{1,0\}$ represents the selection decision, i.e. $d_s = 1$ representing selecting and $d_s = 0$ representing unselecting.

In the case where more than 1 but less than 4 classifiers give the wrong estimation, the selection decision is made based on the average accuracies of emotion recognition. In our method, a sample whose emotion is correctly judged by at least a classifier with a higher average accuracy is more likely to be selected. On the other hand, we will more likely discard samples with emotional states that are wrongly estimated by the classifier with high accuracy of recognition. The probability to decide whether a sample is selected can be expressed as:

$$p_s = \left(\sum_{icrr=1}^{N_{crr}} a_{icrr} + \left(\sum_{iwrg=1}^{N_{wrg}} \left(1 - a_{iwrg}\right) \right) \right) / \left(N_{crr} + N_{wrg}\right), \quad (2)$$

where a_{icrr} for $icrr \in \{1,...N_{crr}\}$ and a_{iwrg} for $iwrg \in \{1,...N_{wrg}\}$ are the accuracies of the classifiers giving correct and wrong predictions, respectively. Here, N_{crr} and N_{wrg} are the number of classifiers that make correct wrong prediction, respectively, and there is

$$N_{crr} + N_{wrg} = 4. \quad (3)$$

III. GMM Based Voice Conversion

After the selection process presented above is completed for all the target speech used for training the transformation model, a GMM-based voice conversion method is applied to generate emotional speech from neutral speech.

In GMM-based voice conversion, a GMM is trained for each of the target emotions using the parallel utterances in the target emotional as well as neutral states. Dynamic Time Warping (DTW) is used to align the parallel data to accommodate the differences in timing between neutral and emotional speech.

The features used here are Line Spectral Frequencies (LSFs) which are extracted from the aligned data. The LSF is a representation of the Linear Predictive Coding (LPC). LPC represents the spectral envelope of the speech signal using the information of a linear predictive model. It is popular in speech processing as it is able to accurately estimate the spectral parameters. Here we use LSFs as they are superior to the direct quantization of the LPC and less sensitive to quantization noise.

At the training stage, the conversion function is modeled by the joint density of the source and target features. At the transformation stage, the features of neutral speech are then transformed using this conversion function so that the synthesized speech possesses the emotion of the target.

IV. Evaluation

A. Emotional Speech Database

The speech database we used in this study was recorded as part of the efforts towards the development of a 3-dimensional emotional talking head. Neumann U87s were used in an acoustically treated room. The speech was recorded together with 8 video feeds and 3-dimensional motion data. Although the speech was recorded in stereo, only one channel was used in this project. An experienced theatrical actress, who is a speaker of native British English, was engaged. Two thousand sentences were recorded in the neutral state as well as each of 6 different emotions, namely, happiness, anger, sadness, fear, disgust as well as surprise. The sampling rate of the audio data is 48 kHz, which is down-sampled to 16 kHz in our experiment.

B. Data Pre-processing

In the experiment, the speech expressed in the emotions of happiness, sadness, anger, and fear were re-synthesized by means of conversion from the neutral speech. Prior to conversion, the target speech samples with the distinct emotions expressed were selected for training the transformation model.

In order to increase data diversity, the utterances that expressed disgust and surprise were included to train the classifiers for emotion recognition, apart from those that expressed happiness, sadness, anger, and fear. Specifically, 200 samples were used for each emotion category. The 200 samples were separated into 2 groups, with each group containing 100 samples. In all, each group included 7 classes and had 700 samples. When one group was used in training the classifiers, the emotions of the speech samples in the other group were recognized. By this way, better samples with distinct emotional characteristic were selected to obtain a refined database based on the classification results. Their roles were exchanged to select data from the two groups. Combining the classification results from both groups, the average accuracies for happiness, sadness, anger and fear using the 4 classifiers are listed below in Table I.

Fifty samples were selected for each targeted emotion from the 200 speech samples altogether. These selected samples and their corresponding neutral samples were used to train the transformation model in voice conversion.

TABLE I
ACCURACIES (%) OF EMOTION RECOGNITION.

Accuracy	PNN	KNN	LDA	SVMs
Anger	76	59	80	77
Fear	82	63	88	89
Happy	74	43	68	56
Sad	70	72	82	72

C. Objective Tests

For each of the 4 emotions, 100 sentences were synthesized, which were used in the objective tests. For the purpose of comparison, we also synthesized these sentences using the original 200 samples for each emotion, which were used in the tests too. There are, hence, 2 sets of emotional speech synthesized using different sizes of training datasets as listed in Table II. In the objective test, the classifiers that were used in data pre-processing were used to recognize the emotional states of these synthesized sentences in both sets. Each of the 4 classifiers was employed individually. The recognition accuracies are tabulated in Table III.

It can be seen from Table III that the emotions of the speech synthesized using the selected samples were more accurately recognized by all classifiers. Although 3 values in

Table III (B) are larger than their corresponding values in Table III (A), all of their average accuracies in Set A are higher than those in Set B. This indicates that the speech synthesized using only the samples selected were perceived to best portray the targeted speech emotion. Although only a quarter of the samples were used for training the transformation model, the degree of emotional expression of the speech synthesized in Set A was improved in comparison with that in Set B. The computational cost is also reduced with the size reduction of the training data.

TABLE II
EMOTIONAL SPEECH SYNTHESIZED USING TWO SIZES OF TRAINING DATA.

Synthesized speech set	No. Sentences	No. Emotions	Size of Training data
Set A	100×4	4	50×4
Set B	100×4	4	200×4

TABLE III
ACCURACIES (%) OF EMOTION RECOGNITION FOR SYNTHESIZED SPEECH.

(A) SET A

Accuracy	PNN	KNN	LDA	SVMs
Anger	66	65	76	67
Fear	55	50	72	72
Happy	52	42	54	49
Sad	47	51	75	68
Avg.	*55.00*	*52.00*	*69.25*	*64.00*

(B) SET B

Accuracy	PNN	KNN	LDA	SVM
Anger	59	50	74	64
Fear	**57**	44	66	**73**
Happy	44	36	51	43
Sad	46	**55**	69	55
Avg.	*51.50*	*46.25*	*65.00*	*58.75*

D. Subjective Tests

TABLE IV
PAIR-WISE PREFERENCE RESULTS IN LISTENING TESTS

Preference	Speech quality	Emotion perceived
Set A	10.7%	40.9%
Set B	13.8%	23.4%
No Preference	75.5%	35.7%

Since the emotional speech is synthesized with the human listener in mind, subjective tests were conducted on the synthesized speech. Forty sentences evenly distributed in 4 different emotions were selected from each of the 2 sets to form 40 testing pairs. Pair-wise preference tests were conducted on 10 listeners. Each listener was tasked to tell which sentence in the pair is better according to both the quality of the synthesized speech and the amount of the particular emotion the listeners perceived it to portray. They are allowed to choose "no preference" if they perceive the two utterances in a pair to be the same. The results are shown in Table IV. The rate on the amount of the emotion perceived is 40.9% for the utterances in Set A, while only 23.4% for those in Set B. Although only a quarter of the samples were used in training, the quality of speech in Set A was comparable to that in Set B. It is shown from the results that our method can effectively improve the naturalness of the synthesized speech by better portraying the targeted emotion.

V. CONCLUSIONS

In the synthesis of emotional speech by means of conversion from neutral speech, samples of speech in the target emotion are used to train a transformation model between the source and target speech. This paper describes a machine-learning based method to pre-process the database in order to select samples with distinct characteristics of the target emotion. Using a smaller, refined, training database, the synthesized speech is able to better portray the target emotion at a lower computational cost while the quality of the generated speech is not reduced.

REFERENCES

[1] Murray, I. R., and Arnott, J. L., "Implementation and testing of a system for producing emotion-by-rule in synthetic speech," *Speech Commun.*, vol. 16, no. 4, pp. 369-390, 1995.

[2] Burkhardt, F., and Sendlmeier, W. F., "Verification of acoustical correlates of emotional speech using formant-synthesis," *ISCA Workshop on Speech & Emotion*, pp. 151-156, Northern Ireland, 2000,.

[3] Iida, A., Campbell, N., Iga, S., Higuchi, F., and Yasumura, M., "A speech synthesis system for assisting communication," *ISCA Workshop on Speech & Emotion*, pp. 167-172, Northern Ireland, 2000.

[4] Montero, J. M., Gutiérrez-Arriola, J., Palazuelos, S., Enríquez, E., Aguilera, S., and Pardo, J. M., "Emotional speech synthesis: from speech database to TTS," *4th International Conference on Spoken Language Processing*, vol. 3, pp. 923-926, 1998.

[5] Tao, J. H., Kang, Y. G., and Li, A. J., "Prosody conversion from neutral speech to emotional speech," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 4, July 2006.

[6] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and Kitamura, T., "Speech parameter generation algorithms for HMM-based speech synthesis," *Proc. of Acoustics, Speech and Signal Processing*, vol.3, pp.1315–1318, June 2000.

[7] Tamura, M., Masuko, T., Tokuda, K., Kobayashi,,T. "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," *Proc of Acoustics, Speech and Signal Processing*, vol.1, pp.1–1, May 2001.

[8] Cen, L., Chan, P., Dong, M. H., Li, H. Z., "Generating emotional speech from neutral speech," *Proc of Chinese Spoken Language Processing*, pp. 383-386, 2010.

[9] Cen, L., Dong, M. H., Li, H. Z., Yu, Z. L., and Chan, P., *Application of machine learning*, IN-TECH, 2010, ch. Machine learning methods in the application of speech emotion recognition.

[10] Specht, D. F., "Probabilistic neural networks for classification, mapping or associative memory," *Proc Neural Network*, vol. 1, pp. 525-532, Jun. 1988.

[11] Vapnik, V. (1995), *The nature of statistical learning theory*, Springer-Verlag, 1995, ISBN 0-387-98780-0.

[12] Morrison, D, Wang, R., and De Silva, L. C., "Ensemble Methods for Spoken Emotion Recognition in Call-centres," *Speech Communication*, vol. 49, no. 2, pp. 98-112, Feb. 2007.

[13] Nguyen, T. and Bass, "Investigation of Combining SVM and Decision Tree for Emotion Classification," *Proc. 7th EEE International Symp. Multimedia*, pp. 540-544, Dec. 2005.