Journal of Computer Science and Cybernetics, V.33, N.3 (2017), 229–246
DOI 10.15625/1813-9663/33/3/11017

GMM FOR EMOTION RECOGNITION OF VIETNAMESE

DAO THI LE THUY[1,2], TRINH VAN LOAN[2], NGUYEN HONG QUANG[2]

[1]*Faculty of Information Technology, Ha Noi Vocational College of High Technology*
[2]*Ha Noi University of Science and Technology; thuydt@hht.edu.vn*

Abstract. This paper presents the results of GMM-based recognition for four basic emotions of Vietnamese such as neutral, sadness, anger and happiness. The characteristic parameters of these emotions are extracted from speech signals and divided into different parameter sets for experiments. The experiments are carried out according to speaker-dependent or speaker-independent and content-dependent or content-independent recognitions. The results showed that the recognition scores are rather high with the case for which there is a full combination of parameters as MFCC and its first and second derivatives, fundamental frequency, energy, formants and its correspondent bandwidths, spectral characteristics and $F0$ variants. In average, the speaker-dependent and content-dependent recognition scrore is 89.21%. Next, the average score is 82.27% for the speaker-dependent and content-independent recognition. For the speaker-independent and content-dependent recognition, the average score is 70.35%. The average score is 66.99% for speaker-independent and content-independent recognition. Information on $F0$ has significantly increased the score of recognition.

Keywords. GMM, recognition, emotion, Vietnamese, corpus, $F0$.

1. INTRODUCTION

Recognition of emotional speech has been of interest to researchers because it is particularly useful for applications that require a natural interaction between man and machine. There are many studies on recognition of emotional speech available in a number of different languages around the world such as English, German, Chinese, French, Spanish,. . . [1]. The majority of these studies use speech features in four categories [1]: continuous features (pitch, energy, formant), voice quality features (easy or hard listening, stress level, breathing level), spectral features LPC (Linear Prediction Coding), MFCC (Mel Frequency Cepstral Coefficients), LFPC (Log-frequency power coefficients)), TEO features (TEO-Teager-energy-operator) proposed by Teager (TEO-FM-Var (TEO-decomposed FM variation), TEO-Auto-Env (normalized TEO autocorrelation envelope area), TEO-CB-Auto-Env (critical band-based TEO autocorrelation envelope area)). At present, the study of emotional Vietnamese is mainly done in terms of language [2]. In terms of signal processing, there are very few studies on emotional Vietnamese. A number of studies on emotional Vietnamese have been published, often in multi-modal corpus, combining facial expressions, gestures, and voices with major applications for synthesis of Vietnamese. For example, the study in [3, 4] tested the modeling of Vietnamese prosody with multi-modal corpus to synthesize Vietnamese with emotion. The authors in [5] used SVM (Support Vector Machines) for classification with the inputs as brain electrical signals and the results showed that real-time recognition is possible with five states of emotion and the average accuracy is 70.5%. In addition, there are few studies of emotional Vietnamese that are performed abroad and not primarily by Vietnamese [6, 7]. The corpus of [6] contains two male voices and two female voices with six sentences for six emotions: happiness, neutrality, sadness, surprise, anger, and fear. The GMM (Gaussian Mixture Model) model was used with the characteristic parameters such as MFCC, short-term energy, pitch, formants. The highest recognition score is 96.5% for neutrality and the lowest is 76.5% for sadness. The corpus for [7] consists of 6 voices with 20 sentences and the emotions as in [6]. In [7], the recognition score on Vietnamese language using Im-SFLA (Improved Shuffled Frog Leaping Algorithm) SVM (Support Vector Machine) reaches 96.5% for neutrality and has dropped to 84.1% for surprise.

For speech emotion recognition, one can use models such as HMM (Hidden Markov Model), GMM [1, 6, 19], SVM [1, 7], ANN (Artificial Neural Network), KNN (K-Nearest Neighbors) and some other classifiers [1]. In fact, no classifier is the most suitable for emotional recognition. Because each classifier has its own advantages and disadvantages. Nevertheless, GMM is a model that is appropriate for emotion recognition as this is a model that targets the information envelope rather than the detailed content of the information. As it can be seen later in Section 3 of this paper, among the three sets of parameters for determining a GMM model, there are two sets of parameters directly related to the average: mean vectors and covariance matrices. According to [7, 18], the GMM model is a popular and promising model for speech emotion recognition. For the research on this paper, the experiments with GMM were performed on different corpora and different parameters.

The paper consists of 5 sections. Section 2 shows the characteristic parameters and the corpora used for experiments. Section 3 gives an overview of the GMM model. The experiment results using the GMM model with the specific parameters for Vietnamese emotion recognition are given in Section 4. Finally, Section 5 is conclusion.

2. CORPORA AND CHARACTERISTIC PARAMETERS FOR EXPERIMENTS ON EMOTION RECOGNITION

2.1. Corpora for experiments

The corpus of emotional Vietnamese used for the experiments in this paper included 5584 files of the corpus BKEmo [8] with four emotions: neutral, sadness, anger and happiness being spoken by 8 male and 8 female voices. The authors of this paper have performed listening to remove the error files or the files that did not express well emotions, so the remaining files are 5584 files with 22 sentences of different. Among these sentences, there are short, long, exclamation sentences such as "*Got a salary*", "*Oh, that person can not change that*" to analyze the characteristic parameters of emotions. Each sentence is pronounced four times. The number of wave files for each male and female voice is 2792 files, each emotion has 698 files. The ages of artists participating in emotion pronouncing are from 20-58. The voice is recorded with sampling frequency 16 kHz, 16 bits/sample. The recording is conducted in dubbing studio. This corpus used to recognize emotions of Vietnamese for the four experiment cases is shown in Table 1.

For the corpus Test1, the training and testing corpus have the same content and the same speakers and sentences with the same content are pronounced at different times. For the corpus Test2, the training and testing corpus have the same speakers but 22 sentences are divided into two parts, the contents of the 11 sentences used for the training and the remaining used for testing. For the corpus Test3, the number of speakers is divided into two parts. For the corpus Test4, number of sentences and number of speakers are divided by 2.

Table 1. Vietnamese emotional corpus for experiments with GMM model

Corpus	**Experiment Corpus**	**Total Number of Files**	**Number of Training Files**	**Number of Testing Files**
Test1	Speaker-dependent and content-dependent	5584	2792	2792
Test2	Speaker-dependent, content-independent	5584	2793	2791
Test3	Speaker-independent, content-dependent	5584	2794	2790
Test4	Speaker-independent, content-independent	2803	1403	1400

2.2. Characteristic parameters

The characteristic parameters used for the experiments include 87 parameters as shown in Table 2. These parameters were extracted from the speech signals in the corpus using Praat[1] and Alize toolkits [9]. Formants and its correspondent bandwidths are determined by Praat and based on LPC. Fundamental frequency $F0$ is calculated by Praat and based on cross-correlation analysis. The range for determining $F0$ depends on the gender. For female voices, the maximum $F0$ value is 350 Hz, and this value is 200 Hz for male voices.

In Table 2, according to Praat harmonicity represents the degree of acoustic periodicity also known as the Harmonics-to-Noise Ratio (HNR) and can be used as a measure for voice quality. If $S\left(f\right)$ is complex spectrum, where f is the frequency, the centre of gravity is given by formula

$$\frac{\int_0^\infty f\left|S\left(f\right)\right|^p df}{\int_0^\infty \left|S\left(f\right)\right|^p df}, \tag{1}$$

where $\int\limits_0^\infty \left|S\left(f\right)\right|^p df$ is energy. So, the centre of gravity is the average of frequency f over the entire frequency domain, weighted by $\left|S\left(f\right)\right|^p$. For $p=2$, the weighting is done by the power spectrum, and for $p=1$, the weighting is done by the absolute spectrum. The commonly used value is $p=2/3$. If $S\left(f\right)$ is a complex spectrum, the n^{th}central moment is given by (2) where f_c is the spectral centre of gravity

$$\frac{\int_0^\infty \left(f-f_c\right)^n \left|S\left(f\right)\right|^p df}{\int_0^\infty \left|S\left(f\right)\right|^p df}. \tag{2}$$

The n^{th} central moment is the average of $\left(f-f_c\right)^n$ over the entire frequency domain, weighted by $\left|S\left(f\right)\right|^p$. Moment is related to n^{th} order in formula (2). If $n=2$ we have the

[1] www.praat.org

variance of the frequencies in the spectrum. Frequency standard deviation is the square root of this variance.

If $n = 3$ we will have the third-order central moment, which is also the non-normalized skewness of the spectrum. Skewness indicates the deviation of the dataset relative to the standard distribution, if the deviation is below the mean, the data is more concentrated than that when the deviation is above the mean. The higher the absolute value of skewness is, the more unbalanced the distribution is. A symmetric distribution will have a skewness of 0.

With $n = 4$, we have the kurtosis of the non-normalized spectrum. For normalization, divide by the square of the second central moment and subtract 3. Kurtosis is an index to evaluate the shape characteristics of a probability distribution. Specifically, kurtosis compares the central portion of a distribution to its normal distribution. The greater and the sharper the center of the distribution is, the greater the kurtosis of the distribution is. The kurtosis of the normal distribution is equal to 3.

Table 2. The characteristic parameters for the Vietnamese emotional corpus

Index	**Characteristic parameters**	**Number of parameters**
(1)	*MFCC*	19
(2)	*The 1st-order derivatives of MFCC*	19
(3)	*The 2nd-order derivatives of MFCC*	19
(4)	*Energy, the 1st-order and the 2nd-order derivatives of energy*	3
(5)	*Fundamental frequency* $F0$	1
(6)	*Speech intensity*	1
(7)	*Formants and its correspondent bandwidths*	8
(8)	*Harmonicity*	1
(9)	*Centre of gravity*	1
(10)	*Central moment*	1
(11)	*Skewness*	1
(12)	*Kurtosis*	1
(13)	*Frequency standard deviation*	1
(14)	*LTAS (Long Term Average Spectrum) mean*	1
(15)	*Slope and standard deviation of LTAS*	2
(16)	$difF0\,(t)$	1
(17)	$F0NormAver\,(t)$	1
(18)	$F0NormMinMax\,(t)$	1
(19)	$F0NormAverStd\,(t)$	1
(20)	$difLogF0\,(t)$	1
(21)	$LogF0NormMinMax\,(t)$	1
(22)	$LogF0NormAver\,(t)$	1
(23)	$LogF0NormAverStd\,(t)$	1

The average value of the spectrum is related to the standard deviation of the spectrum.

With the classification problem, when a set of values of data tends to be near the average, the concentration of data is better than that when the data set tends to be far from the average. Thus, the average can be useful to describe the set of values of the correlated data. The average of the values $x_1, ..., x_N$ is

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j. \tag{3}$$

The variants of $F0$ shown in Table 1 are as follows.

Derivative of $F0\,(difF0\,(t))$

$$difF0\,(t) = dF0\,(t)\,/dt. \tag{4}$$

Normalization of $F0$ by the average value $F0$ for each file $(F0NormAver\,(t))$

$$F0NormAver\,(t) = F0\,(t)\,/\overline{F0\,(t)}. \tag{5}$$

Normalization of $F0$ by min value $\min F0\,(t)$ and max value $\max F0\,(t)$ for each file $(F0NormMinMax\,(t))$

$$F0NormMinMax\,(t) = \frac{F0\,(t) - \min F0\,(t)}{\max F0\,(t) - \min F0\,(t)}. \tag{6}$$

Normalization of $F0$ by average and standard deviation of $F0\,(F0NormAverStd\,(t))$

$$F0NormAverStd\,(t) = \frac{F0\,(t) - \overline{F0\,(t)}}{\sigma F0\,(t)}. \tag{7}$$

Derivative of $LogF0\,(t)$ $(difLogF0\,(t))$

$$difLogF0\,(t) = dLogF0\,(t)\,/dt. \tag{8}$$

Normalization of $LogF0\,(t)$ by min value min $LogF0\,(t)$ and max value max $LogF0\,(t)$ for each file $(LogF0NormMinMax\,(t))$

$$LogF0NormMinMax\,(t) = \frac{LogF0\,(t) - \min LogF0\,(t)}{\max LogF0\,(t) - \min LogF0\,(t)}. \tag{9}$$

Normalization of $LogF0\,(t)$ by average of $LogF0(t)$ for each file $(LogF0NormAver\,(t))$

$$LogF0NormAver\,(t) = LogF0\,(t)\,/\overline{LogF0\,(t)}. \tag{10}$$

Normalization of $LogF0\,(t)$ by average and standard deviation of $LogF0\,(t)$ for each file $(LogF0NormAverStd\,(t))$

$$LogF0NormAverStd\,(t) = \frac{LogF0\,(t) - \overline{LogF0\,(t)}}{\sigma LogF0\,(t)}. \tag{11}$$

The characteristic parameters in Table 2 are divided into six sets for experiments as shown in Table 3. The reason for using the parameters as in Table 2 and how to divide the parameter sets as in Table 3 for emotion recognition of Vietnamese can be explained as follows. Since the MFCC have been proposed [16, 17], these characteristic parameters have been used commonly in systems such as speech recognition, speaker recognition, speech emotion recognition etc. [1]. Thus, the MFCC are considered as the basic characteristic parameters of these systems. One can say that MFCC are the basic parameters related to the speech signal spectra which have been condensed and based on the sensibility of the auditory system. In addition, the characteristic parameters from (8) to (15) are also parameters related to the speech signal spectrum that are statistically determined. In particular, the characteristic parameters (11) and (12) are closely related to the standard distribution that GMM has used. Vietnamese is a tonal language. $F0$'s variation rules will determine the tones of the Vietnamese. On the other hand, the $F0$'s variation rules of a word or a sentence also contribute to the emotion expression [8]. Therefore, $F0$ and the parameters from (16) to (23) are very closely related to Vietnamese and the emotions of the voice.

Table 3. Establishment of characteristic parameter sets used for experiments

Set of parameters	**Name**	**Characteristic parameters for the indexes in Table 2**	**Number of parameters**
1	MFCC	(1)	19
2	MFCC+Delta1	(1), (2)	38
3	MFCC+Delta12	From (1) to (3)	57
4	prm60	From (1) to (4)	60
5	prm79	From (1) to (15)	79
6	prm87	From (1) to (23)	87

3. VIETNAMESE EMOTIONAL RECOGNITION USING GMM

From the statistical aspects of pattern recognition, each class is modeled by probability distribution based on available training data. Statistical classifiers have been used in many speech recognition applications such as HMM, GMM. The GMM model is a probability model for density estimation using a convex combination of multivariate normal distributions. GMM can be considered as a special continuous HMM containing only one state [12]. GMM is very effective when modeling multimodal distributions and training requirements are far less than the requirements of a general continuous HMM. Thus, GMM is preferable to HMM for speech emotion recognition as only the general features are extracted from the speech used for training. However, GMM can not model the time structure of training data because equations for training and recognition are based on the assumption that all vectors are independent. In fact, GMM has been used quite commonly for speaker identification [9], language identification [10], dialect recognition [11] or classification of music genres [13]. In the case of emotion recognition, each emotion will be modeled by a GMM model and the set of parameters will be determined through training on the set of learning patterns.

Suppose that with a statement of the emotion j corresponding to K speech frames, each speech frame extracts the feature vector x_i with the dimension D. Thus, a statement of emotion j corresponds to the set X containing K feature vectors $X = \{x_1, x_2, \ldots, x_K\}$. Assume the feature vectors are consistent with the Gaussian distribution in which the distribution is determined by the mean and the deviation from the mean. From there, the distribution of the features of emotion j. can be modeled by the mixture of Gaussian distributions. The mixture model of the Gaussian distribution λ_j of emotion j will be the weighted sum of M component distributions determined by the probability

$$p(X|\lambda_j) = \sum_{m=1}^{M} g_m N(X; \mu_m, \Sigma_m). \tag{11}$$

In (1), g_m is the weight of the mixture that satisfies condition $\sum\limits_{m=1}^{M} g_m = 1$, $N\left(X; \mu_m, \Sigma_m\right)$ are component density functions with the multivariate Gaussian distributions as follows

$$N\left(X; \mu_m, \Sigma_m\right) = \frac{1}{\left(2\pi\right)^{D/2} \left|\Sigma_m\right|^{1/2}} e^{-\frac{1}{2}(X-\mu_m)^T \Sigma_m^{-1} (X-\mu_m)}. \tag{12}$$

In (2), μ_m is the mean vector $\mu_m \in \mathrm{R}^D$ and Σ_m is the covariance matrix $\Sigma_m \in \mathrm{R}^{D \times D}$. Thus, the GMM model λ_j for emotion j is defined by the triple: mean vectors, covariance matrices, and weights for M components: $\lambda_j = \{\mu_m, \Sigma_m, g_m\}_j, m = 1, 2, \ldots, M$. In fact, the determination of the GMM model λ_j of the emotion j will be done according to the expectation-maximization algorithm. This algorithm will determine the maximum likelihood of log likelihood $log(p\left(X|\lambda_j\right)$ [14]. In this paper, the Alize toolkit [19] has been used to evaluate models λ_j and perform emotion recognition experiments. Although using the Alize and Praat toolkits, by using MatLab as the intermediate programming language to connect, coordinate, compute and set up appropriate configuration files, the emotion recognition of Vietnamese for our research has been performed completely automatically.

4. EXPERIMENT RESULTS

This section presents the recognition experiments based on the GMM model for the four basic emotions of Vietnamese: neutral, sadness, anger and happiness, which correspond to the four sets of corpora in Table 1. Each experiment was conducted with the number of Gauss components M increasing from 16 to 8192 by the power of 2 ($M = 2^n, n = 4, 5, \ldots, 13$) and six sets of parameters in Table 3. The following are four recognition experiments performed in the same way that evaluates recognition results.

4.1. 4.1 Experiment 1: Speaker-dependent and content-dependent corpus

Figure 1 is the result of emotion recognition with six sets of parameters. The results show that, in general, the recognition score increases as M increases. When using the prm87 to identify, the average recognition score was 98.96%, the highest in comparison with the remaining cases and ranged from 98.96% to 99.97%. In the case of MFCC only, the recognition score was lowest and ranged from 72.96% to 88.90%, and the mean was 82.96%. The other four sets of parameters have an approximate recognition score ranging from 87.43% to 89.19%. When $M > 128$, the parameter set MFCC + Delta1 gives higher recognition scores than the parameter sets MFCC+Delta12, prm60, prm79.

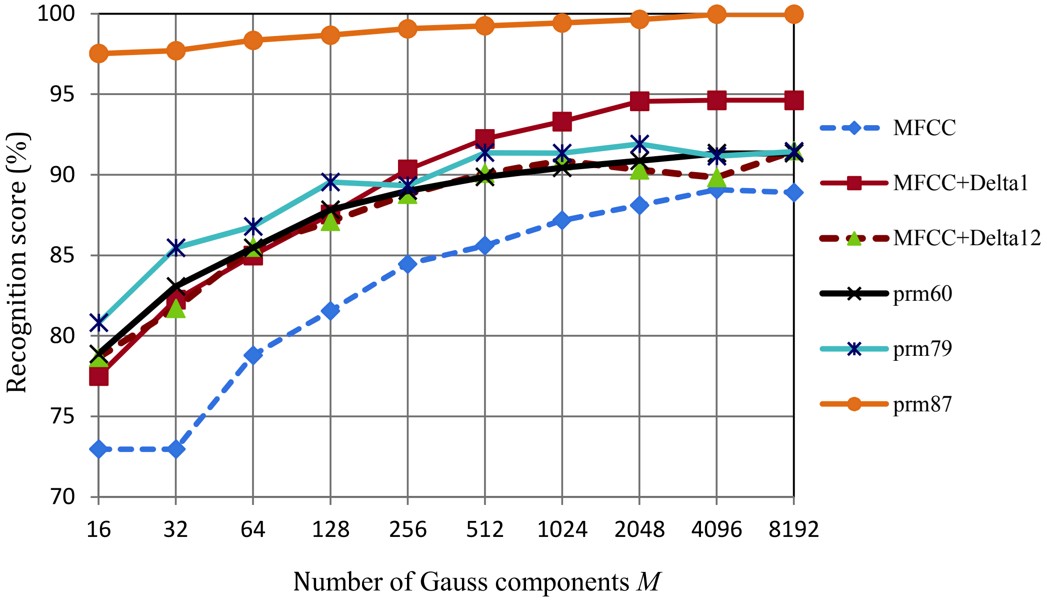

Figure 1. Experiment results with the corpus Test1

Figure 2 is the mean of the recognition results for each emotion for each set of parameters and for all values of M. Statistical results showed that the average recognition score of sad emotions was lowest (83.69%). The recognition score for happy emotions is higher than sad emotions (86.57%). The other two emotions had higher average recognition scores and these scores were approximately equal, in which angry emotion had the recognition score of 89.06% and neutral emotion was 89.08%. All four emotions had the highest recognition scores using the prm87 parameter set with the mean recognition scores of 99.66%, 98.77%, 97.7%, 90.64%, respectively for neutral, happy, angry and sad emotions.

In Experiment 1, when using the parameter set prm87 with $M = 4096$, the confusion recognition scores among the emotions were lowest. The confusion recognition scores (%) of emotions are shown in Table 4. Table 4 shows that the recognition scores are the highest and the wrong recognition scores are the smallest. The average recognition score of four emotions is 99.965%, in which happy, sad, neutral emotions are 100% and angry emotions are 99.86%. The confusion score from angry emotion to happy emotion is only 3.15%. The rest, among other couples, has a confusion recognition scores which were less than or equal to 1%.

Table 4. Confusion recognition scores (%) among emotions using Test1

***M*=4096**	**Happy**	**Sad**	**Angry**	**Neutral**
Happy	**100**	0	1	0
Sad	0	**100**	0	0.72
Angry	3.15	0.57	**99.86**	0
Neutral	0	0.43	0	**100**

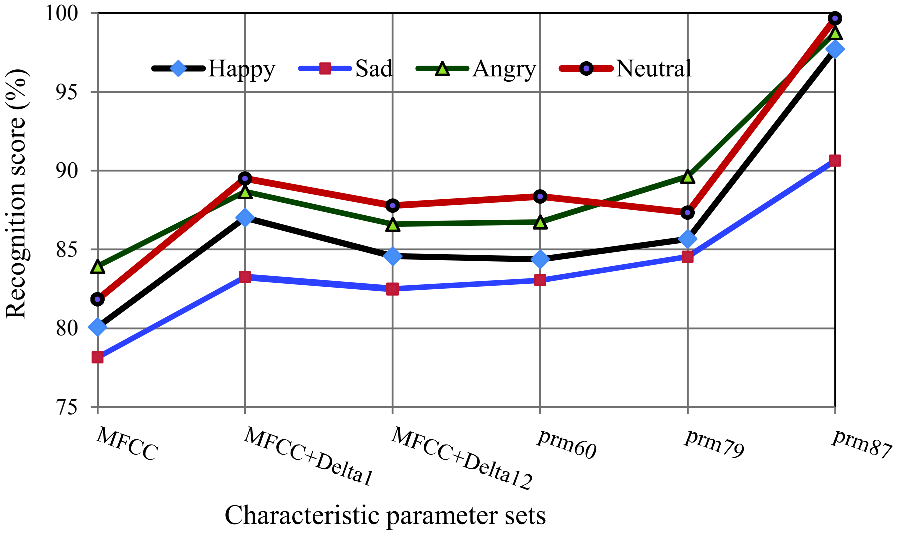

Figure 2. Average of recognition scores for four emotions with characteristic parameter sets in Experiment 1

4.2. Experiment 2: Speaker-dependent and content-independent corpus

The averages of the recognition scores for four emotions with each set of parameters are shown in Figure 3. Figure 3 shows that, when using the prm87 parameter set, the recognition score remains the highest in comparison with the remaining parameter sets and ranges from 93% to 99.11%. With the remaining 5 sets of parameters, the recognition scores range from 72.29% to 85.71%. For MFCC parameter set, the recognition score remains the lowest. Two cases using MFCC+Delta12 and prm60 have the recognition scores almost equal.

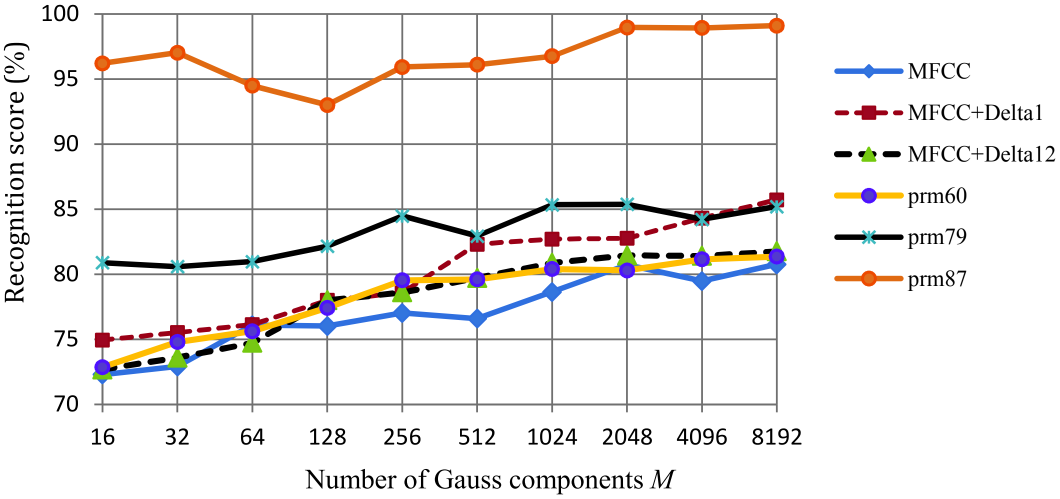

Figure 3. Experiment results with the corpus Test2

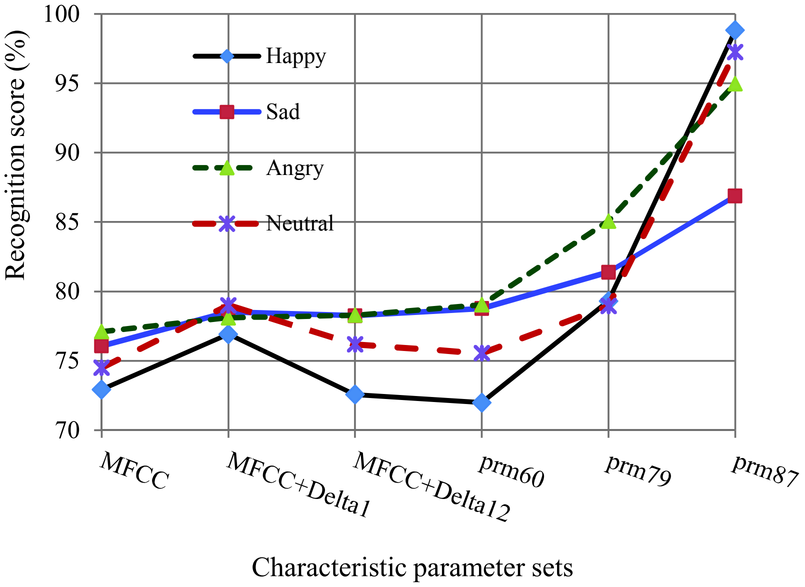

Figure 4. Average of recognition scores for four emotions with characteristic parameter sets in Experiment 2

Figure 4 shows that, with happy emotion, the recognition score is lower than that of the other three emotions when using the MFCC+Delta1, MFCC+Delta12, prm60, prm79 parameter sets. However, the average score of recognition with this emotion increases more strongly than the other three emotions when using the parameter set prm87. With the prm87 parameter set, sad emotion has the lowest exact score compared to the other three emotions. The recognition scores for angry and neutral emotion increase when using parameter prm79 and prm87.

For Experiment 2, if the parameter set prm87 and $M = 128$ are used, the confusion scores for the emotions are the lowest. The confusion scores are summarized in Table 5.

The highest recognition score is 97.98% for happy emotion, and the lowest is 85.09% for angry emotion. Confusion score from sad to neutral emotions is highest and equals to 3.01%. The remaining cases of confusion have a confusion score less than 1%. On average, the recognition score of four emotions is 93% and the confusion score is 0.418%.

Table 5. Confusion recognition scores (%) among emotions using Test2

M=128	**Happy**	**Sad**	**Angry**	**Neutral**
Happy	**97.98**	0	0.29	0
Sad	0	**93.83**	0	3.01
Angry	0	0.85	**85.09**	0
Neutral	0	0.86	0	**95.11**

4.3. Experiment 3: Speaker-independent and content-dependent corpus

Figure 5 is the results of recognition with the Test3 corpus. The results indicate that the set of parameters prm 87 still gives the highest recognition score and the average is 85.44%. Especially, in this expriment, the highest score is 90.14% with $M = 16$ and the lowest is 80.54% with $M = 256$.

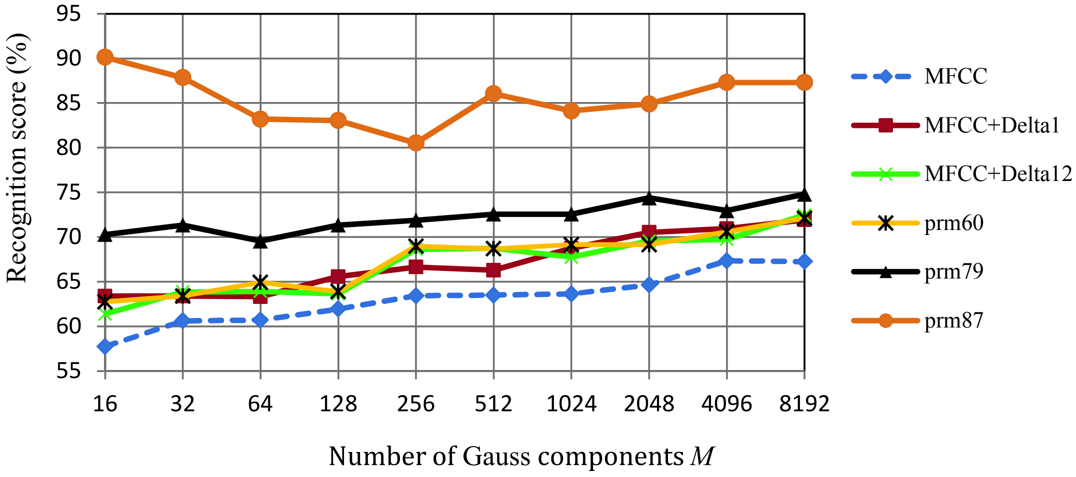

Figure 5. Experiment results with the corpus Test3

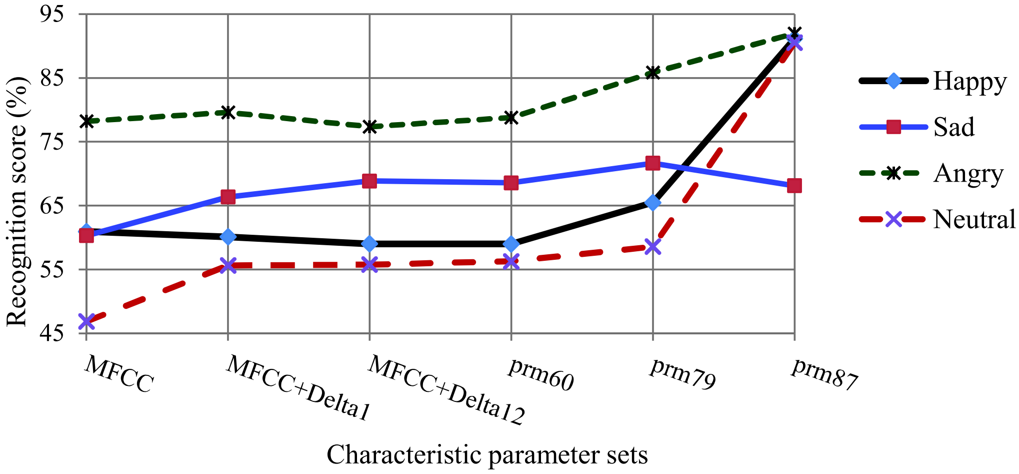

Figure 6. Average of recognition scores for four emotions with characteristic parameter sets in Experiment 3

In the remaining cases, the recognition scores range from 57.75% to 74.79%. As M. increases, the recognition scores for these parameter sets are increased but not much from

66.97% to 67.37% only.

Figure 6 shows that, with angry emotion, the recognition score is higher than the other three emotions when using the corresponding parameter sets. The average of the recognition scores of happy, angry and neutral emotions increase with the use of parameter set prm87. Sad emotion has a declining recognition score with the prm87 parameter set. If only MFCC are used, the recognition score for neutral emotion is minimal.

The confusion score from neutral to sad emotions is 23.42% and is the highest. The average recognition score of four emotions in Experiment 3 is 80.54%, and the average confusion score is 2.7%. The confusion scores (%) are shown in Table 6.

Table 6. Confusion recognition scores (%) among emotions using Test3

***M*=256**	**Happy**	**Sad**	**Angry**	**Neutral**
Happy	**89.6**	0	0.28	0
Sad	0	**63.71**	3	0.29
Angry	5.49	0	**79.19**	0
Neutral	0	23.42	0	**89.66**

4.4. Experiment 4: Speaker-independent and content-independent corpus

With Experiment 4, the recognition score for the prm87 parameter set is significantly higher than the remaining parameter sets. When $M = 1024$, this score is highest at 94.22% while the average recognition score is 90.76%. The remaining sets of parameters have lower recognition scores ranging from 52.69% to 69.40%.

Figure 8 is the recognition scores for each emotion. The recognition score of angry emotion is the best for all parameter sets. Next, the recognition score decreases respectively with the sad, happy and neutral emotions. In general, the recognition scores for emotions are less varied when using the MFCC +Delta1, MFCC+Delta12, prm60 and prm79 parameter sets: happy (52.60% - 62.52%), sad (58.96% - 67.61%), angry (74.21% - 87.44%), neutral (40.55% - 45.09%).

However, when using the prm87 parameter set, the recognition scores of the emotions increase: 97.17% (happy), 98.15% (angry), 97.08% (neutral) except for sad emotion, this score dropps (64.33%) compared with the remaining three emotions.

The confusion score among the emotions is lowest when using the parameter set prm87 with $M = 16$. The confusion scores are summarized in Table 7. Table 7 shows that the highest recognition score is 97.44% for happy emotion, the lowest is 48.01% for sad emotion. Confusion score from neutral to sad emotions is the highest and equal to 25.43% and the confusion score from angry to happy emotions is only 1.14%. The other pair of emotions has a confusion score of 0%. The average recognition score for four emotions is 84.42% and the average confusion score is 2.21%.

4.5. Comparison of experiment results

The average of the recognition scores for four experiments mentioned above is shown in Figure 9. Figure 9 shows that the average recognition score for the emotions in the

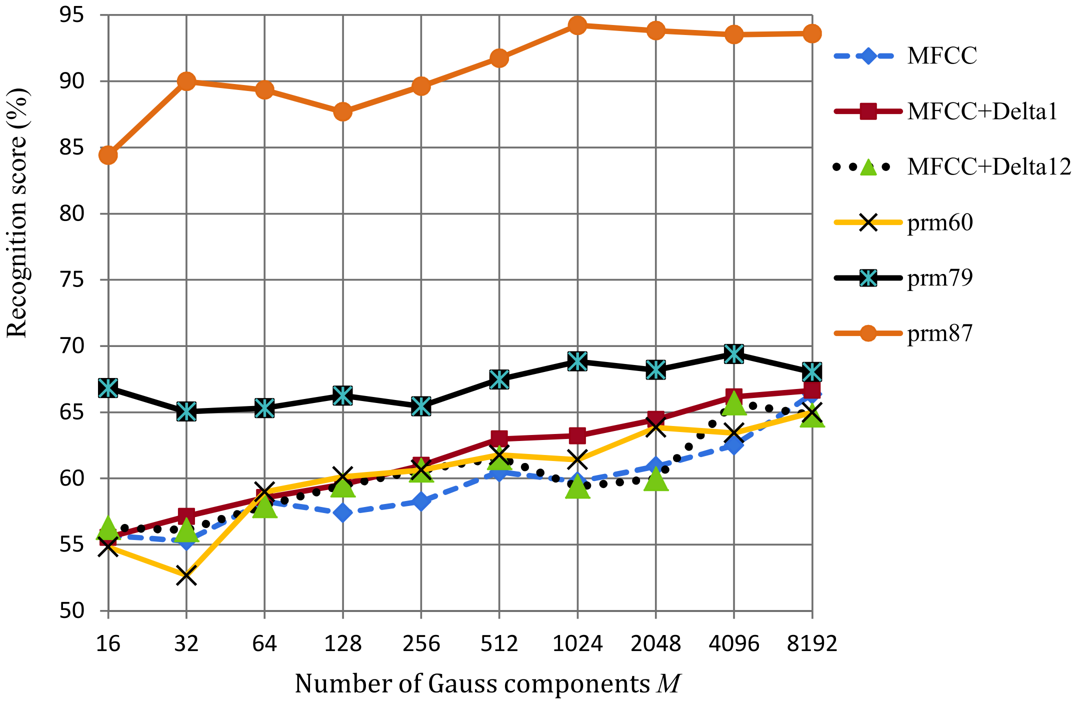

Figure 7. Experiment results with the corpus Test4

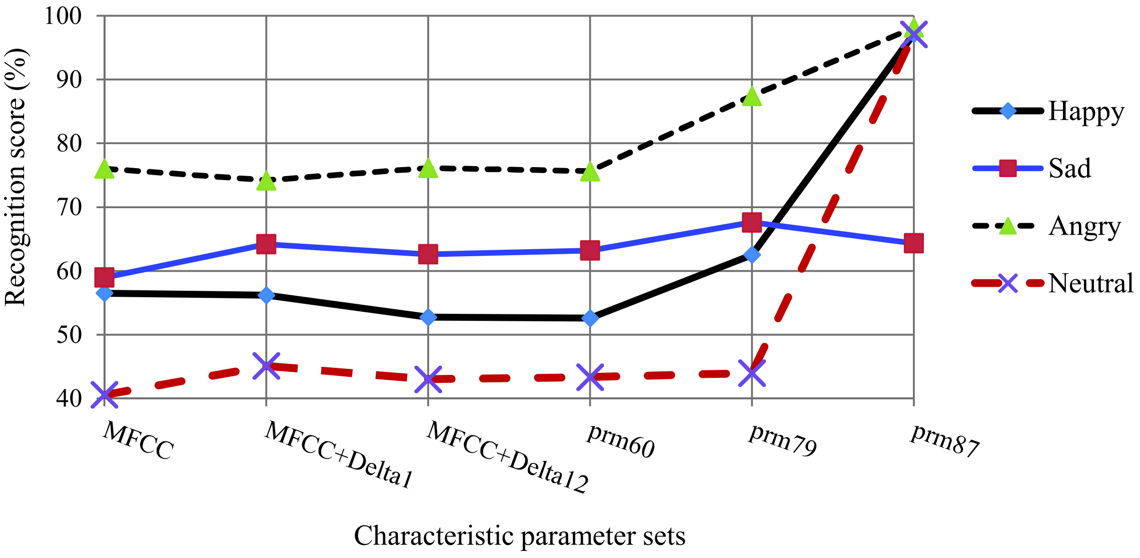

Figure 8. Average of recognition scores for four emotions with characteristic parameter sets in Experiment 4

Experiment 1 is the highest and equal to 89.21% and this score is 82.27%, 70.35% and

Table 7. Confusion recognition scores (%) among emotions using Test4

***M*=16**	**Happy**	**Sad**	**Angry**	**Neutral**
Happy	**97.43**	0	0	0
Sad	0	**48.01**	0	0
Angry	1.14	0	**97.44**	0
Neutral	0	25.43	0	**94.8**

66.99%, respectively for Experiments 2, 3, and 4. This is appropriate because in Experiment 1, the training and recognition phase have the same speaker and the content is the same, only the moments of pronunciation are different, so the recognition scores will reach the highest. For Experiment 4, speakers and content are different for the training and recognition phases. As a result, the recognition scores for this case will be the lowest.

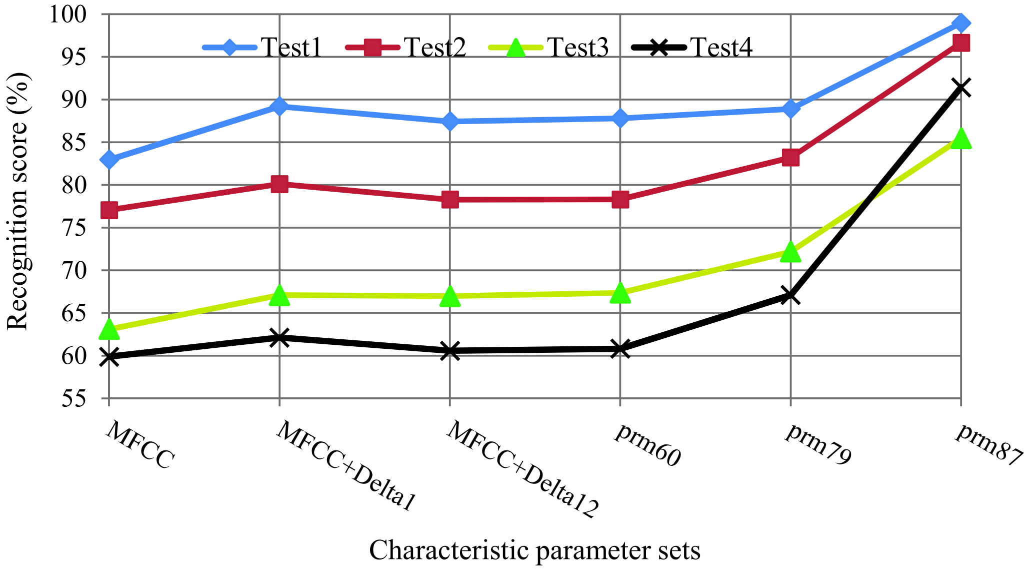

Figure 9. Average of recognition scores for four Experiments

4.6. Number of Gauss components *M*

Figure 10 is the relationship between the number of Gauss components M and the average of recognition scores of the four Experiments mentioned above. Figure 10 shows that with a low M value (between 16 and 512), the recognition scores increase significantly. When M increases from 512 to 8192, the average of recognition scores increases very little.

It can be seen that when M increases sufficiently (over 512), the GMM model almost reaches the approximation of the emotion modeling, the average of recognition scores increases in the form of saturation with increasing M. The optimal determination of the

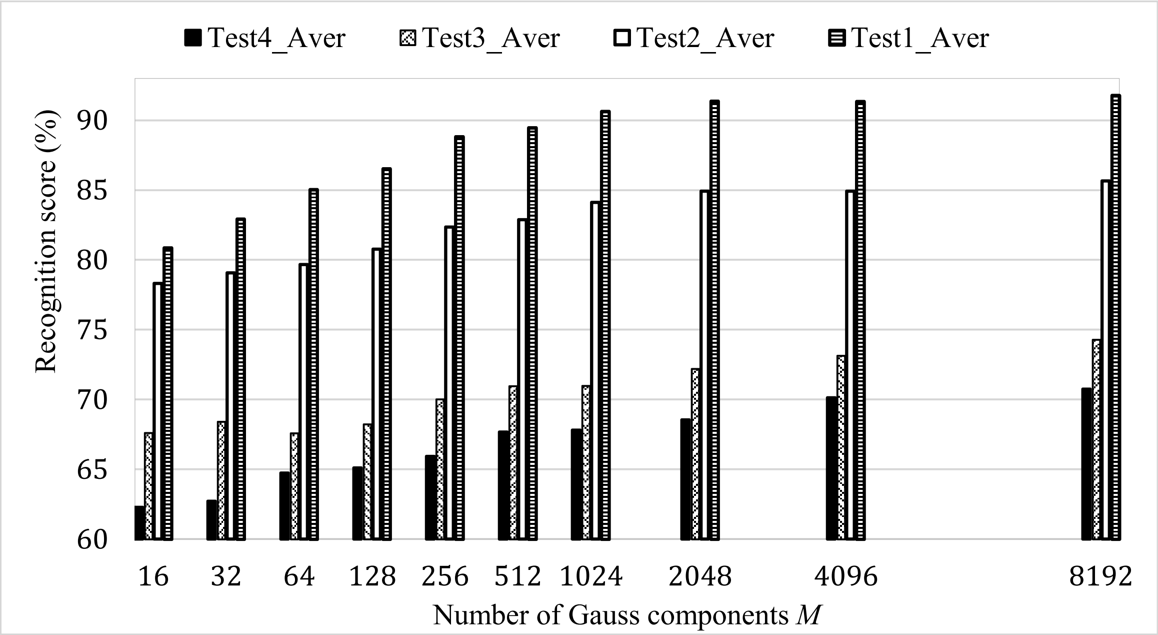

Figure 10. Relationship between the number of Gauss components M and the average of recognition scores

Gauss components M is important but it is also a difficult problem [1]. M increases, the computation time increases as well. Depending on the set of parameters to be included in the recognition, the optimum value of M should be chosen appropriately according to the time required to calculate and the desired recognition accuracy.

4.7. Influence of fundamental frequency on Vietnamese emotional recognition

Our recent study [15] on the individual effects of each spectral feature from (7) to (15) shows that spectral features (11) and (12) have a greater effect on the recognition score than the rest because characteristic parameters (11) and (12) are based on the standard distribution that the GMM uses. In the following, the effect of each $F0$ variant on the score of Vietnamese emotional recognition will be presented. As it can be seen from Fig. 9, when the parameters directly related to $F0$ are added, the recognition score increases significantly compared to the addition of parameters directly related to the spectrum. When adding variants of $F0$ (from prm79 to prm87), the average of recognition scores increases the most strongly for Test4 and the increase was 24.32%. The smallest increase is 10.05% for Test1. However, the smallest increase in this case is still greater than the maximum increase in the case of addition of spectral features (6.29% for Test4). To consider the individual effects of each variant $F0$, all other parameters except the $F0$ variants were preserved.

The number of Gaussian components $M = 512$ was chosen to conduct the evaluation. From 4.6, this value of M can be considered as belonging to the range of the fast-rising recognition score to the slower-rising recognition score when increasing M. Figure 11 is the

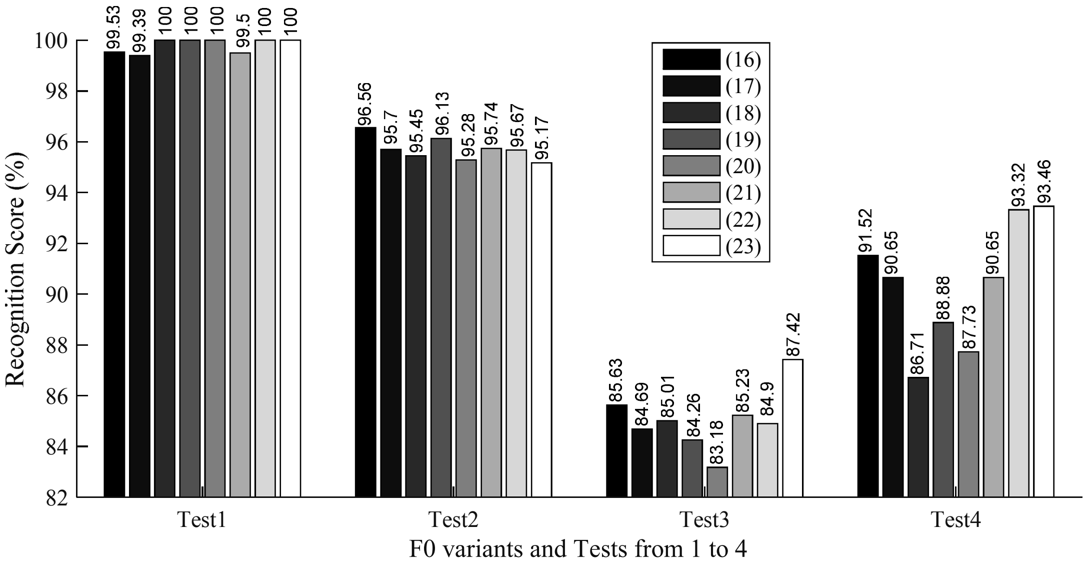

Figure 11. Average of recognition scores for four emotions depends on $F0$ variants, other 79 parameters, and Tests from 1 to 4 with $M = 512$

average of recognition scores for four emotions when one of the eight $F0$ variants was added while the remaining parameters remained unchanged and $M = 512$. The effect of the $F0$ variants is not quite the same for the different Tests. Among four Tests, with the addition of $F0$ variants, Test3 had a lower recognition score. With Test1, $F0$ variants (18), (19), (20), (22) and (23) increased the maximum score of recognition and this score reached 100%. Similar to Test1, Test3 and 4 had the highest score of recognition by adding $F0$ variant (23) and the recognition score is 87.42% and 93.46%, respectively. Meanwhile, $F0$ variant (23) had the least effect on Test2 compared to the other three Tests. When adding $F0$ variant (16), Test2 had the highest score of recognition and this score was 96.56%. Corresponding to Test1, Test2, Test3, and Test4, respectively, $F0$ variants (17), (23), (20), and (18) have the least effect. The significant increase in Vietnamese emotional recognition scores when supplementing the $F0$ variants is perfectly reasonable because $F0$ plays a very important role in the tonal language such as Vietnamese, while $F0$ also participates actively on the emotional expression.

5. CONCLUSION

The paper presents the recognition experiment results for four basic emotions of Vietnamese such as happiness, sadness, neutral, and anger with four different corpora depending on the independence or dependence of the speaker and the content. These experiments were also conducted with six different parameter sets based on the GMM model. The results show that the recognition scores are the highest when speaker-dependent and content-dependent corpus is used. The recognition score is the lowest in the case of speaker-independent and content-independent corpus. With speaker-dependent but content-independent corpus or speaker-independent but content-dependent corpus, the recognition scores are intermediate between the two cases with the highest and lowest recognition scores. In all four experiments, the prm87 parameter set always gave the highest recognition scores. In general, information on fundamental frequency has significantly increased the score of Vietnamese emotional recognition.

REFERENCES

[1] Moataz El Ayadi, Mohamed S. Kamel, Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", *Pattern Recognition*, vol.44, pp. 572–587, 2011.

[2] Do Tien Thang, "Primary examination of Vietnamese intonation", Ha Noi National University Publishing House, 2009.

[3] Dang-Khoa_Mac, Eric Castelli, Véronique Aubergé, "Modeling the prosody of Vietnamese attitudes for expressive speech synthesis", Workshop of Spoken Languages Technologies for Under - resourced Languages (SLTU 2012), Cape Town, South Africa, May 7-9, 2012.

[4] Dang-Khoa Mac, Do-Dat Tran, "Modeling Vietnamese speech prosody: a step-by-step approach towards an expressive speech synthesis system ", *Springer, Trends and Applications in Knowledge Discovery and Data Mining,* vol. 9441, Springer, pp. 273–287, 2015.

[5] Viet Hoang Anh, Manh Ngo Van, Bang Ban Ha, Thang Huynh Quyet, "A real-time model based support vector machine for emotion recognition through EEG", *International Conference on Control, Automation and Information Sciences (ICCAIS)*, Ho Chi Minh city, Vietnam, Nov 26-29, 2012.

[6] La Vutuan, Huang Cheng-Wei, Ha Cheng, Zhao Li, "Emotional feature analysis and recognition from vietnamese speech ", *Journal of Signal Processing,* China, vol.29, no.10, pp 1423–1432, Oct 2013.

[7] Jiang Zhipeng, Huang Chengwei, "High-order markov random fields and their applications in cross-language speech recognition ", *Cybernetics and Information Technologies*,Sofia, volume 15, no. 4, pp 50–57, 2015.

[8] Le Xuan Thanh, Dao Thi Le Thuy, Trinh Van Loan, Nguyen Hong Quang, "Speech emotions and statistical analysis for vietnamese emotion corpus", *Journal of Vietnam Ministry of Information and Communication*, no. 15 (35), pp 86–98, 2016.

[9] Jean-Franois Bonastre, Frédéric Wils, "Alize, a free toolkit for speaker recognition", *IEEE International Conference*, In ICASSP (1), pp. I 737 – I 740, 2005.

[10] Torres-Carrasquillo, P. A., Singer, E., Kohler, M. A., Greene, R. J., Reynolds, D. A., and Deller Jr., J. R., "Approaches to language identification using gaussian mixture models and shifted delta cepstral features ", *In Proc. International Conference on Spoken Language Processing in Denver*, CO, ISCA, pp. 33-36, 82-92, September, 2002.

[11] Bin MA, Donglai ZHU and Rong TONG, "Chinese dialect identification using tone features based on pitch ", *ICASSP,* pp 1029–1032, 2006.

[12] D. Reynolds, C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", *IEEE Trans, Speech Audio Process*, vol. 3, no. 1, 72–83, 1995.

[13] Bacı U., Erzin E., "Boosting Classifiers for Music Genre Classification", In: Yolum., Güngör T., Gürgen F., Özturan C. (eds) Computer and Information Sciences – ISCIS, Lecture Notes in Computer Science, vol 3733. Springer, Berlin, Heidelberg, 2005.

[14] J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models", *Technical Report TR-97-021. International Computer Science Institute (ICSI)*, Berkeley, CA, pp 1–13, 1998.

[15] J.-F. Bonastre, F. Wils, and S. Meignier, "Alize, a free toolkit for speaker recognition", *ICASSP* (1), pp. 737 – 740, 2005.

[16] Dao Thi Le Thuy, Trinh Van Loan, Nguyen Hong Quang, Le Xuan Thanh, "Influence of spectral features of speech signal on emotion recognition of Vietnamese", *Fundamental and Applied IT Research (FAIR)*, pp 36–43, 2017.

[17] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," in *Pattern Recognition and Artificial Intelligence*, C. H. Chen, Ed., Academic, New York, pp. 374–388, 1976.

[18] S.B. Davis, and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[19] Marcel Kockmann, Luka's". Burget, Jan "Honza" C"ernocky, "Application of speaker- and language identification state-of-the-art techniques for emotion recognition", *Speech Communication*, vol. 53, pp 1172–1185, 2011.

[20] R. Subhashree1, G. N. Rathna, "Speech emotion recognition: performance analysis based on fused algorithms and gmm modelling ", *Indian Journal of Science and Technology*, vol 9(11), doi: 10.17485/ijst/2016/v9i11/88460, March 2016.

Received on December 26, 2017
Revised on February 08, 2018