Deep Convolutional Neural Networks for Emotion Recognition of Vietnamese

Thuy Dao Thi Le, Loan Trinh Van, and Quang Nguyen Hong

Abstract—**Human emotions play a very important role in communication. Emotional speech recognition research brings human–machine communication closer to human-to-human communication. This paper presents the evaluation using ANOVA and T-test for the Vietnamese emotional corpus and using deep convolutional neural networks to recognize four basic emotions of Vietnamese based on this corpus: neutrality, sadness, anger, and happiness. Five sets of characteristic parameters were used as inputs of the deep convolutional neural network in which the mel spectral images were taken and attention was paid to the fundamental frequency, F0, and its variants. Experiments were conducted for these five sets of parameters and for four cases, depending on dependent or independent content and dependent or independent speakers. On average, the maximum recognition accuracy achieved was 97.86% under speaker-dependent and content-dependent conditions. The results of the experiments also show that F0 and its variants contribute significantly to the increased accuracy of Vietnamese emotional recognition.**

Index Terms—**Corpus, deep convolutional neural network, emotion, T-test, ANOVA, recognition, fundamental frequency, mel spectrum, Vietnamese.**

I. Introduction

Emotional expression through speech can be regarded as sophisticated human behavior. In addition to speech, people can express emotions through gestures and facial expressions. Understanding the message that the communicator wants to convey through his or her emotional expression plays a very important role in communication between people in everyday life. Likewise, for human–machine interaction to achieve the desired effect (i.e., human-to-human communication), the proper identification of the object of communication during conversation is critical to successful communication. If we look at Reference [1] as early research on speech emotion based on signal processing, it can be said that this field of study has gone through 20 years and has achieved encouraging results [2]. However, the expression of human emotions through speech is very diverse, rich, and sophisticated; hence, research in this area has several difficulties to overcome.

According to statistics [3], we can have up to 300 emotional forms. The difficulties for an emotional recognition system are that the same form of emotion, sadness for example, expresses this state through the same sentence, which may differ according to the context: the reason for sadness, level of sadness, etc. In Vietnamese, for example, the same sad state can be expressed in various ways: *buồn* (sad), *buồn rầu* (sad and melancholy), *buồn rũ rượi* (sad and very tired), *buồn mênh mang* (sadness stretches), *buồn da diết* (sad and tormenting), *buồn cười* (sad and funny), *buồn chán* (sad and bored), *buồn bã* (sad and tired), *buồn tênh* (sad and empty feeling), *buồn phiền* (sadness sorrow), *buồn bực* (sad and angry), *buồn đau* (desolate), or *buồn tuyệt vọng* (sad and depressed). A slight variation in the intonation, intensity, pronunciation, or duration of a word can be interpreted in terms of various emotional states.

To recognize the emotions, besides the models that use EEG signals as input data [4]–[7], there are several models and classifiers that use the characteristic parameters as input data calculated from the speech signal: for example, HMM [8]–[12], GMM [13]–[22], SVM [23]–[29], hierarchical classifiers [30]–[32], kNN [33]–[38], and ANN [26], [29], [39]–[55]. Among these models and classifiers, ANN has been used quite regularly and is now achieving satisfactory results through deep learning and the convolutional neural network (CNN) [39], [43], [53], [44]–[47], [50]–[51], [56]–[58]. In this paper, the model for emotional recognition in Vietnamese is implemented in this direction.

The majority of models, classifiers, and speech features were recapitulated in Reference [59]. For the speech emotion corpus, there are three ways in which to build a corpus: corpus of emotion speech based on simulation, corpus of emotion speech based on elicitation, and corpus of speech based on natural emotion. For the corpus of emotion speech based on simulation, experienced and well-trained artists express various emotions on demand for linguistically neutral sentences. This is an easy way to build an emotional corpus in terms of volume and quality. However, this method tends to express emotions in the corpus more strongly than real emotions. A corpus of speech emotions based on elicitation simulates artificial expressions in which the speaker is unaware of the emotion expressed in the first way. The speaker will naturally express emotions in the context of dialog that has been built. In this case, the quality of the emotional expression is less appropriate if the speaker knows that the dialog is being recorded. In the third method, the corpus can be collected through natural dialog in everyday life. In this case, it would be difficult to meet the volume of the corpus and the manifold expressions of emotion.

Up to now, research on emotions in Vietnamese has been conducted at the linguistic level [60], but little research has been conducted on signal processing. It can be said that the first corpus on Vietnamese emotions was that by Le Thi Xuyen [61]. In her Ph.D. thesis, the Vietnamese emotional corpus consists of five sentences and two speakers (one male and one female). The corresponding sentences in French are

Manuscript received July 9, 2020 (At that time, IJMLC was indexed by Scopus); revised February 17, 2020.

Thuy Dao Thi Le is with Hanoi University of Science and Technology, Vietnam (e-mail: thuydtl@soict.hust.edu.vn).

doi: 10.18178/ijmlc.2020.10.5.992

also spoken by two French speakers. Speakers had to familiarize themselves with the sentences, study them, and somehow repeat their performance before the final recording. Among the five sentences, four are expressed with 12 emotions: neutral*, deception, surprise*, joy*, anger*, contentment (satisfaction), confirmation, boredom*, advice, doubt*, irony*, and regret. The remaining sentence is expressed in seven emotions (the emotions are marked with *). On the basis of this corpus, Le Thi Xuyen studied acoustic signals representing psychological and expressive attitudes, the relationship between acoustic facts and the results of perception tests, and crossed experience in both languages. Khoa M. D. *et al*. in [62]–[64] tested the modeling of Vietnamese prosodies with a multimodal corpus to synthesize expressive Vietnamese. The corpus of this research was built with one male voice. Duyen N. T. and Duy B. T. [65] proposed a modified Vietnamese-speaking model to create an emotional expression in the voice channel for a Vietnamese-speaking virtual person. In this study, emotional vocabulary included the Vietnamese pronunciation of a male artist and a female artist pronouncing 19 sentences in five emotions: natural, happy, sad, angry, and very angry. For emotional recognition in Vietnamese, the study [66] used SVM to classify emotions with the input using the EEG signal. The results show that five emotional states can be identified in real time with an average accuracy of 70.5%. For the research in Reference [67], the corpus contains two male voices and two female voices with six sentences for six emotions: happiness, neutrality, sadness, surprise, anger, and fear. The characteristic parameters were MFCC, short-term energy, pitch, and formants, which were used with the GMM model. The highest recognition score was 96.5% for a neutral emotion, and the lowest was 76.5% for sadness. The corpus for Reference [68] consists of six voices with 20 sentences and the emotions as in Reference [67]. In Reference [68], with Im-SFLA SVM, the recognition score on the Vietnamese language reached 96.5% for neutrality and dropped to 84.1% for surprise. For the first time, the emotional corpus for the Vietnamese language BKEmo based on the simulation method by recording artists' voices expressing the desired emotions was built at Hanoi University of Science and Technology. See References [69], [70] for more details on this corpus. With the corpus BKEmo, we performed the recognition of four emotions using the GMM model: happiness, neutrality, sadness, and anger [70]. The parameters for our GMM model are MFCC and its first and second derivatives, fundamental frequency, energy, formants and the corresponding bandwidths, spectral characteristics, and F0 variants. The average score of recognition was 77.21%.

In this paper, the deep CNN model will be used to recognize the four emotions mentioned above with the same BKEmo corpus. However, as described below, the use of input characteristic parameters for deep CNN is different from the GMM model in Reference [70].

The rest of the paper is organized as follows. Section II describes the additional evaluation of the corpus for the emotional recognition of Vietnamese. Section III describes the deep CNN configuration for training experiments. Section IV details the parameter sets used for the experiments. The corpus used for the experiments is described in Section V, and the experimental results are presented and discussed in Section VI. Section VII concludes the paper.

II. Evaluation of the Corpus

A. ANOVA Analysis

In Reference [69], using ANOVA and T-test, we evaluated the corpus BKEmo with nine characteristic parameters of the speech signal spectrum: harmonicity, center of gravity, standard deviation, skewness, kurtosis, central spectral moment, mean, slope, and standard deviation of LTAS (long-term average spectrum). The results show that the above parameters affect the differentiation of four emotions. In this paper, we perform additional evaluation of the remaining 18 characteristic parameters, including intensity, four formants and the corresponding bandwidths, and F0 and F0 variants (dF0, F0NormAver, F0NormMinMax, F0NormAverStd, LogF0NormAver, dLogF0, LogF0NormMinMax, and LogF0NormAverStd). See Reference [70] for details on the F0 variants. All the parameters of the speech signal used in our experiments were calculated using Praat [71].

Next, this section presents the results of using one-way ANOVA to evaluate the influence of the 18 characteristic parameters of the speech signal mentioned above. Table I presents the ANOVA results with statistical values of F and P-value for the 18 characteristic parameters.

TABLE I: P-Value of ANOVA for Characteristic Parameters with Pairs of Emotions

No.	Parameters	F - value	P-value
1	Intensity	877.2506	0.0
2	Formant1	414.7949	3.008715E-243
3	Band1	423.7496	5.269112E-248
4	Formant2	1066.9129	0.0
5	Band2	417.9563	6.265008E-245
6	Formant3	418.7876	2.265692E-245
7	Band3	342.0899	6.349E-204
8	Formant4	198.3841	2.954376E-122
9	Band4	321.2435	2.071804E-192
10	F0	453.9051	6.902E-264
11	dF0	25.6372	1.8217E-16
12	F0NormAver	3241.4807	0.0
13	F0NormMinMax	257.7462	1.1788E-156
14	F0NormAverStd	3241.481	0.0
15	dLogF0	91.85127	5.104602E-58
16	LogF0NormMinMax	186.8734	1.792956E-115
17	LogF0NormAver	1168.8149	0.0
18	LogF0NormAverStd	3001.683	0.0

Table I shows that the P-value is very small (≅0), which means that the probability of the expected values being the same for emotions is very low. Therefore, the hypothesis of nondistinction between emotions is rejected, and one can assert that emotions can be differentiated based on these parameters. The P-values in the tables of this paper are calculated with MatLab using double precision and are retained in the format rendered by MatLab. In fact, very small values such as xxxE−243, xxxE−248, and so on can be considered as a value of 0.

B. T-Test Results

Because of the T-test, we can identify which pairs of emotions can be distinguished from each other based on the above parameters.

TABLE II: P-VALUE OF T-TEST FOR CHARACTERISTIC PARAMETERS WITH PAIRS OF EMOTIONS

No.	Parameters	P-value					
		Neutrality-Sadness	Neutrality-Anger	Neutrality-Happiness	Sadness-Anger	Sadness-Happiness	Anger-Happiness
1	Intensity	3.7683E-09	0.4524	3.7683E-09	3.7683E-09	3.7683E-09	3.7683E-09
2	Formant1	3.7683E-09	0.0032784	3.7683E-09	3.7683E-09	3.7683E-09	3.7683E-09
3	Band1	3.7683E-09	3.7683E-09	3.7694E-09	4.0303E-09	3.7683E-09	3.7683E-09
4	Formant2	3.7683E-09	3.9206E-09	3.7683E-09	3.7683E-09	3.7683E-09	3.7683E-09
5	Band2	3.7683E-09	3.7683E-09	0.00011336	3.7683E-09	3.7683E-09	3.7683E-09
6	Formant3	3.7683E-09	3.7683E-09	7.2108E-05	3.7683E-09	3.7683E-09	3.7683E-09
7	Band3	3.7683E-09	3.7683E-09	3.7683E-09	3.7683E-09	3.8514E-09	3.7683E-09
8	Formant4	3.7683E-09	3.7683E-09	3.7683E-09	3.7683E-09	0.015535	3.7683E-09
9	Band4	3.7683E-09	0.0013824	3.7683E-09	3.7683E-09	3.7683E-09	3.7683E-09
10	F0	3.7683E-09	0.018149	3.7683E-09	3.7683E-09	3.7683E-09	3.7683E-09
11	dF0	2.8295E-07	0.0092189	0.99973	3.7683E-09	1.6984E-07	1.6984E-07
12	F0NormAver	3.7683E-09	3.7683E-09	3.7683E-09	3.7683E-09	3.7683E-09	3.7683E-09
13	F0NormMinMax	3.7683E-09	0.014222	3.7683E-09	3.7683E-09	3.7683E-09	3.7683E-09
14	F0NormAverStd	3.7683E-09	3.7683E-09	3.7683E-09	3.7683E-09	3.7683E-09	3.7683E-09
15	dLogF0	3.7683E-09	0.18898	0.42845	3.7683E-09	3.7683E-09	0.96305
16	LogF0NormMinMax	4.4826E-07	3.84E-09	3.7683E-09	3.7683E-09	3.7683E-09	3.7683E-09
17	LogF0NormAver	3.7683E-09	3.7683E-09	3.7683E-09	3.7683E-09	3.7683E-09	3.7683E-09
18	LogF0NormAverStd	3.7683E-09	3.7683E-09	3.7683E-09	3.7683E-09	3.7683E-09	3.7683E-09

Table II shows that the P-values are very small in the evaluation of each parameter for each pair of emotions. It is known that in the majority of cases, a P-value of 0.05 is used as the cutoff for significance [72]. If the P-value is less than 0.05, the null hypothesis that there is no difference between the emotion pairs will be rejected and a significant difference does exist. The emotion pairs sad–happy and sad–angry are best distinguished for most of the 18 characteristic parameters. This distinction is in line with the fact that these two pairs of emotions are easy to distinguish by hearing them. The remaining emotion pairs are also distinctive; however, the mean, standard deviation of LTAS, and intensity parameters are less affected in the neutral–angry pair. The parameters dF0 and dLogF0 affect the angry–happy pair less. For this pair, it is also difficult to distinguish using the dLogF0 parameter.

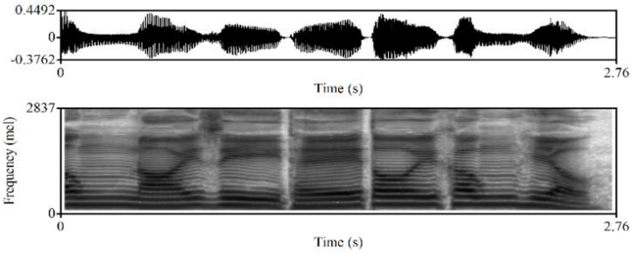


Fig. 1. Example of the speech signal's mel spectrum as the input image for layer 1 in the case of the baseline model.

III. DEEP CNN CONFIGURATION FOR EXPERIMENTS

For the emotional recognition of Vietnamese, the full configuration of the deep CNN for training is described in Table III in the case of the baseline model with 260 parameters. For models with parameter numbers greater than 260, the network configurations can be easily deduced in a similar way.

For layer 1, the input image is 260 ×260 (260 mel spectrum coefficients ×260 frames). The chosen frame number is 260 because for the files used for experiments, if the frame width is 0.025 s and the frame shift is 0.01 s, the total number of frames will be around 260 for most files. An example of the mel spectrum of the speech signal as the input image for layer 1 in the case of the baseline model is given in Fig. 1. After convolution using a 3 ×3 moving filter with padding, there are 64 feature maps with a size of 260 × 260. The next operation of layer 1 is batch normalization, followed by a nonlinear activation function ELU, max pooling with 2 ×2 windows, and finally dropout with a coefficient of 0.5. The operation of the convolution was carried out in a turn from left to right, top to bottom. The purpose of the convolution operation is to determine the probability of occurrence of samples at certain locations in the image. Formula (1) describes the convolution operation at a particular location:

$$y = \sum_i w_i x_i + b \tag{1}$$

Here, x_i consists of spectral pixels in the scanned window (i.e., 3 ×3 spectral pixels), w_i are weights to learn, and b is bias. If K is the number of input pictures, M is the number of feature maps, and the number of parameters for a convolution operation is $M \times (K \times (\text{moving filter size}) + 1)$. For each layer, the goal of batch normalization is to achieve a stable distribution of activation values throughout training and thereby yield a substantial speedup in training [73]. ELU speeds up learning in deep neural networks and leads to higher classification accuracies [74]. Max pooling reduces the number of model parameters, also known as downsampling or subsampling, while also making the detection of features invariant to orientation changes or scale [75]. Finally, dropout is considered as a means of preventing neural networks from overfitting [76].

Layers 2, 3, and 4 perform the same operations, convolution using a 3 ×3 moving filter with padding, and the outputs of these convolutions are 128 feature maps with the dimensions 130 × 130, 65 × 65, and 32 × 32, respectively. After the convolution, these layers also undergo batch normalization, ELU, max pooling, and dropout. Finally, for layer 5, after batch normalization, ELU, max pooling, and dropout, we have a fully connected layer with 256 inputs and four outputs corresponding to four emotions. The transfer function of the fully connected layer is Softmax, representing the probability distribution for each emotion.

Usually neural networks with three or more hidden layers can be considered as deep neural networks. Reference [77] gives an example of CNN architecture that also has five layers, but this deep CNN is used for image classification.

TABLE III: CNN CONFIGURATION FOR EMOTION RECOGNITION OF VIETNAMESE

Layer Index	Layer (type)	Output Shape	Param #
1	BatchNormalization	(260, 260, 1)	1040
	Convolution2D (3x3)	(260, 260, 64)	640
	BatchNormalization	(260, 260, 64)	256
	ELU	(260, 260, 64)	0
	MaxPooling2D (2x2)	(130, 130, 64)	0
	Dropout (0,5)	(130, 130, 64)	0
2	Convolution2D(3x3)	(130, 130, 128)	73856
	BatchNormalization	(130, 130, 128)	512
	ELU	(130, 130, 128)	0
	MaxPooling2D(2x2)	(65, 65, 128)	0
	Dropout(0,5)	(65, 65, 128)	0
3	Convolution2D(3x3)	(65, 65, 128)	147584
	BatchNormalization	(65, 65, 128)	512
	ELU	(65, 65, 128)	0
	MaxPooling2D(2x2)	(32, 32, 128)	0
	Dropout(0,5)	(32, 32, 128)	0
4	Convolution2D(3x3)	(32, 32, 128)	147584
	BatchNormalization	(32, 32, 128)	512
	ELU	(32, 32, 128)	0
	MaxPooling2D(2x2)	(10, 10, 128)	0
	Dropout(0,5)	(10, 10, 128)	0
5	Convolution2D(3x3)	(10, 10, 64)	73792
	BatchNormalization	(10, 10, 64)	256
	ELU	(10, 10, 64)	0
	MaxPooling2D(2x2)	(2, 2, 64)	0
	Dropout (0.5)	(2, 2, 64)	0
	Fully connected	(4)	1028

TABLE IV: FIVE SETS OF PARAMETERS

Number of parameters	Parameters
260	260 Mel spectrum coefficients
264	260 Mel spectrum, F0 and three F0 variants: F0NormMinMax, LogF0NormAver, LogF0NormMinMax
267	264 parameters and three F0 variants: F0NormAver, F0NormAverStd, LogF0NormAverStd
294	260 Mel spectrum, intensity, F0 and five F0 variants: F0NormMinMax, LogF0NormMinMax, F0NormAver, F0NormAverStd, LogF0NormAverStd, 4 formants and its corresponding bandwidths, harmonicity, center of gravity, central moment, skewness, kurtosis and 14 coefficients of inverse filter of the vocal tract
296	294 parameters and two F0 variants: dF0, LogF0NormAver

IV. PARAMETER SETS USED FOR EXPERIMENTS

The parameters are divided into five sets and detailed in Table IV.

V. TESTS FOR EXPERIMENTS

The method of dividing the corpus to perform the tests in this paper is similar to that used in Reference [70], which means that there are four sets of corpus for the four tests. The tests are denoted by *Ti* (with *i* = 1, 2, 3, 4).

Test1 (T1): Speaker-dependent and content-dependent corpus.

Test2 (T2): Speaker-dependent and content-independent corpus.

Test3 (T3): Speaker-independent and content-dependent corpus.

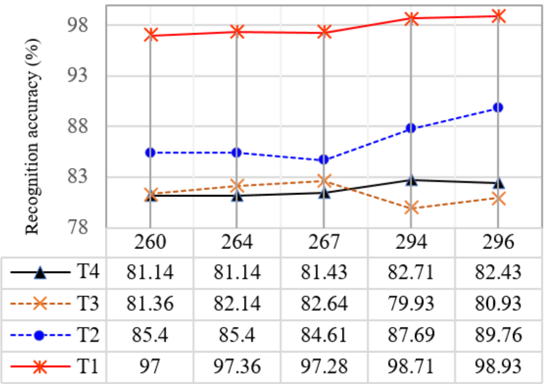
Test4 (T4): Speaker-independent and content-independent corpus.

For Test1, speaker-dependent and content-dependent mean that the same speaker expresses the same content but expresses at different times. This still has a practical meaning because of the variability of the speech signal. The same person speaks the same sound at different times; however, the speech signals of different times are not quite the same [78].

In this paper, the number of voice files per test is divided into a 2–1–1 proportion corresponding to the training-valid-test.

VI. EXPERIMENTAL RESULTS FOR FOUR EMOTIONS WITH DIFFERENT PARAMETERS AND TESTS

Fig. 2 shows the experimental results for four emotions with various parameters and tests. For the experiments $\mathrm{Ti} - 260$ and $\mathrm{Ti} - 264$ $(i = 1, 2, 3, 4)$, the recognition accuracy increases from T4 to T1. In the case of 260 parameters, the largest increase in recognition accuracy was 15.86% ($\mathrm{T1} - 260$ vs. $\mathrm{T4} - 260$), and the smallest increase in recognition accuracy was 0.22% ($\mathrm{T3} - 260$ vs. $\mathrm{T4} - 260$). In the case of 264 parameters, the largest increase in recognition accuracy was 16.22% ($\mathrm{T1} - 264$ vs. $\mathrm{T4} - 264$), and the smallest increase in recognition accuracy was 1.0% ($\mathrm{T4} - 264$ vs. $\mathrm{T3} - 264$). When we increased the number of parameters from 260 to 264, the T4 and T2 tests did not increase the recognition accuracy, but the T1 and T3 tests did.



	260	264	267	294	296
▲ T4	81.14	81.14	81.43	82.71	82.43
× T3	81.36	82.14	82.64	79.93	80.93
● T2	85.4	85.4	84.61	87.69	89.76
✱ T1	97	97.36	97.28	98.71	98.93

Fig. 2. Recognition accuracy average for four emotions with different parameters and tests.

For the $\mathrm{Ti} - 267$ tests $(\mathrm{i} = 1, 2, 3, 4)$, the $\mathrm{T1}, \mathrm{T2}$, and T4 tests all increased the recognition accuracy. The highest increase in recognition accuracy was 15.85% ($\mathrm{T1} - 267$ vs. $\mathrm{T4} - 267$), and the smallest increase in recognition accuracy was 1.21% ($\mathrm{T3} - 267$ vs. $\mathrm{T4} - 267$). In the case of $\mathrm{T2} - 267$, the recognition accuracy decreased by 0.79% compared with $\mathrm{T2} - 264$.

In the case of 294 parameters, the $\mathrm{T1}, \mathrm{T2}$, and T4 tests all increased the recognition accuracy. In particular, the recognition accuracy in test $\mathrm{T2} - 294$ increased significantly (3.08%) compared with that in $\mathrm{T2} - 267$. In contrast, the recognition accuracy in test $\mathrm{T3} - 294$ was reduced (2.71%)

from that in test $T3-267$. As the number of parameters increased to 296, the T1 and T2 tests had the highest recognition accuracy compared with the other cases. The recognition accuracy reached 98.93% for test T1 and 89.76% for test T2.

Therefore, in the four tests with five parameter sets, the T1 and T2 tests achieved the highest recognition accuracy with 296 parameters. The highest recognition accuracy of the T3 test is with 267 parameters; however, the highest recognition accuracy of the T4 test is with 294 parameters. The average recognition accuracy of the four tests and the four emotions for each set of parameters is given in Fig. 3.

Fig. 3 shows that the average recognition accuracy of the tests was highest when using 296 parameters and lowest when using 260 parameters. This result also shows that the addition of energy, spectral characteristics, the fundamental frequency F0, the variants of F0, the four formants, and the corresponding bandwidths increased the recognition accuracy.

These characteristics are parameters that have a strong influence on the ability to differentiate emotions and have been analyzed by one-way ANOVA and T-test, as reported in the above section. The influence of the fundamental frequency through the dF0 and LogF0NormAver parameters as they are used in the 296 parameters set has improved the recognition accuracy (from 87.26% to 88.01%).

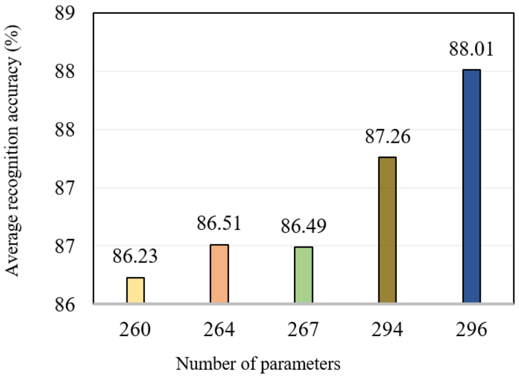


Fig. 3. Average accuracy of recognition for four tests and four emotions.

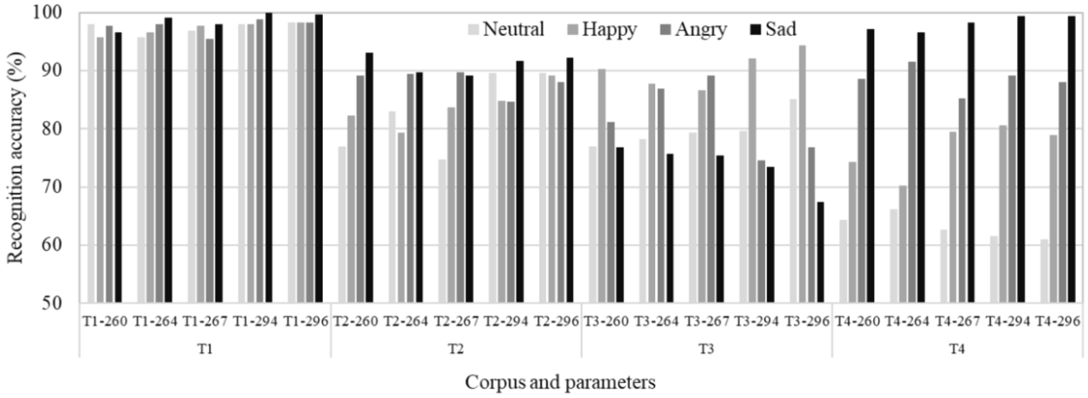


Fig. 4. Highest recognition accuracy of each emotion for each experiment.

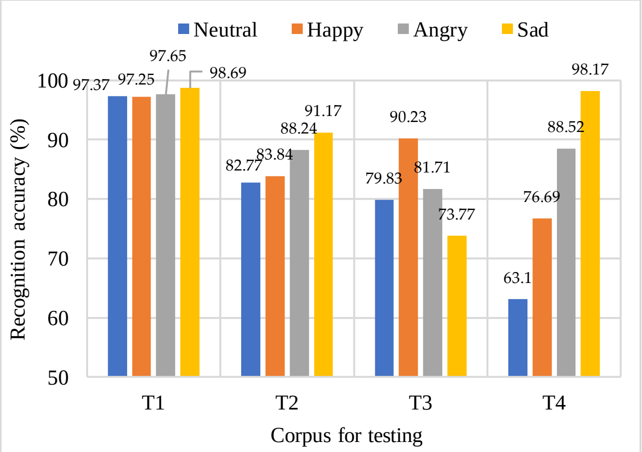


Fig. 5. Average recognition accuracy of each test for each emotion.

The highest recognition accuracy of each emotion for each experiment is given in Fig. 4. Fig. 4 shows that, in general, sadness has the highest recognition accuracy compared with the other emotions in the T1, T2, and T4 tests but was low for the T3 test. This result is shown in Fig. 5, which demonstrates each emotion's average recognition accuracy for the tests. For the T1 test, the highest recognition accuracy was achieved, and the percentage was not significantly different (95.42%–100%). With the T4 test, neutral emotion has a greater disparity of recognition accuracy than sadness.

In terms of the average recognition accuracy of all the tests for each set of parameters, the mean percentages of sadness were highest (90.54%), followed by anger (89.03%), happiness (87%), and neutral emotion (80.77%). The average recognition accuracy of each test for each emotion is shown in Fig. 5.

For the case using 260 parameters, these are universal parameters for different languages. F0 is an important parameter for both tonal language and language without tone. Furthermore, depending on the language, specific language parameters such as voice quality can be exploited.

For 296 parameters, 14 coefficients of the inverse filter of the vocal tract take up a significant amount. These coefficients carry information about the vocal tract. However, it can be seen that the significant increase in these coefficients does not significantly increase the recognition accuracy when compared with increasing the number of parameters related to F0. This shows the importance of F0 for emotional recognition and is also worth noting when considering the trade-off between number of parameters and average accuracy, in terms of training time and costs.

The experiments were performed on two machines with Intel Core i7-4790 CPU @ 3.60 GHz × 8, 16 GB of RAM, GPU NVIDIA GeForce GTX 1050 Ti/PCIe/SSE2, 4 GB of RAM. The average training time for this deep CNN is about

6 days for one of four tests (*Ti*, *i* = 1, 2, 3, 4) with one of five parameter sets. Among the characteristic parameters used, the effect of spectral parameters on emotional recognition accuracy for the GMM model was studied in Reference [69]. The influence of parameters directly related to F0 on emotional recognition accuracy for the GMM model was presented in Reference [70]. These might represent the information required to decide on the optimum number of parameters.

VII. Conclusion

The results of ANOVA and T-test showed the ability to distinguish emotions from the BKEmo corpus by using the speech signal's characteristic parameters. The recognition results of four emotions using deep CNN are also consistent with these evaluation results of the ANOVA and T-test for the corpus. The feature parameter set for our deep CNN includes the characteristic parameters for the sound source and vocal tract. For the characteristic parameters, the fundamental frequency F0 is the sound source parameter and is important for Vietnamese because it is a tonal language. The variation law of fundamental frequency also depends on the emotion that needs to be expressed. Overall, the accuracy of Vietnamese emotional recognition achieved with deep CNN is very positive, and the experiment's results show that information on the fundamental frequency F0 improved the accuracy in emotional recognition. The corpus BKEmo is shared for the deep CNN model of this paper and for the GMM model in Reference [70]. However, it would be unreasonable if the comparison of the recognition scores of these two models is carried out to conclude which model has the higher recognition score. In fact, the characteristic parameters used by these two models are different in our case. However, both the deep CNN and GMM models have improved the recognition scores by adding more information on F0 and the F0 variants. For a preliminary comparison between the GMM and ANN models, it can be said that besides learning ability, the ANN model has promising prospects toward building recognition systems in general because of diversity in the architecture of neural networks. Our next work is to extend the corpus to other forms of emotion in Vietnamese and perform recognition for these forms of emotion. One problem for recognition systems including emotional recognition is that data in the real environment may not belong to a training or a testing set. In such cases usually, the recognition accuracy may be reduced. To approach this issue in emotional recognition, there have been studies using transfer learning [79]–[82], and this will also be one of the upcoming research directions for us to recognize emotions.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Dao Thi Le Thuy, Trinh Van Loan conducted the research; Dao Thi Le Thuy, Nguyen Hong Quang analyzed the data; Dao Thi Le Thuy, Trinh Van Loan wrote the paper; all authors had approved the final version.

Acknowledgment

The authors of the paper would like to thank the reviewers for their valuable comments to help our paper improve.

References

[1] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proc. ICSLP 3*, 1996, pp. 1970–1973.

[2] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90-99, May 2018.

[3] M. Schubiger, "English intonation: Its form and function," *Language*, vol. 36, no. 4, pp. 544-548, 1960.

[4] L. J. M. Rothkantz, R. Horlings, and Z. Dharmawan, "Recognition of emotional states of car drivers by EEG analysis," *Neural Network World*, no. 1, pp. 119-128, 2009.

[5] N. Zhuang, Y. Zeng, L. Tong, C. Zhang, H. M. Zhang, and B. Yan, "Emotion recognition from EEG signals using multidimensional information in EMD domain," *Biomed Research International*, 2017.

[6] M. Li, H. P. Xu, X. W. Liu, and F. Lu, "Emotion recognition from multichannel EEG signals using K-nearest neighbor classification," *Technology and Health Care*, vol. 26, no. S1, pp. 509-519, 2018.

[7] K. Giannakaki, G. Giannakakis, C. Farmaki, and V. Sakkalis, "Emotional state recognition using advanced machine learning techniques on EEG data," in *Proc. IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, Thessaloniki, 2017, pp. 337-342.

[8] T. L. New, S. W. Foo, and L. C. D. Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, no. 4, pp. 603-623, 2013.

[9] B. Vlasenko and A. Wendemuth, "Tuning hidden Markov model for speech emotion recognition," *Fortschritte der Akustik*, vol. 33, no. 1, pp. 317-318, 2007.

[10] C. Lee, S. Yildrim, M. Bulut, A. Kazemzadeh *et al.*, "Emotion recognition based on phoneme classes," in *Proc. Int. Conf. on Spoken Language Processing*, 2004, pp. 889-892.

[11] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proc. IEEE International Conference on Acoustics Speech and Signal Processing*, 2003, vol. 2, II 1- II 4.

[12] D. Ozkan, S. Scherer, and L.-P. Morency, "Step-wise Emotion Recognition Using Concatenated-HMM," in *Proc. the 14th ACM International Conference on Multimodal Interaction*, 2012, Santa Monica, California, USA - October 22 – 26, pp. 477-484.

[13] N. Vlassis and A. Likas, "A kurtosis-based dynamic approach to gaussian mixture modeling - Part A: Systems and humans," *IEEE Trans. Man, and Cybernetics*, vol. 29, no. 4, 1999.

[14] B. Robert, D. A. D. Reynolds *et al*., "Approaches to speaker detection and tracking in conversational speech," *Digital Signal Processing*, vol. 10, pp. 93-112, 2000.

[15] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[16] M. M. H. Elayadi, M. S. Kamel, and F. Karray, "Speech emotion recognition using Gaussian mixture vector autoregressive models," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007, pp. IV 957 - IV 960.

[17] K. K. Inakollu and S. Kocharla, "Gender dependent and independent emotion recognition system for Telugu speeches using Gaussian mixture models," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, issue 11, pp. 4172-4175, 2013.

[18] R. Subhashree and G. N. Rathna, "Speech emotion recognition: Performance analysis based on fused algorithms and GMM modelling," *Indian Journal of Science and Technology*, vol. 9, no. 11, pp. 1-8, 2016.

[19] A. Utane and S. Nalbalwar, "Emotion recognition through speech using Gaussian Mixture model and hidden Markov model," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, issue 4, pp. 742-746, 2013.

[20] V. Dimitrios and C. Kotropoulos, "Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm," in *Proc. IEEE International Conference on Multimedia and Expo*, 2005, pp. 1500-1503.

[21] B. Yang and M. Lugger, "Emotion recognition from speech signals using new harmony features," *Signal Processing*, vol. 90, pp. 1415–1423, 2010.

[22] L.-C. Ramón, J. Silovsky, and M. Kroul, "Enhancement of emotion detection in spoken dialogue systems by combining several

information sources," *Speech Communication*, vol. 53, pp. 1210–1228, 2011.

[23] A. Batliner *et al*, "Whodunnit – Searching for the most important feature types signaling emotion-related user states in speech," *Computer Speech and Language*, vol. 25, pp. 4–28, 2011.

[24] H. W. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Computer Speech and Language*, vol. 29, pp. 186–202, 2015.

[25] W. H. Dai, D. M. Han, Y. H. Dai, and D. R. Xu, "Emotion recognition and affective computing on vocal social media," *Information & Management*, vol. 52, pp. 777–788, 2015.

[26] X. X. Ke, Y. J. Zhu, L.Wen, and W. Z. Zhang, "Speech emotion recognition based on SVM and ANN," *International Journal of Machine Learning and Computing*, vol. 8, no. 3, pp. 198-202, June 2018.

[27] R. Gajšek, F. Mihelič, and S. Dobrišek, "Speaker state recognition using an HMM-based feature extraction method," *Computer Speech and Language*, vol. 27, no. 1, pp. 135-150, January 2013.

[28] T. Iliou and C.-N. Anagnostopoulos, "Classification on speech emotion recognition — a comparative study," *International Journal on Advances in Life Sciences*, vol. 2, pp. 18-28, 2010.

[29] R. Jan and A. Janicki, "Comparison of speaker dependent and speaker independent emotion recognition," *Applied Mathematics and Computer Science*, vol. 23, pp. 797-808, 2013.

[30] E. M. Albornoz, D. H. Milonea, and H. L. Rufiner, "Spoken emotion recognition using hierarchical classifiers," *Computer Speech & Language*, vol. 25, pp. 556-570, 2011.

[31] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, issues 9–10, pp. 1162-1171, November–December 2011.

[32] R. Fernandez and R. Picard, "Recognizing affect from speech prosody using hierarchical graphical models," *Speech Communication*, vol. 53, pp. 1088–1103, 2011.

[33] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vo. 49, issues 10–11, pp. 787-800, October–November 2007.

[34] S. L. Jing, X. Mao, and L. J. Chen, "Prominence features: Effective emotional features for speech emotion recognition," *Digital Signal Processing*, vol. 72, pp. 216–231, 2018.

[35] C. M. Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2001, pp. 240-243.

[36] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293-303, 2005.

[37] C. M. Lee, S. S. Narayanan, and R. Pieraccini, "Classifying emotions in human-machine spoken dialogs," in *Proc. IEEE International Conference on Multimedia and Expo*, 2002, vol.1, pp. 737-740.

[38] C. M. Lee *et al*., "Combining acoustic and language information for emotion recognition," in *Proc. the 7th International Conference on Spoken Language Processing*, Denver, Colorado, USA, September 16-20, 2002, pp. 873-876.

[39] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *InterSpeech*, pp. 1263-1267, 2017.

[40] H. Kaya and A. A. Karpov, "Efficient and effective strategies for cross-corpus acoustic emotion recognition," *Neurocomputing*, vol. 275, pp. 1028–1034, 2018.

[41] G. H. Wen, H. H. Li, J. B. Huang, D. Y. Li, and E. Y. Xun, "Random deep belief networks for recognizing emotions from speech signals," *Computational Intelligence and Neuroscience*, pp. 1- 9, 2017.

[42] J. Deng *et al*, "Semi-supervised autoencoders for speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, October 2017, pp. 31-43.

[43] E. Frant *et al*, "Voice based emotion recognition with convolutional neural networks for companion robots," *Romanian Journal of Information Science and Technology*, vol. 20, no. 3, pp. 222–240, 2017.

[44] T. George *et al*, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2016, pp. 5200-5204.

[45] B. A. Malik *et al*, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *Proc. International Conference on Platform Technology and Service* (PlatCon), 2017, pp. 1-5.

[46] S. Andr éet al, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2011, pp. 5688-5691.

[47] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proc. INTERSPEECH 2017*, August 20–24, 2017, Stockholm, Sweden, pp. 1089-1093.

[48] T. Raquel *et al*, "Emotional space improves emotion recognition," in *Proc. 7th Int. Conf. on Spoken Language Processing*, (Denver), 2002, pp. 2029-2032.

[49] I. Guoth, M. Rusko, M. Ritomsk ý, M. Trnka, and D. Sakhia, "Identifying tense arousal in speech using phase based features," in *Proc. Meetings on Acoustics, Acoustical Society of America*, Boston, Massachusetts, 25-29 June 2017, pp. 30, 1-9.

[50] J. Chang and S. Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2017, pp. 2746- 2750.

[51] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. INTERSPEECH*, 2014, pp. 223-227.

[52] D. Q. Luo, Y. X. Zou, and D. Y. Huang, "Speech emotion recognition via ensembling neural networks," in *Proc. APSIPA Annual Summit and Conferenc*e, 2017, Malaysia, pp. 1351-1355.

[53] Y. Zhao, X. Y. Jin, and X. L. Hu, "Recurrent convolutional neural network for speech processing," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2017, pp. 5300-304.

[54] S. Tripathi *et al*., "Using deep and convolutional neural networks for accurate emotion classification on DEAP dataset," in *Proc. the Twenty-Ninth AAAI Conference on Innovative Applications (IAAI-17),* 2017, pp. 746-752.

[55] B. Nakisa, M. N. Rastgoo, A. Rakotonirainy, F. Maire, and V. Chandran, "Long short term memory hyperparameter optimization for a neural network based emotion recognition framework," *IEEE Access* p. 99, September 2018.

[56] R. I. Bendjillali, M. Beladgham, K. Merit, and A. Taleb-Ahmed, "Improved Facial expression recognition based on DWT feature for deep CNN," *Electronics*, vol. 8, no. 3, p. 324, 2019.

[57] J. Espinosa-Aranda, N. Vallez *et al*., "Smart doll: Emotion recognition using embedded deep learning," *Symmetry*, vol. 10, no. 9, p. 387, 2018.

[58] W. H. Abdulsalam, R. S. Alhamdani, and M. N. Abdullah, "Facial Emotion Recognition from Videos Using Deep Convolutional Neural Networks," *International Journal of Machine Learning and Computing*, vol. 9, no. 1, pp. 14-19, 2019.

[59] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, pp. 572-587, 2011.

[60] D. T. Thang, "Primary examination of Vietnamese intonation," *Hanoi National University Publishing House*, 2009.

[61] L. T. Xuyen, "Etude contrastive de l'intonation expressive en français et en vietnamien," Thèse de Doctorat "Nouveau Régime", Université Paris 3, 1989.

[62] D.-K. Mac and D.-D. Tran, "Modeling Vietnamese speech prosody: A step-by-step approach towards an expressive speech synthesis system," *Trends and Applications in Knowledge Discovery and Data Mining*, Springer, vol. 9441, pp. 273-287, 2015.

[63] D.-K. Mac, E. Castelli, and V. Aubergé, "Modeling the prosody of vietnamese attitudes for expressive speech synthesis," in *Proc. Workshop of Spoken Languages Technologies for under - Resourced Languages (SLTU 2012)*, Cape Town, South Africa, May 7-9, 2012, pp. 114-118.

[64] M. D. Khoa, "Génération de parole expressive dans le cas des langues àtons," MICA, INPG, 2012.

[65] T. D. Ngo and T. D. Bui, "A study on prosody of Vietnamese emotional speech," in *Proc. the Fourth International Conference on Knowledge and Systems Engineering*, IEEE, Danang city, Vietnam, Aug 17-19, 2012.

[66] V. H. Anh, M. N. Van, B. B. Ha, and T. H. Quyet, "A real-time model based support vector machine for emotion recognition through EEG," in *Proc. International Conference on Control, Automation and Information Sciences* (ICCAIS), Ho Chi Minh City, Vietnam, Nov. 26-29, 2012, pp. 191-196.

[67] V. T. La, C.-W. Huang, H. Cheng, and Z. Li, "Emotional feature analysis and recognition from Vietnamese speech," *Journal of Signal Processing*, China, vol. 29, no. 10, pp. 1423-1432, Oct. 2013.

[68] Z. P. Jiang and C.-W. Huang, "High-order Markov random fields and their applications in cross-language speech recognition," *Cybernetics and Information Technologies*, vol. 15, no. 4, Sofia, pp. 50-57, 2015.

[69] D. T. L. Thuy, T. V. Loan, N. H. Quang, and L. X. Thanh, "Influence of spectral features of speech signal on emotion recognition of

Vietnamese," in *Proc. the 10th National Conference on Fundamental and Applied Information Technology Research* (FAIR'10), Da Nang, 2017, pp. 36-43.

[70] D. T. L. Thuy, T. V. Loan, and N. H. Quang, "GMM for emotion recognition of Vietnamese," *Journal of Computer Science and Cybernetics*, vol. 33, no. 3, pp. 229-246, 2017.

[71] Praat. [Online]. Available: http://www.praat.org

[72] J. L. Devore, *Probability and Statistics for Engineering and the Sciences*, 8th Edition, *Brooks/Cole Edition*, 2010.

[73] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. the 32nd International Conference on International Conference on Machine Learning*, France, July 2015, vol. 37, pp. 448-456.

[74] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *Proc. 4th International Conference on Learning Representations*, 2016.

[75] M. D. Zeiler and R. Fergus, "Stochastic pooling for regularization of deep convolutional neural networks," in *Proc. the International Conference on Learning Representation* (ICLR), 2013.

[76] N. Srivastava *et al.*, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929-1958, 2014.

[77] A. Marios, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1207-1216, 2016.

[78] J. Picone, "Fundamentals of speech recognition: A short course," *Inst. for Signal and Info. Process*, Department of Electrical and Computer Engineering, Mississippi State University, 1995.

[79] G. John, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Progressive neural networks for transfer learning in emotion recognition," *arXiv preprint arXiv*: 1706.03256, 2017.

[80] D. Jun, S. Frühholz, Z. X. Zhang, and B. Schuller, "Recognizing emotions from whispered speech based on acoustic feature transfer learning," *IEEE Access*, vol. 5, pp. 5235-5246, 2017.

[81] L. Siddique, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," *arXiv preprint arXiv:1801.06353*, 2018.

[82] L. Margaret, M. Stolar, R. Bolia, and M. Skinner, "Amplitude-frequency analysis of emotional speech using transfer learning and classification of spectrogram images," *Advances in Science, Technology and Engineering Systems Journal*, vol. 3, no. 4, pp. 363-371, 2018.

Thuy Dao Thi Le was born in 1976. She graduated from the Military Technical Academy in 2008. She is currently a Ph.D student at the School of Information and Communication Technology, Hanoi University of Science and Technology. Her research areas are signal processing, speech processing, software engineering.

Loan Trinh Van was born in 1956. He graduated from the Hanoi University of Science and Technology (HUST) in 1978. He received a DEA in 1988 and received a doctor diploma in 1992 at the Grenoble INP (INPG), France. He is currently an Assoc. Prof. at the School of Information and Communication Technology, HUST.

His research areas are in signal processing, speech processing, embedded systems.

Quang Nguyen Hong was born in 1978. He graduated from the Hanoi University of Science and Technology in 2000. He received a doctor diploma at the Avignon University, France in 2008. He is currently a lecturer at the School of Information and Communication Technology, HUST.

His research areas are in speech processing, statistical machine learning.