Computer Systems Science & Engineering
DOI:10.32604/csse.2022.021635
Article

Tech Science Press

Emotion Recognition with Capsule Neural Network

Loan Trinh Van[1], Quang H. Nguyen[1,*] and Thuy Dao Thi Le[2]

[1]School of Information and Communication Technology, Hanoi University of Science and Technology, Hanoi, 10000, Vietnam
[2]Faculty of Information Technology, University of Transport and Communications, Hanoi, 10000, Vietnam
*Corresponding Author: Quang H. Nguyen. Email: quangnh@soict.hust.edu.vn
Received: 08 July 2021; Accepted: 13 August 2021

Abstract: For human-machine communication to be as effective as human-to-human communication, research on speech emotion recognition is essential. Among the models and the classifiers used to recognize emotions, neural networks appear to be promising due to the network's ability to learn and the diversity in configuration. Following the convolutional neural network, a capsule neural network (CapsNet) with inputs and outputs that are not scalar quantities but vectors allows the network to determine the part-whole relationships that are specific 6 for an object. This paper performs speech emotion recognition based on CapsNet. The corpora for speech emotion recognition have been augmented by adding white noise and changing voices. The feature parameters of the recognition system input are mel spectrum images along with the characteristics of the sound source, vocal tract and prosody. For the German emotional corpus EMO-DB, the average accuracy score for 4 emotions, neutral, boredom, anger and happiness, is 99.69%. For Vietnamese emotional corpus BKEmo, this score is 94.23% for 4 emotions, neutral, sadness, anger and happiness. The accuracy score is highest when combining all the above feature parameters, and this score increases significantly when combining mel spectrum images with the features directly related to the fundamental frequency.

Keywords: Emotion recognition; CapsNet; data augmentation; mel spectrum image; fundamental frequency

1 Introduction

Today, robots are present in many places and different areas where people gather. In mass production lines, robots ensure uniformity of manufactured products. It can be said that robots think no less than humans when playing chess. However, for robots, the ability to express emotions through body language, facial expressions, and especially through voice is also limited. Emotional expression is a very subtle human behavior, but at present, robots are very inferior. To achieve the goal of human-machine interaction similar to human-human interaction, there is clearly considerable research needed. Emotional recognition of speech is a research aspect that needs to be considered to achieve that goal.

Neural networks, in addition to being designed to simulate the activity of human neurons, are even more special because of their ability to mimic human perception of the world around them. To identify objects through the visual system, people can quickly and accurately capture information through the part-whole relationship of the object. Therefore, for humans, it is not difficult to exactly identify one object in different poses. The part-whole relationship exists in different objects and is also a specific feature for those objects. The capsule neural network (CapsNet) was proposed to focus on exploiting this feature [1–3]. When explaining the capsule neural network exploiting this feature, some authors often take the example of human face descriptions. The relative position of the eyes, nose and mouth on the human face is that the eyes and nose are positioned above the mouth, and the nose and mouth are arranged on the vertical symmetrical axis of the face. This is one of the characteristics of the spatial relationships of objects to correctly identify the human face, and these relationships are equivariant. This paper is based on a capsule neural network used to identify images to perform emotion recognition of speech. The input of the recognition system is mel spectrum. A part-whole relationship to the mel spectrum of the speech signal can be taken as an example for the case of the syllable, including a fricative consonant [s] preceding a voiced sound [ε], for example. The mel spectrum of syllable [sε] is given in Fig. 1.

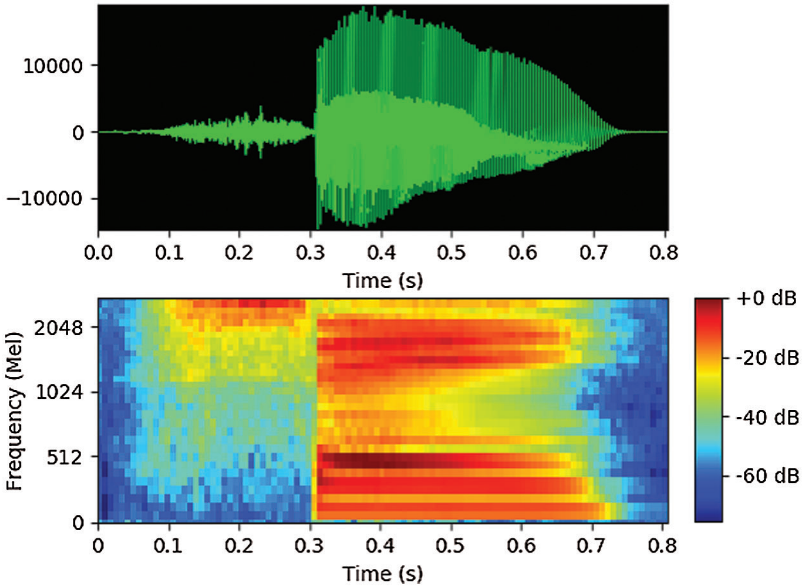

Figure 1: Mel spectrum of syllable [sε]

From Fig. 1, the characteristic of the mel spectrum for a fricative sound is that spectral energy is concentrated mainly in the high-frequency domain (the upper part of the mel spectrum), whereas for voiced sounds, the spectral energy is focused more on low-order formants (the lower part of the mel spectrum). Thus, it can be said that in the mel spectrum, the spatial relationship of the energy concentration sections between the fricative sound and the voiced sound is the upper left-the lower right. This also means that the part-whole relationship that CapsNet exploited also exists in the mel spectrum of the speech signal.

The rest of the paper is organized as follows. Section 2 describes related work. The emotional corpora used in this paper, and data augmentation are presented in Section 3. Section 4 details the configuration of the capsule neural network for emotion recognition. The experimental results are provided in Section 5. Finally, Section 6 presents the discussion and conclusion.

2 Related Work

If only considering the field of research, such as speech processing, there have been some CapsNet-based studies. In [4], the capsule network was applied to capture the spatial relationship and pose information of speech spectrogram features in both frequency and time axes. The authors showed that the end-to-end speech recognition system with capsule networks on one-second speech commands dataset achieves better results on both clean and noise-added tests than baseline convolutional neural network models. A capsule network for low resource spoken language understanding was proposed for command-and-control applications in [5]. For small quantities of data, the proposed model is shown to significantly outperform the previous state-of-the-art model.

The literature [6–9] provided an overview of speech emotion recognition, including models, used classifiers and corpus, specific parameters and corresponding recognition accuracy. The different classifiers may be Gaussian mixture models (GMM), support vector machines (SVM), artificial neural networks (ANN), k-nearest neighbor classifier, Bayes classifier, linear discriminant analysis with Gaussian probability distribution, and hidden Markov models (HMM). The feature parameters can be classified into 3 groups. The first group includes parameters directly related to the sound source. The second group is the parameters of the vocal tract, while the third group is related to the prosody. For the sound source, the feature parameters may be LP (linear prediction) residual energy or LP residual, glottal excitation signal. The parameters of the vocal tract include MFCC (mel frequency cepstral coefficients), LPCC (linear predictive cepstral coefficients), and RASTA (Relative Spectra) PLP (perceptual linear predictive) coefficients, formants and their bandwidth and spectral features. The prosodic parameters consist of pitch, energy, duration and voice quality features. In addition to these parameters, statistical features such as mean, StdDev, min, and max have been used. Until recently, GMM, HMM, and SVM were still used to recognize emotional speech [10–17] or GMM and DNN, and GMM and SVM have been combined [18–20]. Sequential minimal optimization (SMO), J48, and random forest have been used for testing the adaptive data boosting (ADB) technique [21]. For ANN, it can be seen that the models used are ANN with 3 layers [22], DNN [23], progressive neural network [24], recurrent neural network [12,25], backpropagation neural network [26], deep convolutional recurrent network [27], coupled deep convolutional neural network (CDCNN) [28], deep belief networks (DBNs) [29], combination of SVM and belief networks [30], CNN [31], convolutional recurrent neural network (CRNN) [32], deep learning [33,34], a combination of convolutional and recurrent layers for reusing ASR (automatic speech recognition) network [35] and LSTM (long short-term memory) network [36]. A number of issues have also been raised for emotion recognition of speech, such as transfer learning [37–39], using cross-corpus [27,40], and adversarial training [41].

For Vietnamese emotion recognition, SVM was used in [42] to classify emotions using the EEG signal. An average accuracy of 70.5% was achieved in real-time for five emotional states. Research in [43] used the GMM model to recognize 6 emotions: happiness, neutrality, sadness, surprise, anger, and fear. In this research, two male voices and two female voices expressed 6 emotions for 6 different sentences. The feature parameters were MFCC, short-term energy, pitch, and formants. The highest recognition score was 96.5% for neutral emotion, and the lowest was 76.5% for sad emotion. In [44], the corpus included 6 voices and 20 sentences and the same number of emotions as in [43]. The recognition score on the Vietnamese language was 96.5% for neutrality and dropped to 84.1% for surprise using SVM with Im-SFLA (improved shuffled frog leaping algorithm). The authors in [45,46] used GMM and CNN to recognize 4 emotions with Vietnamese emotional corpus BKEmo. Details of this study and the corpus BKEmo are briefly presented in the following sections.

In this paper, the EMO-DB corpus was also used to perform emotion recognition using CapsNet. Since its appearance, there have been many emotional speech recognition studies using this corpus. In the context of this paper, we review most of the research conducted in the last more 10 years using EMO-DB (Tab. 4). From the review, most studies with EMO-DB also use models, classifiers and feature parameters, as mentioned above. In addition, EMO-DB is also used in cross-corpus and transfer learning studies [47].

3 Proposal Method

Overall architecture of our proposal method is shown in Fig. 2 which has composed: data augmentation module, parameter extraction module, CapsNet module. These modules are presented as following.

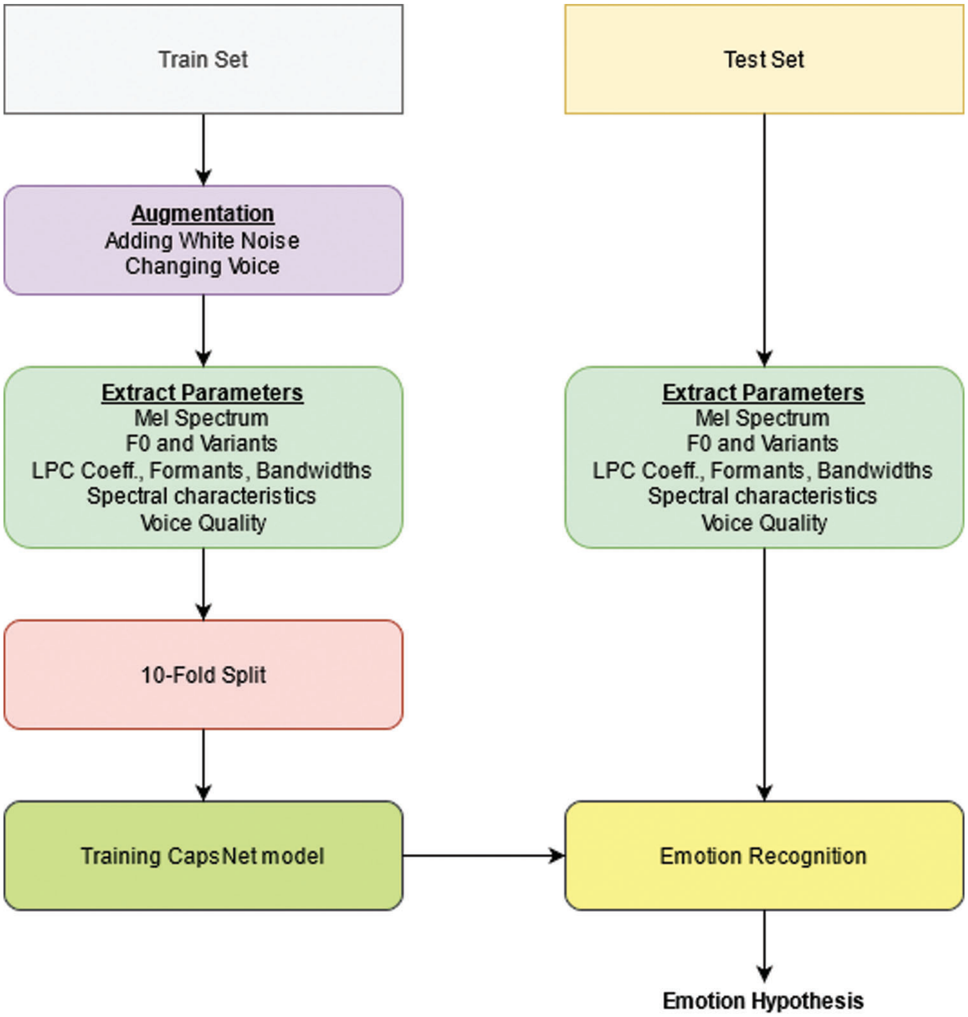

Figure 2: Overall architecture of our proposal method

3.1 Data Augmentation

It is well known that for the classification problem or in machine learning in general, the more available data, the better the classification performance. Therefore, if data are insufficient, data augmentation is necessary. In addition, data enhancement is one method to avoid overfitting. Ocquaye et al. [48] implemented data augmentation for EMO-DB by adding background noise as proposed in [49]. For data augmentation in our case, adding white noise and changing voices were made for BKEmo and specifically for EMO-DB because the existing EMO-DB is not a corpus of sufficiently large size.

3.1.1 Adding White Noise

Fig. 3 illustrates the addition of white noise to the augmented corpus and the magnified noise.

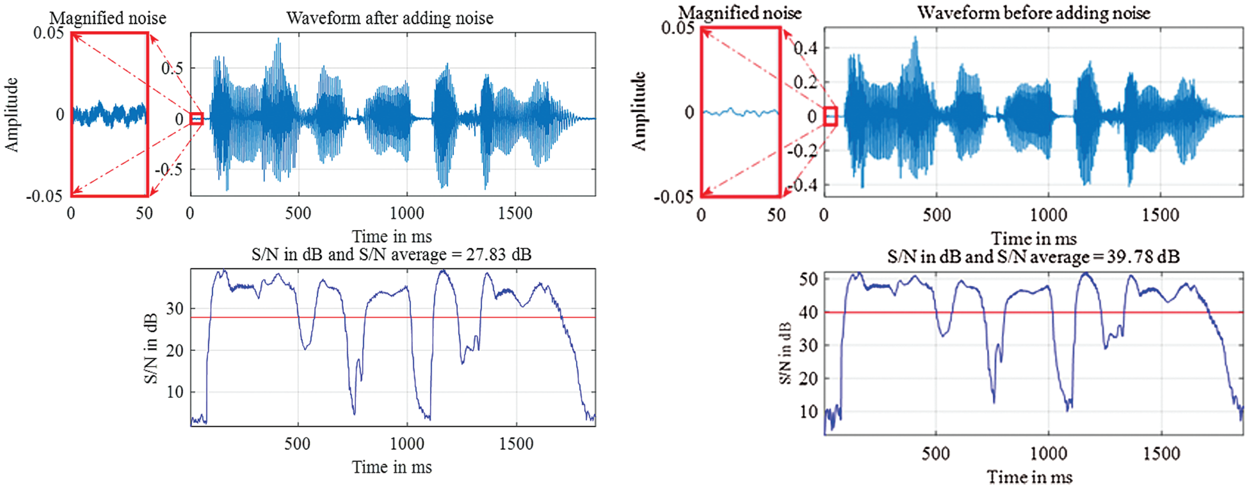

Figure 3: Example to illustrate the addition of white noise for data augmentation

The example illustrates that after adding white noise, the average signal-to-noise ratio decreases by 11.95 dB. The signal-to-noise ratio is calculated using the Eq. (1):

$$S/N(dB) = 10log_{10}\frac{P_S}{P_N} \tag{1}$$

where P_S is the power of signal and P_N is the power of noise. Since the signal power before adding noise remains the same as the signal power after adding noise, we have $10log_{10}\frac{P_{N(after\ adding\ noise)}}{P_{N(before\ adding\ noise)}} = 11.95$ dB.

That means on average $P_{N(after\ adding\ noise)} \approx 15.67 \times P_{N(before\ adding\ noise)}$. We used Praat toolkit [50] for this process.

3.1.2 *Changing Voice*

Fig. 4 illustrates changing the voices for the augmented corpus. If it is a male voice, the formant is raised towards the high frequency so that the male voice is closer to a female voice. If it is a female voice, the formant is lowered towards the low frequency to be closer to a male voice. Translation of formant is performed with Praat toolkits [50]. For the formant lifting case, the lift coefficient used in Praat is 1.1, while for the formant reduction, the reduction factor used in Praat is 0.909.

3.2 Parameter Extraction

The mel spectrum image is extracted from sound file with the fixed size 260 × 260 corresponding to 260 mel spectral coefficients × 260 frames. The number of frames is taken by 260 because this is the average number of frames for WAV files in both EMO-DB and BKEmo corpus. This parameter set is named MELSPEC.

Beside 260 mel spectral coefficients such as baseline, we added 8 parameters related to fundamental frequency F_0. This parameter set is named MELSPEC_F0. These 8 parameters include: F_0 + 7 F_0 variants:

- Derivation of F_0.
- Normalization of F_0 according to the average value of F_0 for each file.
- Normalization of F_0 according to $minF_0$ and $maxF_0$ for each file.
- Normalization of F_0 by the mean and the standard deviation of F_0 for each file.

- Normalization of $LogF_0(t)$ by the average of $LogF_0(t)$ for each file.
- Normalization of $LogF_0(t)$ by the minimal value $minLogF_0(t)$ and the maximal value $maxLogF_0(t)$ for each file.
- Normalization of $LogF_0(t)$ by the mean and the standard deviation of $LogF_0(t)$ for each file.

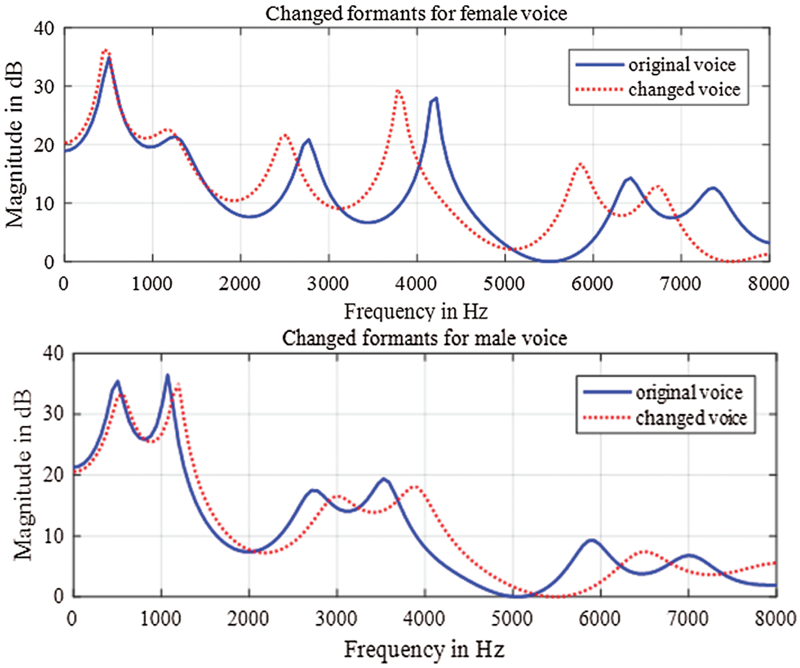

Figure 4: Illustrating changing voices for data augmentation

We also added 28 other parameters related to the vocal tract, spectral characteristics and voice quality, so there are 296 parameters in total. These 28 parameters are listed in Tab. 1. This parameter set is named MELSPEC_F0_OTHER.

Table 1: Parameters related to the vocal tract, spectral characteristics and voice quality

Parameters	Number of parameters
Intensity	1
Formants and its correspondent bandwidths	8
Harmonicity	1
Center of gravity	1
Central moment	1
Skewness	1
Kurtosis	1
LPC coefficients	14
Sum	**28**

For the basic discrete-time model for speech production, the vocal tract's transfer function is $H(z)$. $H(z)$ is a pth-order all-pole rational function of the form [51] by Eq. (2) as following:

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}} \tag{2}$$

$A(z)$ is an inverse filter for the vocal tract system and is often called the prediction error polynomial or the linear predictive coding (LPC) polynomial. $\{\alpha_k, k = 1, 2, \ldots, p\}$ are vocal tract system parameters, and in this paper, they are LPC coefficients.

Other parameters, such as harmonicity, center of gravity, central moment, skewness, and kurtosis, were explained based on Praat and are presented in [45]. ANOVA and T-test were used in [52] to evaluate the corpus BKEmo. The P-value = 0.05 is used in the majority of cases [53], and this value is also used as the cutoff for significance in our case. If the P-value is less than 0.05, a significant difference for a pair of emotions does exist. The T-test results from [52] showed that the emotion pairs of BKEmo are best distinguished for most of the above feature parameters. All feature parameters are calculated using Praat toolkits [50].

3.3 CapsNet Based Model

3.3.1 Capsule Neural Network

A capsule is a group of neurons in which the inputs and outputs of the capsule are vectors [1]. To illustrate the basic activity of the capsule neural network to be used in the paper, we take an example of a capsule neural network consisting of M capsules in the higher level and N capsules in the lower level, as denoted in Fig. 5. Capsule 1 in the higher level has an output vector $\vec{v}_1$. This vector encodes existence and pose of object 1, for example. Capsule 1 has N input vectors corresponding to N outputs of N capsules 1,…, i,…, N in the lower level. Assume that the output vectors corresponding to N capsules in the lower level are $\vec{u}_1, \ldots, \vec{u}_i, \ldots, \vec{u}_N$. Assume that the output vector $\vec{u}_i$ of the capsule i in the lower level encodes existence and pose of part i and part i belongs to the object j described by capsule j in the higher level.

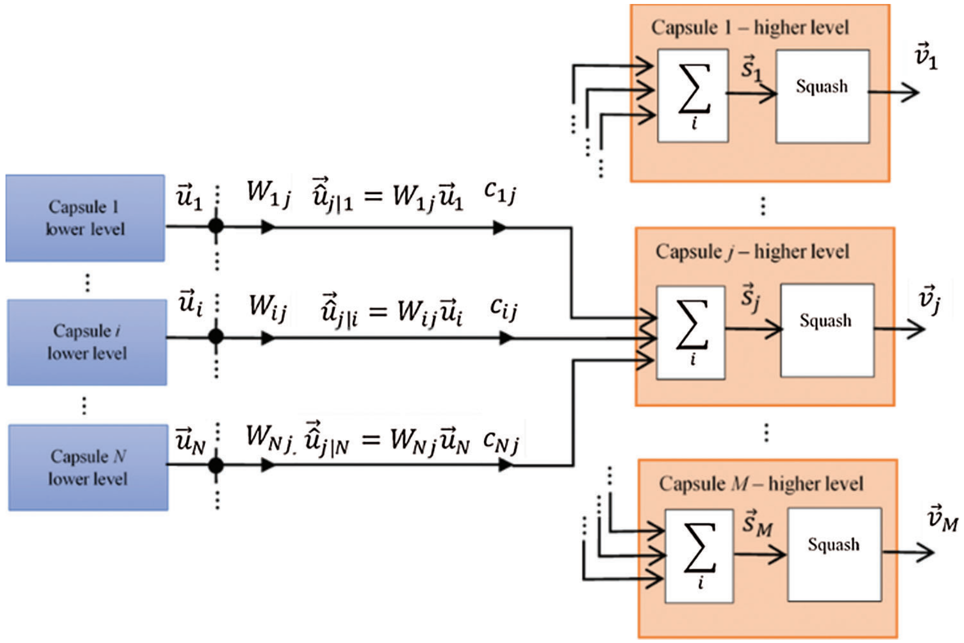

Figure 5: A capsule neural network consisting of M capsules in the higher level and N capsules in the lower level

Before entering capsule *j* at the higher level, the output vector $\vec{u}_i$ is multiplied by the weight matrix W_{ij} that encodes the spatial relationship between part *i* and object *j* and then becomes the vector $\overrightarrow{\hat{u}}_{j|i} = W_{ij}\vec{u}_i$. The vector $\vec{u}_{j|i}$ is multiplied by scalar weight c_{ij} to become $c_{ij}\vec{u}_{j|i}$ before actually entering capsule *j*. Scalar weight c_{ij} is determined by a routing algorithm. Similarly, capsule 1 and capsule *N* at the lower level provide vectors entering capsule *j*, respectively, $c_{1j}\vec{u}_{j|1}$ and $c_{Nj}\vec{u}_{j|N}$, where $\overrightarrow{\hat{u}}_{j|1} = W_{1j}\vec{u}_1$ and $\overrightarrow{\hat{u}}_{j|N} = W_{Nj}\vec{u}_N$.

Capsule *j* in the higher level performs the sum:

$$\vec{s}_j = \sum_i c_{ij}\vec{u}_{j|i} \tag{3}$$

The output vector $\vec{s}_j$ of capsule *j* is passed through the squash function:

$$\vec{v}_j = \frac{\|\vec{s}_j\|^2}{1 + \|\vec{s}_j\|^2}\frac{\vec{s}_j}{\|\vec{s}_j\|} \tag{4}$$

Vector output $\vec{v}_j$ of capsule *j* encodes the existence and pose of object *j*. The coupling coefficients (routing coefficients) c_{ij} between capsule *i* and all the capsules in the layer above sum to 1. These coefficients are determined by a "routing softmax":

$$c_{ij} = \frac{\exp\left(b_{ij}\right)}{\sum_k \exp\left(b_{ik}\right)} \tag{5}$$

where b_{ij} are the log prior probabilities that capsule *i* should be coupled to capsule *j*.

The basis of dynamic routing algorithms proposed in [2] is that the capsule in the lower level sends its input to a higher-level capsule that agrees with that input. The result of the algorithm with a certain number of routing iterations (usually equal to 3) gives a set of routing coefficients that best match the output from the capsule in the lower level with the output of the capsules at a higher level.

CapsNet computes the margin loss for class *k* as following:

$$\mathcal{L}_k = T_k \max(0, m^+ - \|\vec{v}_k\|)^2 + \lambda(1 - T_k)\max(0, \|\vec{v}_k\| - m^-)^2 \tag{6}$$

where T_k = 1 if an entity of class *k* is present and $m^+ = 0.9$ and $m^- = 0.1$. The weight λ = 0.5.

3.3.2 Configuration of Capsule Neural Network for Emotion Recognition

The neural network used to recognize the four emotions in this paper consists of two parts: the first part is 5 CNN layers, and the second part is a capsule neural network. Take the configuration example of the neural network for the case of 296 parameters × 296 frames. Five CNN layers include the following:

- Layer 1: Convolution 2D, input (296, 296,1), output (148, 148, 64), kernel (3 × 3) × 64, stride = 2, parameter #: 640, activation function: ReLU, Dropout (rate = 0.5).
- Layer 2: Convolution 2D, input (148, 148, 64), output (74, 74, 16), kernel (2 × 2) × 16, stride = 2, parameter #: 4112, activation function: ReLU, Dropout (rate = 0.5).
- Layer 3: Convolution 2D, input (74, 74, 16), output (37, 37, 16), kernel (2 × 2) × 16, stride = 2, parameter #: 1040, activation function: ReLU, Dropout (rate = 0.5), MaxPooling2D (pool size = (2, 2)).
- Layer 4: Convolution 2D, input (37, 37, 16), output (19, 19, 16), kernel (2 × 2) × 16, stride = 2, parameter #: 1040, activation function: ReLU, dropout (rate = 0.5), MaxPooling2D (pool size = (2, 2)).

- Layer 5: Convolution 2D, input (19, 19, 16), output (10, 10, 16), kernel (2 × 2) × 16, stride = 2, parameter #: 1040, activation function: ReLU, Dropout (rate = 0.5), MaxPooling2D (pool size = (2, 2)).

The output of layer 5 is the input of the primary capsule in Fig. 6. The configuration of the CapsNet is basically inspired by the CapsNet configuration proposed by [2], but the parameters have been changed to suit our case.

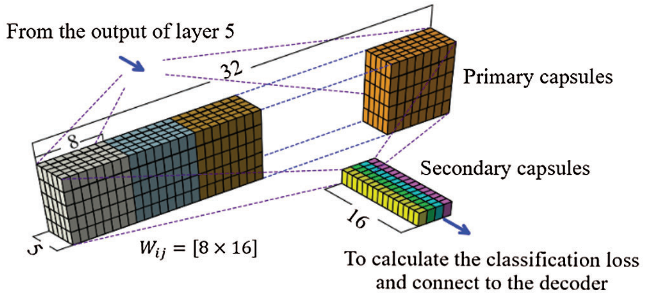
From the output of layer 5
32
8
5
Primary capsules
Secondary capsules
16
W_{ij} = [8 × 16]
To calculate the classification loss
and connect to the decoder

Figure 6: Illustration of primary capsules and secondary capsules

Fig. 6 is an illustration of the primary and secondary capsules (capsule layer). In nature, the primary capsule layer is similar to the convolutional layer. This layer reduces the spatial dimension from 10 × 10 to 5 × 5 by using kernel 9 × 9 with stride 2 and no padding. The primary capsule layer uses 8 × 32 kernels to generate 32 8-D capsules, i.e., 8 output neurons are grouped together to form a capsule. The output of the primary capsule layer is reshaped to (800 (=5 × 5 × 32), 8). Next, the capsule layer applies a transformation matrix W_{ij} with shape 8 × 16 to convert the 8-D capsule from the output of the primary capsule layer to a 16-D capsule for one of four emotions. W_{ij} is a weight matrix between each $\vec{u}_i,\ i\ \in (1, 32 \times 5 \times 5)$ in primary capsules and $\vec{v}_j, j\ \in (1, 4)$. Dynamic routing is performed between the primary capsules and the capsule layer [2].

The above configuration does not change for the remaining 2 cases (260 and 268 feature parameters). Of course, the number of corresponding parameters to be calculated varies depending on 260 and 268 feature parameters.

Emotion hypothesis is determined as following:

$$emotion = \arg\max_j \|\vec{v}_j\| \tag{7}$$

4 Results and Discussions

4.1 EMO-DB and BKEmo Datasets

EMO-DB is a German emotional corpus [54]. The corpus was built using a simulation method with 10 professional artists (5 male artists and 5 female artists) and includes 7 emotions: neutral, anger, fear, happiness, sadness, disgust and boredom. There are 10 sentences for the artists to express different emotions. Each emotion is expressed 1 to 6 times. Along with EMO-DB, a Vietnamese emotional corpus BKEmo is also used in this paper. The Vietnamese emotional corpus used for recognition is extracted from the BKEmo corpus developed at Hanoi University of Science and Technology. BKEmo is built according to the simulation method for four emotions: neutral, sadness, anger and happiness. EMO-DB's emotions, neutral, boredom, anger and happiness, are chosen because these are emotions with the largest number of files. The number of emotions in BK-Emo is also equal to four, and thus, the architecture of the emotion recognition system remains the same for both languages. The total number of files for these

4 emotions of EMO-DB is 358, of which anger has 127 files, neutral has 79 files, happiness has 71 files and boredom has 81 files.

Using data augmentation, we obtained 2,148 files from 358 files. Of the 2,148 files, the subset has 195 files used for testing. This subset of 195 files includes 42 files for neutral, 42 files for boredom, 41 files for happiness, and 70 files for anger. After taking 195 files for testing, the remaining files are split into 10 subsets (these 10 subsets have slightly different file numbers for each subset) for 10-fold cross-validation.

From the BKEmo corpus, the authors of the paper listened and selected 5,584 files for 4 emotions with 22 sentences, 8 male and 8 female voices, and these 5,584 files were used for emotion recognition. More details about the corpus can be found in [45,46]. The 5,584 files are divided into 11 parts, and 1/11 parts (507 files) are data used for testing, the remaining 5,077 files are used for training and validation. By using data augmentation, these 5077 files were augmented into 20,308 files. The set of files for the test, in any case, does not contain the files used for training and validation. The augmented corpus is split into 10 subsets for 10-fold cross-validation, and these 10 subsets have slightly different file numbers for each subset. Data distribution for each emotion is depicted in Tab. 2.

Table 2: Number of audio files for each emotion in EMO-DB and BKEmo dataset

Dataset	Neural	Sadness	Anger	Happiness	Total
EMO-DB train	432	444	692	385	1953
EMO-DB validation	42	42	70	41	195
EMO-DB test	42	42	70	41	195
BKEmo train	4572	4568	4568	4568	18277
BKEmo validation	508	508	508	508	2031
BKEmo test	126	127	127	127	507

4.2 Experiment Results

The experiments were performed on a machine with the configuration as following:

- CPU: an Intel Core i7-7700 CPU @ 3.60 GHz × 8
- RAM: 32 GB
- GPU: GeForce GTX 1080 Ti/PCIe/SSE2 with 11 GB of RAM
- Hard-disk: SSD 512 GB

For EMO-DB, the average training time for one fold is approximately 3.3 min, while for BKEmo, this time is approximately 30 min. The accuracy score (%) for each emotion for EMO-DB and BKEmo are given in Tab. 3. The results in this table are the average accuracy of 10 experiments corresponding with 10-folds.

4.3 Discussions

At first, for both corpora EMO-DB and BKEmo, the average accuracy score increased when the number of parameters increased from 260 to 268 and 296, respectively. So beside mel spectrum, the parameters related to F_0 and variant, vocal tract, spectral characteristics and voice quality have contributed to increase the accuracy of the speech emotion recognition system.

Table 3: The accuracy score (%) for each emotion with German EMO-DB *corpus* and vietnamese BKEmo *corpus*

Dataset	Parameter type	Neural	Boredom/Sadness	Anger	Happiness	Average
EMO-DB	MELSPEC	97.62	100	99.57	98.29	98.87
	MELSPEC_F0	98.81	100	99.71	99.02	99.39
	MELSPEC_F0_OTHER	99.76	100	99.86	99.13	99.69
BKEmo	MELSPEC	97.17	95.00	96.95	78.57	91.92
	MELSPEC_F0	97.01	97.54	94.92	86.35	93.96
	MELSPEC_F0_OTHER	96.77	96.19	95.94	88.02	94.23

If only comparing the accuracy scores for the EMO-DB corpus of the studies listed in Tab. 4, in general, the average accuracy score in our case is superior to the accuracy score of the vast majority of available studies (except for [55], 99.8% *vs*. 99.69%). German is not a tonal language. The addition of parameters directly related to F_0 increased accuracy because the law of variable F_0 contributes significantly to emotional expression.

Table 4: Comparison of our results with the state-of-the-art accuracy scores on German EMO-DB *corpus*

References	Year	Model, classifier	Parameters	Accuracy score (%)
Mishra et al. [56]	2009	GMM	MFCCs and energy	63.78
Luengo et al. [57]	2010	k-means clustering	Prosody, voice quality, spectral and segmental features	78.60
Amarakeerthi et al. [58]	2011	HMM	TLCSF-CC features (two-layered cascaded subband filter-cepstral coefficient)	72.85
Shen et al. [59]	2011	SVM	Energy, pitchLPCC, MFCC, linear prediction coefficients and mel cepstrum coefficients (LPCMCC)	82.50
Stuhlsatz et al. [60]	2011	Generalized discriminant analysis (GerDA) based on DNN	Zero crossing rate, signal energy logarithmic pitch, voice quality spectral, mel spectrum, cepstral	85.10
Pan et al. [61]	2012	SVM	MFCC and mel-energy spectrum dynamic coefficients (MEDC) + energy	95.10
Jin et al. [62]	2014	SVM	Intense, loudness, MFCC, LSP (line spectral pairs), ZCR (zero crossing rate), probability of voicing, F_0	83.10
Gjoreski et al. [63]	2014	SVM	400 features extracted by OpenSmile	87.00

(Continued)

Table 4 (continued)

References	Year	Model, classifier	Parameters	Accuracy score (%)
Revathi et al. [55]	2018	VQ/Fuzzy/MHMM/ SVM	Gamma tone filters spaced in equivalent rectangular bandwidth (ERB), MEL and BARK scale	99.80
Ocquaye et al. [48]	2019	Dual exclusive attentive transfer (DEAT) convolutional neural network	Spectrogram	67.79
Mao et al. [64]	2019	SGMM-HMM (SGMM-Subspace based GMM)	15-dimensional MFCCs with the first- and second-order derivatives + pitch + voicing probability	88.15
Seo et al. [65]	2020	VACNN (Visual attention convolutional neural network)	Log-mel spectrogram	86.92
Lech et al. [66]	2020	AlexNet (real-time SER)	Spectrograms converted into RGB	82.00
Haider et al. [67]	2021	SVM	eGeMAPs (a total of 88 features)	76.90
Chauhan et al. [68]	2021	CNN	Log-mel spectrograms	72.02
MELSPEC (ours)	2021	Capsule network	Melspectrogram	97.62
MELSPEC_F0 (ours)	2021	Capsule network	Melspectrogram, F_0	98.81
MELSPEC_F0_OTHER (ours)	2021	Capsule network	Melspectrogram, F_0, vocal tract, spectral characteristics and voice quality	99.76

Vietnamese is a tonal language. There are 6 tones of Vietnamese. For Vietnamese, changing the tone of a syllable changes the meaning of the syllable. The variable rule of fundamental frequency F_0 determines the tone among 6 tones. Not only for Vietnamese but also for other languages such as German mentioned above, the law of variable F_0 of a sentence participates in determining the intonation of that sentence, and the intonation of a sentence is closely related to emotional expression. The result of recognition of the set of 268 parameters in which F_0 and 7 variants of F_0 are added shows a significantly higher score than the baseline model with 260 parameters for both corpora. This fits perfectly with the comment on the importance of F_0 mentioned above. For the set of 296 parameters, the accuracy score increases compared to the set of 268 parameters but does not increase considerably, which also reinforces the role of parameter F_0 for emotion recognition for Vietnamese in particular and for other languages (such as German in our case) in general.

In [45,46], the GMM and DCNN models were used to recognize Vietnamese emotions with the same corpus BKEmo containing only 5,584 original files, which means that there was no data augmentation for the corpus. The maximal number of parameters in [45] was 87. Therefore, the corpus of these two models (GMM and CapsNet) is not exactly the same, and the number of parameters of the two models is also different. The average accuracy score for [45] is 93.12% *vs*. 94.23% for this CapsNet model. Also with the same set of 296 parameters without data augmentation, the DCNN showed the average recognition accuracy of the 4 emotions was 88.01% *vs*. 94.23% for this CapsNet model. The common point of the two models is that the recognition score increases significantly when adding parameters related to F_0. The results of this paper also allow us to say that the CapsNet model is also suitable for emotion recognition of speech in which the input parameters can be viewed as corresponding to a large-sized image.

5 Conclusions

In summary, our experiments of emotion recognition using a capsule neural network with parameters related to mel spectrum, F_0 and variants, vocal tract, spectral characteristics showed an overwhelming advantage in recognition scores compared to many other models and classifiers. A problem for the recognition systems, in general, is that in real environments, the recognition score may be reduced because the actual data are not quite similar to the trained data. To approach this issue of emotion recognition, there is a research direction such as transfer learning. We will then apply our CapsNet-based model to solve the multi-lingual speech emotion recognition. This is also our upcoming research direction of emotion recognition.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] G. E. Hinton, A. Krizhevsky and S. D. Wang, “Transforming auto-encoders,” in *Proc. ICANN*, Espoo, Finland, pp. 44–51, 2011.

[2] S. Sabour, N. Frosst and G. E. Hinton, “Dynamic routing between capsules,” in *Proc. NIPS*, Long Beach, CA, USA, pp. 3859–3869, 2017.

[3] G. E. Hinton, S. Sabour and N. Frosst, “Matrix capsules with EM routing,” in *Proc. ICLR*, Vancouver, Canada, pp. 1–15, 2018.

[4] J. Bae and D. S. Kim, “End-to-end speech command recognition with capsule network,” in *Proc. Interspeech*, Hyderabad, India, pp. 776–780, 2018.

[5] J. Poncelet and V. Renkens, “Low resource end-to-end spoken language understanding with capsule networks,” *Computer Speech & Language*, vol. 66, no. 101142, pp. 1–21, 2021.

[6] S. G. Koolagudi and K. S. Rao, “Emotion recognition from speech: A review,” *International Journal of Speech Technology*, vol. 15, no. 2, pp. 99–117, 2012.

[7] M. E. Ayadi, M. S. Kamel and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[8] T. Thanapattheerakul, K. Mao, J. Amoranto and J. H. Chan, “Emotion in a century: A review of emotion recognition,” in *Proc. IAIT*, Bangkok, Thailand, pp. 17–24, 2018.

[9] B. Basharirad and M. Moradhaseli, “Speech emotion recognition methods: A literature review,” *AIP Conference Proceedings*, vol. 1891, no. 020105, pp. 1–7, 2017.

[10] I. Shahin, “Emotion recognition based on third-order circular suprasegmental hidden Markov model,” in *Proc. JEEIT*, Amman, Jordan, pp. 800–805, 2019.

[11] M. Jain, S. Narayan, P. Balaji, A. Bhowmick and M. R. Gurdaspur, “Speech emotion recognition using support vector machine,” in *Proc. ICICES*, India, pp. 1–6, 2018.

[12] J. Han, Z. Zhang, G. Keren and B. Schuller, “Emotion recognition in speech with latent discriminative representations learning,” *Acta Acustica United with Acustica*, vol. 104, no. 5, pp. 737–740, 2018.

[13] M. Abdelwahab and C. Busso, “Supervised domain adaptation for emotion recognition from speech,” in *Proc. ICASSP*, Queensland, Australia, pp. 5058–5062, 2015.

[14] J. Deng, Z. Zhang, F. Eyben and B. Schuller, “Autoencoder-based unsupervised domain adaptation for speech emotion recognition,” *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, 2014.

[15] C. Huang, R. Liang, Q. Wang, J. Xi, C. Zha *et al.,* “Practical speech emotion recognition based on online learning: from acted data to elicited data,” *Mathematical Problems in Engineering*, vol. 2013, no. 265819, pp. 1–9, 2013.

[16] H. Hu, M. X. Xu and W. Wu, “Gmm supervector based svm with spectral features for speech emotion recognition,” in *Proc. ICASSP*, Honolulu, USA, pp. 413–416, 2007.

[17] W. Zehra, A. R. Javed, Z. Jalil, H. U. Khan and T. R. Gadekallu, “Cross corpus multi-lingual speech emotion recognition using ensemble learning,” *Complex & Intelligent Systems*, vol. 7, pp. 1845–1854, 2021.

[18] I. Shahin, A. B. Nassif and S. Hamsa, “Emotion recognition using hybrid gaussian mixture model and deep neural network,” *IEEE Access*, vol. 7, pp. 26777–26787, 2019.

[19] I. Shahin, A. B. Nassif and S. Hamsa, “Novel cascaded Gaussian mixture model-deep neural network classifier for speaker identification in emotional talking environments,” *Neural Computing and Applications*, vol. 32, no. 7, pp. 2575–2587, 2020.

[20] A. S. Utane and S. L. Nalbalwar, “Emotion recognition through speech using Gaussian mixture model and support vector machine,” *International Journal of Scientific & Engineering Research*, vol. 4, no. 5, pp. 1439–1443, 2013.

[21] J. Bang, T. Hur, D. Kim, J. Lee, Y. Han *et al.,* “Adaptive data boosting technique for robust personalized speech emotion in emotionally-imbalanced small-sample environments,” *Sensors*, vol. 18, no. 11, pp. 1–21, 2018.

[22] D. Czerwinski and P. Powroznik, “Human emotions recognition with the use of speech signal of polish language,” in *Proc. EPMCCS*, Kielce, Poland, pp. 1–6, 2018.

[23] W. Han, H. Ruan, X. Chen, Z. Wang, H. Li *et al.,* “Towards temporal modelling of categorical speech emotion recognition,” in *Proc. Interspeech*, Hyderabad, India, pp. 932–936, 2018.

[24] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis and E. M. Provost, “Progressive neural networks for transfer learning in emotion recognition,” in *Proc. Interspeech*, Stockholm, Sweden, pp. 1098–1102, 2017.

[25] E. Tzinis and A. Potamianos, “Segment-based speech emotion recognition using recurrent neural networks,” in *Proc. ACII*, San Antonio, TX, USA, pp. 190–195, 2017.

[26] A. J. Kayal and J. Nirmal, “Multilingual vocal emotion recognition and classification using back propagation neural network,” *AIP Conference Proceedings*, vol. 1715, no. 20054, pp. 1–7, 2016.

[27] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou *et al.,* “Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network,” in *Proc. ICASSP*, Shanghai, China, pp. 5200–5204, 2016.

[28] O. E. Nii Noi, M. Qirong, G. Xu and Y. Xue, “Coupled unsupervised deep convolutional domain adaptation for speech emotion recognition,” in *Proc. BigMM*, Xi’an, China, pp. 1–5, 2018.

[29] E. M. Schmidt and Y. E. Kim, “Learning emotion-based acoustic features with deep belief networks,” in *Proc. WASPAA*, NY, USA, 65–68, 2011.

[30] B. Schuller, G. Rigoll and M. Lang, “Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture,” in *Proc. ICASSP*, Montreal, Canada, pp. 577–580, 2004.

[31] J. Liu, W. Han, H. Ruan, X. Chen, D. Jiang *et al.,* “Learning salient features for speech emotion recognition using cnn,” in *Proc. ACII Asia*, Beijing, China, pp. 1–5, 2018.

[32] D. Luo, Y. Zou and D. Huang, “Investigation on joint representation learning for robust feature extraction in speech emotion recognition,” in *Proc. Interspeech*, Hyderabad, India, pp. 152–156, 2018.

[33] H. M. Fayek, M. Lech and L. Cavedon, “Evaluating deep learning architectures for speech emotion recognition,” *Neural Networks*, vol. 92, pp. 60–68, 2017.

[34] C. P. Latha and M. Priya, “A review on deep learning algorithms for speech and facial emotion recognition,” *APTIKOM Journal on Computer Science and Information Technologies*, vol. 1, no. 3, pp. 92–108, 2016.

[35] S. Wermter, C. Weber, S. Magg and E. Lakomkin, “Reusing neural speech representations for auditory emotion recognition,” in *Proc. IJCNLP*, Taipei, Taiwan, pp. 423–430, 2017.

[36] S. Latif, R. Rana, J. Qadir and J. Epps, “Variational autoencoders for learning latent representations of speech emotion: A preliminary study,” in *Proc. Interspeech*, Hyderabad, India, pp. 3107–3111, 2018.

[37] S. Latif, R. Rana, S. Younis, J. Qadir and J. Epps, “Transfer learning for improving speech emotion classification accuracy,” in *Proc. Interspeech*, Hyderabad, India, pp. 257–261, 2018.

[38] J. Deng, Z. Zhang, E. Marchi and B. Schuller, “Sparse autoencoder-based feature transfer learning for speech emotion recognition,” in *Proc. ACII*, Geneva, Switzerland, pp. 511–516, 2013.

[39] P. Song, “Transfer linear subspace learning for cross-corpus speech emotion recognition,” *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 265–275, 2017.

[40] S. Latif, J. Qadir and M. Bilal, “Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition,” in *Proc. ACII*, Cambridge, United Kingdom, pp. 732–737, 2019.

[41] M. Abdelwahab and C. Busso, “Domain adversarial for acoustic emotion recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, 2018.

[42] V. A. Hoang, V. M. Ngo, H. B. Ban and Q. T. Huynh, “A real-time model-based support vector machine for emotion recognition through eeg,” in *Proc. ICCAIS*, HoChiMinh city, Vietnam, pp. 191–196, 2012.

[43] L. Vutuan, H. Chengwei, Z. Cheng and Z. Li, “Emotional feature analysis and recognition from Vietnamese speech,” *Journal of Signal Processing*, vol. 20, no. 10, pp. 1423–1432, 2013.

[44] J. Zhipeng and H. Chengwei, “High-order Markov random fields and their applications in cross-language speech recognition,” *Cybernetics and Information Technologies*, vol. 15, no. 4, pp. 50–57, 2015.

[45] T. L. T. Dao, V. L. Trinh and H. Q. Nguyen, “GMM for emotion recognition of Vietnamese,” *Journal of Computer Science and Cybernetics*, vol. 33, no. 3, pp. 229–246, 2017.

[46] T. L. T. Dao, V. L. Trinh and H. Q. Nguyen, “Deep convolutional neural networks for emotion recognition of Vietnamese,” *International Journal of Machine Learning and Computing*, vol. 10, no. 5, pp. 692–699, 2020.

[47] Z. Xiao, D. Wu, X. Zhang and Z. Tao, “A cross-corpus recognition of emotional speech,” in *Proc. ISCID*, Hangzhou, China, pp. 42–46, 2016.

[48] E. N. N. Ocquaye, Q. Mao, H. Song, G. Xu and Y. Xue, “Dual exclusive attentive transfer for unsupervised deep convolutional domain adaptation in speech emotion recognition,” *IEEE Access*, vol. 7, pp. 93847–93857, 2019.

[49] M. Papakostas, E. Spyrou, T. Giannakopoulos, G. Siantikos, D. Sgouropoulos *et al.,* “Deep visual attributes *vs*. hand-crafted audio features on multidomain speech emotion recognition,” *Computation*, vol. 5, no. 2, pp. 1–15, 2017.

[50] P. Boersma, “Praat, a system for doing phonetics by computer,” *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[51] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*. Hoboken, NJ: Prentice Hall Press, 2010.

[52] T. L. T. Dao, V. L. Trinh, H. Q. Nguyen and X. T. Le, “Influence of the spectral characteristics of the signal speech to emotion recognition of Vietnamese,” in *Proc. FAIR*, Danang, Vietnam, pp. 36–43, 2017.

[53] J. L. Devore, *Probability and Statistics for Engineering and the Sciences*, 8th ed., California, USA: Brooks/Cole, 2010.

[54] B. Felix, A. Paeschke, M. Rolfes, W. F. Sendlmeier and B. Weiss, “A database of German emotional speech,” in *Proc. EUROSPEECH*, Lisbon, Portugal, pp. 1517–1520, 2005.

[55] A. Revathi, N. Sasikaladevi, R. Nagakrishnan and C. Jeyalakshmi, “Robust emotion recognition from speech: Gamma tone features and models,” *International Journal of Speech Technology*, vol. 21, no. 3, pp. 723–739, 2018.

[56] H. K. Mishra and C. C. Sekhar, “Variational Gaussian mixture models for speech emotion recognition,” in *Proc. ICAPR*, Kolkata, India, pp. 183–186, 2009.

[57] I. Luengo, E. Navas and I. Hernáez, “Feature analysis and evaluation for automatic emotion identification in speech,” *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 490–501, 2010.

[58] S. Amarakeerthi, T. L. Nwe, L. C. D. Silva and M. Cohen, “Emotion classification using inter-and intra-subband energy variation,” in *Proc. INTERSPEECH*, Florence, Italy, pp. 1569–1572, 2011.

[59] P. Shen, Z. Changjun and X. Chen, “Automatic speech emotion recognition using support vector machine,” in *Proc. EMEIT*, Heilongjiang, China, pp. 621–625, 2011.

[60] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier *et al.,* “Deep neural networks for acoustic emotion recognition: Raising the benchmarks,” in *Proc. ICASSP*, Prague, Czech Republic, pp. 5688–5691, 2011.

[61] Y. Pan, P. Shen and L. Shen, “Speech emotion recognition using support vector machine,” *International Journal of Smart Home*, vol. 6, no. 2, pp. 101–108, 2012.

[62] Y. Jin, P. Song, W. Zheng and L. Zhao, “A feature selection and feature fusion combination method for speaker-independent speech emotion recognition,” in *Proc. ICASSP*, Florence, Italy, pp. 4808–4812, 2014.

[63] M. Gjoreski, H. Gjoreski and A. Kulakov, “Machine learning approach for emotion recognition in speech,” *Informatica*, vol. 38, no. 4, pp. 377–384, 2014.

[64] S. Mao, D. Tao, G. Zhang, P. C. Ching and T. Lee, “Revisiting hidden Markov models for speech emotion recognition,” in *Proc. ICASSP*, Brighton, UK, pp. 6715–6719, 2019.

[65] M. Seo and M. Kim, “Fusing visual attention cnn and bag of visual words for cross-corpus speech emotion recognition,” *Sensors*, vol. 20, no. 19, pp. 1–21, 2020.

[66] M. Lech, M. Stolar, C. Best and R. Bolia, “Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding,” *Frontiers in Computer Science*, vol. 2, no. 14, pp. 1–14, 2020.

[67] F. Haider, S. Pollak, P. Albert and S. Luz, “Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods,” *Computer Speech & Language*, vol. 65, no. 101119, pp. 1–10, 2021.

[68] K. Chauhan, K. K. Sharma and T. Varma, “Speech emotion recognition using convolution neural networks,” in *Proc. ICAIS*, Coimbatore, India, pp. 1176–1181, 2021.