Speech Emotion Recognition using Convolutional Long Short-Term Memory Neural Network and Support Vector Machines

Nattapong Kurpukdee*‡, Tomoki Koriyama†, Takao Kobayashi†,
Sawit Kasuriya‡, Chai Wutiwiwatchai‡, and Poonlap Lamsrichan*
* TAIST Tokyo Tech, ICTES Program, Kasetsart University, Thailand,
† School of Engineering, Tokyo Institute of Technology, Yokohama, 226-8502 Japan,
‡ NECTEC, National Science and Technology Development Agency (NSTDA),
112 Pahonyothin Road, Pathumthani, 12120, Thailand
E-mail : nattapong.kurpukdee@nectec.or.th, {koriyama, takao.kobayashi}@ip.titech.ac.jp,
{sawit.kasuriya, chai.wutiwiwatchai}@nectec.or.th, fengpll@ku.ac.th

***Abstract*—In this paper, we propose a speech emotion recognition technique using convolutional long short-term memory (LSTM) recurrent neural network (ConvLSTM-RNN) as a phoneme-based feature extractor from raw input speech signal. In the proposed technique, ConvLSTM-RNN outputs phoneme-based emotion probabilities to every frame of an input utterance. Then these probabilities are converted into statistical features of the input utterance and used for the input features of support vector machines (SVMs) or linear discriminant analysis (LDA) system to classify the utterance-level emotions. To assess the effectiveness of the proposed technique, we conducted experiments in the classification of four emotions (anger, happiness, sadness, and neutral) on IEMOCAP database. The result showed that the proposed technique with either of SVM or LDA classifier outperforms the conventional ConvLSTM-based one.**

Index Terms: speech emotion recognition, convolutional long short-term memory neural network (ConvLSTM), support vector machines (SVMs), linear discriminant analysis (LDA)

I. Introduction

In human-machine interaction (HMI), we are still far from being able to fully communicate with machines because it is difficult for machines to interpret some paralinguistic information appearing in the spoken language such as emotions. Speech emotion recognition (SER), which aims to classify speaker's emotional states through speech signals, is one of the essential tasks for making HMI more natural and realistic. Although SER has been widely studied and attracting researchers' attention, the performance of SER systems developed so far remains relatively low, especially for spontaneous conversational speech. Consequently, improvement of SER performance is a crucial problem to be solved in HMI research area.

Many researchers have proposed various feature extraction and classification techniques [1] for SER at frame-level or utterance-level. Many successful cases of the frame-level approach were shown based on the powerful classification methods such as Gaussian mixture models (GMMs) [2], hidden Markov models (HMMs) [3], support vector machines (SVMs) [4] and linear discriminant analysis (LDA) [5]. These techniques used low-level features for constructing the model of each emotional state directly. In the utterance-level approach, the global statistical features derived from low-level features were used for SER. For example, Naive Bayes classifier with statistical features of pitch, energy, zero crossing rate, and mel frequency cepstral coefficients (MFCCs) was used for classifying the emotional state of each utterance [6].

Recently, it has been demonstrated that the combination of multiple features such as acoustic and lexical features [7] or the combination of multiple techniques such as GMMs and SVMs [8] outperforms the case using only single feature or technique. More recently, deep learning techniques have been applied to SER for obtaining statistical feature representations from low-level acoustic features in combination with SVMs or extreme learning machine (ELM), and it has been shown that they can achieve state-of-the-art performance [9], [10]. Moreover, convolutional neural networks (CNNs) have been applied to SER [11]–[14] with the promising results. However, the performance of CNN-based approach to SER on IEMOCAP database remains far from that of humans [15].

In this paper, we examine the ability of CNN-based technique combined with long short-term memory neural networks (LSTMs) on English speech emotion database, IEMOCAP database. Then, we propose a new technique for improving the performance of convolutional LSTM (ConvLSTM) recurrent neural network (RNN)-Based SER. A key idea of the proposed technique is to use ConvLSTM-RNN as a phoneme-based feature extractor from raw input speech signal. Most of the conventional SER techniques have not considered phoneme-based information. By contrast, in the proposed technique, ConvLSTM-RNN outputs phoneme-based emotion probabilities to every frame of an input utterance. Although IEMOCAP database has utterance-level emotion label information, every word in one utterance does not always have the same emotion information as the utterance-level label. Thus we integrate frame-level phoneme-based emotion probabilities to determine

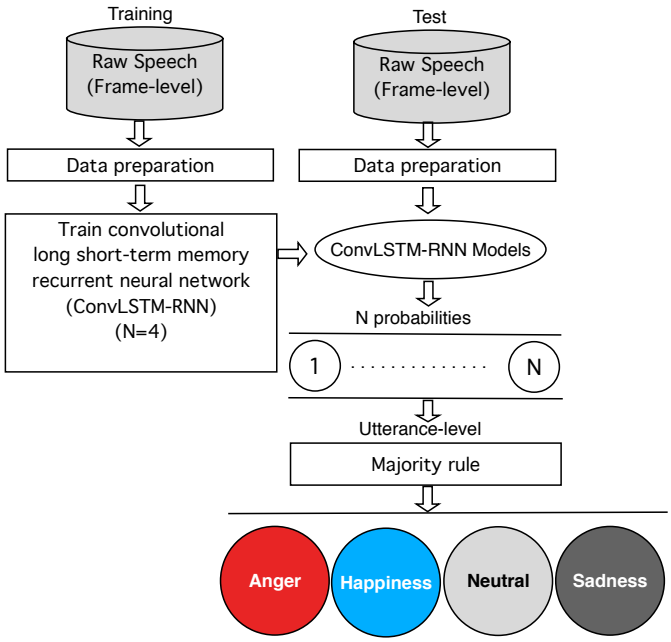


Fig. 1: Block diagram of the conventional framework of speech emotion recognition using ConvLSTM-RNN (baseline).

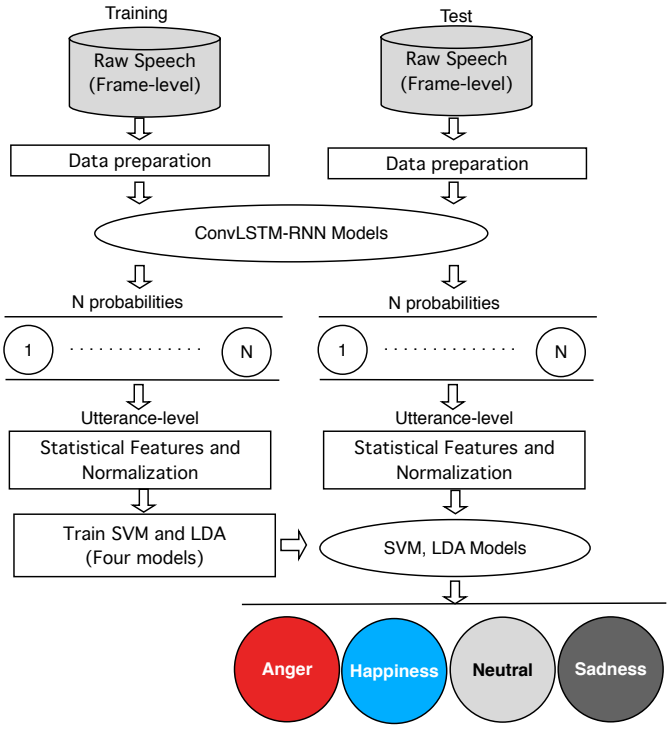


Fig. 2: Block diagram of proposed framework.

the emotional state in each utterance. For this purpose, we convert the ConvLSTM-RNN outputs into statistical features of the input utterance and use them for the input features of SVMs or LDA system to classify the utterance-level emotional states.

This paper is organized as follows. The next section describes an overview of the conventional techniques using CNNs. Section 3 describes the proposed framework, training algorithm, and feature extraction. Experimental setting and results will be explained in section 4, and the last section is the conclusion.

II. Related Work

Recently, CNNs have shown the significant performance in pattern recognition area. The strength of CNN is convolution layer, which works like a feature extraction process. When CNN is applied to SER, appropriate choice of effective input features becomes one of importance issues.

In this context, many attempts have been reported using CNNs to perform SER [11]–[15], [18]. In these techniques, selected input features include spectrogram [12], [13], [15], short-time Fourier transform (STFT) [14], and raw speech signal [11], [18]. Although the spectrogram is commonly used in CNN-based approach, an alternative way is to use raw speech signal with 1-dimensional long sequences [18] or 2-dimensional [11] data format, and this approach has shown an improvement of recognition performance. An overview of the CNN-based SER system using raw speech signal as the input features is shown in Figure 1.

In the system with 2-dimensional data format, CNN is replaced with ConvLSTM-RNN. Data preparation process segments 1-dimensional input speech signal into frames and then reshapes them to 2-dimensional data array with 100 rows and 100 columns. In the training phase, ConvLSTM-RNN models corresponding to respective emotions are trained so that each of them outputs probability of being the labeled emotion to the training input. In the testing phase, each ConvLSTM-RNN model outputs its own emotion probability to a testing input. Since the model output is frame-level emotion information, a majority voting rule is used to determine the utterance-level emotion for each input utterance.

III. ConvLSTM-RNN with Phoneme-based Emotion Information for SER

A. Overview of Proposed technique

In this section, we describe the details of our proposed framework. Although the CNN-based technique showed a good result in SER [11], we found that its performance is not so high when we apply it to IEMOCAP database. One of the possible reasons is that it uses relatively long duration input data regardless of whether a certain emotion is consistently expressed in whole utterance duration. In fact, although IEMOCAP database has emotion labels annotated at utterance level, it does not mean all the portions in one utterance have the consistent emotional expression. We found that the phone-level information is useful in such a case, and we also found that LSTM-RNN for representing emotional features [10] gives improved performance. From these facts, in this paper, we examine an alternative approach using ConvLSTM-RNN with phone-level alignment.

Basically, the framework of the proposed technique is very similar to the baseline system shown in Figure 1. However,

we use 1-dimensional short sequences as the input data instead of 2-dimensional one. Then ConvLSTM-RNN models learn phoneme-based emotion probabilities instead of simple emotion probabilities. Since ConLSTM-RNN models output frame-level information, we convert emotion probabilities into utterance-level statistical features and normalize them by calculating z-scores. After that, we apply SVM- or LDA-based classifier to obtain utterance-level emotional states from the normalized statistical features.

B. Convolutional Recurrent Neural Network Target

We utilize ConvLSTM-RNN as a model which converts raw speech data into frame-level emotion features. In the training speech corpus, the emotion labels are often annotated at utterance-level but it does not always mean that every phrase, word, or frame has the same emotion. To treat the emotion labels at frame-level, we initially assume that every frame in one utterance has the same emotion label. Then we examine three scenarios as follows.

First, ConvLSTM-RNN with 1-dimensional input data will learn four emotion probabilities: anger, happiness, neutral and sadness. We compare the effectiveness of 1-dimensional input data for SER with conventional framework using 2-dimensional input data.

Secondly, the corpus that will be used in experiments contains 52 phoneme categories covering silent, laughter, garbage, lip smack, and breathing, and we do not use lip smack category. We use silent, laughter, garbage, and breathing categories together with the remaining 47 phoneme categories. Then ConvLSTM-RNN with 1-dimensional input data will learn phoneme-based emotion probabilities, specifically, 4 emotion probabilities for 47 phoneme categories, $N = (4 \times 47) + 4 = 192$ emotion probabilities in total.

Thirdly, we group 47 phonemes into 8 phoneme classes that include vowels, stops, affricates, fricatives, aspirates, liquids, semivowels, and nasals together with silent, laughter, garbage, and breathing categories. Then ConvLSTM-RNN with 1-dimensional input data will learn phoneme-class-based emotion probabilities, 4 emotion probabilities for 8 phoneme classes, namely $N = (4 \times 8) + 4 = 36$ emotion probabilities.

C. Statistical Features

The statistical features are computed from statistics of ConvLSTM-RNN's output of all frames in one utterance. Specifically, let $F_s(P_k)$ denote the output value of the k-th emotion probability at frame s. We compute the statistical features for each utterance as

$$a_1^k = \max_{s \in U} F_s(P_k) \tag{1}$$

$$a_2^k = \min_{s \in U} F_s(P_k) \tag{2}$$

$$a_3^k = \frac{1}{|U|} \sum_{s \in U} F_s(P_k) \tag{3}$$

$$a_4^k = \sqrt{\frac{1}{|U|} \sum_{s \in U} (F_s(P_k) - a_3^k)^2} \tag{4}$$

TABLE I: Number of Emotion Utterances and Duration Per Emotion Category.

Emotion	Utterance	Duration (hours)
Anger	1,103	1.38
Happiness	595	0.71
Excitement	1,040	1.38
Neutral	1,708	1.85
Sadness	1,084	1.65
Total	**5,530**	**6.89**

where U denotes the set of all frames. These statistical features a_1^k, a_2^k, a_3^k, and a_4^k correspond to the maximum, minimum, mean, and standard deviation of the k-th output of ConvLSTM-RNN, respectively.

D. Feature Normalization

We normalize the statistical features mentioned above to obtain z-scores as follows.

$$z_i^k = \frac{a_i^k - \mu_i}{\sigma_i}, \quad i = 1, 2, 3, 4 \tag{5}$$

where

$$\mu_i = \frac{1}{N} \sum_{e=1}^{N} a_i^e \tag{6}$$

$$\sigma_i = \sqrt{\frac{1}{N} \sum_{e=1}^{N} (a_i^e - \mu_i)^2} \tag{7}$$

The values of μ and σ correspond to the mean and standard deviation of the i-th statistical feature in each utterance, respectively.

E. Emotion Recognition Techniques

We use two classification techniques to obtain utterance-level emotion: SVMs and LDA.

1) Support Vector Machines (SVMs): SVM is one of powerful classification methods and known to be simple and efficient in high dimensional spaces. Even if the training set is small, SVMs can provide high performance. In this paper, we investigate three kernels of SVMs including linear, radial basis function (RBF), and polynomial kernels. Regarding the parameter setting in RBF and polynomial kernels, we define γ (gamma) parameter equal to $\frac{1}{N}$. For the polynomial kernel, the degree parameter is determined to be three.

2) Linear Discriminant Analysis (LDA): LDA is a simple classification technique with a linear decision boundary. There is no more parameter to setup. We can use LDA for classifier or dimensionality reduction. In this paper, we use LDA as a classifier. For the parameter setting in LDA, we define only solver parameter to singular value decomposition.

IV. Experiments

A. Experimental Setup

We evaluated the performance of the proposed technique on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [16]. The database is an acted, multimodal and

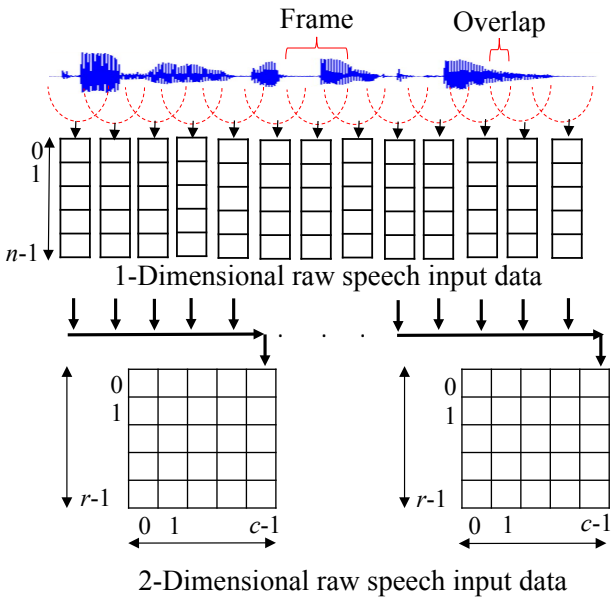


Fig. 3: Data preparation process.

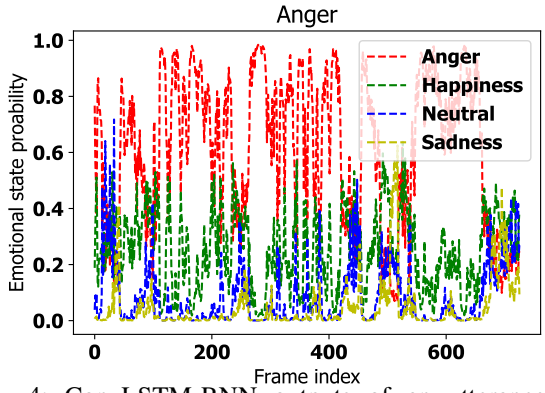
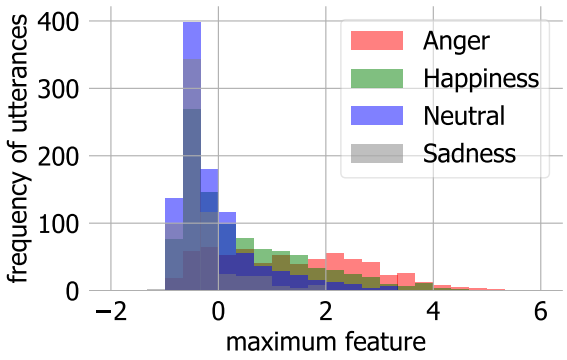


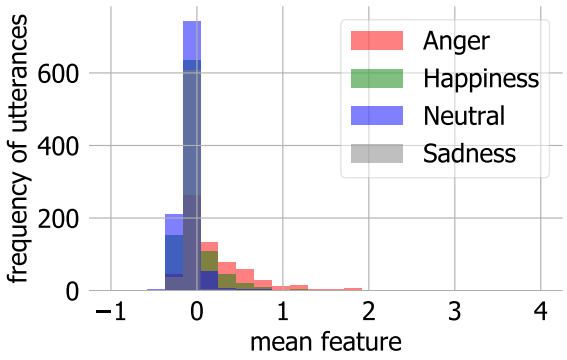
Fig. 4: ConvLSTM-RNN outputs of an utterance with 4 emotional state.

multispeaker one under spoken communication scenarios. It contains audiovisual data from 10 different actors (5 males, and 5 females) covering 10 emotional categories of neutral, happiness, sadness, anger, surprise, fear, disgust, frustration, excited, and others. In the experiments, we focused on four basic emotions, that is anger, happiness, sadness, and neutral. Table I shows the numbers of utterances and durations of speech data labeled with 4 emotion categories. As shown in the table, the number of happiness utterances is less than other emotions significantly. To avoid unbalanced data in the training and testing, we increased the utterances of happiness category by adding those of excitement category, because we believe that excitement and happiness share more similar characteristics in emotional definition than others.

A five-fold leave-one-out cross-validation (LOOCV) with speaker independent scheme was applied in all experiments. It implies that all utterances in each emotion category are divided into five groups. We used three groups from each emotion for the training set, and used next group for development set, then



(a) maximum of AA-phoneme.



(b) mean of AA-phoneme.

Fig. 5: Histogram of statistical features of AA-phoneme with anger emotion label (AA-anger).

TABLE II: Setup of 1-Dimensional ConvLSTM-RNN in the proposed framework.

Layer	Input
Input	400*1
Convolution layer 1	200 feature map, convolution window size: 5*1, pooling window size: 2*1
Convolution layer 2	100 feature map, convolution window size: 5*1, pooling window size: 2*1
Convolution layer 3	50 feature map, convolution window size: 5*1, pooling window size: 2*1
Full connection layer 1	LSTM neurons:1024
Output layer	**4, 36, 192 output**

used the remaining group for the test set. All the results from each fold were averaged at the end.

The input signal was sampled at 16 kHz and converted into frames using a 25ms window with a 10ms frame shift. For the baseline system (N=4), the input data was reshaped from the sequence of frames to 100 $\times$ 100 2-dimensional data. Since each frame contains 400 samples, we split it into 4 sub-frames, each of which contains 100 samples. Then we formed 100 $\times$ 100 2-dimensional sample array (r=100, c=100) by using 100 sub-frames obtained from 25 consecutive frames as shown in Figure 3. In the baseline system, ConvLSTM-RNN used two

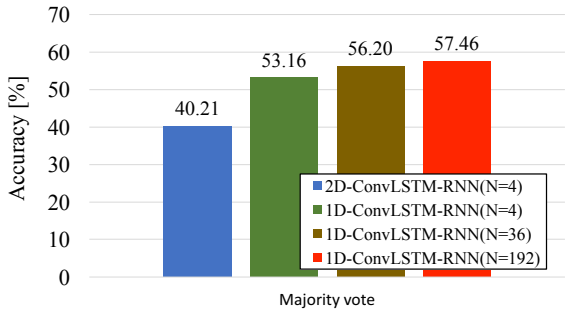


Fig. 6: Accuracy comparison in the conventional framework using majority voting rule.

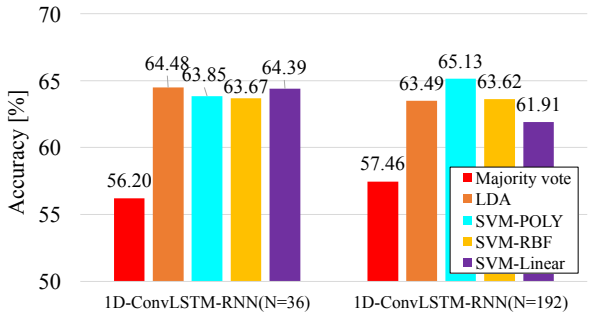


Fig. 7: Accuracy comparison among classification approaches in the proposed framework.

TABLE III: Confusion Matrix Of Majority Vote In 192 Phoneme-Based Emotions

		Output (Accuracy %)			
	Emotion	Anger	Happiness	Neutral	Sadness
Input	Anger	**65.00%**	19.67%	12.15%	3.17%
	Happiness	19.69%	**52.48%**	17.00%	10.83%
	Neutral	5.50%	23.24%	**54.39%**	16.86%
	Sadness	3.69%	15.87%	18.27%	**62.18%**

TABLE IV: Confusion Matrix Of SVM With Polynomial Kernel In 192 Phoneme-Based Emotions

		Output (Accuracy %)			
	Emotion	Anger	Happiness	Neutral	Sadness
Input	Anger	**68.72%**	15.59%	13.06%	2.63%
	Happiness	9.85%	**61.04%**	20.24%	8.87%
	Neutral	4.33%	17.10%	**67.04%**	11.53%
	Sadness	2.21%	11.90%	21.22%	**64.67%**

convolutional layers with a setting of 200 and 50 feature maps, pooling size = (2, 2), and filter shape = (5, 5). Moreover, ConvLSTM-RNN had two LSTM layers with 500 nodes each, and mini-batch of 128 samples were used to learn.

In the proposed SER system shown in Figure 2, raw speech signal of each frame is considered to be an input vector of one dimension with 400 data points for ConvLSTM-RNN. The networks contained three convolution layers with valid border mode and activation function was tanh. ConvLSTM-RNN had single LSTM layer consisting of 1024 nodes with sigmoid activation function. A dropout rate of 0.25 was applied to all hidden layers and Adam optimization method with large-batch of 1024 samples were used to learn the models. In the output layer, softmax activation function was used for obtaining probabilities of 4 emotions, 36 phoneme-class based emotions, and 192 phoneme-based emotions in accordance with three scenarios as described in 3.1. In addition, to avoid overfitting of the network, we selected ConvLSTM-RNN models by considering the tendency of training loss rate and validation loss rate with the development set in the training process. Details of the ConvLSTM-RNN setting are shown in Table II. The number of dimensions of statistical features for SVMs and LDA input was $36 \times 4 = 144$ in the phoneme-class-based emotion case and $192 \times 4 = 768$ in the phoneme-based case, respectively.

The ConvLSTM-RNN was implemented using Keras toolkit [19] with Theano backend and Scikit-learn [17] machine learning in Python was used to train SVMs and LDA.

B. Experimental Results

In order to assess the performance of the proposed technique, we used the conventional ConvLSTM-RNN-based SER system with 2-dimensional raw speech input as the baseline system. The accuracy of the baseline system was 40.21% and it is noted that this result is quite similar to that of the study applying CNN to SER [15] with the same database.

Figure 4 shows an example of 1D-ConvLSTM-RNN(N=4) outputs for an utterance labeled as anger. From figure 4, the probability of each emotional state changes within one utterance. In this example, emotional state proability of anger gives the highest value among 4 emotions in most frames. Thus the emotional state for this utterance can be classified into anger.

Figure 5 shows an example of histograms of normalized statistical features, specifically, the maximum and mean of the output from 1D-ConvLSTM-RNN(N=192) for AA-phoneme samples included in anger utterances (AA-anger). From this figure, we can see that statistical features of anger emotion have outliers and they do not overlap with those of other emotions. This fact could lead to an advantage of using combination with other features to improve the performance of the proposed SER system.

Figure 6 shows the accuracies of the proposed technique, where 1D-ConvLSTM-RNN with N=4, 36, and 192 represent the systems based on the three scenarios with 1-dimensional raw speech input. In the figure, the result of the baseline system, 2D-ConvLSTM RNN(N=4), is also shown. It can be seen that the proposed technique with/without using phoneme-class-/phoneme-based emotion probabilities gave better performance than the baseline. The best accuracy was 57.46% using 192 phoneme-based emotion probabilities. The performance for each emotion is shown in Table III. The highest emotion recognition accuracy was 65.00% for anger, followed by sadness and neutral, and happiness was the lowest.

In the experiment mentioned above, to obtain utterance level emotion category, we applied a simple majority voting rule. To confirm the effectiveness of a machine learning approach to obtaining utterance-level emotion category from

frame-level information, we conducted further performance comparison of machine learning techniques, SVMs and LDA with ConvLSTM-RNN model outputs.

In this experiment, we compared the accuracies of SVMs and LDA classifiers with 192 phoneme-based and 36 phoneme-class-based emotion probabilities. The results are shown in Figure 7. From the figure, it is seen that the proposed techniques applying SVMs and LDA outperformed significantly the baseline. More specifically, the accuracies became 4.45% to 7.67% higher than the baseline in 192 phoneme-based emotions, and 7.47% to 8.28% higher in 36 phoneme-class-based emotions. The best performance 65.13% was achieved by SVM with the polynomial kernel in 192 phoneme-based emotions and the performance for individual emotion is shown in Table IV. From the result, we can see that anger emotion recognition accuracy was highest, followed by neutral and sadness, and happiness was the most confused emotion.

V. Conclusions

In this paper, we have proposed a speech emotion recognition technique using convolutional long short-term memory recurrent neural network. In the proposed technique, frame-level phoneme-based emotion probabilities were obtained from raw input speech signal with ConvLSTM-RNN and converted into statistical features of the input. Then, we utilized SVMs or LDA classifier to recognize the emotional state at the utterance level. Experimental results on IEMOCAP database showed that the proposed technique could improve the performance of speech emotion recognition.

In future work, we will apply SVM parameter tuning to achieve a better result. Then we will find feature selection techniques for dimensionality reduction, as well as performance improvement and processing time reduction. Applying the proposed technique to other databases and evaluating the performance of different deep learning neural network techniques and different features for SER are also our future work.

References

[1] S. Lugovi, I. Dun, and M. Horvat, “Techniques and applications of emotion recognition in speech,” in Proc. MIPRO, pp. 1278–1283, 2016.

[2] D. Neiberg, K. Elenius, and K. Laskowski, “Emotion recognition in spontaneous speech using GMMs,” in Proc. INTERSPEECH - ICSLP, pp. 809–812, 2006.

[3] B. Schuller, G. Rigoll, and M. Lang, “Hidden Markov model-based speech emotion recognition,” in Proc. ICASSP, pp. 401–404, 2003.

[4] N. R. Kanth, “Efficient speech emotion recognition using binary support vector machines & multiclass SVM,” in Proc. ICCIC, pp. 1–6, 2015.

[5] M. Murugappan, N. Q. I. Baharuddin, and S. Jerritta, “DWT and MFCC based human emotional speech classification using LDA,” in Proc. ICoBE, pp. 203–206, 2012.

[6] S. K. Bhakre, “Emotion recognition on the basis of audio signal using naive bayes classifier,” in Proc. ICACCI, pp. 2363–2367, 2016.

[7] Q. Jin, C. Li, S. Chen, and H. Wu, “Speech emotion recognition with acoustic and lexical features,” in Proc. ICASSP, pp. 4749–4753, 2015.

[8] H. Hu, M.-X. Xu, and W. Wu, “GMM supervector based SVM with spectral features for speech emotion recognition,” in Proc. ICASSP, pp. 413–416, 2007.

[9] K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” in Proc. INTERSPEECH, pp. 223–227, 2014.

[10] J. Lee and I. Tashev, “High-level feature representation using recurrent neural network for speech emotion recognition,” in Proc. INTERSPEECH, pp. 1537–1540, 2015.

[11] B. Zhang, C. Quan, F. Ren, “Study on CNN in the recognition of emotion in audio and images,” in Proc. ICIS, pp.1–5, 2016.

[12] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, “Learning salient features for speech emotion recognition using convolutional neural networks,” IEEE Trans. Multimedia., vol. 16, no. 8, pp. 2203–2213, 2014.

[13] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, “Speech emotion recognition using CNN,” in Proc. ACMMM, pp. 801–804, 2014.

[14] W. Lim, D. Jang, T. Lee, “Speech emotion recognition using convolutional and recurrent neural networks,” in Proc. APSIPA, pp. 1–4, 2016

[15] W. Q. Zheng, J. S. Yu, and Y. X. Zou, “An experimental study of speech emotion recognition based on deep convolutional neural network,” in Proc. ACII, pp. 827–831, 2015.

[16] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP: Interactive emotional dyadic motion capture database,” Lang. Resour. Eval., vol. 42, no. 4, pp. 335–359, 2008.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B.Thirion, et al., “Scikit-learn: Machine learning in Python,” J. Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[18] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, Mihalis A Nicolaou, B. Schuller, S. Zafeiriou, “Adieu Features? End-To-End Speech Emotion Recognition Using A Deep Convolutional Recurrent Network” in Proc. ICASSP, pp. 5200–5204, 2016.

[19] Chollet, François and others, “Keras", https://github.com/fchollet/keras, 2015