INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Multi-stream Attention-based BLSTM with Feature Segmentation for Speech Emotion Recognition

Yuya Chiba[1], *Takashi Nose*[1], *Akinori Ito*[1]

[1]Graduate School of Engineering, Tohoku University, Japan
{yuya@spcom.ecei, tnose@m, aito@spcom.ecei}.tohoku.ac.jp

Abstract

This paper proposes a speech emotion recognition technique that considers the suprasegmental characteristics and temporal change of individual speech parameters. In recent years, speech emotion recognition using Bidirectional LSTM (BLSTM) has been studied actively because the model can focus on a particular temporal region that contains strong emotional characteristics. One of the model's weaknesses is that it cannot consider the statistics of speech features, which are known to be effective for speech emotion recognition. Besides, this method cannot train individual attention parameters for different descriptors because it handles the input sequence by a single BLSTM. In this paper, we introduce feature segmentation and multi-stream processing into attention-based BLSTM to solve these problems. In addition, we employed data augmentation based on emotional speech synthesis in a training step. The classification experiments between four emotions (i.e., anger, joy, neutral, and sadness) using the Japanese Twitter-based Emotional Speech corpus (JTES) showed that the proposed method obtained a recognition accuracy of 73.4%, which is comparable to human evaluation (75.5%).

Index Terms: emotion recognition, segmental feature, multi-stream emotion recognition, data augmentation

1. Introduction

In the last two decades, speech emotion recognition has been studied actively [1–3] to enhance the performance of speech applications, such as call-center systems [4], academic information systems [5], and tutor systems [6]. The speech emotion recognition performance has gradually improved thanks to many efforts examining many kinds of features and classifiers. For example, the best-known feature for speech emotion recognition is the statistics of low-level descriptors (LLDs) calculated from the entire speech (e.g., [7, 8]).

The LLDs contain features that represent the acoustic signal's characteristics, such as energy, fundamental frequency (F_0), and Mel-frequency cepstral coefficients (MFCC). Therefore, statistics of the LLDs can capture the suprasegmental characteristics of the speech comprehensively. In terms of the classifiers, many recent studies employ a deep neural network, including Multilayer Perceptron (MLP) [9], Convolutional Neural Network (CNN) [10, 11] and Long Short-Term Memory (LSTM) [12, 13]. In particular, an attention-based Bidirectional LSTM (BLSTM) has attracted attention because it can focus on a particular temporal region of speech that contains strong emotional expressions [12, 14]. These studies employ the LLDs [14] and spectrogram [12, 15] as the input sequence. The networks of these methods could focus on emotionally salient parts of an utterance.

One problem of approaches using attention-based BLSTM [12, 14] is that they cannot sufficiently capture the suprasegmental characteristics of the speech, which were proved to be effective in conventional studies. Marsamadi et al. [14] used velocity coefficients of the LLDs, but they were not enough to capture the statistics of features like maximum or minimum. One solution to this problem is feature segmentation. For example, Mao et al. [16] conducted feature segmentation by calculating statistics of the LLDs and using them for segment-level emotion recognition. They showed the effectiveness of aggregating segment-level decisions for utterance-level recognition. Such feature segmentation seems to be also effective for attention-based BLSTM to capture the suprasegmental characteristics of speech features. Another problem with the conventional methods [12, 14] is that the network cannot train individual attention parameters for different descriptors because a single BLSTM handles the input sequence. The property of speech is represented by the three aspects of intensity, pitch, and timbre, and humans control these parameters individually to some extent when making an emotional speech. Therefore, the speech emotion recognition model should be constructed to focus on different temporal regions for individual speech properties.

In view of the above, this paper proposes a speech emotion recognition technique using a multi-stream attention-based BLSTM with feature segmentation. In this method, the network can consider both suprasegmental characteristics and the temporal change of the input sequence by feature segmentation. In addition, the proposed method can focus on the different temporal regions according to speech features by training the individual attention parameters for respective streams. Here, training the LSTM-based network requires large-scale emotional speech data. In this paper, we conducted data augmentation based on emotional speech synthesis [17] in the training step.

2. Speech Emotion Recognition using Attention-based BLSTM

2.1. Conventional Single-stream Speech Emotion Recognition with LLD Feature

This study focuses on 4-class emotion recognition. The upper panel of Fig. 1 shows the conventional method [14]. It estimates the emotion category from the standard LLD sequence (i.e., power, F_0, and MFCC). When the LLD sequence $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T)$ is input, the BLSTM outputs vector sequence $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_T)$. The attention-based BLSTM calculates the attention weights α_t based on the inner-product between the attention parameter vector $\mathbf{u}$ and $\mathbf{y}_t$ at each time frame t. Then, the network takes the weighted sum of $\mathbf{Y}$ using α_t and obtains the utterance-level representation $\mathbf{z}$. $\mathbf{z}$ is input to the succeeding layer and the network obtains the posterior probability to the target emotion. The attention-based BLSTM is assumed to focus on the particular temporal region of the speech that contains strong emotional characteristics.

http://dx.doi.org/10.21437/Interspeech.2020-1199

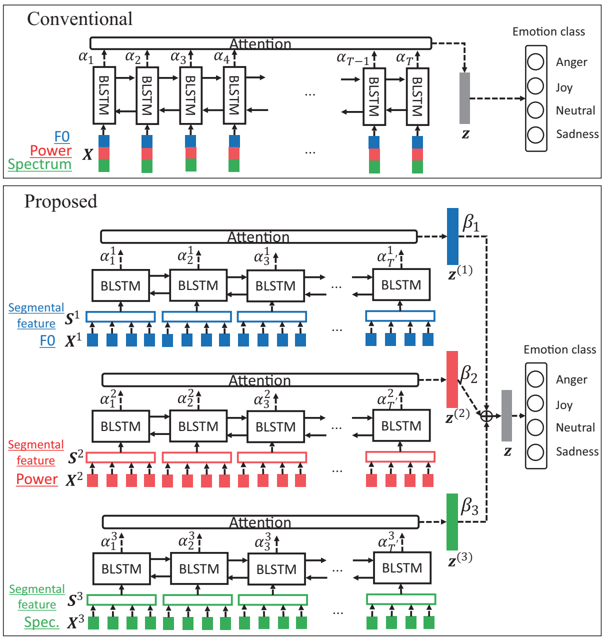

Figure 1: *Overview of speech emotion recognition using attention-based BLSTM. The upper panel shows the conventional method, and the lower panel shows the proposed multi-stream speech emotion recognition model using segmental features.*

2.2. Proposed Multi-stream Speech Emotion Recognition with Segmental Feature

The lower panel of Fig. 1 shows the proposed method. First, we calculate the statistics of the LLD sequence $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_{T'})$ from $\mathbf{X}$ at a fixed interval, and use them for the input of the network. In this approach, the network can capture the suprasegmental characteristics at a local region and the temporal change of the input sequence. We calculated the statistics of LLDs to construct the segmental features. The segmental features were separated into the features related to the power, F_0, and spectrum and input to the different attention-based BLSTMs. Each attention-based BLSTM processes the speech parameter separately. Let the segmental feature of the k-th speech parameter be $\mathbf{S}^{(k)} = (\mathbf{s}_1^{(k)}, \mathbf{s}_2^{(k)}, \cdots, \mathbf{s}_{T'}^{(k)})$ and the output vector sequence of the k-th BLSTM be $\mathbf{Y}^{(k)} = (\mathbf{y}_1^{(k)}, \mathbf{y}_2^{(k)}, \cdots, \mathbf{y}_{T'}^{(k)})$. Each BLSTM has its own attention parameter $\mathbf{u}^{(k)}$ and the attention weights $\alpha_t^{(k)}$ are calculated by processing the inner-product between the attention parameter and output vector. This process is represented as

$$\alpha_t^{(k)} = \frac{\exp\left(\mathbf{u}^{(k)\top}\mathbf{y}_t^{(k)}\right)}{\sum_{\tau=1}^{T}\exp\left(\mathbf{u}^{(k)\top}\mathbf{y}_\tau^{(k)}\right)} \tag{1}$$

Then, the network takes the weighted sum of $\mathbf{Y}^{(k)}$ using $\alpha_t^{(k)}$ and obtains the utterance-level representation of the k-th stream $\mathbf{z}^{(k)}$. Therefore, each attention-based BLSTM network obtains a single representation vector from the respective input sequence. Finally, the representation vectors are added with stream weight β_k as:

$$\mathbf{z} = \beta_k \mathbf{z}^{(k)} \tag{2}$$

where, β_k represents the weight of the k-th stream. This paper used the same stream weight for all streams. Therefore, the process is equal to taking the average of the representation vectors. The utterance-level representation $\mathbf{z}$ is input to succeeding layers, and the final layer outputs the posterior probability to target emotions. The multi-stream attention-based BLSTM can focus on the temporal regions with strong emotional characteristics for individual speech features by processing the input sequence with the multiple streams.

3. Experimental Data

3.1. Japanese Twitter-based Emotional Speech

In the experiments, we used the Japanese Twitter-based Emotional Speech corpus (JTES) [18], a publicly available emotional speech database. JTES is a well-controlled corpus in terms of the balance of gender and emotion. Moreover, it contains many samples and speakers compared with IEMOCAP [19], a commonly-used evaluation corpus. Therefore, we selected this dataset for the evaluation dataset. The dataset contains emotional speech uttered by 100 Japanese amateur speakers (50 females and 50 males). Each speaker read 200 different sentences (50 sentences × 4 emotions), and so the total number of samples is 20,000 utterances. The contained emotions are anger, joy, neutral, and sadness. The sentences differ for the different emotions, but are the same among speakers. The utterances were recorded at 48 kHz sampling rate with 16-bit quantization, and then down-sampled to 16 kHz. When recording, the speakers were instructed to speak as if they were conveying their emotions to a robot.

Lee [20] obtained an average recall of 81.4% with this dataset under speaker-independent and sentence-closed conditions. However, the speakers of the JTES had no acting experience, and it is not easy to recognize their emotions under speaker-independent and sentence-independent conditions. For example, Takeishi et al. [18] reported that the average recall is 54.0% when using Support Vector Machine (SVM). In the present study, we conducted experiments under speaker-independent and text-independent conditions.

3.2. Dataset Construction Including Augmented Synthetic Speech

The amount of training samples is usually insufficient for speech emotion recognition because collecting emotional speech is a time-consuming and costly task. One way to overcome the lack of training samples is data augmentation. This study employs emotional speech synthesis based on DNN-based parametric speech synthesis [17] for data augmentation. Emotional speech synthesis can create a natural and arbitrary emotional speech that retains the target emotion's prosodic and spectral characteristics. In this study, we conducted speaker and emotion adaptation to achieve emotional speech synthesis. We first trained speaker-independent DNNs using multiple target speakers contained in ATR Japanese speech database set B [21] and then fine-tuned the network using the training data of JTES in the speaker and emotion adaptation steps. The sentences for synthesis were selected from the Japanese Newspaper Article Sentences corpus (JNAS) [22]. We randomly assigned 100 sentences to each speaker without overlapping. Each sentence was used for the speech synthesis of four emotions. Thus, the sentences were different from speaker to speaker, but common between emotions.

3.3. Summary of Dataset

The JTES was separated so that the training, validation, and test sets did not include the same sentences and the same speakers.

Table 1: *Numbers of samples in the dataset.*

Original dataset

Emotion	Train	Valid.	Test
Anger	2,400	100	100
Joy	2,400	100	100
Neutral	2,400	100	100
Sadness	2,400	100	100
Total	9,600	400	400

Augmented dataset

Emotion	Train	Valid.	Test
Anger	10,400	100	100
Joy	10,400	100	100
Neutral	10,400	100	100
Sadness	10,400	100	100
Total	41,600	400	400

The training set contained 30 utterances of 80 speakers per emotion, and the validation and test sets contained 10 utterances of 10 different speakers per emotion. Therefore, the total numbers of samples were 9,600, 400, and 400 for the training, validation, and test sets, respectively. In the data augmentation process, we synthesized 100 utterances for the respective 80 speakers and 4 emotions. Therefore, the total number of the training data after data augmentation was 41,600 (9,600 utterances $+$ 4 emotions $\times$ 80 speakers $\times$ 100 utterances).

Table 1 shows a summary of the dataset. The table shows the original separation of the JTES and the augmented dataset. In this study, we also examined the performance when using the original dataset for comparison because data augmentation significantly improved recognition accuracy.

4. Experimental Conditions

4.1. Experimental methods

We compared the performance of the four methods. The first one is the LLD_Single, which uses the attention-based BLSTM with 32-dimensional LLD sequence. The network is trained using the original JTES dataset in LLD_Single. The second approach is LLD_Single_DA, which uses the same network and features as LLD_Single, but is trained by the augmented dataset. The third approach is Seg_Single_DA, which uses the attention-based BLSTM with the segmental feature. The final one is the proposed method called Seg_Multi_DA, that uses the multi-stream attention-based BLSTM with the segmental feature. The networks of Seg_Single_DA and Seg_Multi_DA are trained by using the augmented dataset. The overview of the method is described in Section 2.

4.2. Conditions of Features

For the baseline method, the LLDs extracted from OpenSMILE were used as the input sequence. The LLDs contain the 12-dimensional MFCC, F_0, power, voice probability, and zero-crossing rate, and their delta features. The dimension of the feature sequence is 32. The frame shift of the feature extraction is 10 ms. To construct the segmental feature, we computed the statistics of the LLD features every five frames. The kinds of statistics are the same as the feature set used in The INTERSPEECH 2009 Emotional Challenge [7]. Therefore, the segmental feature has 384 dimensions. In the multi-stream condition, we input 24-dimensional features related to power, the 24-dimensional feature related to F_0, and 336-dimensional features related to spectrum to different streams. The features of the training set were scaled with zero mean and unit variance. The validation set and the test set were scaled using the scaling parameters obtained from the training data.

4.3. Conditions of Training and Evaluation

The network was trained to minimize the cross-entropy loss. We investigated the classification performance while changing the number of nodes of the hidden layers as 16, 32, 64, 128, 256, 512, and 1024. The condition that yielded the best accuracy for the validation set was used for the definitive evaluation. The network had dropout layers after each layer excluding the output layer. The dropout rate was set to 0.3. The other parameters were determined by preliminary experiments. The activation functions of the hidden layer and output layer were ReLU and softmax functions. The optimization method was Adam with a learning rate of 0.0005. The minibatch size was 32 and the maximum number of epochs was 30. The parameters for the training network were the same for all methods. For the multi-stream attention-based BLSTM, the numbers of hidden nodes of the three BLSTMs were fixed to the same number.

In the following sections, we show the recognition results for the test set. The estimation results sometimes changed depending on the initial value of the network parameters. Therefore, we repeated the training and evaluation process 10 times and compared the average of them to improve the reliability of the experimental results.

5. Experimental Results

5.1. Recognition Accuracy

Table 2 shows the experimental results. The results of the LLD_Single and LLD_Single_DA methods show that data augmentation significantly improves recognition accuracy. These results indicate that training the attention-based BLSTM requires a large amount of data, and the statistically generated data is useful to compensate for the shortage of emotional speech samples. In addition, the feature segmentation improved the recognition accuracy by 1.1 points, and the multi-stream network improved it by a further 1.6 points. The proposed method's total improvement was 2.7 points, and the obtained definitive average recognition accuracy was 73.4%. For the JTES, Yamanaka et al. [23] annotated the perceived emotion labels using crowdsourcing. They reported that the concordance rate between reference labels and the annotators' perceived emotion labels was 75.5%. They instructed the annotators to ignore the linguistic content and evaluate the acoustic emotional expressions. Therefore, this concordance rate is assumed to be one of the performance limits of emotion recognition with JTES only using acoustic information. The recognition accuracy of the proposed method is comparable to this human evaluation, suggesting that it is necessary to consider other modalities to improve the recognition performance further.

Next, we compared the LLD_Single_DA, Seg_Single_DA, and Seg_Multi_DA methods by statistical tests. We conducted one-way layout ANOVA factoring the methods and obtained a significant difference ($p < 0.05$). Then, we conducted the Tukey-Kramer test to investigate the differences between the methods. From the results, a significant difference was observed between the LLD_Single_DA and Seg_Multi_DA methods ($p < 0.05$). This result suggests that the combination of the segmental feature and multi-stream network contributes to the improvements in recognition performance. However, a significant difference was not observed between LLD_Single_DA and Seg_Single_DA. One of the reasons is that the segment length employed in this study is not long enough to capture the emo-

Table 2: *Accuracy of emotion recognition [%]. The table shows the average accuracy and standard error of 10 trials.*

Method	Data augmentation	Features	Stream	Accuracy (Avg.±SE)
LLD_Single [14]	None	LLD	Single	62.1± 0.82
LLD_Single_DA	Used	LLD	Single	70.7±0.63
Seg_Single_DA	Used	Segmental	Single	71.8±0.68
Seg_Mult_DA	Used	Segmental	Multi	**73.4±0.53**
Human [23]	–	–	–	75.5

Table 3: *Confusion matrix of recognition results for each feature set [%] (underlines show the misclassification ratio over 15.0%).*

LLD_Single_DA

	Anger	Joy	Neutral	Sadness
Anger	**65.7**	16.8	1.1	0.2
Joy	21.3	**59.7**	6.1	3.8
Neutral	7.5	8.2	**69.1**	7.7
Sadness	5.5	15.3	23.7	**88.3**
Total	100	100	100	100

Seg_Single_DA

	Anger	Joy	Neutral	Sadness
Anger	**63.0**	17.0	1.0	0.8
Joy	27.6	**62.3**	8.0	7.7
Neutral	6.8	9.2	**75.0**	4.8
Sadness	2.6	11.5	16.0	**86.7**
Total	100	100	100	100

Seg_Multi_DA

	Anger	Joy	Neutral	Sadness
Anger	**66.1**	19.0	2.3	1.1
Joy	25.3	**62.5**	6.7	4.7
Neutral	5.1	6.5	**74.0**	3.1
Sadness	3.5	12.0	17.0	**91.1**
Total	100	100	100	100

tional characteristics. The conventional studies [16,24,25] modeled the speech section using 250 ms or longer sections and confirmed the effectiveness. Therefore, extending the segment length is expected to be effective in improving recognition performance. In future work, we are going to conduct additional experiments to find the optimum segment length.

5.2. The Tendency of Recognition Results

Table 3 shows the confusion matrices of the recognition results. The results in the tables are the averages of 10 trials. The tables show that sadness is clearly separated from other emotions in all conditions. Besides, the proposed method improved the accuracy of neutral by 4.9 points. On the other hand, the accuracy of anger and joy was around 60%, and these two emotions tended to be confused frequently. The proposed method can alleviate the confusion of these classes, but 20% of the samples are still misrecognized. It is known that these two emotions are distributed at similar positions on the valence axis of the circular emotion model (e.g. [26]) and are intrinsically difficult to discriminate from acoustic information [27,28]. This result also indicates that it is important to combine other information, such as linguistic information, to improve accuracy.

5.3. Analysis of the Attention Weights

Finally, the attention weights obtained from the examined methods are shown in Fig. 2. The trends of the attention weights are slightly different between LLD_Single_DA and other methods. The attention weight of LLD_Single_DA only changes according to the power envelope of the waveform, but Seg_Single_DA and Seg_Multi_DA focus on a more particular segment, such as the section from 1.4 s to 1.6 s. This result suggests that feature segmentation successfully captures information other than the change of the amplitude. In addition, the Seg_Multi_DA method trained different attention weights for individual speech features. In particular, the Seg_Multi_DA method focuses on the section from 0.7 s to 1.3 s for which the attention weight of Seg_Single_DA is small. From this analysis, it is confirmed that the attention weights considering individual speech features contribute to improving the recognition accuracy.

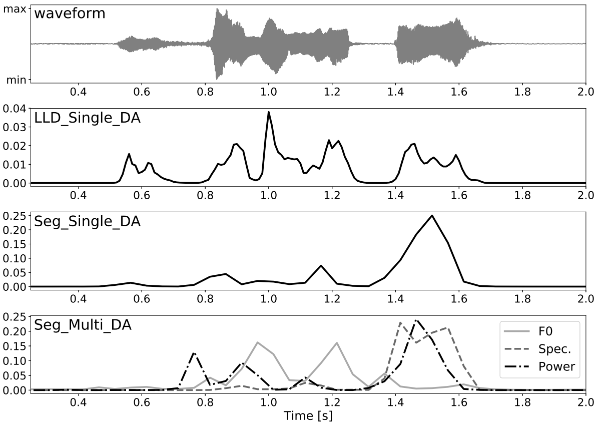

Figure 2: *Attention weights obtained from the examined models. The figure shows the attention weights of LLD_Single_DA, Seg_Single_DA, and Seg_Multi_DA in order from the top.*

6. Conclusion

This study proposed a speech emotion recognition technique based on the multi-stream attention-based BLSTM using segmental features. In this method, the LLDs of the local region are aggregated and input to the network. This method can consider the suprasegmental characteristics and the temporal change of the speech features. In addition, we modified the network topology to handle the LLD sequence by multiple streams. The proposed multi-stream network can obtain individual attention parameters for different descriptors. The experiments showed that feature segmentation improves the recognition accuracy by 1.1 points, and the multi-stream network improves it by a further 1.6 points. Finally, the recognition accuracy reached 73.4%, which is equivalent to human evaluation (75.5%).

In future studies, we will investigate the optimum segment length and examine a method that uses linguistic information. We will also compare the proposed method with the other speech emotion recognition techniques, such as CNN-based networks or x-vector-based approaches [29].

7. Acknowledgments

Funding was provided by Grants-in-Aid for Scientific Research (Grant No. JP16K13253, JP17H00823, JP18K18136).

8. References

[1] M. 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.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, “Emotion recognition in human-computer interaction,” *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[3] M. Swain, A. Routray, and P. Kabisatpathy, “Databases, features and classifiers for speech emotion recognition: a review,” *International Journal of Speech Technology*, vol. 21, no. 1, pp. 93–120, 2018.

[4] P. Gupta and N. Rajput, “Two-stream emotion recognition for call center monitoring,” in *Proc. INTERSPEECH*, 2007, pp. 2241–2244.

[5] J. Adelhardt, R. Shi, C. Frank, V. Zeißler, A. Batliner, E. Nöth, and H. Niemann, “Multimodal user state recognition in a modern dialogue system,” in *Proc. Annual Conference on Artificial Intelligence*, 2003, pp. 591–605.

[6] H. Ai, D. J. Litman, K. Forbes-Riley, M. Rotaru, J. Tetreault, and A. Purandare, “Using system and user performance features to improve emotion detection in spoken tutoring dialogs,” in *Proc. INTERSPEECH*, 2006, pp. 797–800.

[7] B. Schuller, S. Steidl, and A. Batliner, “The INTERSPEECH 2009 emotion challenge,” in *Proc. INTERSPEECH*, 2009, pp. 312–315.

[8] B. W. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, “The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language,” in *Proc. INTERSPEECH*, 2016, pp. 2001–2005.

[9] 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.

[10] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, “Attention based fully convolutional network for speech emotion recognition,” in *APSIPA-ASC*. IEEE, 2018, pp. 1771–1775.

[11] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, “Emotion recognition using a hierarchical binary decision tree approach,” *Speech Communication*, pp. 1162–1171, 2011.

[12] A. Ando, R. Masumura, H. Kamiyama, S. Kobashikawa, and Y. Aono, “Speech emotion recognition based on multi-label emotion existence model,” in *Proc. INTERSPEECH*, 2019, pp. 2818–2822.

[13] B. Schuller, G. Rigoll, and M. Lang, “Hidden Markov model-based speech emotion recognition,” in *Proc. International Conference on Multimedia and Expo*, no. II, 2003, pp. 1–4.

[14] S. Mirsamadi, E. Barsoum, and C. Zhang, “Automatic speech emotion recognition using recurrent neural networks with local attention,” in *Proc. ICASSP*, 2017, pp. 2227–2231.

[15] A. Satt, S. Rozenberg, and R. Hoory, “Efficient emotion recognition from speech using deep learning on spectrograms,” in *Proc. INTERSPEECH*, 2017, pp. 1089–1093.

[16] S. Mao, P. Ching, and T. Lee, “Deep learning of segment-level feature representation with multiple instance learning for utterance-level speech emotion recognition,” in *Proc. INTERSPEECH*, 2019, pp. 1686–1690.

[17] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in *Proc. ICASSP*, 2013, pp. 7962–7966.

[18] E. Takeishi, T. Nose, Y. Chiba, and A. Ito, “Construction and analysis of phonetically and prosodically balanced emotional speech database,” in *Proc. O-COCOSDA*, 2016, pp. 16–21.

[19] 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,” *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[20] S.-w. Lee, “The generalization effect for multilingual speech emotion recognition across heterogeneous languages,” in *Proc. ICASSP*, 2019, pp. 5881–5885.

[21] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, “ATR Japanese speech database as a tool of speech recognition and synthesis,” *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.

[22] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, “JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research,” *Journal of the Acoustical Society of Japan*, vol. 20, no. 3, pp. 199–206, 1999.

[23] M. Yamanaka, T. Nose, Y. Chiba, and A. Ito, “Labeling and analysis of perceived emotion for japanese large-scale emotional speech database JTES,” in *Proc. RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing*, 2020, pp. 230–233.

[24] Y. Kim and E. M. Provost, “Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions,” in *Proc. ICASSP*, 2013, pp. 3677–3681.

[25] E. Mower and S. Narayanan, “A hierarchical static-dynamic framework for emotion classification,” in *Proc. ICASSP*, 2011, pp. 2372–2375.

[26] H. Schlosberg, “Three dimensions of emotion,” *Psychological Review*, vol. 61, no. 2, pp. 81–88, 1954.

[27] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, “Emotion recognition from variable-length speech segments using deep learning on spectrograms,” in *Proc. INTERSPEECH*, 2018, pp. 3683–3687.

[28] ——, “Speech emotion recognition with emotion-pair based framework considering emotion distribution information in dimensional emotion space,” in *Proc. INTERSPEECH*, 2017, pp. 1238–1242.

[29] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, “X-vectors meet emotions: A study on dependencies between emotion and speaker recognition,” in *Proc. ICASSP*, 2020, pp. 7169–7173.