Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model

Bagus Tris Atmaja
Japan Adv. Inst. of Science & Tech. and
Institut Teknologi Sepuluh Nopember
Nomi, Japan and Surabaya, Indonesia
bagus@ep.its.ac.id

Masato Akagi
School of Information Science
Japan Adv. Inst. of Science & Tech. (JAIST)
Nomi, Japan
akagi@jaist.ac.jp

***Abstract*—Automatic speech emotion recognition has become popular as it enables natural interaction between human-machine interaction. One modality of recognizing emotion is speech. However, the speech also contains silence that may not relevant to emotion. Two ways to improve performance is by removing silence and/or paying more attention to speech segment while ignoring the silence. In this paper, we propose both, a combination of silence removal and attention model to improve speech emotion recognition performance. The results show that utilizing combination silence removal and attention model outperforms the use of either noise removal only or attention model only.**

***Index Terms*—voice segments, silence removal, speech emotion recognition, attention model**

I. Introduction

Recognition of human emotion by machine gives advantages in such areas like virtual reality, games, robotics, and call center application. In call center application, knowing the caller's emotion will help call center operator to take further response suitable for the caller [1]. In virtual reality and games, it can be used to recognize a player's emotional distress. Recognizing human emotion by a robot can improve natural interaction between them. Another example is the analysis of emotion on voice mail messages. In these cases, speech modality plays an important role to attain human emotion.

To obtain emotion from speech, a set of acoustic features can be extracted from speech. Then, a classifier can be used to map those features to a given emotion category. For dimensional emotion, regression method is used instead of classification. Both classification and regression currently can be performed simply using deep neural networks. This paper presents our works on categorical emotion recognition from speech modality using speech parts of an utterance, segments of utterance that contain speech parts only by removing silence.

Speech is unique, it is generated by communicative intentions of the speaker to whom they talk to [2]. The speakers cannot hide their emotion from their voice except they pretend to do it. Therefore, the emotion of the speaker theoretically can be recognized. The interlocutor, one who takes part in dialog, can perceive the speaker emotion from the speech. To make the computer do this is the task of speech emotion recognition (SER), a computational method to recognize human emotion from the speech sound. Voice segments rich of information including emotion cue presented by the speaker. However, speech not only consists of speech part of the speakers but also silence and noise, as the speakers pause between words and the juncture between syllables of the words. This makes speech contains voice and non-speech for every utterance. Accordingly, SER, in some cases, will be difficult to train speech features for emotion recognition. Some result shows low performance on speech emotion recognition using whole speech ([3], [2]).

To deal with unnecessary information in non-speech voice segment, two approaches are proposed. The first one is to use a voice activity detector (VAD) to remove the silence part of the speech. Mirsamadi et al. [5] reported the use of a segment-based approach to recognize emotion in speech. Nonetheless, they also showed the utterance level (whole speech) achieved higher performance than a segment-based (speech) approach. To improve accuracy, they utilize both segment-based and utterance-based approaches and obtain improvement by using those combinations. The second approach to deal with silence voice is to ignore it instead of removing it. The attention model in deep learning is one that tries to ignore silence frames and other parts of the utterance which do not carry emotional content [6]. While the first approach is done in the feature extraction or before it (preprocessing), the second approach is performed on the classification stage using a deep neural network.

We propose to use those two approaches in this paper to deal with silence part in speech. First, we remove silence inside speech sound, we do feature extraction from those speech segments. The obtained features than are trained with neural networks to map to given emotion label. Attention model is added in those networks to find the important output only on the previous layer. Finally, the evaluation is given by comparing performance from whole speech and speech only segment, with and without attention model.

II. Dataset

The IEMOCAP (interactive emotional dyadic motion capture) database developed by the University of Southern California was used in this research [7]. Although the researchers who made the dataset provide multimodal emotion measurement from facial expression, speech, head, and hand motion,

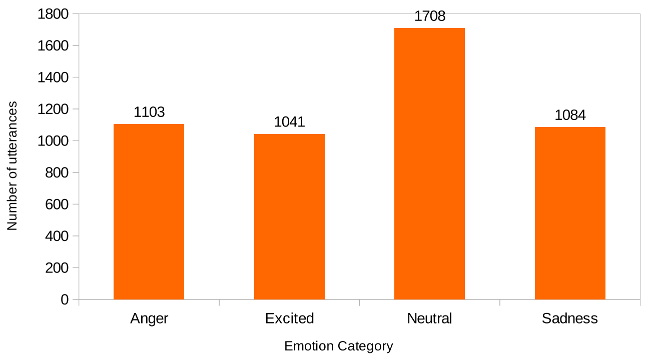

Fig. 1. Distribution of used categorical emotion in dataset

we only used speech modal for emotion recognition. The language used in that database is American English. This database consists of 10,039 turns from nine categories: neutral state; happiness; sadness; anger; surprise; fear; disgust; frustration; and excitement. From those, to balance the dataset for each category, we only used 4,936 utterances from four emotions: neutral, sad, angry and excitement. The distribution of the used dataset is shown in Fig. 1.

We preserve the original sampling rate of the emotional speech dataset (16000 Hz) when processing audio files. While the original file has two channels, we only use one channel as they have the same waveform.

III. Proposed Method

The typical main blocks of pattern recognition workflow including speech emotion recognition task consist of two steps: feature extraction and classification. The contribution of the proposed method is the addition of silence removal before feature extraction and the use of the attention model with bidirectional long short-term memory network (LSTM) for classification. The proposed system is shown in Figure 2.

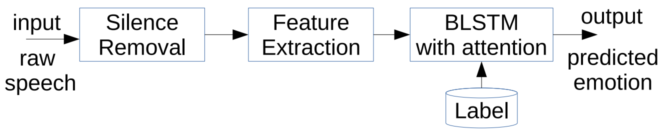

Fig. 2. Proposed speech emotion recognition system with silence removal and bidirectional LSTM classifier

A. Silence removal

The first step when working with audio data is to read the audio file as vector or matrix. After each sound file on the dataset is read as a vector, we do silence removal on each file based on the threshold and the minimum number of samples. Those two parameters are evaluated over each speech utterance in speech dataset. The output of this silence removal method is filtered speech. The complete algorithm for silence removal is presented in Algorithm 1.

Algorithm 1 Algorithm for silence removal

```
Require: speech_dataset
Ensure: filtered_speech
minimum_threshod = threshold
minimum_samples = n_i_min
n_i = 0
for speech in speech_dataset do
    for i in speech do
        if abs(amplitude[i]) < threshold then
            n_i = n_i + 1
        end if
        if n_i = n_i_min then
            remove n_i samples
        end if
    end for
end for
```

From the reference [10], it is known that the pause duration in speech can be divided into three: brief, medium and long. For a brief or short pause, the duration is less than 200 milliseconds. Here, also based on that paper, we use the minimum duration of 60 milliseconds beside 10 and 100 milliseconds (ms). For the threshold of amplitude, we varied a small number of 0.01, 0.07, and 0.1% due to the big fluctuation of speech inside the dataset. While we do not perform normalization to capture the dynamic of speech sound, the normalization of each speech frame is performed within the feature extraction step.

B. Feature Extraction

We follow feature extraction steps in [4]. First, each speech utterance is split into window frames and moved with overlapped each other. The feature extraction steps conducted over each frame within each utterance. A total number of features extracted for each utterance is 34 features consisting of the values of the following variables:

1) 3 time-domain features (zero crossing rate, energy, entropy of energy)
2) 5 spectral-domain features (spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll off)
3) 13 MFCCs
4) 13 Chroma (12-dimensional chroma vector + standard deviation of chroma vector)

To reduce computation complexity, we reduce the length of data for each utterance to be 100 numbers of window frames. Hence, the size of the feature vector for each utterance is (100, 34). These features are fed to bidirectional LSTM networks.

C. Bidirectional LSTM

The idea of using LSTM network comes from an approach that human has the persistence to keep memory long in short-term period [9]. Humans do not start their thinking from scratch every second. As we read this paper, we understand each word based on our understanding of previous words. We do not throw everything away and start thinking from scratch again. Our thoughts have persistence [8].

Using LSTM only preserves information in the previous time, although it holds information within the layer. It means, we can reserve information from the past, but not from the future. Using bidirectional address this issue. Both information, from the past to the future, and from the future to the past, can be accommodated to improve network performance. In most case ([3], [6]), the bidirectional LSTMs outperforms unidirectional LSTMs. Using bidirectional LSTM as described in [11], we implement speech emotion recognition system using feature extraction from speech segment. The speech segments then processed to generate features which are trained with bidirectional LTSMs.

The LSTM networks used in this research consists of two layers, the first layer contains 256 units and the second layer contains 512 units. For the batch size, we use the size of 32 with default 0.001 learning rate. Although the use of dropout can increase the performance by preventing overfitting in some cases, here we do not use the dropout parameter as it does not change the result even make a worse result. We limit our experiments to 50 iterations as we found that the result is stable

D. Attention Model

As additional processing to the Bidirectional LSTMs, we incorporate attention model. As explained previously that BLSTMs can pass information from the past to the future, and to the future to the past, but which information is the most relevant remains unknown. In this case, the attention model is proposed to address this issue. Attention model can work to choose relevant information (e.g. ignore noises). For the speech segments, the attention model will choose relevant information from the previous layer to the next layer. For example, given speech segments, which parts of those speeches' segments contribute more to emotion recognition.

The attention model consists of three parts: encoder, attention manipulation, and decoder. On the encoder side, A BLSTM network is used containing the input feature $x = (x_1, x_2, \dots, x_T)$ and output an encoded sequence of $h = (h_1, h_2, \dots, h_T)$, where T is a number of input feature for each sequence. Before h is passed to the decoder, some manipulation can be done. In this case, we choose the last encoding only as it reflects the summary of the whole state. On the decoder, for any input feature at position t, output decoder accepts encoded sequence of $h = (h_1, h_2, \dots, h_T)$ also from the previous state s_{t-1} which is shared within the decoder cell and emotion label y_{t-1} which is one of four emotions. The final output now is also one of four emotion $y = (y_1, y_2, y_3, y_4)$ in binary form, 1 for predicted emotion and 0 for the rest emotions.

The attention mechanism created context vector. First, attention probabilities $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_T)$ is calculated based on the encoded sequence (Eq. (1)) of the internal hidden state of the decoder cell, s_{t-1}. This probability calculation shown in Eq. (2) is a softmax function. Then the context vector c_t (Eq. (3)) is calculated as weighted sum of the encoded sequence with attention probabilities.

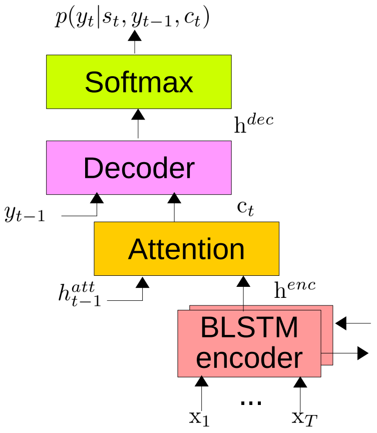

Fig. 3. Schematic diagram of attention model from Bidirectional LSTM encoder

$$e_{(j,t)} = V_a \tanh(W_a s_{t-1} + U_a h_j) \tag{1}$$

$$\alpha_{(j,t)} = \frac{\exp(e_j)}{\sum_{k=1}^{T} \exp(e_k)} \tag{2}$$

$$c_t = \sum_{k=1}^{T} \alpha_{(k,t)h_k} \tag{3}$$

The detail equation of attention mechanism can be found in [12]. The schematic diagram of this attention model can be shown in Fig 3.

IV. Result and Discussion

A. Accuracy of proposed method

To compare our proposed method, we make some variation of the system for the baselines. First, we choose the whole speech fed to LSTM networks which consist of two layers with 512 and 256 units. The second architecture is the same as the first with speech segmented preprocessing as explained in silence removal subsection in section III. The third architecture is similar system consisting of Bidirectional LSTM with 256 units and attention model at the second layer (the first layer also acts as an encoder). This third architecture is fed with whole speech while the last fourth architecture is fed with speech segments. The performance result of each method is shown in Table I. It is clearly shown that our proposed method outperforms the rest of methods significantly both in average accuracy and maximum accuracy. The average accuracy is the mean of accuracy from all epochs within one experiment, while the maximum accuracy is the highest accuracy achieved from one running experiments, which a model can be saved based on this accuracy. Both scores are obtained from running a number of five experiments.

For all four methods, we use the same parameter value. A number of 34 features are used for the input which is calculated on 0.2 s of window size and moved within 0.1

TABLE I
PERFORMANCE COMPARISON IN TERM OF ACCURACY (ACC) FROM IEMOCAP DATASET

Method	mean acc	max acc
Whole speech + BLSTM	55.0	57.70
Speech segment + BLSTM	56.46	58.87
Whole speech + BLSTM + Attention	63.52	65.80
Speech segment + BLSTM + Attention	66.54	70.34

TABLE II
PERFORMANCE COMPARISON IN TERM OF ACCURACY (ACC) BY DIFFERENT MINIMUM DURATION OF SILENCE REMOVAL

Minimum duration (ms)	mean acc	max acc
10	65.54	70.34
60	62.08	64.52
100	60.48	62.92

second step. The sampling frequency of the speech signal is 16000 Hz. For each utterance, we only use 100 frames of feature resulting input size (100, 34) for LSTM/BLSTM networks. For each method, a dense layer with 512 units is added after the second LSTM/BLSTM layer with ReLU activation function. Finally, an output layer is added with 4 units of a dense layer with softmax activation. This reflects one of the predicted emotion categories. All network used RMSProp [14] as optimization function with categorical entropy as loss function. The accuracy metrics show the percentage of true predicted emotion over all validation data.

For training data, we also use the same parameter value including the same batch size=25, the number of epochs=50 and validation split=0.33 (equal to 1629 utterances). Finally, we reach the best performance with silence removal after some configuration. With a minimum duration of silence is 0.01 s and threshold of 0.001 %, the performance beats other configurations. The following subsection discusses the used value for those variables.

B. Effect of minimum duration

Although the paper [10] classify that brief pause is less than 200 ms, it is still unclear how long the minimum duration of silence/pause in utterances. The paper found that in a spontaneous speech the distribution of brief pause has peaks at 78 ms and 426 ms, while read speech (acted) has the first peak in distribution at 100-150 ms and a second peak at 500-600 ms. The IEMOCAP dataset used here consists of spontaneous and acted speech, therefore both values can be used. To find the best minimum duration of silence for each utterance, the lower value should be used i.e the minimum pause duration from spontaneous speech. In this case, we use 60 ms to represent spontaneous speech and 100 ms to represent acted speech. As predicted, the 60 ms of minimum duration gives a better result than 100 ms of minimum pause duration. After some experimentation, we found that the best minimum duration is 10 ms. Table II shows the comparison of accuracy among three different minimum duration.

TABLE III
PERFORMANCE COMPARISON IN TERM OF ACCURACY (ACC) BY DIFFERENT % OF MINIMUM THRESHOLD FOR SILENCE REMOVAL.

Threshold (%)	mean acc	max acc
0.01	65.54	70.34
0.07	56.47	58.69
0.1	45.58	50.03

C. Effect of threshold

The second parameter for silence removal beside minimum duration is the threshold. It represents the minimum amplitude of a sample point to be removed. If the amplitude of the sample at sample t below the threshold and the number of those samples exceeding the minimum duration, then those samples will be categorized as silence and will be removed. For this parameter, we use a small value due to the high fluctuation of the amplitude of each utterance in the dataset. Three values are varied, 0.01%, 0.07%, and 0.1%. For the last value of 0.1%, no part in some utterances (wav file) is under the threshold. In this case, we use the whole speech although those data consist of noises only or noise dominant data. This noisy data might be the source of the lowest performance obtained by the system using that parameter.

For this threshold parameter, after some experimentation, we found the best value of the threshold is 0.01%. The comparison of each threshold is given in term of mean accuracy and maximum accuracy for consistency. Please note, in that data, the speech signal is not normalized to capture the dynamics of a signal resulting big gap of amplitude. Hence, the used threshold is very small.

V. CONCLUSION

An architecture of speech emotion recognition based on the speech segment using BLSTM network and attention model is presented in this paper. For comparison, we choose some baselines with whole speech and varying it with LSTM/BLSTM and attention model. The result shows that our proposed method outperforms other methods in term of mean and maximum accuracy. We also found, after some experimentation, that the best performance is achieved by using a minimum duration of 10 ms and a threshold of 0.01% for silence removal. To check the consistency of the result, the proposed method can be applied to another dataset or other languages for future research. For reproducing this research result, a repository is made[1] where all codes to obtain the result are provided.

REFERENCES

[1] Petrushin, Valery. ”Emotion in speech: Recognition and application to call centers.” In Proceedings of artificial neural networks in engineering, vol. 710, p. 22. 1999.

[2] O’Callaghan, Casey. ”Is speech special?.” UBCWPL (2009): 57.

[3] Tripathi, Samarth, and Homayoon Beigi. ”Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning.” arXiv preprint arXiv:1804.05788 (2018).

[1] https://github.com/bagustris/SER_ICSigSys2019

[4] Chernykh, Vladimir, Grigoriy Sterling, and Pavel Prihodko. ”Emotion recognition from speech with recurrent neural networks.” arXiv preprint arXiv:1701.08071 (2017).

[5] Shami, Mohammad T., and Mohamed S. Kamel. ”Segment-based approach to the recognition of emotions in speech.” In 2005 IEEE International Conference on Multimedia and Expo, pp. 4-pp. IEEE, 2005.

[6] Mirsamadi, Seyedmahdad, Emad Barsoum, and Cha Zhang. ”Automatic speech emotion recognition using recurrent neural networks with local attention.” In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2227-2231. IEEE, 2017.

[7] C. Busso et al., IEMOCAP: Interactive emotional dyadic motion capture database, Lang. Resour. Eval., vol. 42, no. 4, pp. 335359, 2008.

[8] Christopher Olah, Understanding LSTM Networks — colahs blog, 2015.

[9] Sepp Hochreiter and J J Urgen Schmidhuber, LONG SHORT-TERM MEMORY, MEMORY Neural Computation, vol. 9, no. 8, pp. 1735 — 1780, 1997

[10] Campione, Estelle, and Jean Vronis. ”A large-scale multilingual study of silent pause duration.” In Speech prosody 2002, international conference. 2002.

[11] Graves, Alex, and Jrgen Schmidhuber. ”Framewise phoneme classification with bidirectional LSTM and other neural network architectures.” Neural Networks 18, no. 5-6 (2005): 602-610.

[12] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. ”Neural machine translation by jointly learning to align and translate.” arXiv preprint arXiv:1409.0473 (2014)

[13] Sarma, Mousmita, Pegah Ghahremani, Daniel Povey, Nagendra Kumar Goel, Kandarpa Kumar Sarma, and Najim Dehak. ”Emotion Identification from raw speech signals using DNNs.” Proc. Interspeech 2018 (2018): 3097-3101.

[14] Hinton. ”Neural networks for machine learning.” Coursera, video lectures, 2012. S.