High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition

Jinkyu Lee[1] *and Ivan Tashev*[2]

[1]Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
[2]Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA

shuya@dsp.yonsei.ac.kr, ivantash@microsoft.com

Abstract

This paper presents a speech emotion recognition system using a recurrent neural network (RNN) model trained by an efficient learning algorithm. The proposed system takes into account the long-range context effect and the uncertainty of emotional label expressions. To extract high-level representation of emotional states with regard to its temporal dynamics, a powerful learning method with a bidirectional long short-term memory (BLSTM) model is adopted. To overcome the uncertainty of emotional labels, such that all frames in the same utterance are mapped into the same emotional label, it is assumed that the label of each frame is regarded as a sequence of random variables. Then, the sequences are trained by the proposed learning algorithm. The weighted accuracy of the proposed emotion recognition system is improved up to 12% compared to the DNN-ELM based emotion recognition system used as a baseline.

Index Terms: Speech emotion recognition, recurrent neural network, deep neural network, long short-term memory

1. Introduction

In speech enabled Human-Machine Interfaces (HMI) the context plays important role for improving the user interface. One of the important components of the context is the emotion in speaker's voice. Emotion recognitions provide important priors, making it possible to add human-like features to the HMI, such as empathy and responding with proper emotion in the text to speech engine.

Emotion detection systems can form the decision on frame level (or short segment) or on utterance level. In the first approach, low-level features are directly used to generate the distribution of each emotional state using Gaussian mixture model (GMM) [1] or hidden Markov model (HMM) [2]. On the other hand, the second approach applies statistical functions into the low-level features to obtain the global characteristic of each utterance, then these global features are used for discriminative classifiers such as support vector machine (SVM) [3]. Similar to other recognition systems, it is very important to choose efficient low-level features. However, it is hard to find the features which represent each emotional state well. Recently, deep learning techniques have been applied to obtain high-level representations from low-level acoustic features, and those features combined with SVM or extreme learning machine (ELM) show state-of-the-art performance [4].

This work was performed while the first author was worked as an intern at Microsoft Research, Redmond.

In this paper, we consider more effective high-level features which are more robust to the long-range contextual effect adopting a recurrent neural network, a powerful learning model for sequential data. Furthermore, we propose a new learning algorithm for speech emotion recognition with the uncertainty of emotional labels. In conventional algorithms, all frames in the utterance are mapped to the same label. However, since the labels are annotated for utterances, not for frames, it does not mean all the frames in the same utterance should be mapped to the same label. Therefore, it is reasonable to assume the emotional state as a random variable, and we propose the corresponding learning algorithm which internally decides the importance of each frame using the EM algorithm with the efficient dynamic programming.

The rest of paper is organized as follows. In Section 2, an overview of the conventional emotion recognition algorithm using DNN and ELM. In Section 3, the proposed structure and training algorithm is described in details. The performance evaluation results are given in Section 4, and conclusion follows in Section 5.

2. Related Work: DNN-ELM based Speech Emotion Recognition

One of the main issues on developing a speech emotion recognition system is to find an efficient feature set that well represents emotional state. Typically, the features are extracted based on acoustic characteristics, such as pitch-related features, intensity, and spectral information. Although earlier studies tried to find relation between the acoustic features and each emotion class [5], still it is hard to find efficient features.

Recently, deep learning techniques have been applied to obtain high-level representation for speech and emotion recognition [4, 6, 7, 8]. Among them, a DNN-ELM based emotion recognition system shows a state-of-the-art performance [4]. The DNN-ELM based system consists of two main functional modules: high-level feature representation and utterance-level classification. Note that conventional utterance-level emotion recognition systems use the statistical characteristics of the acoustic features only to model the characteristics of each emotion. This DNN-ELM system uses for classification only high-energy frames from each utterance, and only those frames are used to extract high-level features. The next step is to find high-level representation from the selected low-level features. In the system, the softmax output of the network is regarded as a probability of each emotion over time. Since the network tries to minimize the cross-entropy between target labels and network outputs, its outputs can be good measures for representing emotional states.

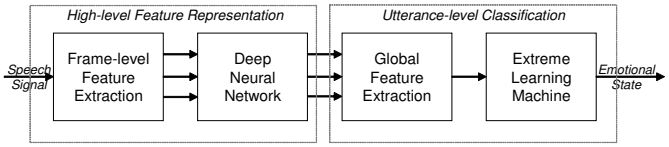

Figure 1: Block diagram of the conventional speech emotion recognition system based on DNN and ELM.

Since one emotional state is mapped into one utterance in many applications, classification is usually performed on utterance-level. It is also reported that the utterance-level systems give better accuracy as well as lower computational cost. From the high-level features derived by DNN, the utterance-level global characteristics are extracted by applying a various statistical functions such as mean, maximum, minimum, median, quantile, etc. Then the characteristics of each emotion are trained by the discriminative classifier such as K-nearest neighbor (KNN), or support vector machine (SVM) [3]. Likewise, extreme learning machine (ELM) [9], a type of single-hidden-layer neural network, is introduced as the utterance-level classifier. Due to its simple structure and closed-form solution, the training mechanism is very efficient. The input data are projected into high-dimensional space by random projection or kernel function, then the weight matrix in the hidden layer is trained by pseudo inverse operation as follows:

$$\mathbf{W}_k = \left(\mathbf{H}\mathbf{H}^T\right)^{-1}\mathbf{H}\mathbf{T}^T, \tag{1}$$

where $\mathbf{H}$ and $\mathbf{T}$ is projected data and target label matrix respectively.

3. Proposed RNN-based Emotion Recognition Framework

In the DNN-based system, the estimation of the probability for the current frame uses a few past and future frames, which are not sufficient to cover the long time contextual effect in emotional speech. This problem can be addressed by adopting a recurrent neural network (RNN) model. In addition, since the segment selection algorithm uses only frames with highest energy, it may not fully utilize the information for measuring emotional status. The frame selection problem can be also addressed by the proposed RNN learning algorithm which automatically estimates the importance of each frame.

3.1. Recurrent Neural Network: Deep Bidirectional LSTM

The feed-forward DNN, a learning model for non-sequential data, can be extended for sequential data using so called context features which stack few neighborhood frames into the current frame. It is assumed that the contextual effect can be covered by the long window whose length is typically 150-250 ms. However, this hypothesis does not hold well for emotional speech because its long and variable context effect is not sufficiently modeled well by the fixed-range of long window.

A recurrent neural network (RNN) is a powerful tool for modeling sequential data. The only difference between a standard RNN and DNN is that RNN has recurrent connections

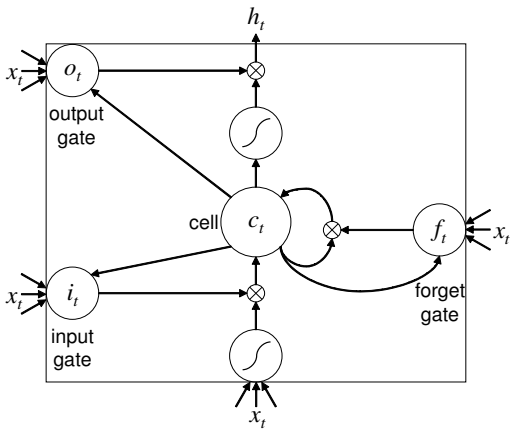

Figure 2: A graphical representation of Long Short Term Memory

$\mathbf{W}_{hh}$ represented in the following equation:

$$\mathbf{h}_t = f\left(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1}\right), \tag{2}$$

$$\mathbf{y}_t = \mathrm{softmax}\left(\mathbf{W}_{hy}\mathbf{h}_t\right), \tag{3}$$

where $\mathbf{W}_{ij}$ indicates weight matrix from layer i to j; $\mathbf{x}_t$, $\mathbf{h}_t$, and $\mathbf{y}_t$ indicate the input, hidden, and output layers at time t respectively. The bias terms are not represented in the above equations for simplicity. Although those recurrent connections have a capability of memorizing previous information, RNN still has a limitation to cover the long context information like emotion because of the gradient vanishing problem. To overcome this problem, a long short-term memory (LSTM) network was proposed [10], which consists of recurrently connected memory blocks as shown in Figure 2. In this work, we use bidirectional long short-term memory (BLSTM) network, which keeps well the temporal dynamic characteristics non-causally.

3.2. Proposed Learning Framework

In conventional neural network frameworks for classification, the target class labels are assumed to be deterministic values in the training phase. Therefore, minimizing mean square error, or cross-entropy, is a typical learning criterion for the network. However, in emotion corpus cases, the label is annotated for the entire utterance, not for every frame. This is why it is reasonable to assume that there is uncertainty in frame-level emotional labels. For example, in the utterance labeled as *Happy*, not all frames contain emotional information related to *Happy*.

To represent the uncertainty of emotional labels, in this paper we adopt an additional class for non-emotional frames - *Null*. Then, we represent the emotional label as a random variable between two states, one is the given emotion class and the other one is the additional class *Null*. Based on this assumption, we design a new training criterion for RNN to maximize the sum of log-probabilities of all possible sequences over the training data. Basically, there are 2^T possible sequences, where T is the number of frames in the given utterance. Among them, some sequences can be reasonable, but majority of sequences are not meaningful. For example, it is obvious that silence regions do not contain any emotional information.

Thus, it is better to reduce the number of the possible sequences using a prior knowledge. First we divide each utterance into small segments with voiced region, then we assume

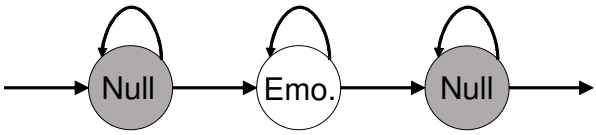

Figure 3: Markov chain for each speech segment

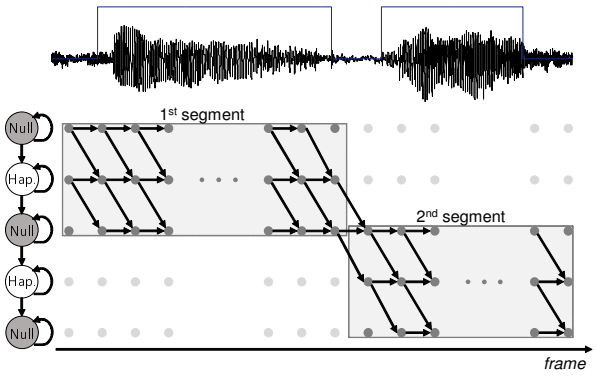

Figure 4: All possible emotional states when the label is annotated as *Happy*.

that the label sequences of each segment follows the Markov chain shown in Figure 3. It means that the sequence from each segment starts from the *Null* state and goes through the relevant emotional state and finally goes back to the *Null* state. Then, we concatenated the label sequences of each segment to generate the sequences for the entire utterance. To be applicable for continuous emotion recognition, the last state of the current segment is merged with the first state of the next segment. Figure 4 shows the reduced possible paths with the assumed prior knowledge where the label of utterance is annotated as *Happy*, and there are two voiced regions in the utterance.

As a result, a learning criterion that maximizes the log-probability of all possible cases in the reduced set can be written as follows:

$$L_{new} = -\sum_{\mathbf{s}\in\mathbf{S}} \ln p\left(\mathbf{s}|\mathbf{x}\right), \tag{4}$$

where $\mathbf{x}$, $\mathbf{s}$ and $\mathbf{S}$ indicate given input features, label sequence and its possible set, respectively.

During the back-propagation through time (BPTT) process [11], the derivative of (4) for each moment of time is needed and can be calculated efficiently by using a dynamic programming, forward-backward algorithm [12][13]. Since we consider the target labels as random variables, the proposed training method is performed in the EM process framework. In the expectation step, the sum of log probabilities of all possible cases is obtained from the current network, and in the maximization step, weight matrices are updated based on the importance of target emotion obtained in the expectation phase.

Figure 5 shows the process of changing the importance of the target emotion during the training stage. Because of the randomly initialized weights, the neural network cannot estimate it properly at the beginning. However, after several iterations, when the network becomes stable, it starts gradually to learn the characteristic of each emotion.

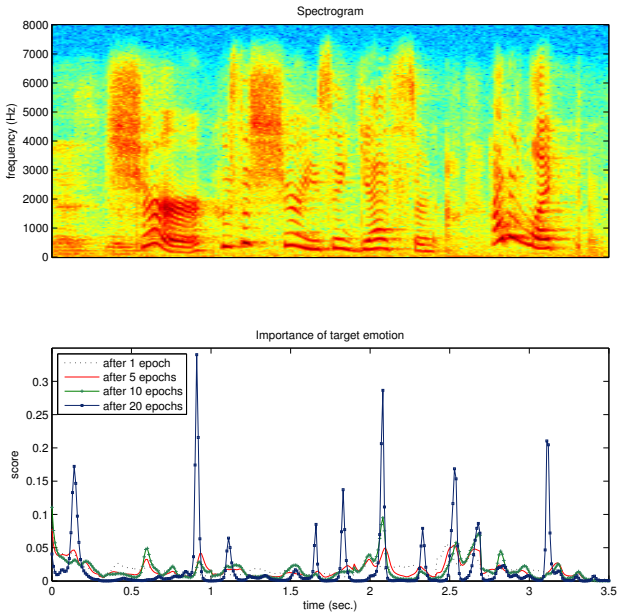

Figure 5: Spectrogram of an utterance labeled as *Happy* (Top), and the importance of each emotion over time from different network models (Bottom).

4. Experimental Evaluation

4.1. Experimental Setup

To evaluate the performance of the proposed framework, we used Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [14], which contains audio-visual data with transcription performed by ten different actors. In the corpus, we have five sessions and in each session, a pair of speakers (male-female) talk to each other. For training and evaluation, we used four categorical emotions *Angry*, *Happy*, *Sad* and *Neutral*, which represent the majority of the emotion categories in the database. For context-independent situation, we used only *improvised* data which are recorded in a pre-defined situation without given scripts.

For low-level acoustic features, we extract 32 features for every frame: F0 (pitch), voice probability, zero-crossing rate, 12-dimensional Mel-frequency cepstral coefficients (MFCC) with log energy, and their first time derivatives. In the DNN-based framework, we used as a baseline, those 32-dimensional vectors are expanded to 800-dimensional vectors using the context window with the size of 250 ms. The network contains 3 hidden layers and each hidden layer has 256 nodes, and the weights were trained by back-propagation algorithm using stochastic gradient descent with mini-batch of 128 samples. In the RNN-based system, the 32-dimensional vectors are directly used for input. The network contains 2 hidden layers with 128 BLSTM cells (64 forward nodes and 64 backward nodes). Later experiments showed that the performance did not improve with higher number of hidden layers and nodes in both DNN-based and RNN-based systems. The reason is most probably overfitting caused by data insufficiency.

To extract global characteristics, we apply statistical functions to the output of each network, then the utterance-level features are fed into an ELM network. In order to consider the non-linearity, we use a radial basis function as a kernel instead

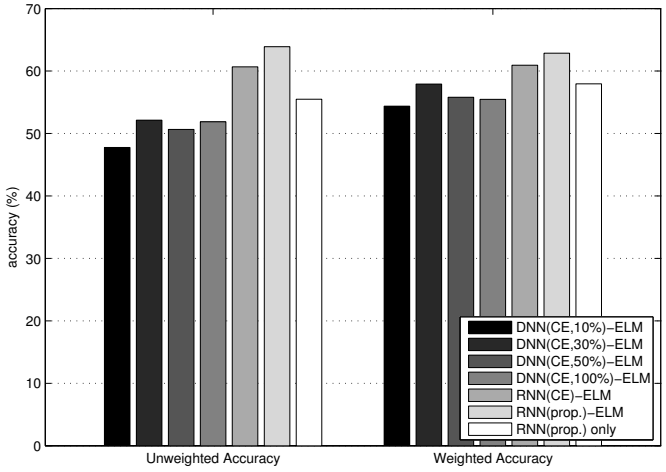

Figure 6: Comparison of different emotion recognition systems in terms of weighted and un-weighted accuracies

of random projection method. The kernel-based ELM gives not only better performance but also has the advantage that the number of hidden nodes does not need to be specified.

In order to measure the performance in the speaker-independent manner, we used 5-fold cross-validation technique among the five sessions. In each evaluation, four sessions were used for training the network, then the remaining session was divided into two sub-sessions depending on the gender. We used one for the parameter setting, and the other for measurement. For evaluation, we use the following two measures: weighted accuracy (WA) and un-weighted accuracy (UA). Weighted accuracy is the classification accuracy on the entire test data set, and un-weighted accuracy is an average of the classification accuracy for each emotion. Here all parameters, such as the number of epochs, the network structure, and kernel parameters, are tuned to maximize un-weighted accuracy, because of the imbalanced data set.

4.2. Experimental Results

Figure 6 shows the recognition accuracy of the evaluated systems. The first four indicate the performance of the performance of the DNN-based systems described in [4]. There are four results from different systems depending on how much data selected in both training and test phase. In the experiment, 30% data with highest energy shows the best result.

Following three bars show the results of proposed RNN-based systems. *RNN(CE)-ELM* means where the DNN network is substituted to BLSTM-RNN with cross-entropy (CE) training. *RNN(prop.)-ELM* indicates the system where the proposed learning algorithm is used. *RNN(prop.) only* means only high-level features obtained from RNN are used for classification without the utterance-level classifier. Since the temporal dynamic is explicitly considered in RNN-based system, *RNN(prop.) only* shows similar performance without the utterance-level classifier whose function is to consider temporal dynamics in the conventional system. The proposed system *RNN(prop.)-ELM* shows around 12% (from 52.13% to 63.89%) and 5% (from 57.91% to 62.85%) absolute improvements in UA and WA measures, respectively.

5. Conclusion and Future Work

In this paper, we proposed an RNN-based speech emotion recognition framework with efficient learning approach, which allows to account for long contextual effect in emotional speech and the uncertainty of emotional labels. The proposed approach provides insight on how recurrent neural network and maximum-likelihood based learning process can be combined into emotion recognition.

6. References

[1] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using GMMs." in *Interspeech*, 2006.

[2] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mariño, "Speech emotion recognition using hidden Markov models." in *INTERSPEECH*, 2001, pp. 2679–2682.

[3] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1057–1070, 2011.

[4] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. Interspeech*, 2014.

[5] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression." *Journal of personality and social psychology*, vol. 70, no. 3, p. 614, 1996.

[6] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1798–1828, 2013.

[7] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[8] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3687–3691.

[9] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[12] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[13] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[14] 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.