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 contextual 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) structure is adopted. To overcome the uncertainty of emotional labels, such that all frames in the same utterance are mapped to the same emotional label, it is assumed that the label of each frame is regarded as a sequence of random variables. The sequences are then 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, long short-term memory

1. Introduction

In speech enabled human-machine interfaces (HMI), context plays an important role in improving the user interface, and one of the critical components of context is the emotion in the speaker's voice. Emotion recognition provides important priors, making it possible to add human-like features to the HMI such as empathy and the capability of responding with the appropriate emotion to the text produced by the speech recognition engine.

Speech emotion detection systems can be realized at the frame-level (short segment) or at the utterance-level. In the frame approach, low-level features are used directly to generate the distribution of each emotional state by using the Gaussian mixture model (GMM) [1] or the hidden Markov model (HMM) [2]. The utterance approach, on the other hand, applies statistical functions to the low-level features to obtain the global characteristics of each utterance; these global features are then used for discriminative classifiers such as support vector machines (SVM) [3]. Similar to other recognition systems, it is very important to choose efficient low-level features. However, it is hard to identify features that represent each emotional state well. Recently, deep learning techniques have been applied to obtain high-level representations from low-level acoustic features, and these features, in combination with SVM or extreme learning machine (ELM), show state-of-the-art performance [4].

This work was conducted while the first author was an intern in Microsoft Research.

In this paper, we consider more effective high-level features which are robust in terms of long-range contextual effects by adopting a recurrent neural network (RNN), which is a powerful learning model for sequential data. Furthermore, we propose a new learning algorithm for speech emotion recognition, which addresses 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 the entire utterance, all frames in the utterance do not necessarily carry the same emotion and should not necessarily be mapped to the same label. Therefore, it is reasonable to assume the emotional state as a random variable, and we propose a corresponding learning algorithm that can internally decide the importance of each frame using the expectation-maximization (EM) algorithm in combination with efficient dynamic programming.

The rest of the paper is organized as follows. Section 2 provides an overview of the conventional emotion recognition algorithm using a deep neural network (DNN) and an ELM. In Section 3, the proposed structure and training algorithm are described in detail. The performance evaluation results are discussed in Section 4, and the conclusions follow in Section 5.

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

One of the main issues involved in developing a speech emotion recognition system is finding an efficient feature set that provides an accurate representation of the emotional state. Typically, the features are extracted based on acoustic characteristics such as pitch-related features, intensity, and spectral information. Although earlier studies have attempted to identify links between the acoustic features and each emotion class [5], it is still difficult to find efficient features.

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

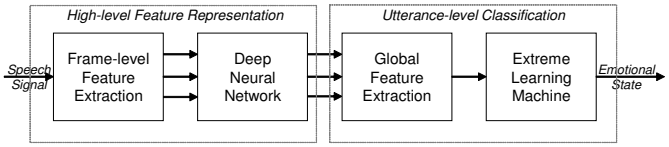

Figure 1: Block diagram of the conventional speech emotion recognition system based on DNN and ELM.

labels and network outputs, its outputs are good measures for representing emotional states.

Since one emotional state is mapped onto one utterance in many applications, classification is usually performed at the utterance-level. It is also reported that utterance-level systems involve less computational cost. From the high-level features derived by the DNN, the utterance-level global characteristics are extracted by applying various statistical functions such as mean, maximum, minimum, median, and quantiles. The characteristics of each emotion are then trained by a discriminative classifier such as K-nearest neighbor (KNN) or SVM [3]. Likewise, an ELM [9], which is 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 the ELM employs is very efficient. During the training process, the input data are projected into high-dimensional space by means of random projection or a kernel function, then the weight matrix in the hidden layer is calculated using a pseudo-inverse operation, as follows:

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

where $\mathbf{W}$, $\mathbf{H}$, and $\mathbf{T}$ represent the weight matrix in the hidden layer, the projected data, and the 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 few past and future frames that are not sufficient to cover the long-range contextual effect in emotional speech. This problem can be addressed by adopting an RNN structure. In addition, since the segment selection algorithm uses only the frames with the highest energy, it may not fully utilize the information for the emotional state. The frame selection problem can also be addressed using the proposed RNN learning algorithm, which automatically estimates the importance of each frame.

3.1. Recurrent Neural Network: Bidirectional Long Short-Term Memory

The feed-forward neural network, 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 a long window whose length is typically 150-250 ms. However, this hypothesis can not be applied effectively to emotional speech because its long and variable context effect is not sufficiently modeled by the fixed range of the window.

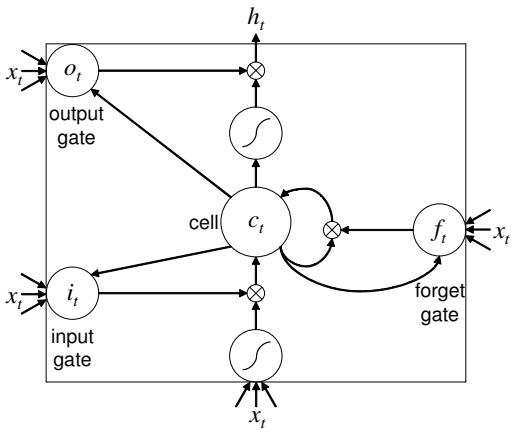

Figure 2: A graphical representation of long short-term memory.

An RNN is a powerful tool for modeling sequential data. The main difference between a standard feed-forward and an recurrent neural network is that the RNN has recurrent connections, which are represented as $\mathbf{W}_{hh}$ 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 the weight matrix from layers i to j and $\mathbf{x}_t$, $\mathbf{h}_t$, and $\mathbf{y}_t$ indicate the input, hidden, and output layers at time t, respectively. To ensure simplicity, the bias terms are not represented in the above equations. Although these recurrent connections are capable of memorizing previous information, the RNN still has a limitation with regard to covering long context information like emotions, because of the vanishing gradient problem.

To overcome this problem, a long short-term memory (LSTM) network was proposed. Different from a traditional RNN network, each block of the LSTM network contains self-connected memory cells with three multiplicative gates (input, output, and forget) as shown in Figure 2 above [10]. By these three gates, the network allows to learn the characteristic of data over long periods of time avoiding the vanishing gradient problem. In this work, we used a bidirectional long short-term memory (BLSTM) network, which effectively retains the temporal dynamic characteristics non-causally.

3.2. Proposed Learning Framework

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

To represent the uncertainty of emotional labels, in this paper, we adopted an additional class for non-emotional frames, namely, *Null*. We then represented the emotional label as a random variable between two states, one of which is the given emotion class and the other, the additional class *Null*. Based on this assumption, we designed a new training criterion for RNNs to

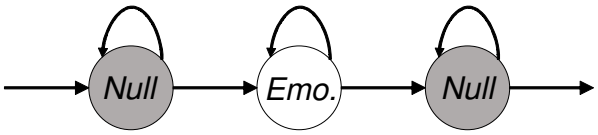

Figure 3: Markov chain for each speech segment.

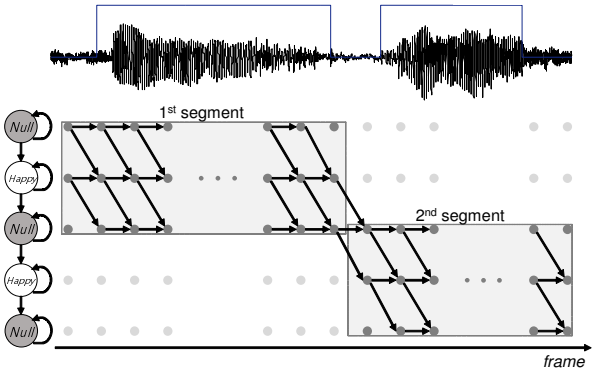

Figure 4: All possible emotional states when the label is annotated as *Happy*.

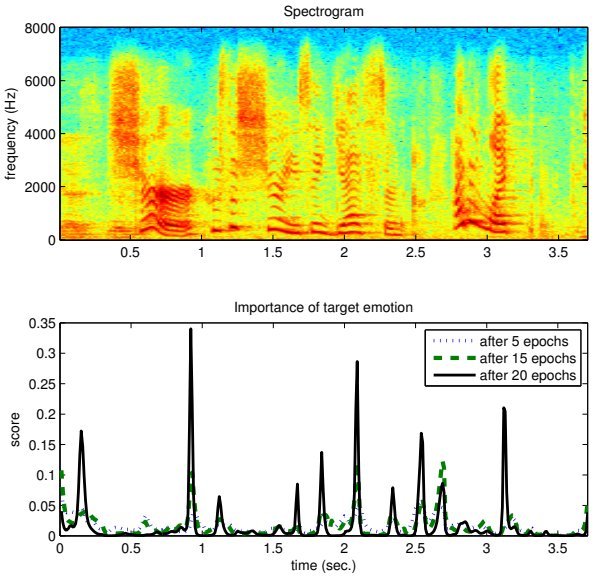

Figure 5: Spectrogram of an utterance labeled *Happy* (top) and the importance of each emotion over time from different network models (bottom).

maximize the sum of the log probabilities relating to 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. Of those sequences, some may be reasonable, but most are not. For example, it is obvious that silence regions do not contain any emotional information.

Thus, it is better to reduce the number of possible sequences using prior knowledge. First, we divided each utterance into small segments with voiced region. We then assumed that the label sequences for each segment followed the Markov chain shown in Figure 3. This means that the sequence from each segment starts from the *Null* state, goes through the relevant emotional state (*Emo.*) and finally goes back to the *Null* state. Subsequently, we concatenated the label sequences of each segment to generate the sequences for the entire utterance. To render it applicable for continuous emotion recognition, the last state of the current segment was 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 the 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 the given input features, the 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 considered the target labels to be random variables, the proposed training method was performed in the EM framework. In the expectation step, the sum of the log probabilities for all possible cases was obtained from the current network, and in the maximization step, weight matrices were updated based on the importance of the target emotion obtained during 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 network cannot estimate the emotion properly at the beginning. However, after several iterations leading by degrees to network stability, it gradually starts to learn the characteristics of each emotion.

4. Experimental Evaluation

4.1. Experiment Setup

To evaluate the performance of the proposed framework, we used the interactive emotional dyadic motion capture (IEMOCAP) database [14], which contains audio-visual data (with transcription) performed by ten different actors. There were five sessions in the corpus, and in each session, a pair of speakers (male and female) talked to each other. For training and evaluation purposes, we used four categorical emotions, *Angry*, *Happy*, *Sad*, and *Neutral*, to represent the majority of the emotion categories in the database. For context-independent scenarios, we used only *improvised* data, which are recorded in a pre-defined situation without specific scripts.

To obtain the low-level acoustic features, we extracted 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 were expanded to 800-dimensional vectors using a context window with a size of 250 ms. The network contained three hidden layers, each of which had 256 nodes. The weights were trained by a back-propagation algorithm using a stochastic gradient descent with a mini-batch of 128 samples. In the RNN-based system, 32-dimensional vectors were used directly for input. The network contained two hidden layers with 128 BLSTM cells (64 forward nodes and 64 backward nodes). Later experiments showed that the performance did not improve

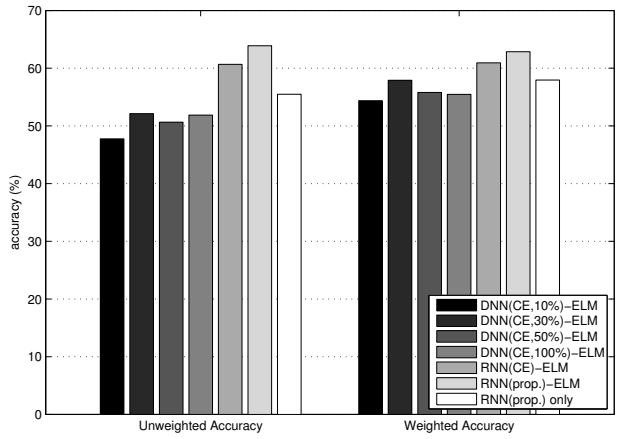

Figure 6: Comparison of different emotion recognition systems in terms of weighted and unweighted accuracies

using a higher number of hidden layers and nodes in either the DNN-based or the RNN-based systems. The reason for this was most probably over-fitting caused by data insufficiency.

To extract the global characteristics, we applied statistical functions to the output of each recursive network; the utterance-level features were then fed into an ELM network. To take non-linearity into account, we used a radial basis function as a kernel instead of the random projection method. The kernel-based ELM not only gives 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 a speaker-independent manner, we used a five-fold cross-validation technique. In each evaluation, four sessions were used for training the network, and the remaining session was divided into two sub-sessions depending on the gender. We used one for parameter setting, and the other for measurement. For evaluation purposes, we used the following two measures: weighted accuracy (WA) and unweighted accuracy (UA). Weighted accuracy is the classification accuracy for the entire test data set, and unweighted accuracy is an average of the classification accuracy for each emotion. Here, all parameters including the number of epochs, the network structure, and the kernel parameters, were tuned to maximize the unweighted accuracy value because of the imbalanced data set.

4.2. Experiment Results

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

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

5. Conclusion

In this paper, we proposed an RNN-based speech emotion recognition framework using an efficient learning approach, which enable us to account for the long-range contextual effect in emotional speech and the uncertainty of emotional labels. The proposed approach provides an insight into how RNNs and maximum-likelihood based learning process can be combined to address the shortcomings in emotion recognition systems.

6. References

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

[2] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mariño, "Speech emotion recognition using hidden Markov models." in *Proceedings of Interspeech*, 2001.

[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 *Proceedings of 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 *Proceedings of IEEE ICASSP*, 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.