EFFECTIVE ATTENTION MECHANISM IN DYNAMIC MODELS FOR SPEECH EMOTION RECOGNITION

Po-Wei Hsiao and Chia-Ping Chen

National Sun Yat-sen University, Kaohsiung, Taiwan

ABSTRACT

We propose to integrate the attention mechanism into deep recurrent neural network models for speech emotion recognition. This is based on the intuition that it is beneficial to emphasize the expressive part of the speech signal for emotion recognition. By introducing attention mechanism, we make the system learn how to focus on the more robust or informative segments in the input signal. The proposed recognition model is evaluated on the FAU-Aibo tasks as defined in Interspeech 2009 Emotion Challenge. Our baseline deep recurrent neural network model achieves 37.0% unweighted averaged (UA) recall rate, which is on par with the official HMM baseline system for dynamic modeling framework. The proposed integration of attention mechanism on top of the baseline deep RNN model achieves 46.3% UA recall rate. As far as we know, this is the best UA recall rate ever achieved on FAU-Aibo tasks within the dynamic modeling framework.

Index Terms— speech emotion recognition, deep recurrent neural network, attention mechanism

1. INTRODUCTION

A speech emotion recognition (SER) system takes a speech waveform as input and outputs one of the target emotion categories. Research activities in SER can be traced back to the 1980s [1][2]. Speech database annotated with emotion labels, such as EMO-DB [3] and FAU-Aibo [4], have been release to the research community. On the signal processing side, researches have focused on seeking informative features for emotion classes [5]. In Table 1, we list examples of low-level descriptors (LLD) and functionals commonly used for SER. On the machine learning side, Gaussian Mixture models (GMM) [6], hidden Markov models (HMM) [7], support vector machines (SVM) [8], multi-layer perceptrons (MLP) [9], and recurrent neural networks (RNN) [10] have all been used as recognition models for speech emotion recognition. Some techniques have also been proposed to boost the performance of the SER system, e.g. i-vector framework [11].

The Interspeech 2009 Emotion Challenge [12] was a very important event for SER. It allowed fair comparison between different systems via a common corpus of FAU-Aibo and a standard front end for feature extraction. Recognition results with basic static and dynamic recognition models were released as baselines. In the five-class task, the baseline for the static model and the dynamic model were respectively 38.2% and 35.9% unweighted averaged (UA) recall rate. According to Schuller et al. [13], the best performance in Interspeech 2009 Emotion Challenge five-class task was 41.65% UA, achieved by Marcel Kockmann et al. [14]. Improved results have been achieved since the Challenge. Emily Provost et al. [15] achieved 45% UA through a combination of deep belief network (DBN) and hidden Markov model (HMMs). More recently, Shih et al. [16] proposed a skew-robust training method, achieving 45.3% UA.

Table 1: Common low-level descriptors (LLDs) and functionals for speech emotion recognition.

LLDs	fundamental frequency (F0), jitter, voicing probability, zero-crossing rate, formant amplitude/position, energy, harmonics-to-noise ratio, MFCCs
Functionals	mean, min, max, range, quartile, standard deviation, skewness, kurtosis, linear regression coefficients

In this work, we study a system in the dynamic modeling framework for emotion recognition on FAU-Aibo. We introduce attention mechanism to a deep recurrent neural network model. The proposed system achieves 46.3% UA recall rate, a performance level that we are aware of no other single system has ever achieved.

This paper is organized as follows. In Section 2, we introduce the proposed method. In Section 3, we describe the experiment settings and present the evaluation results. Finally, we draw conclusion in Section 4.

2. METHOD

2.1. Attention Mechanism

Attention mechanism in neural networks is inspired by the biological visual attention mechanism found in nature. For example, human beings are able to focus on a certain region of an image with high resolution while perceiving the sur-

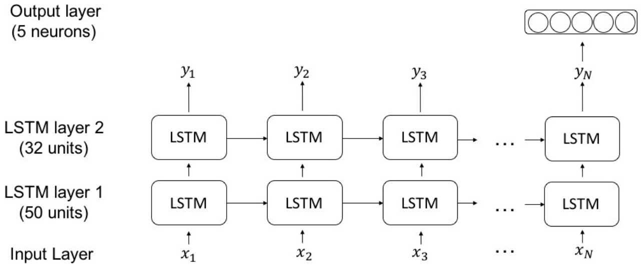


Fig. 1: LSTM model

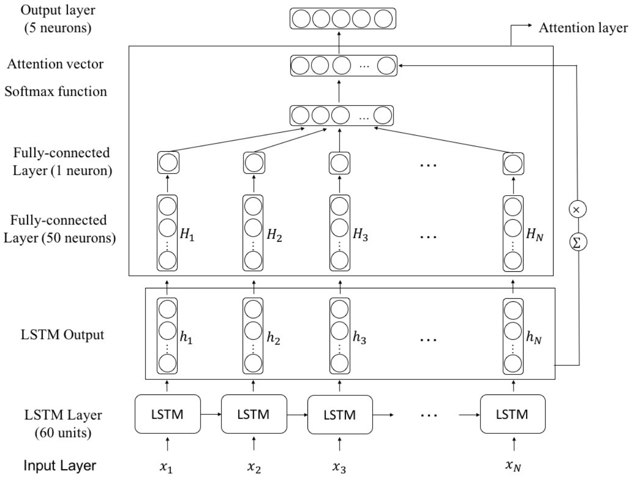


Fig. 2: LSTM-attention model

rounding regions in low resolution. Furthermore, the region of focus can be shifted dynamically in a seemingly effortless manner.

With attention mechanism, each element in the output sequence is conditional on selective elements in the input sequence. This increases the computational burden of the model, but results in a more accurate and better performing model. In most implementations, attention is realized as a weight vector (often as the output of a softmax function), the dimension of which is equal to the length of the input sequence. A larger component (weight) at a position indicates more importance of the input at the corresponding position. Besides the normal attention mechanism, a special type of attention mechanism called the structural attention mechanism [17] has been proposed. It is different from the normal attention mechanism in that it uses a memory matrix to store the contextual structural information.

Attention mechanism has been successfully applied in image recognition [18] and natural language processing [19]. In machine translation, attention mechanism allows the decoder to attend to different parts of the source sentence at each step of the output generation based on input and current partial output hypothesis.

Table 2: Class weights for data balance

	Angry	Emphatic	Neutral	Positive	Rest
weight	1.1	0.5	0.2	1.5	1.4

2.2. Recurrent Neural Networks

In our system, attention mechanism is applied on top of a recurrent neural network (RNN) model. RNN is a type of artificial neural network for modeling sequential data. Thus, RNN is a good model for natural language processing (sequence of words) or speech signal processing (sequence of audio signal). RNN often has feedback loops that allow information to flow through time steps. Currently, an RNN often uses long short-term memory (LSTM) cells [20] or gated recurrent units (GRUs) [21]. Bi-directional RNN (BiRNN) [22], which allows two-way information flow, has recently become a common practice.

2.3. Data Balance and Feature Normalization

To deal with the issue of unbalanced data, we apply class weights

$$r_k = \frac{N}{N_k} \propto \frac{1}{N_k}$$

where N is the total number of the training examples and N_k is the number of the training examples of each class.

When doing mini-batch accumulation of back propagation for parameter updates [16]. The idea is similar to data balancing, through emphasizing the errors of small-class examples and de-emphasizing the errors of large-class examples. The class weights are listed in Table 2.

Since emotion expression of different speakers can be quite diverse, we apply speaker normalization to reduce the variance due to speaker variation, and to keep variance due to emotion variation. For each feature dimension, the values of each speaker are normalized to zero mean and unit variance.

3. EXPERIMENT

3.1. Data

FAU-Aibo is a speech corpus in German. Utterances are annotated with emotion labels. It consists of 9 hours of German speech from 51 children when they are interacting with Sony's pet robot Aibo. It consist of a training set of 9959 speech chunks, and a test set of 8257 speech chunks. Interspeech 2009 Emotion Challenge uses FAU-Aibo as the competition corpus. In this work, we focus on the 5-class classification task with the emotion categories of Anger, Emphatic, Neutral, Positive, and Rest. The data of each emotion category is summarized in Table 3.

Table 3: Data sizes of FAU-Aibo corpus

	Angry	Emphatic	Neutral	Positive	Rest
train	881	2093	5590	674	721
test	611	1508	5377	215	546

Table 4: HMM baseline

# states	UA recall rate
1	32.6%
3	33.9%
5	**36.6%**

3.2. Feature Extraction

The acoustic features used in this experiment are extracted based on [12]. The dynamic modeling framework is based on 16 low-level descriptors (LLDs) including 12 mel-frequency cepstral coefficients, root-mean-square energy, zero-crossing rate, harmonics-to-noise ratio, fundamental frequency and the corresponding delta features. For each frame, a feature vector has 32 features.

3.3. Results of Evaluated Dynamic Models

We evaluate three speech emotion recognition system in the dynamic modeling framework in our experiments.

3.3.1. Hidden Markov Model

We build a baseline hidden Markov (HMM) model for this evaluation, with

- left-to-right Markov chain
- one HMM per emotion
- varying number of states
- 2 Gaussian mixture components per state
- 6+4 Baum-Welch re-estimation iterations

Experiment results with baseline HMM are shown in Table 4. The 5-state HMM model achieves 36.6% UA. As a sanity check, the baseline results agree with [12].

3.3.2. RNN model

Next, we evaluate an emotion recognition model based on RNN with LSTM cells. The hyper-parameters/settings of the network is provided in Table 5. Our RNN model consists of two LSTM layers with 50 units and 32 units respectively. At the last time step, we connect the output of the second LSTM layer to an output layer of 5 units for the target emotion classes. The structure of RNN model is depicted in Figure 1.

Table 5: Settings and hyper-parameters of RNN model

batchsize	100
learning rate	0.002
learning rate decay	0.00001
optimizer	Adam
loss function	cross-entropy

Table 6: Confusion matrix of the RNN model (A: Anger, E: Emphatic, N: Neutral, P: Positive, R: Rest)

	A	E	N	P	R	UA
A	56	275	201	65	14	9.1%
E	110	838	439	84	37	55.5%
N	277	1086	2666	1186	162	49.5%
P	3	3	61	140	8	65.1%
R	24	84	238	169	31	5.6%
					Avg.	**37.0%**

The results of RNN are shown in Table 6, along with the confusion matrix. Using this deep RNN model, we achieved a UA recall rate of 37.0%. The performance is close to (slightly better than) the 5-state HMM model.

3.3.3. LSTM-Attention Model

Finally, we evaluate the proposed LSTM-attention model for emotion recognition. The structure of the LSTM-attention model is depicted in Figure 2. The implemented LSTM-attention model has one LSTM layer with 60 units and one attention layer. The attention layer is inserted between the LSTM layer and the output layer of the original RNN. The input to the attention layer is the output of the LSTM layer. Within the attention layer, there is a layer of 50 units which are fully connected to the LSTM output. The 50 units in one frame are connected to one unit in the same frame. The one-unit outputs from all frames go through a softmax function to compute attention weights. Using these weights, the LSTM output vectors are linearly combined. Finally, the weighted sum of vectors is converted to a conditional probability for the emotion classes.

The results of these different conditions are shown in Table 7. The proposed LSTM-attention model achieves 46.3% UA, which is significantly better than RNN model. The results clearly show that attention mechanism is effective for speech emotion recognition, as we have anticipated. Also included in the table is the results from our implementation of a BiLSTM-attention model with 60 LSTM cells in each direction. The results show that using bi-directional information does not help to achieve better attention or recognition.

Table 7: Results of LSTM-attention model

	UA recall rate
LSTM-attention model	**46.3%**
BiLSTM-attention model	45.9%

Table 8: Confusion matrix of the LSTM-attention model (A: Anger, E: Emphatic, N: Neutral, P: Positive, R: Rest)

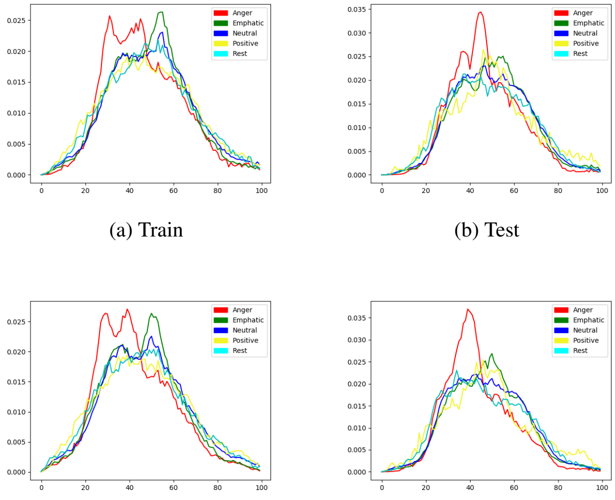
	A	E	N	P	R	UA
A	349	141	51	37	33	57.1%
E	264	899	229	35	81	59.6%
N	781	1221	2178	683	514	40.5%
P	9	9	53	127	17	59.1%
R	105	109	115	133	84	15.4%
					Avg.	**46.3%**

The confusion matrix of the best result achieved by the proposed LSTM-attention model is shown in Table 8. It is interesting to compare this matrix to the confusion matrix of the RNN model shown in Table 6. The results show that the Anger class benefits the most from attention mechanism. In addition, the Rest class also benefits, but it remains very difficult to recognize.

The time-normalized distribution of the attention weights of each class is shown in Figure 3. We can see that the middle part of utterance often gets larger weights than the marginal parts of both sides, on average. This shows that emotion content of children's utternace is often focused at the middle of speech, an interesting discovery. Finally, the t-SNE distribution graph of the data shown in Figure 4 indicates that data distributions of different classes are significantly overlapped. To disentangle the class manifolds appear to be a very challenging task.

4. CONCLUSION

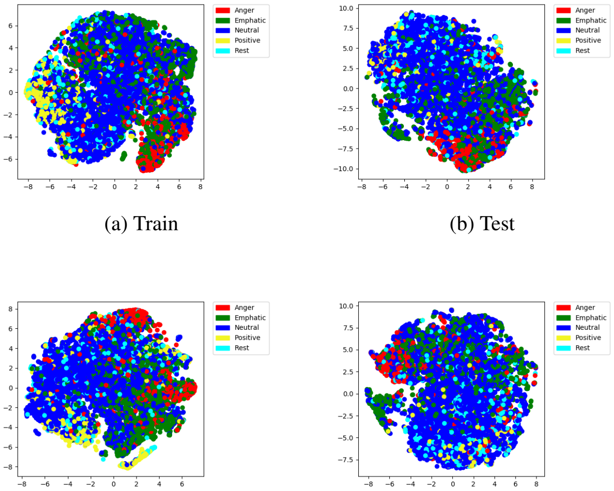
According to the experiment results, the effect of using the attention mechanism is remarkable. Performance measured by UA recall rate improves from 37.0% of RNN model to 46.3% of LSTM-attention model. The main reason for the difference is the ability to locate and focus on the salient or reliable parts of the signal. From the distribution of the attention weights, we can see that the middle part of an utterance is often more important than the beginning/ending parts. Thus, the attention mechanism allows the system to be less vulnerable to noises in the input. The performance of 46.3% is among the best performance ever achieved in the dynamic modeling framework of FAU-Aibo tasks.



(a) Train (b) Test

(c) Train (Bidirectional) (d) Test (Bidirectional)

Fig. 3: Attention weight distribution of each class



(a) Train (b) Test

(c) Train (Bidirectional) (d) Test (Bidirectional)

Fig. 4: t-SNE distribution of each class

Although we obtain good overall results by using attention mechanism, it is still difficult to recognize data of the Rest class. The Rest class is a catch-all label for the data not belonging to the other four classes. Therefore, the manifold(s) for the Rest class in the representation space is highly twisted, as we can see from the t-SNE figures. It is possible to achieve better performance with a deeper network architecture. We hope to further look into this class in the future.

5. REFERENCES

[1] Renée Van Bezooijen, Stanley A Otto, and Thomas A Heenan, "Recognition of vocal expressions of emotion: A three-nation study to identify universal characteristics," *Journal of Cross-Cultural Psychology*, vol. 14, no. 4, pp. 387–406, 1983.

[2] Rosalind W. Picard, *Affective Computing*, MIT Press, Cambridge, MA, USA, 1997.

[3] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss, "A database of german emotional speech.," in *Proceedings of Interspeech*, 2005, vol. 5, pp. 1517–1520.

[4] Stefan Steidl, *Automatic classification of emotion related user states in spontaneous children's speech*, University of Erlangen-Nuremberg Erlangen, Germany, 2009.

[5] Björn Schuller, Anton Batliner, Dino Seppi, Stefan Steidl, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Noam Amir, Loic Kessous, et al., "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals," in *Proceedings of Interspeech*, 2007.

[6] Xianglin Cheng and Qiong Duan, "Speech emotion recognition using gaussian mixture model," in *Proceedings of the 2nd International Conference on Computer Application and System Modeling*, 2012.

[7] Angeliki Metallinou, Athanasios Katsamanis, and Shrikanth Narayanan, "A hierarchical framework for modeling multimodality and emotional evolution in affective dialogs," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 2401–2404.

[8] Hao Hu, Ming-Xing Xu, and Wei Wu, "GMM supervector based svm with spectral features for speech emotion recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2007, vol. 4, pp. IV–413.

[9] Norhaslinda Kamaruddin and Abdul Wahab, "Emulating human cognitive approach for speech emotion using mlp and gensofnn," in *Proceedings of IEEE International Conference on Information and Communication Technology for the Muslim World (ICT4M)*, 2013, pp. 1–5.

[10] John J Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the national academy of sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.

[11] Rui Xia and Yang Liu, "Using i-vector space model for emotion recognition," in *Proceedings of the Thirteenth Annual Conference on International Speech Communication Association*, 2012.

[12] Björn Schuller, Stefan Steidl, and Anton Batliner, "The interspeech 2009 emotion challenge," in *Proceedings of Interspeech*, 2009, pp. 312–315.

[13] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Journal of Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[14] Marcel Kockmann, Lukáš Burget, and Jan Černockỳ, "Brno university of technology system for interspeech 2009 emotion challenge," in *Proceedings of Interspeech*, 2009.

[15] Duc Le and Emily Mower Provost, "Emotion recognition from spontaneous speech using hidden markov models with deep belief networks," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013, pp. 216–221.

[16] Po-Yuan Shih, Chia-Ping Chen, and Hsin-Min Wang, "Speech emotion recognition with skew-robust neural networks," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2751–2755.

[17] Kai Lin, Dazhen Lin, and Donglin Cao, "Sentiment analysis model based on structure attention mechanism," in *Proceedings of UK Workshop on Computational Intelligence*. Springer, 2017, pp. 17–27.

[18] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al., "Recurrent models of visual attention," in *Proceedings of advances in neural information processing systems*, 2014, pp. 2204–2212.

[19] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

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

[21] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[22] Mike Schuster and Kuldip K Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.