LEARNING UTTERANCE-LEVEL REPRESENTATIONS FOR SPEECH EMOTION AND AGE/GENDER RECOGNITION USING DEEP NEURAL NETWORKS

Zhong-Qiu Wang[1] *and Ivan Tashev*[2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Microsoft Research, One Microsoft Way, Redmond, USA
wangzhon@cse.ohio-state.edu, ivantash@microsoft.com

ABSTRACT

Accurately recognizing speaker emotion and age/gender from speech can provide better user experience for many spoken dialogue systems. In this study, we propose to use deep neural networks (DNNs) to encode each utterance into a fixed-length vector by pooling the activations of the last hidden layer over time. The feature encoding process is designed to be jointly trained with the utterance-level classifier for better classification. A kernel extreme learning machine (ELM) is further trained on the encoded vectors for better utterance-level classification. Experiments on a Mandarin dataset demonstrate the effectiveness of our proposed methods on speech emotion and age/gender recognition tasks.

Index Terms—pooling, deep neural networks, kernel extreme learning machine, speech emotion recognition, speech age/gender recognition

1. INTRODUCTION

Speech emotion recognition is becoming more and more important for many applications related to human computer interactions, especially for spoken dialogue systems. With emotion recognition from users' speech, a better user experience can be achieved. Speech emotion recognition is essentially a sequence classification problem, where the input is a variable-length sequence and the output is one single label. It is similar to many other tasks, such as action or gesture recognition in videos [1], speaker identification in speech [2], and text categorization or sentiment analysis in natural language processing [3]. However, speech emotion recognition itself is a difficult problem, as different people express their emotions in different ways. Even human annotators sometimes cannot agree on the exact emotion labels. When the linguistic information is unavailable, it becomes even harder to determine the exact emotion. In addition, the training data for building emotion recognizers is normally highly imbalanced, making it harder for the classifiers to make predictions. This research area is emerging and there is still no consensus on what features we should use for the task. In many studies, people just combine a lot of features for classification [4]. In recent years, there are some studies trying to approach this problem by predicting continuous dimensions according to the arousal/valence space [5], [6]. Their goal is to predict one continuous value for every frame. In this study, we focus on predicting one emotion label for each utterance, which better aligns with the abilities to label the utterances.

There are two major categories of methods for this problem. The first one is to first obtain fixed-length utterance-level statistics and then use a classifier for classification, without accounting for temporal dynamics information. In [4] and [7], low level features, such as pitch, energy, zero-crossing rate, MFCC, voice probability etc., are extracted per frame, and then utterance-level statistics, such as minimum, maximum, average, median etc., are calculated before final classification. In [8], Han *et al.* use a deep neural network (DNN) to obtain frame-level posteriors for utterance-level statistics computation. In [9], Ghosh *et al.* employ deep auto-encoders and recurrent auto-encoders to learn better frame-level features, and then average the learned features to obtain utterance-level statistics. The second category is to use sequence models for classification. The most conventional one is to train one GMM-HMM system for every emotion, and assign the emotion label according to the one giving largest likelihood at the test stage [10]. Recently, recurrent neural networks with long-short term memory (LSTMs) are utilized for this task. In [11], the same utterance-level label is assigned to every frame for LSTM training. At the test stage, frame-level predictions are averaged to make final predictions. We point out here that assigning the same utterance-level to every frame, especially to silence frames, may incur some problems. One of them is that the datasets for emotion recognition are normally imbalanced. In such cases, many silence frames would be labeled as the majority classes, and therefore the predictions on new silence frames would be highly biased towards majority classes. Another problem is that even if the utterance is labeled as one emotion, it does

This work was performed while the first author was a research intern at Microsoft Research Labs in Redmond, WA.

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

not necessarily mean that every frame should be labeled as that emotion. To deal with this problem, Lee and Tashev [12] propose an RNN-CTC approach, in which they assume that different frames should have different labels, and the label sequence should be alternating between the utterance-level label and a newly-introduced NULL state. In their study, expectation maximization (EM) algorithms are used for inferring the uncertainties in the label sequence. Here, we point out that adding one more NULL state may make the data more imbalanced, as silence frames normally take up around 30%~40% of all the frames in a dataset. And in addition, the EM algorithm would probably end up in assigning the NULL states to silence frames.

In this context, we propose to use DNNs to directly predict utterance-level labels, rather than problematic frame-level labels as in many other studies, for better sequence classification. The utterance-level representations are learned jointly with the utterance-level classifiers, rather than in a two-stage way. In addition, a more powerful kernel extreme learning machine is further applied to the learned utterance-level representations to improve utterance-level classification. Furthermore, we use a voice activity detector to filter out silence frames with no emotional information. In our study, we also apply the same proposed method to recognize age/gender from speech.

The rest of this paper is organized as follows. We describe our methods in more details in Section 2. Experimental setup and results are presented in Section 3 and 4. We conclude this paper in Section 5.

2. SYSTEM DESCRIPTION

We use a voice activity detector to select active frames, with which we learn the utterance-level representations jointly with the utterance-level classifiers. Finally, a kernel ELM is used for better utterance-level classification.

2.1. Voice Activity Detection

As discussed earlier, giving a new silence label to silence frames may make the data more imbalanced, and assigning the same utterance-level label to silence frames may make the predictions on new silence frames biased towards majority classes. In this study, we use a voice activity detector to filter out silence frames, and only consider frames with voice activity for our tasks. The voice activity detector used in our study is a statistical activity classifier with HMM smoothing of decisions [13], [14]. To refine the VAD results, we apply a hangover scheme and throw away the segments with less than five consecutive active frames to remove sudden bursts and the sound of puff of air in the beginning and end of each utterance. Another advantage of using VAD is that even if there is long silence in the beginning or in the end of an utterance, the behavior of the classifier would not be negatively influenced. We believe that using a voice activity detector to select frames is better than only using a fixed percentage of high-energy frames [8] or all the frames [9], [12], [11], because only these active frames contains information for emotion and age/gender recognition.

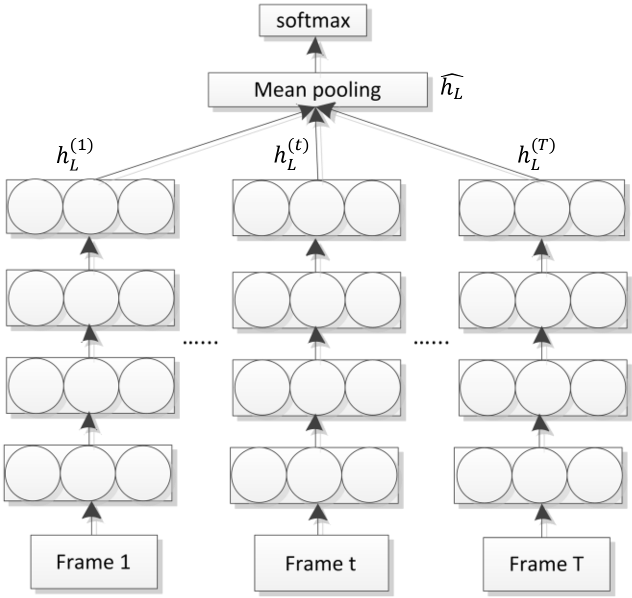

Figure 1. Overall diagram of the proposed utterance-level DNN.

2.2. Utterance-level DNN for Sequence Classification

The overall diagram of our utterance-level DNN is presented in Fig. 1. The key idea is to directly optimize the utterance-level target, and jointly learn the utterance-level classifier with the learned utterance-level representations.

In our system, every utterance is a mini-batch for DNN training. Therefore, we can feed all the frames in one utterance into the DNN, and then perform mean-pooling to the hidden activations of the last hidden layer to obtain utterance-level representations. Finally, the mean-pooled vector is sent to a softmax classifier for final classification. Mathematically,

$$\widehat{h_L} = \frac{1}{T}\sum\nolimits_{t=1}^{T} h_L^{(t)} \tag{1}$$

$$pred = softmax(W_{L+1} * \widehat{h_L} + b_{L+1}) \tag{2}$$

where $h_L^{(t)}$ is the activation of the last hidden layer at frame t, $\widehat{h_L}$ is the mean-pooled utterance-level representation, and W_{L+1} and b_{L+1} are the parameters between the last hidden layer and the output layer.

Here we only use mean-pooling to obtain utterance-level representations. We tried to use min-pooling, max-pooling, and a combination of them, but mean-pooling consistently gave us the best performance in our experiments.

Here we would like to point out that as every utterance is a mini-batch for DNN training, data shuffling may not be as good as training DNNs frame-wisely. In addition, when we back-propagate the error gradient through the mean pooling layer, the error gradient is equally dispatched to different frames as in Eq. (3).

$$\frac{\partial Loss}{\partial h_L^t} = \frac{1}{T} \frac{\partial Loss}{\partial \widehat{h_L}} \tag{3}$$

As a result, the network could not be well optimized. To speed up the training process, a simple yet very effective strategy we use is to initialize the utterance-level DNN with a DNN trained with frame-wise labels. In our experiments, we not only observe much faster convergence, but also obtain better results.

We emphasize that the utterance-level representations are learned jointly with the utterance-level classifier. This strategy would be better than first representing an utterance as a fixed-length vector and then training an utterance-level classifier in a two-stage way [8], or optimizing frame-wise targets.

2.3. Kernel Extreme Learning Machines

The goal of the softmax classifier is to perform utterance-level classification. One potential problem is that most of the datasets in emotion recognition consists of only several thousand utterances. As a result, the softmax classifier may not be well trained because of the limited number of training examples. To deal with this problem, we replace the softmax classifier with a more powerful kernel ELM. We train one kernel ELM after every training epoch of the utterance-level DNN. The kernel ELM [15] is basically a variant of kernel methods, which performs better on small datasets [16]. Several other methods, such as support vector machines (SVMs) and ordinary ELMs, were compared in this study and our previous studies [8], [12], and we find that kernel ELMs consistently give us better performance over the other two models.

3. EXPERIMENTAL SETUP

17,408 real-traffic Mandarin utterances have been collected from a Microsoft spoken dialogue system. Each utterance is labeled by five crowdsourcing judges using the Microsoft UHRS labeling system. All of the judges are native Mandarin speakers. There are four meta-categories for the emotion recognition task, i.e. neutral (no clear emotion), happy (excited, interested, happy, funny, and flirting), sad (depressed, bored, tired, sad, and frustrated) and angry (disgust, impatient, offended, and angry). We have finer categories for each meta-category when prompting the crowdsourcing judges to label the data, mainly because speech emotion itself is very fuzzy.

Although a large number of labeled utterances can be quickly obtained using crowdsourcing judges, the labels are less reliable compared with professional and serious annotators. Therefore, utterance selection is of great importance for building good emotion recognition classifiers. In our study, we only retain utterances with labels like “AAABC”, “AAAAB”, and “AAAAA”, as the annotations with less than three agreements are considered unreliable. For these retained utterances, we use majority voting to label each utterance. We can also compute the performance of human judges using these labels. In this case, it is 82.18%. Note that utterances with labels like “AAABB” are not considered, because the underlying emotion could be “A” and it also could be “B”. If we incorporate these, the performance of human judges drops to around 75%.

Each judge is requested to give the age/gender label for every utterance as well. There are three categories for the age/gender recognition task, i.e. Children, Female adults, and Male adults. The performance of human judges is 91.36% for this task using the selected utterances above.

From the initial set of 17,408 utterances, 10,527 utterances (~10 hours) are left after filtering using the aforementioned criterion. As we can see from Fig. 2, the class distributions for both tasks are very imbalanced, which is typical for many other datasets for this task. Nonetheless, we think that both distributions are reasonable from a practical perspective. Many users are excited to talk with the dialogue system, so there are a lot of happy utterances. There are also many angry utterances. This may be because the dialogue system sometimes cannot understand or does not satisfy the users. There are only a small number of sad utterances in the dataset. This is reasonable because people do not want to talk with a chatbot when they are in a sad mood. For the age/gender distribution, there are a lot of utterances from male adults. This is likely because the dialogue system is designed to be a young female. There are only a small number of utterances from children. This may be because many children in China do not have their own cell phones at an early age.

In our experiments, we randomly choose 70% of the data for training, 15% for validation and the rest 15% for testing. We use the validation set to perform hyper-parameter tuning and early stopping.

We use weighted accuracy and un-weighted accuracy to measure the performance, as is done in many other studies. The weighted accuracy is just the classification accuracy over the entire test set. The un-weighted accuracy is the average of the classification accuracy for each class, which accounts for the imbalanced nature of the data. In this study, we focus on the weighted accuracy, as it represents the percentage of users we can satisfy and it is self-weighted from the usability standpoint. We have tried to give larger weights to minority classes so that un-weighted accuracy can be optimized, but weighted accuracy drops in such cases.

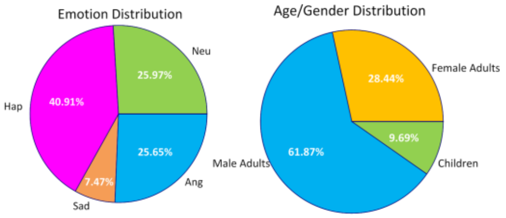

Fig. 2. Class distribution for the emotion recognition task and age/gender recognition task.

4. EVALUATION RESULTS

We compare our approach with the state of the art DNN-ELM approach proposed in [8] by Han *et al*. The key idea of their approach is to first select a fixed percentage of frames with high energy for each utterance, and then a DNN is trained to obtain frame-level posteriors, with which utterance-level statistics are calculated. Finally, an ELM is trained on the utterance-level statistics to perform utterance-level classification. It was shown in their study that much better performance can be obtained over the methods proposed in [10], [17], [4] on the IEMOCAP dataset [18]. When applying their methods to our tasks, we also use the voice activity detector to select frames for every utterance rather than using problematic fixed percentage of high-energy frames. This leads to large improvements in our studies.

There are four hidden layers in our DNNs, each with 512 rectified linear units. The network is trained with mini-batch SGD with momentum to minimize the cross-entropy criterion. The features used in our study are energy, pitch, voice probability, and 26-dimensional log Mel-spectrogram features extracted from each frames. For the log Mel-spectrogram features, we perform utterance-level mean normalization as this can alleviate the channel effects of different microphones. The delta component of all the features mentioned above is added to consider the dynamics of these features. For each frame, we concatenate a symmetric 25-frame context window to obtain our final frame-level features. The input feature dimension is therefore 1,450.

The results are shown in Fig. 3. The proposed approach gives absolute 3.8% (from 59.40% to 63.20%) better weighted accuracy over the DNN-ELM approach on the emotion recognition task, and absolute 2.16% (from 90.56% to 92.72%) better weighted accuracy on the age/gender recognition task. On the age/gender recognition task, the proposed approach even gets better results than the performance of human judges (92.72% compared to 91.36%).

The confusion matrices for both tasks are presented in Table 1 and 2, respectively. For the emotion recognition task, a lot of confusion is concentrated between Neutral and Happy, and between Happy and Angry. For the age/gender recognition task, there is a small amount of confusion between female and male adults. There is some confusion between Children and Female adults as well. We think this is because there is no distinguishing between boys and girls in the labeling process and in addition, the utterances of Children and Female adults are quite similar, especially in terms of pitch features. The classification accuracies of majority classes are generally higher than minority classes for both tasks. This may be because we are focusing on improving weighted accuracy.

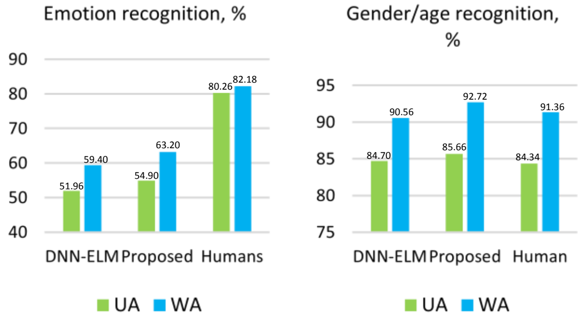

Fig. 3. Comparison of the proposed method with the DNN-ELM approach proposed in [8] on the speech emotion recognition task and age/gender recognition task.

Table 1. Confusion matrix of the emotion recognition task

	Neu	Hap	Sad	Ang	Total (percentage)
Neu	227	115	21	47	410 (25.97%)
Hap	66	511	7	62	646 (40.91%)
Sad	35	33	35	15	118 (7.47%)
Ang	47	125	8	225	405 (25.65%)

Table 2. Confusion matrix of the age/gender recognition task

	Children	Female	Male	Total (percentage)
Children	107	45	1	153 (9.69%)
Female	33	400	16	449 (28.44%)
Male	1	19	957	997 (61.87%)

5. CONCLUDING REMARKS

We have proposed an utterance-level DNN that can directly optimize utterance-level targets, and jointly learn utterance-level representations with utterance-level classifiers. The proposed approach gives us 3.8% weighted accuracy and 2.94% un-weighted accuracy improvement over a strong DNN-ELM approach on the emotion recognition task. In addition, the performance on the age/gender recognition task is close to or better than the performance of human judges. Finally, many previous studies on speech emotion recognition are focused on English. To our best knowledge, this is the first attempt to perform emotion detection for Mandarin in a spoken dialogue system. Our findings in this study suggest that language differences may not pose a big challenge for speech emotion recognition with the selected features set.

6. ACKNOWLEDGEMENTS

We would like to thank Malone Ma for collecting the data.

7. REFERENCES

[1] R. Poppe, “A Survey on Vision-based Human Action Recognition,” *Image Vis. Comput.*, 2010.

[2] J. Hansen and T. Hasan, “Speaker Recognition by Machines and Humans: A Tutorial Review,” *IEEE Signal Process. Mag.*, pp. 74–99, 2015.

[3] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of Tricks for Efficient Text Classification,” in *arXiv preprint arXiv: 1607.01759*, 2016.

[4] F. Eyben, M. Wöllmer, and B. Schuller, “OpenEAR—Introducing the Munich Open-source Emotion and Affect Recognition Toolkit,” in *Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–6.

[5] F. Weninger, F. Ringeval, E. Marchi, and B. Schuller, “Discriminatively Trained Recurrent Neural Network for Continuous Dimensional Emotion Recognition from Audio,” in *Proceedings of IJCAI*, 2016, pp. 2196–2202.

[6] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, “ADIEU Features? End-to-end Speech Emotion Recognition using A Deep Convolutional Recurrent Network,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 5200–5204.

[7] A. Stuhlsatz, C. Meyer, F. Eyben, and T. ZieIke, “Deep Neural Networks for Acoustic Emotion Recognition: Raising the Benchmarks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011.

[8] K. Han, D. Yu, and I. Tashev, “Speech Emotion Recognition using Deep Neural Network and Extreme Learning Machine,” in *Proceedings of Interspeech*, 2014.

[9] S. Ghosh, E. Laksana, L. Morency, and S. Scherer, “Learning Representations of Affect from Speech,” in *arXiv preprint arXiv:1511.04747*, 2015.

[10] B. Schuller, G. Rigoll, and M. Lang, “Hidden Markov Model-based Speech Emotion Recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2003.

[11] G. Keren and B. Schuller, “Convolutional RNN: an Enhanced Model for Extracting Features from Sequential Data,” in *arXiv preprint arXiv:1602.05875*, 2016.

[12] J. Lee and I. Tashev, “High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition,” in *Proceedings of Interspeech*, 2015.

[13] J. Sohn, N. Kim, and W. Sung, “A Statistical Model-based Voice Activity Detection,” in *IEEE signal processing letters*, 1999, pp. 1–3.

[14] I. Tashev, A. Lovitt, and A. Acero, “Unified Framework for Single Channel Speech Enhancement,” in *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, 2009, pp. 883–888.

[15] G. Huang, Q. Zhu, and C. Siew, “Extreme Learning Machine: Theory and Applications,” *Neurocomputing*, pp. 489–501, 2006.

[16] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[17] L. Li, Y. Zhao, D. Jiang, Y. Zhang, and F. Wang, “Hybrid Deep Neural Network--Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition,” in *Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 312–317.

[18] C. Busso, M. Bulut, C. Lee, and A. Kazemzadeh, “IEMOCAP: Interactive Emotional Dyadic Motion Capture Database,” in *Language resources and evaluation*, 2008, pp. 335–359.