INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Speech Emotion Recognition with Discriminative Feature Learning

Huan Zhou, Kai Liu

Artificial Intelligence Application Research Center, Huawei Technologies
Shenzhen, PRC

zhou.huan@huawei.com, liukai89@huawei.com

Abstract

The performance of a speech emotion recognition (SER) system heavily relies on the deep feature learned from the speeches. Most state of the art has focused on developing various deep architectures for effective feature learning. In this study, we make the first attempt to explore feature discriminability instead. Based on our SER baseline system, we propose three approaches, two on loss functions and one on combined attentive pooling, to enhance feature discriminability. Evaluations on IEMOCAP database consistently validate the effectiveness of all our proposals. Compared to the baseline system, the proposed three systems demonstrated at least +4.0% absolute improvements in accuracy, with no increment in the total number of parameters.

Index Terms:speech emotion recognition, discriminative feature learning, attentive pooling

1. Introduction

Speech emotion, as a kind of meta data of speech, plays an important role in human communication by conveying the underlying psychology and preferences of speakers. In human-computer interaction field, there has been increasing interest in developing natural interactions with the ability to recognize, interpret and respond to the emotions expressed by speakers.

However, it remains a challenge to recognize emotion states from speech signals because emotions are usually conveyed in subtle and complex ways. The relevant research, called speech emotion recognition (SER), has been dedicated to extract relevant information from speeches that can effectively represent emotion states.

Traditionally, a basic SER system has a two-stages pipeline, feature extraction and emotion classification. A number of handcrafted LLD (Low-Level Descriptors) features, built upon prior knowledge, are primarily used to represent emotions. In the 1st stage, a typical LLD feature set includes both acoustic and prosodic features, such as pitch, energy, intensity, Mel-frequency cepstral coefficients, speaking rate and so on. Since feature information in a speech follows a chronological sequence, a series of statistical aggregation functions (e.g., mean, max, variance, etc) are typically applied on the LLD feature set to to produce a statistically aggregated feature vector. In the 2nd stage, some popular machine learning methods (like hidden Markov model, support vector machines) are applied on the feature vector to identify associated emotion state.

With the development of deep learning technology, the traditional two-stage SER approaches have evolved to end-to-end deep neural networks (DNNs), where task-oriented deep features for SER can be automatically learned via supervised network training.

To learn the deep SER features, various network architectures, associated with different front-end inputs, have been investigated. In [1], a three-dimensional attention-based convolutional recurrent neural network (CRNN) was proposed, to extract deep feature from inputs of mel-spectrogram with first-order and second-order derivatives. Using the same kind of input, the authors in [2] employed a CNN-BLSTM (bidirectional long short-term memory) model, and enhanced by multi-head attention and multi-task learning, to joint learn deep feature and speaker gender information. In [3], dilated Residual Network (DRN), combined with multi-head self-attention were employed, from LLD feature inputs. An interleaved time-delay neural network with unidirectional LSTM (TDNN-LSTM) was explored in [4], with front end of Fourier log-energies. And a CNN LSTM network was constructed on top of log-mel spectrogram [5].

For DNN feature training, a classification layer is applied on the deep feature subnet, that serves to map the deep feature into different emotion state. Since classification of emotion states is mutually exclusive, the softmax based cross entropy (CE) loss is unanimously adopted in above listed works, as an optimization criterion to guide the deep feature learning process.

We note that while SER has been extensively studied, most prior art focused to investigate DNN architecture of feature learning subnet, little attention is paid on the classifier. In fact, as pointed out in many studies, the softmax loss has insufficient discrimination ability: only good at separating different classes, but not good at making features of the same class compact. To maximize the discrimination ability, several discriminative loss functions were proposed to by optimizing both intra-class compactness and inter-class separability. In research fields such as face recognition and speaker recognition, performance improvements were reported with adoption of these loss functions.

Motivated by the this, in this study, three discriminative feature learning approaches are explored for SER feature learning with the goal of improvement performance. To authors' best knowledge, it is the first attempt to investigate SER from this research objective.

The remainder of this paper is organized as follows: Section 2 briefly introduces the loss functions to be explored. Section 3 details our proposed SER system. Section 4 presents experimental results and discussions. Finally, our conclusions are given in Section 5.

2. Preliminaries for proposed system

In general, currently available discriminative loss functions can be divided into two categories. One category is margin-based by directly increasing the feature margin between different classes. The other category is mining-based by focusing on optimizing hard examples. In the study, we choose to explore one popular loss function from each category, with brief introductions given

http://dx.doi.org/10.21437/Interspeech.2020-2237

in this section.

2.1. Softmax loss

To facilitate our description, the softmax loss is reviewed first. Its formulation is given by:

$$L^i_{Softmax} = log\frac{e^{W^T_{y_i} f_i} + b_{y_i}}{\sum_{j=1}^{c} e^{W^T_j f_i} + b_j} \tag{1}$$

Here c denotes the number of class; $f_i \in \mathbf{R}^d$ denotes feature vector of the i-th input sample; y_i is the corresponding ground-truth label. $W_j \in \mathbf{R}^d$ denotes the jth column of the weights $W \in \mathbf{R}^{d*c}$ and $b \in \mathbf{R}^c$ is the bias, which are trainable network parameters.

From (1), we can see that the softmax only penalizes on classification error, and does not explicitly encourage intra-class compactness.

2.2. Additive margin softmax loss

As one typical margin based loss function, the additive margin softmax loss[6] (AMS) was proposed for deep face recognition. Its idea is to directly enhance the feature discrimination by importing an angular margin into the softmax loss, with formulation of:

$$L^i_{AMS} = log\frac{e^{s*(cos\theta_{y_i}-m)}}{e^{s*(cos\theta_{y_i}-m)} + \sum_{j=1,j\neq y_i}^{c} e^{s*cos\theta_j}} \tag{2}$$

where $cos\theta_{y_i} = \frac{W^T_{y_i} f_i}{||W_{y_i}||||f_i||}$. Compared to the softmax loss in (1), AMS introduces two hyper-parameters. The parameter s is a scaling factor used to ensure the gradient not too small during the training, and m is a hard fixed margin. Note that both feature f_i and weights w_i are normalized and bias term is ignored in the AMS definition.

2.3. Focal loss

As a typical mining based loss function, Focal loss (FL) [7] was designed to focus training on a sparse set of hard, misclassified examples.

The FL is a dynamically scaled CE, expressed by:

$$L_{FL} = -\alpha_t(1-p_t)^{\gamma} log(p_t) \tag{3}$$

where $log(p_t)$ is a convenient representation of CE. FL introduces two hyperparameter, γ and α. The focusing parameter γ controls the modulation strength to the standard CE loss. When $\gamma = 0$, FL is equivalent to the CE loss; as γ increases, it smoothly reduces the loss contribution of easy examples and in turn focuses on the hard negative examples. The balancing parameter α addresses class imbalance, e.g. to place less emphasis on the positives as easy negatives are down-weighted.

Although intra-class compactness is not explicitly optimized in FL, by focusing training on informative examples, FL usually results in more discriminative features.

3. Proposed SER system

In this section, a baseline SER system is proposed first. Then we propose three approaches to enhance the baseline with an intention of discriminative feature learning.

3.1. Proposed SER baseline system

As described in the Sec.1, a lot of network architectures, such as CNNs, RNNs, LSTMs and their variants have been successfully employed in the recent SER works. Along the line, a baseline SER system is built with combined components of CNNs, LSTM and attention. The overall system architecture is illustrated in Fig.1, as per the details given below.

- *Feature Extractor:* as a front-end processor, a LLD feature sequence is extracted from a given raw speech;
- *CNN:* Up to 6 convolutional sub-modules are employed to extract temporal feature sequence from the front-end. Each sub-module includes operations of convolution, batch normalization, average pooling and dropout. Specifically, the first two layer have 128 convolution filters, while remaining 64 filters. The convolution kernel has the same size 3×3, the same pooling size 2×2, with dropout factor as 0.25. In addition, different from top 5 sub-modules, pooling operation in the last sub-module is skipped;
- *BLSTM:* bidirectional LSTM, with 64 memory cells on each direction, generates high-level feature sequence by learning contextual dependencies among long-term temporal features both forward and backward;
- *Attention:* by applying different attention weights on high-level feature sequence, the feature sequence is selectively aggregated into a compact deep feature with fixed-length of 64;
- FCC: as logits layer, a linear affine transformation is applied to map the deep feature into logits;
- *Classifier:* the logits are projected into different classification spaces using softmax loss, and outputs posterior probabilities for all emotional states, which sum to unity.

3.2. Enhanced SER with discriminative classification losses

To enhance the baseline system, the softmax loss in the classifier is substituted by AMS and FL, respectively. This leads two enhanced SER systems.

Note that alternation of loss functions in the classifier, does not affect the number of trainable parameters.

3.3. Enhanced SER with combined attentive pooling

To ignore pauses or non-emotional part presented within an utterance, in most prior SER work, attentive pooling scheme is used to pool features over time.

Such an attentive pooling scheme can be mathematically summarized by:

$$\begin{aligned} \alpha_t &= \frac{exp(e_t)}{\sum_{i=1}^{T} exp(e_i)}, \qquad t = 1, ..., T \\ \tilde{\mu} &= \sum_{t}^{T} \alpha_t h_t \end{aligned} \tag{4}$$

where $\{h_1, h_2, ...h_T\}$ is a temporal feature sequence as input, $\alpha_t \in [0, 1]$ denotes normalized attention weight, e_t denotes a scalar similarity score and $\tilde{\mu}$ is a sum-pooling feature as pooling output. Clearly, the scheme is advantageous over traditional average or global pooling scheme, by additive adjustable focus over temporal axis.

In one paper on speaker verification [8], another attentive pooling was proposed and outperformed the attentive pooling.

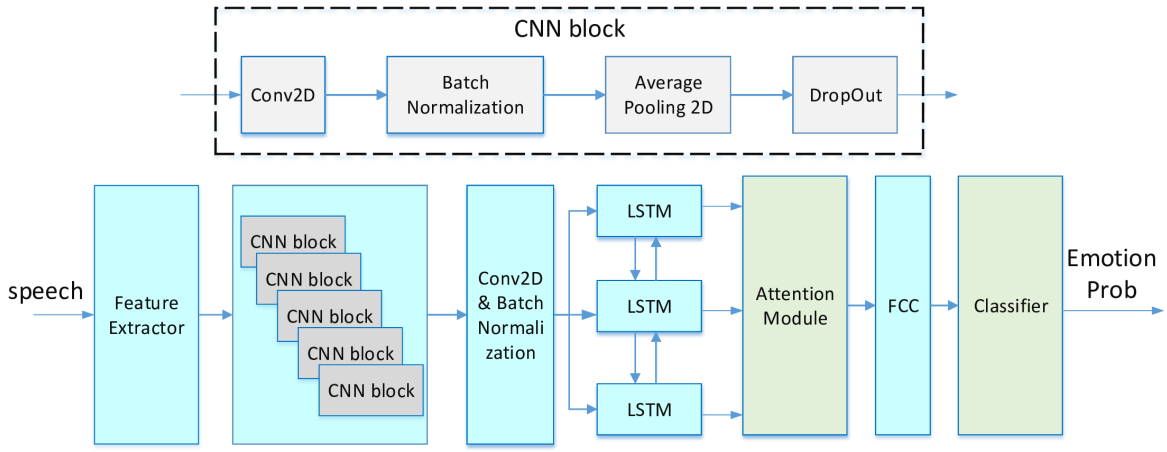


Figure 1: *Framework of our proposed baseline*

The new scheme considered both weighted mean and weighted standard deviation (SD). In our opinion, by capturing feature variations in another dimension, additional SD-pooling can give the pooled feature better degree of discrimination. Motivated by this, we propose to an approach to enhance the baseline by modifying the attention module, namely, combined attentive pooling, as our 3rd proposed SER system.

In detail, let h_t denote BLSTM outputs at each time step and each recurrent direction $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$, the weighted sum pooling is performed following (4), where score e_t is calculated in Luong's multiplicative style [9].

Then additional SD-pooling is formulated as:

$$\tilde{\sigma} = \sqrt{max\left(\sum_{t}^{T} \alpha_t h_t \cdot h_t - \tilde{\mu} \cdot \tilde{\mu}, 0\right)} \tag{5}$$

where $\tilde{\mu}$ is output of sum-pooling from (4). Combining (4) and (5), the final pooled feature is represented by feature concatenation of $[\tilde{\mu}, \tilde{\sigma}]$.

4. Experiments and results

To assess the performance of three SER systems proposed in Sec.3, several experiments were carried out on IEMOCAP [10], a widely databaset used in SER research. Experiment details and results are presented in this section.

4.1. Experiment setup

- **Database** The original IEMOCAP dataset consists 12 hours English conversations, segmented into utterances, associated with labels of 9 classes. In accordance with the reported procedure in prior works, a 4-class subset, with label of *happy, angry, sad, neutral* (where *happy* includes *excited*), is used in our experience, with total 5531 utterances. The dataset is further randomly split into training, validation and test sets in the ratio of $0.55 : 0.25 : 0.2$.
- **Features** The length of all speech utterances are firstly clipped into 7.5 seconds with trimming or zero padding. Then for each utterance, 40-dimensional log Mel-filterbank energy features are extracted for input frames, of which frame window length is 25 ms and hop size is 10 ms.
- **Metrics** For performance assessment, two metrics are employed: weighted accuracy (WA) that is the overall classification accuracy and unweighted accuracy (UA) that is the average recall over the emotion categories. And all the reported experimental results herein are based on 5-fold cross-validation.
- **Implementation Details** The proposed SER system is implemented with TensorFlow toolkit, with initial learning rate of 0.0001 and momentum of 0.9. The system is trained using Adam optimizer, with optimization criterion of categorical cross entropy and maximum epoch as 100. In addition, to overcome overfitting, techniques of cross validation, batch normalization, early stopping are adopted in all our experiments. Validation accuracy is monitored with patience setting of 10 epochs.

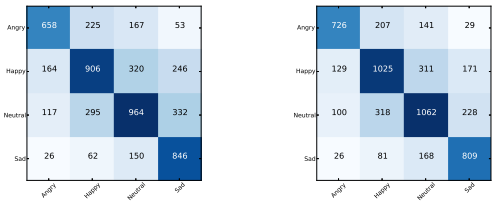
4.2. Experiments on proposed baseline

Although the baseline performance is not our main research objective in this study, it is beneficial to compare it with several recently proposed SER architectures as benchmarks. The comparison results are listed in Tab.1, in terms of UA.

Table 1: *Comparison results with Prior Art*

Method	UA (%)
LSTM [11]	57.1
LSTM + Att [12]	59.6
RNN + Att [13]	59.7
CNN + BLSTM [14]	59.4
TDNN + LSTM [4]	60.7
ResNet + MHA [3]	67.4
3D CRNN [1]	64.7
The proposed baseline	62.3

From the Tab.1, it can be observed that our proposed baseline achieves satisfying results by outperforming most prior art. In addition, it's worth mentioning that our baseline network is very light, with parameter number less than 360K, in contrast



(a) *Softmax* (b) *FL*

Figure 2: *Comparison results on Confusion Matrices*

to previously reported 9-10M parameters in some earlier works [3, 13].

4.3. Experiments on enhanced SER system

Extensive experiments on all 3 proposed SER systems are conducted. Six sets of hyperparameters are tested for either AMS-based or FL-based SER system, respectively. The overall results are reported in Tab.2, in terms of both UA and WA.

Table 2: *Comparison results on IEMOCAP*

system		UA (%)	WA (%)
Baseline		62.3	61.0
AMS	$s = 10, m = 0.1$	65.7	64.7
	$s = 10, m = 0.2$	65.0	63.6
	$s = 10, m = 0.4$	**66.7**	**65.6**
	$s = 30, m = 0.1$	65.1	63.7
	$s = 30, m = 0.2$	63.2	63.5
	$s = 30, m = 0.4$	64.2	63.6
FL	$\gamma = 2, \alpha = 1.0$	66.1	64.9
	$\gamma = 2, \alpha = 2.0$	**66.3**	**65.5**
	$\gamma = 2, \alpha = 4.0$	65.8	65.2
	$\gamma = 5, \alpha = 1.0$	66.5	65.0
	$\gamma = 5, \alpha = 2.0$	65.8	64.6
	$\gamma = 5, \alpha = 4.0$	65.8	64.1
Combined attentive pooling		**66.7**	**65.2**

From the results reported above, we can see that all enhanced systems consistently outperformed the baseline. This validated the effectiveness of our proposals about discriminative SER feature learning. The best performance for AMS is achieved with setting of $s = 10, m = 0.4$ and $\gamma = 2, \alpha = 2.0$ works best for FL. Regarding the complexity in terms of number of trainable parameters, the baseline, AMS and FL-based system all share the same amount of 383K, and the combined attentive pooling approach only introduces negligible increment(around 0.07%).

Lastly, as frequently reported in prior art, *happy* is worst classified. Thus some samples of *happy* can be regarded as hard samples and might be focused by FL. To verify this, a close look is taken on all six FL-related experiments results by comparing categorical accuracy presented with confusion matrices. We observed that, interestingly, all those experiments confirmed that FL is beneficial for *happy* classification. As illustrated in Fig.2, by simply substituting softmax with FL, up to 13% relative improvement was achieved on *happy*.

To sum up, overall at least 4.0% absolute improvements were achieved by all 3 proposed SER systems. A deeper look reveals that each system has its own advantages: 1) the AMS based SER system slightly outperformed the other two, in terms of highest accuracy without increment in parameter cost; 2) the FL based system performed best on classifying *happy*, the class with the poorest performance reported in most prior art; 3) the combined pooling mechanism provided a practically convenient solution without the need to tune additional hyper-parameters.

5. Conclusion

In this study, we investigated the SER problem by focusing on learning discriminative deep feature. We proposed to substitute softmax with more discriminative AMS and FL, and enhance attentive pooling with combined attentive pooling. Experimental results on IEMOCAP dataset showed remarkable performance improvements for all proposals. In our future work we would like to investigate the integration of the proposed schemes.

6. References

[1] M. Chen, X. He, J. Yang, and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, pp. 1–1, 07 2018.

[2] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *Interspeech 2019*, 2019.

[3] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, "Dilated residual network with multi-head self-attention for speech emotion recognition," 05 2019, pp. 6675–6679.

[4] M. Sarma, P. Ghahremani, D. Povey, N. Goel, K. Sarma, and N. Dehak, "Emotion identification from raw speech signals using DNNs," 09 2018, pp. 3097–3101.

[5] J. Zhao, X. Mao, and C. Lijiang, "Speech emotion recognition using deep 1d & 2d CNN LSTM networks," pp. 312–323, Jan. 2019.

[6] F. Wang, W. Liu, H. Liu, and J. Cheng, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, Jul 2018.

[7] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollr, "Focal loss for dense object detection," in *ICCV 2017*, 2017, pp. 2999–3007.

[8] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Interspeech.2018*, 09 2018, pp. 2252–2256.

[9] M.-T. Luong, H. Pham, and C. Manning, "Effective approaches to attention-based neural machine translation," in *EMNLP 2015*, 08 2015, pp. 1412–1421.

[10] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower Provost, S. Kim, J. Chang, S. Lee, and S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 12 2008.

[11] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," 01 2017, pp. 873–883.

[12] G. Ramet, P. Garner, M. Baeriswyl, and A. Lazaridis, "Context-aware attention mechanism for speech emotion recognition," 12 2018, pp. 126–131.

[13] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *ICASSP 2017*, 03 2017.

[14] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Interspeech.2017*, 08 2017, pp. 1089–1093.