HGFM : A HIERARCHICAL GRAINED AND FEATURE MODEL FOR ACOUSTIC EMOTION RECOGNITION

Yunfeng Xu[⋆†] *Hua Xu*[⋆] *Jiyun Zou*[⋆†]

[⋆] State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[†] School of Information Science and Engineering,
Hebei University of Science and Technology, Shijiazhuang 050018, China
hbkd_xyf@hebust.edu.cn, xuhua@tsinghua.edu.cn, zoujiyun@stu.hebust.edu.cn

ABSTRACT

To solve the problem of poor classification performance of multiple complex emotions in acoustic modalities, we propose a hierarchical grained and feature model (HGFM). The frame-level and utterance-level structures of acoustic samples are processed by the recurrent neural network. The model includes a frame-level representation module with before and after information, a utterance-level representation module with context information, and a different level acoustic feature fusion module. We take the output of frame-level structure as the input of utterance-level structure and extract the acoustic features of these two levels respectively for effective and complementary fusion. Experiments show that the proposed HGFM has better accuracy and robustness. By this method, we achieve the state-of-the-art performance on IEMOCAP and MELD datasets.

Index Terms— Emotion Recognition, Hierarchical, GRU

1. INTRODUCTION

With the proliferation of videos posted online (e.g., on YouTube, Facebook, Twitter) for product reviews, movie reviews, political views, and more, affective computing research has increasingly evolved from conventional unimodal analysis to more complex forms of multimodal analysis [1]. Emotion recognition has been widely studied in the textual field. However, research in the acoustics field cannot be ignored. In the expression of different emotions, acoustic information has a unique aspect. For example, surprise and anger tend to contain less textual information, and voice messages are more important and effective in identifying such emotions. This paper focuses on the emotional computation of acoustic modality.

The key to solve the task of acoustic emotion recognition is obtaining effective representation features of original acoustic data and designing robust neural network models. Zhou [2] utilize openSMILE toolkit [3] to extract 1,582 statistic acoustic features, which is the same as the acoustic features used in the INTERSPEECH 2010 Paralinguistic Challenge. And input these features into a multi-path DNN network. Aguilar [4] and Hazarika [5] use the (*ComParE*) feature-set introduced by Schuller [6] for the InterSpeech emotion recognition challenge. The former extracts feature according to the phonetic length of each word and input them into a bi-directional Long short-term Memory network. The latter is a Conversational Memory Network that takes into account the different speakers. Li [7] use the LibROSA [8] speech toolkit. Extracting 41-dimensional frame-level acoustic features. Using a multi-core CNN network to learn these features. However, these methods do not take into account the acoustic characteristics when processing acoustic features. It is easy to ignore the temporal variation of acoustic data when directly extracting the statistical features of the whole utterance. Therefore, it is insufficient to extract the acoustic features of word-level and set the convolution kernel of CNN respectively to learn the acoustic local features.

To tackle these challenges, we proposed a hierarchical grained and feature model. Based on the success of textual modality in the emotion recognition task, we think it is meaningful to explore fine-grained representation in the acoustic field. Therefore, the frame-level and utterance-level structures of acoustic data are modeled. Our model consists of three modules. First, we utilize the features extracted by LibROSA as input and learn the frame-level features that contain temporal sequence through the Gated Recurrent Unit network (F-GRU) as a representation of utterance-level. Then we use utterance GRU (U-GRU) network learning to include a utterance-level representation of contextual information in dialogue. Finally, we fuse this feature with statistical features extracted by openSMILE to obtain the final acoustic data representation feature. Our fusion phase is reasonable because the statistical feature extracted by openSMILE is based on utterance. The validity of the innovative work is verified on the IEMOCAP and MELD dataset.

The rest of this paper is arranged as follows. In the sec-

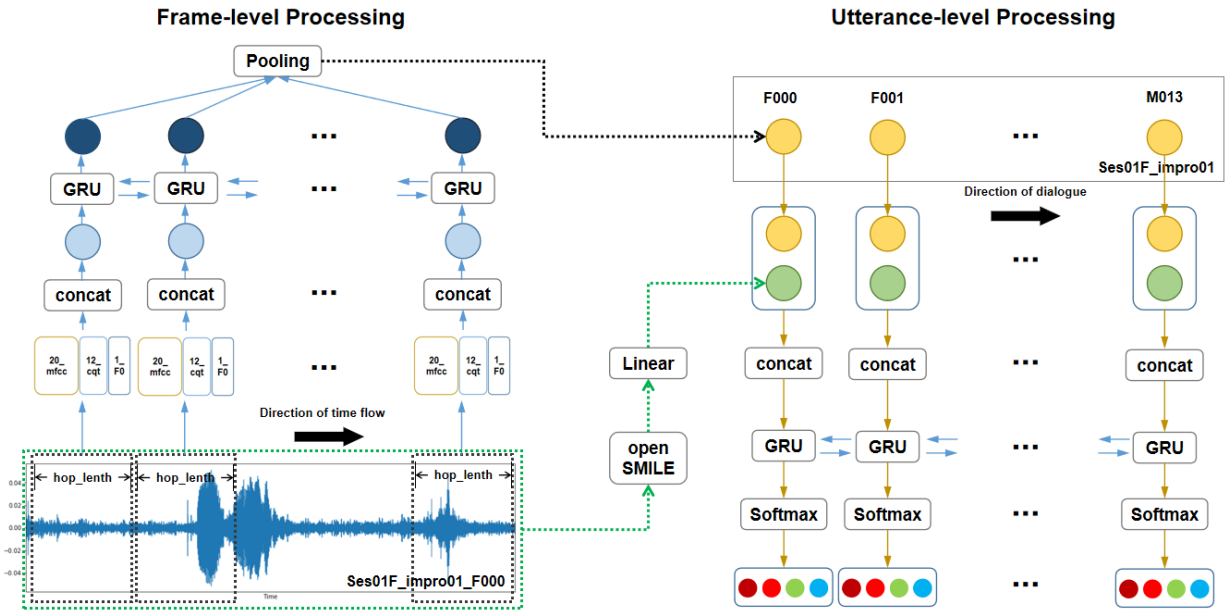


Fig. 1. Overall architecture of the proposed HGFM framework. The black dotted box are used to extracting features by LibROSA, and the green dotted box are used to extracting statistical features by openSMILE.

ond part, we review the related prior work of acoustic modality modeling using recurrent neural networks. The third part explains the definition of the task. The fourth part introduces our hierarchical grained and feature model in detail. The fifth part introduces the experimental steps and results. Finally, in the sixth part, we summarize our conclusions and prospects for future work.

2. RELATION TO PRIOR WORK

In this section, we mainly introduce the approaches most relevant to this paper, including the Context-Dependent method and Hierarchical GRU method. Then we describe how our approach is innovative and different from the previous methods.

Context-Dependent. Poria [9] propose an LSTM-based model that enables utterances to capture contextual information from their surroundings in the same video, thus aiding the classification process. On this basis, Majumder [10] further describes a method based on recurrent neural networks that keeps track of the individual party states throughout the conversation and uses this information for emotion classification. Integrate ideas from the field of conversation into the field of emotion recognition.

Hierarchical GRU. Jiao [11] propose a hierarchical Gated Recurrent Unit (HiGRU) framework with a lower level GRU to model the word-level inputs and an upper-level GRU to capture the contexts of utterance-level embeddings. Inspired by HiGRU, we use a fixed frame-level window to replace word-level inputs in a textual modality. Additional statistical acoustic features are introduced into the U-GRU of the model. It also refers to the progress of context-dependent in emotional recognition tasks. Context features are also fully considered in our model.

We summarize our contributions as follows:

- According to the non-structural acoustic data, we innovatively modeled the frame-level with fixed frame window size as the smallest unit. At the same time, consider a utterance before and after frame information and dialogue context information. Make the model more robust.
- We fusion different acoustic features at different stages of the model. This allows us to retain more valid information in the original audio data through feature engineering.
- We conduct extensive experiments on two dialogue emotion datasets, IEMOCAP, and MELD. The results demonstrate that our proposed HGFM models achieve state-of-the-art methods on each dataset, respectively.

3. PROBLEM DEFINITION

Given a set of dialogues D, $D = [d_1, d_2, \cdots, d_L]$, where L is the number of dialogues. $d_i = [u_{i,1}, u_{i,2}, \cdots, u_{i,N_i}]$, where N_i is the number of utterances for each dialogue. In each utterance, $u_{i,j} = \left\{\left(A_j^O, A_j^L, E_j\right)\right\}_{j=1}^{N_i}$, where $E_j$is the certain

emotion for each utterance, such as anger, happy, sadness, and neutral. A_j^O represents the 1582-dimensional features extracted by openSMILE. $A_j^L \in R^{n \times m}$ represents the features by LibROSA, which n is time dimension, m is feature dimension.

We aim to infer emotions from different granular structures (frame-level and utterance-level) and features (A_j^O and A_j^L).

$$f_{utt}\left(A_j^L\right) \Rightarrow A_j^{utt} \tag{1}$$

$$f_{emotion}(A_j^{utt}, A_j^O) \Rightarrow E_j^{pred} \tag{2}$$

4. METHODOLGY

In this section, we present the general framework of hierarchical grained and feature for the classification of acoustic emotions. We make further improvements based on the hierarchical GRU network [11]. Our general framework is shown in Figure 1, We take *Ses01F_impro01* in IEMOCAP dataset as the sample example. Its main improvement consists of two aspects: (1) Frame-level module: extract acoustic frame-level features, and learn before and after frame information in a utterance through bi-directional GRU model; (2) Utterance-level module: Using frame-level structure outputs fused with statistical features through a BiGRU network learning the final utterance representation containing context information. More details on these two aspects are presented in the following sub-sections.

We utilize Librosa [8] toolkit to extract 33-dimensional framelevel acoustic features (A_j^L), which including 20-dimensional Mel-frequency cepstral coefcients (MFCCs), 1-dimensional logarithmic fundamental frequency (log F0) and 12-dimensional constant-Q transform (CQT) features of the original input acoustic data in time series. Using openSMILE [3] toolkit to extract 1,582 statistic acoustic features (A_j^O).

In the frame-level stage, we utilized bidirectional GRU network [12] to extract feature vectors containing frame-level information. For each A_j^L, $A_j^L = [f_1, f_2, \cdots, f_{M_k}]$, where M_k is the number of frame window (*hop_lenth*) for each A_j^L. With A_j^L as input, learn frame-level embedding in both directions:

$$\overrightarrow{h_k} = \mathrm{GRU}(A_j^L, \overrightarrow{h_{k-1}}) \tag{3}$$

$$\overleftarrow{h_k} = \mathrm{GRU}(A_j^L, \overleftarrow{h_{k+1}}) \tag{4}$$

Based on the work by Jiao [11]. Calculate self-attention of hidden state in each direction. And then concatenate frame-level embedding f_{emb}, h_k^r, $\overrightarrow{h_k}$, h_k^l and $\overleftarrow{h_k}$. Where $f_{emb} \in A_j^L$. Utterance-level embedding u_{emb} is obtained by max-pooling on the contextual frame embeddings. Where $\otimes$ represents the tensor product.

Table 1. Statistics of utterances for IEMOCAP and MELD datasets.

	Emotion						
Dataset	Happy/Joy	Anger	Sadness	Neutral	Surprise	Fear	Disgust
IEMCOAP	1636	1103	1084	1708	-	-	-
MELD	2308	1607	1002	6436	1636	358	361

$$h_k^r = \mathrm{Softmax}(\overrightarrow{h_k} \otimes \overrightarrow{h_k}^T) \otimes \overrightarrow{h_k} \tag{5}$$

$$h_k^l = \mathrm{Softmax}(\overleftarrow{h_k} \otimes \overleftarrow{h_k}^T) \otimes \overleftarrow{h_k} \tag{6}$$

$$u_{emb} = \mathrm{maxpool}(\mathrm{concat}[f_{emb}, h_k^r, \overrightarrow{h_k}, h_k^l, \overleftarrow{h_k}]) \tag{7}$$

We utilize a fully connected layer with the $Tanh$ activation function to control the Statistical feature dimension. Where $u_O \in A_j^O$. We set the number of hidden states of the GRU to be the same as the number of hidden layer neurons in linear transformation. Where u_{emb} and $u_{op} \in \mathbb{R}^{1 \times d}$ have the same dimensions d for each.

$$u_{op} = \mathrm{Tanh}(W_w \cdot u_O + b_w) \tag{8}$$

In the utterance-level stage, We concatenate u_{emb} and u_{op}. Then through BiGRU to learning context information in a dialog.

$$A_j = \mathrm{GRU}(\mathrm{concat}[u_{emb}, u_{op}], (\overrightarrow{h_{k-1}}, \overleftarrow{h_{k+1}})) \tag{9}$$

The Softmax activation function was used to convert this set of the real vector into probability. Finally, employ Cross-Entropy loss as the objective function. We optimize the framework through the objective function as following.

$$E_j^{predA} = \mathrm{Softmax}(W_{ER} \cdot A_j + b_{ER}) \tag{10}$$

$$loss = -\sum_k y_k \log E_j^{predA} \tag{11}$$

5. EXPERIMENTS

The experiments are carried out on two dialogue emotion datasets. Table 1 presents the detail utterance numbers of each dataset. Due to the unbalanced distribution of the MELD dataset, we set weights for each emotion category during the experiment.

5.1. Datasets

IEMOCAP. The IEMOCAP [13] contains following labels: anger, happiness, sadness, neutral, excitement, frustration, fear, surprise and others. To compare with the state of the art as mentioned. Following the work by Zhou [2]: Merging the happiness and excitement categories as the happy category. So we take the four emotions containing happy, angry,

Table 2. The overall performance on IEMOCAP and MELD datasets comparison with the state-of-the-art. The underlined results are derived by us accordingly. Among them, IEMOCAP is a 4-way classification and MELD is a 7-way classification.

Method	IEMOCAP		MELD	
	WA	UWA	WA	UWA
RNN(2017 ICASSP)	63.5	58.8	38.4	20.6
bcLSTM(2017 ACL)	57.1	58.1	39.1	17.2
MDNN(2018 AAAI)	61.8	62.7	34.0	16.9
DialogueRNN(2019 AAAI)	65.8	66.1	41.8	**22.7**
HGFM*(Our Method)	62.6	68.2	41.4	19.9
HGFM(Our Method)	**66.6**	**70.5**	**42.3**	20.3

Table 3. Performance of each emotional category. The underlined results are derived by us accordingly.

Method	Angry	Happy	Sadness	Neutral
bcLSTM(2017 ACL)	58.37	60.45	61.35	52.31
DialogueRNN(2019 AAAI)	88.24	51.69	84.90	47.40
HGFM*(Our Method)	87.98	38.53	75.80	70.54
HGFM(Our Method)	87.84	54.37	72.51	67.36

sadness and neutral. **MELD.** The Multimodal EmotionLines Dataset (MELD) [14] contains about 13,000 utterances from 1,433 dialogues from the TV-series *Friends*. We completely referred to the baseline experimental setup of the paper [14].

5.2. Compared Baselines

We compare the various aspects of the proposed HGFM approach with some state-of-the-art baselines:

RNN [15]: Using a deep recurrent neural network(RNN) and a local attention base feature pooling strategy;

bcLSTM [9]: a bidirectional contextual LSTM with multimodal features extracted by CNNs;

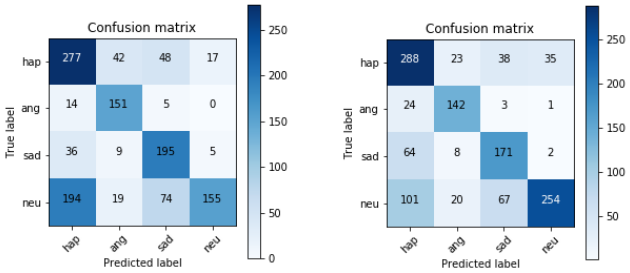
MDNN [2]: a semi-supervised multi-path generative neural network with acoustic features extracted by openSMILE;

DialogueRNN [10]: a recurrent neural network that keeps track of the individual party states throughout the conversation.

5.3. Experimental Results

We evaluate the effectiveness of our proposed hierarchical grained and feature model from two aspects: (1) Performance of Overall accuracy (The weighted accuracy, unweighted accuracy [16] and F1-score). (2) Performance of each emotional category on IEMOCAP dataset. Tabel 2 and 3 report the results. The HGFM* represents only utilize A_j^L feature to predict. All results adopt an average of 10 trials.

We compared four baselines on IEOMCAP and MELD datasets. As presented in Table 2, our proposed HGFM model outperforms the state-of-the-art methods on three evaluation



(a) Single feature (b) Fusion feature

Fig. 2. The confusion matrix of different hierarchical features in emotional categories.

metrics. The UWA is significantly improved on the IEMOCAP dataset. Achieve the 4.4% improvement. Based on the above experimental results, we analyze the performance of the model from two aspects: **(1)** Through hierarchical grained design, our model can learn more effectively the features of emotion recognition in acoustic data. The improved accuracy of experimental results effectively verified this point. **(2)** As expected, the performance of combined features was better than the single features from the experimental results shown in Table 3. Although in terms of the performance of various emotions, only neutral emotions in our model achieved the optimal performance. But through fusion the hierarchical features. Our prediction accuracy in each emotion becomes more balanced, which is also a key to the overall performance improvement. This can be more visually observed in Figure 2. We believe that acoustic data are more subjective, that people can use a relatively peaceful sound to express happy. Our method has some effect in capturing these more implicit informations.

6. CONCLUSION

In this paper, hierarchical grained and feature model is set to solve some defects of acoustic modality emotion recognition. A large number of experiments demonstrate the effectiveness of the new task-setting method for emotion recognition. Hierarchical granular structures help us capture more subtle clues, and hierarchical feature helps us obtain more complete representation from the original acoustic data.

7. ACKNOWLEDGEMENT

This paper is founded by National Natural Science Foundation of China (Grant No: 61673235) and National Key R&D Program Projects of China (Grant No: 2018YFC1707600).

8. REFERENCES

[1] Sidney K D'mello and Jacqueline Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Computing Surveys (CSUR)*, vol. 47, no. 3, pp. 43, 2015.

[2] Suping Zhou, Jia Jia, Qi Wang, Yufei Dong, Yufeng Yin, and Kehua Lei, "Inferring emotion from conversational voice data: A semi-supervised multi-path generative neural network approach," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[3] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[4] Gustavo Aguilar, Viktor Rozgić, Weiran Wang, and Chao Wang, "Multimodal and multi-view models for emotion recognition," *arXiv preprint arXiv:1906.10198*, 2019.

[5] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 2122–2132.

[6] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al., "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[7] Runnan Li, Zhiyong Wu, Jia Jia, Jingbei Li, Wei Chen, and Helen Meng, "Inferring user emotive state changes in realistic human-computer conversational dialogs," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 136–144.

[8] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, vol. 8.

[9] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 873–883.

[10] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 6818–6825.

[11] Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R Lyu, "Higru: Hierarchical gated recurrent units for utterance-level emotion recognition," *arXiv preprint arXiv:1904.04446*, 2019.

[12] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[13] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[14] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[15] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2227–2231.

[16] Viktor Rozgić, Sankaranarayanan Ananthakrishnan, Shirin Saleem, Rohit Kumar, and Rohit Prasad, "Ensemble of svm trees for multimodal emotion recognition," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2012, pp. 1–4.