A FEATURE FUSION METHOD BASED ON EXTREME LEARNING MACHINE FOR SPEECH EMOTION RECOGNITION

Lili Guo[1], Longbiao Wang[1], Jianwu Dang[1, 2], Linjuan Zhang[1], Haotian Guan[1,3]

[1]*Tianjin key Laboratory of Cognitive Computing and Application, Tianjin University, Tianjin, China*

[2] *Japan Advanced Institute of Science and Technology, Ishikawa, Japan*

[3] *Intelligent Spoken Language Technology (Tianjin) Co. Ltd., Tianjin, China*

Abstract

Background: The main flow of current studies utilized convolutional neural network (CNN) directly on spectrograms to extract features, and employed the state-of-the-art models such as the bidirectional long short term memory (BLSTM).

Problems: ① those features did not fully utilize priori knowledge;
② BLSTM is not efficient enough for training small-scale datasets such as the emotional datasets.

Solutions: ① propose a feature fusion method to combine CNN-based features and heuristic-based discriminative features;
② utilize extreme learning machine (ELM) instead of BLSTM to solve the second problem.

Results: our method leads to 40% relative error reduction in F1-score compared to CNN-BLSTM on EmoDB.

Methods

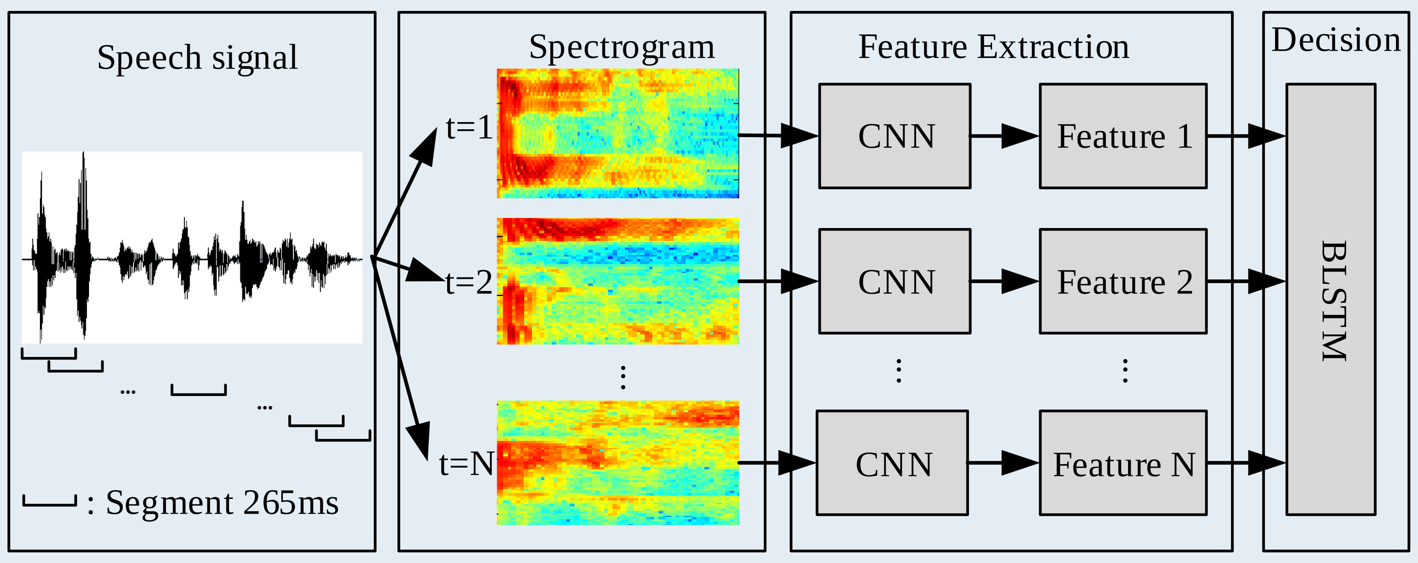

Fig. 1. Baseline: CNN-BLSTM

Problems:

(a) Features: it does not utilize knowledge-based heuristic features (such as MFCC, pitch, energy, etc.);

(b) Models: the framework of BLSTM is complicated, and it needs lots of training data.

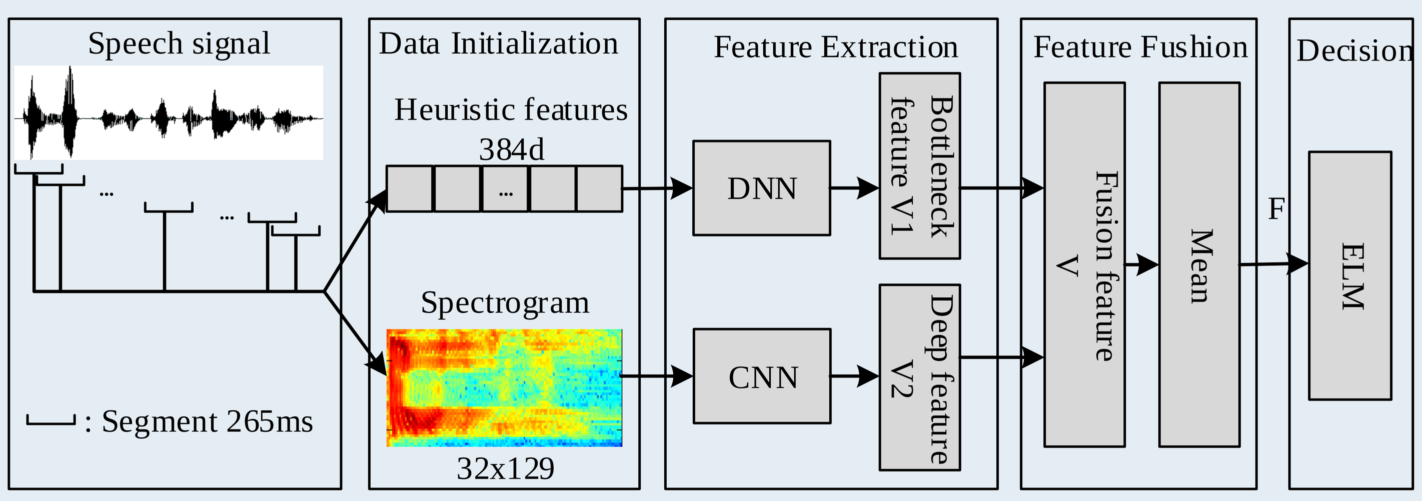

Fig. 2. Our method: Feature Fusion Method based on ELM

Solutions:

(a) propose a feature fusion method that combines CNN-based features and heuristic-based features;

(b) use ELM instead of BLSTM to distinguish emotions.

Experimental Setup

Fig. 3. Emotion distribution.

- **Dataset**: EmoDB consisting of 535 utterances.
- **The structure of CNN**
  Convolutional layer 1: 32@5×5
  Convolutional layer 2: 64@5×5
  Two pooling layers: 2×2
  Full connected layer: 1024 units
  Dropout layer: 0.5 factor.

Validation of Bottleneck Features

Tab. 1. F1 (%) comparison of bottleneck features and heuristic features.

Emo	Heuristic F.	Bottleneck F.	Change
Fea	67.74	66.67	-1.07
Dis	79.07	80.43	+1.36
Hap	60.94	68.66	+7.72
Bor	73.94	76.02	+2.08
Neu	69.82	83.87	+14.05
Sad	84.03	82.26	-1.77
Ang	80.29	85.28	+4.99
Ave	73.69	77.60	**+3.91**

- **Method**: ELM
- There are great improvements when using bottleneck features.
- It is necessity to extract bottleneck features.

Results

Tab. 2. Comparison of different speech emotion recognition models

Model	P (%)	R (%)	F1 (%)
DNN-ELM	85.55	84.09	84.56
CNN-BLSTM	89.41	86.66	87.49
CNN-BLSTM (+ heuristic features)	90.22	89.73	89.68
CNN-ELM	92.64	90.83	91.47
CNN-ELM (+ heuristic features)	**93.30**	**91.97**	**92.50**

- CNN-ELM performs better than CNN-BLSTM in this task.
- CNN-BLSTM(+heuristic features) performs better than CNN-BLSTM alone.
- Our method outperforms CNN-BLSTM by 40% relative error reduction.

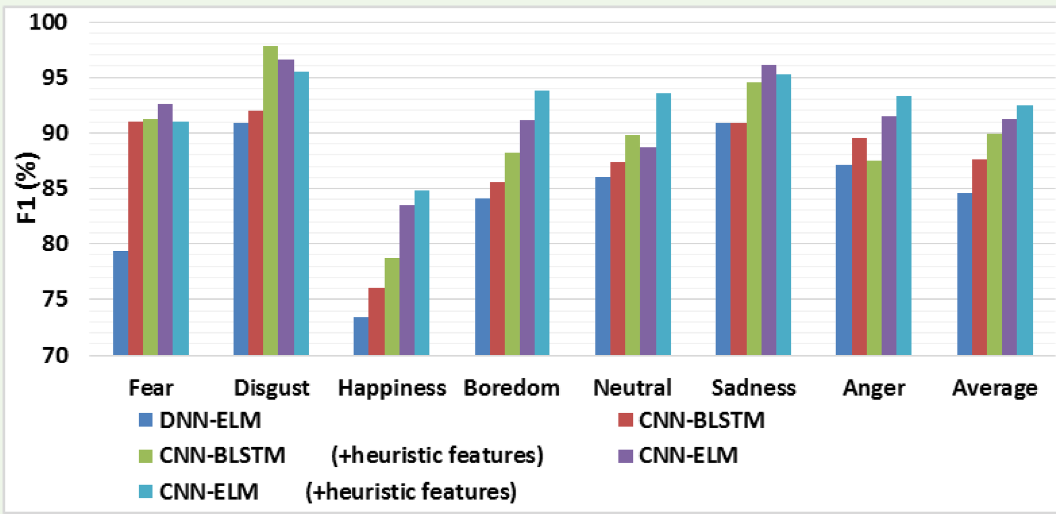

Fig. 4. F1 results for each emotion.

- CNN-ELM(+ heuristic features) achieves best performance in most cases except fear, disgust and sadness.
- The reason might be that the database has less data of disgust and sadness.

	Fea	Dis	Hap	Bor	Neu	Sad	Ang
Fea	61	0	2	0	2	2	2
Dis	0	40	1	1	1	0	3
Hap	3	0	46	0	0	0	22
Bor	0	0	0	68	7	6	0
Neul	1	0	0	7	69	2	0
Sad	0	0	0	2	0	60	0
Ang	0	1	1	0	0	0	125

(a) CNN-BLSTM

	Fea	Dis	Hap	Bor	Neu	Sad	Ang
Fea	61	0	4	0	2	0	2
Dis	1	43	0	0	0	1	1
Hap	2	0	56	0	0	0	13
Bor	0	0	0	76	2	3	0
Neul	1	0	0	4	73	1	0
Sad	0	0	0	1	0	61	0
Ang	0	1	1	0	0	0	125

(a) Our method

Fig. 5. Confusion matrices of CNN-BLSTM and our method.

- **Abscissa**: detected labels
- **Ordinate:** actual labels

Conclusions

✓ We proposed a feature fusion method with ELM, which combines CNN-based features and heuristic-based discriminative features.

✓ It is found that knowledge-based heuristic features have significant contribution although automatically extracted features were good.

✓ The ELM is suitable for small-scale database training for speech emotion recognition.

Future works:

- Taking experiments on a large-scale dataset.
- Taking strict selection about heuristic features.

Acknowledgements

The research was supported by the National Natural Science Foundation of China (No. 61771333 and No. U1736219), JSPS KAKENHI Grant (16K00297) and Didi Chuxing.