Learning Adolescent Mental Health from Speech Data

Yuqi Wang and Shurui Bai

iBHE, Tsinghua University, China

ABSTRACT

- We proposed a model in adolescent mental health identification on speech data.
- Based on LSTM/GRU, this model can learn contextual features among uttrances while using the features of a single speech, so as to judge the mental health of adolescents.
- Through experiments, our model can achieve good classification results.

INTRODUCTION

- **Adolescence**: a critical period for physical and mental development.
- At present, there are **few** deep learning researches in adolescent mental health, and **very few** related studies using speech data.
- **Current** screening method: combination tests of mental health questions (low pertinence, easy to deceive).
- **Deep learning**: to extract audio data features and analyze (accurate, time-saving, meaningful).

METHOD

- **Data Preprocessing** (Figure 1, 2)
- **Feature Extraction of each Uttrance** (Figure 3)
- **Model:** Contextual-LSTM/GRU/BiLSTM/BiGRU (Figure 4)

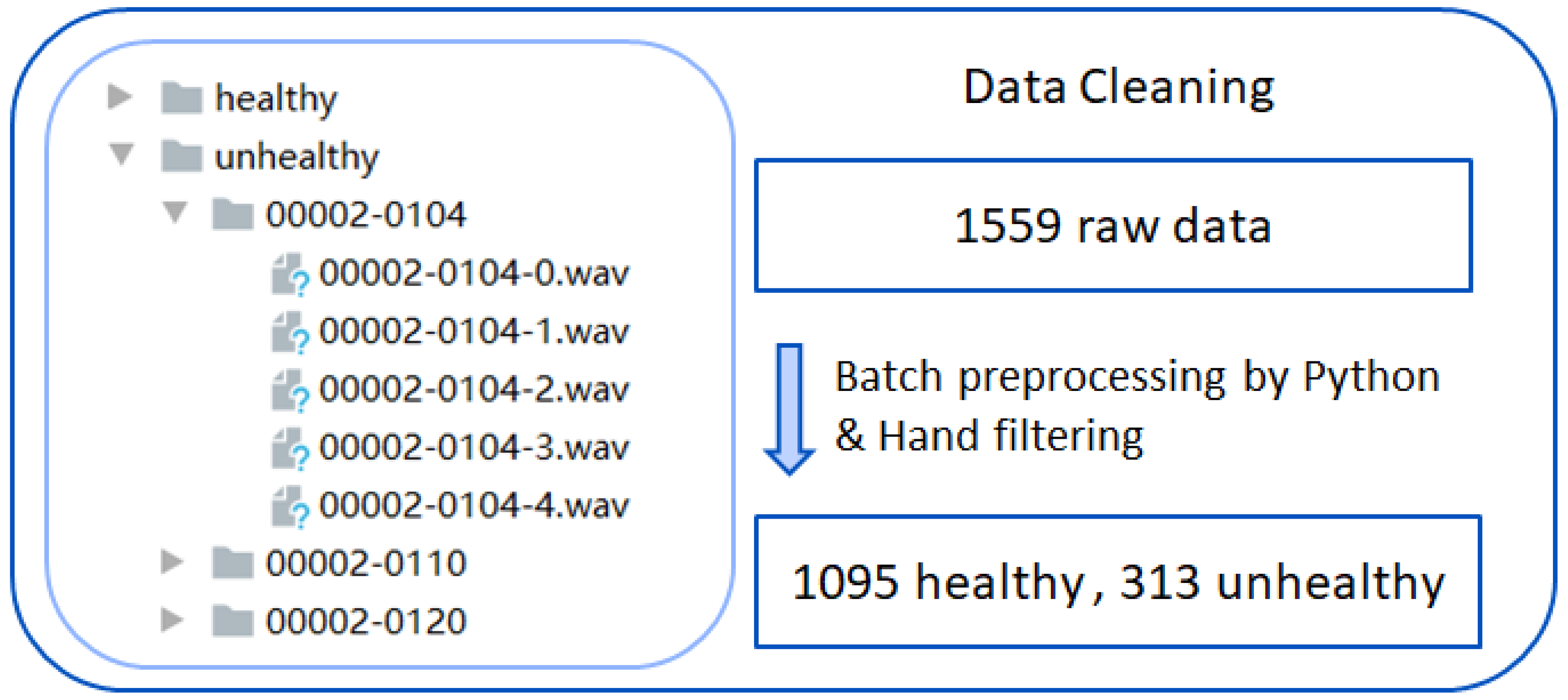
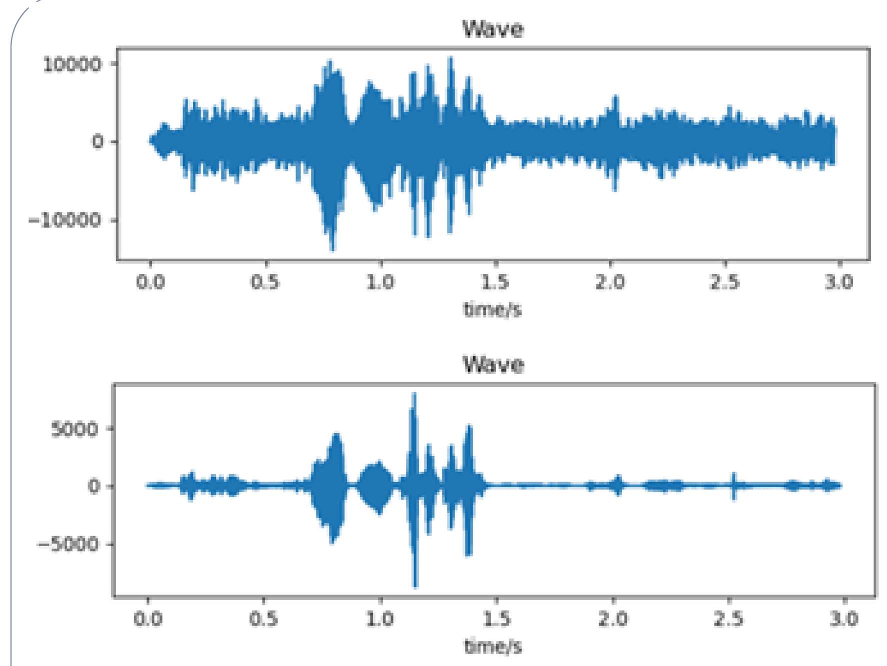


Figure 1. Dataset and Data Cleaning



$$HE = \frac{1}{2} \times \frac{1}{fn} \sum_{i=1}^{fn} E(i)$$

$$LE = \frac{1}{2}\left(\frac{1}{20}\left(\sum_{i=1}^{10} E(i) + \sum_{i=fn-9}^{fn} E(i)\right) + HE\right)$$

$$ZC = 3 \times \frac{1}{20}\left(\sum_{i=1}^{10} Z(i) + \sum_{i=fn-9}^{fn} Z(i)\right)$$

Figure 2. Noise Reduction & Endpoint Detection

a)

	C1	C2	C3	C
1	label	1	2	3
2	0	4.34E+02	5.57E+02	2.2
3	0	2.15E+02	4.14E+02	2.9
4	0	1.04E+02	3.95E+02	1.5
5	0	9.30E+01	1.56E+02	2.4
6	0	9.40E+01	2.18E+02	1.6
7	0	1.34E+02	5.70E+01	3.4
8	0	3.68E+02	2.06E+02	4.1
9	0	3.85E+02	2.32E+02	2.7

b)

C2	C3
1	2
1.873968229	1.449084597
0.20567944	0.457605662
-0.63989159	0.325870699
-0.723686918	-1.331216472
-0.716069161	-0.901344486
-1.423217022	-0.41789907
0.955588572	1.06464763
1.128407782	1.323347054

c)

Figure 3. Feature Extraction of a Single Speech (a) Software tool: openSMILE3.0; b) Extracted IS10 speech features; c) Features after Z-score standardization)

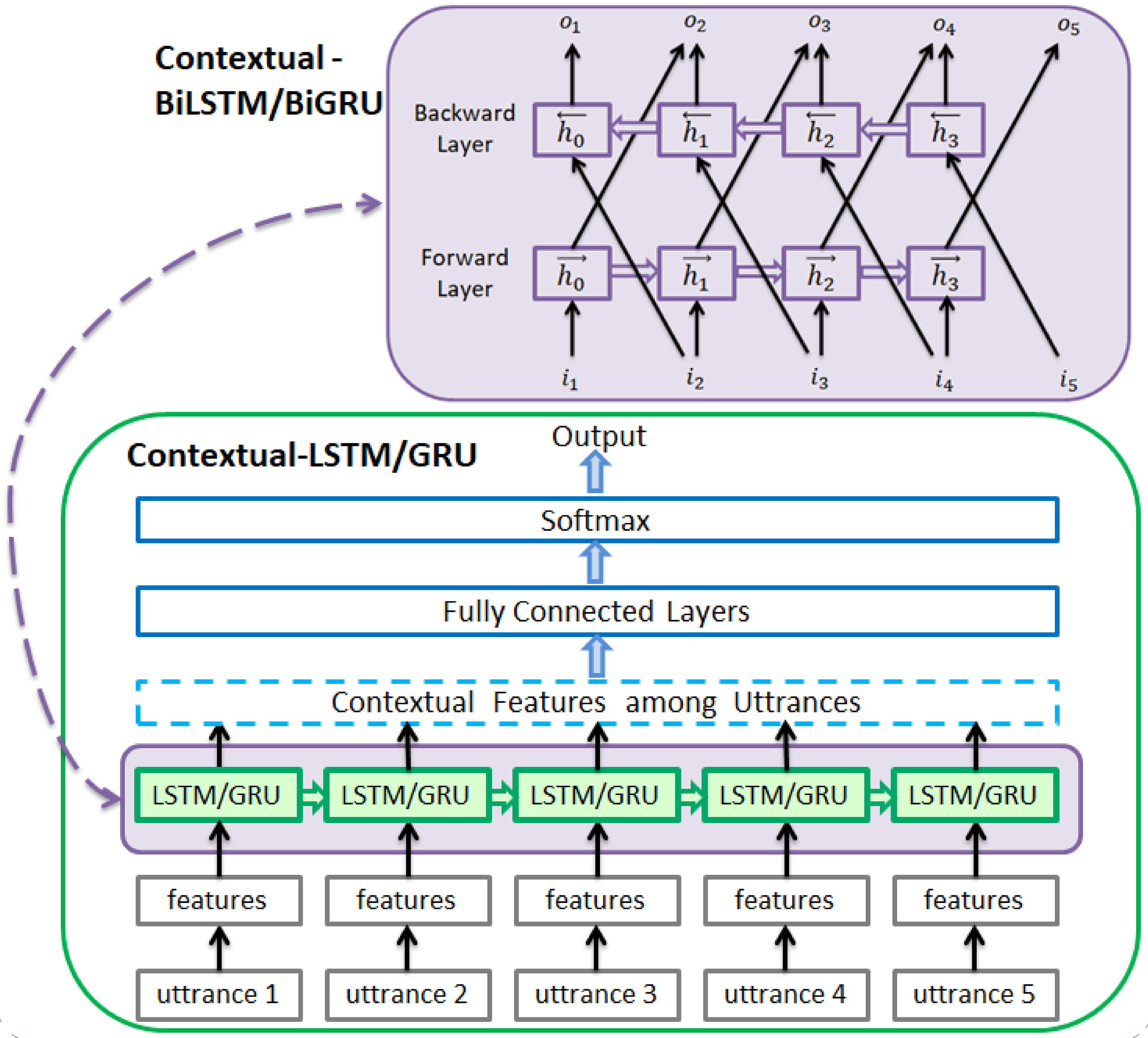


Figure 4. Model: Contextual-LSTM/GRU/BiLSTM/BiGRU

RESULTS

Table 1. Verification of the necessity of data preprocessing and standardization

Data Processing	Standardization	Acc
Raw Data (No Processing)	N	0.6104
	Y	0.6427
Noise Reduction	N	0.6526
	Y	0.7866
Noise Reduction + Endpoint Detection and Separation	N	0.6948
	Y	**0.8313**

Table 2. Experimental results of different models

Model	Feature	Acc	F1_score	Precision	Recall
C-LSTM	IS09	0.8313	0.8384	0.8249	0.8524
	IS10	0.8511	0.8497	0.8682	0.8319
C-GRU	IS09	0.8412	0.8498	0.8388	0.8619
	IS10	0.8437	0.8431	0.8532	0.8333
C-BiLSTM	IS09	0.8462	0.8537	0.8570	0.8504
	IS10	**0.8660**	0.8672	0.8758	0.8590
C-BiGRU	IS09	0.8437	0.8558	0.8498	0.8619
	IS10	0.8561	0.8547	0.8814	0.8295

CONCLUSIONS

- We classify the **speech data** of teenagers, judge their **mental health status**, assist intelligent medical treatment.
- In this task, **bidirectional** contextual models and **LSTM** based models perform slightly better.
- It is a desirable way to make use of **contextual features** among uttrances while using the **acoustic features** of each uttrance.
- **Future**: **multi-class** classification (depression, anxiety, etc.); **multimodal** data fusion (with video, text, etc.).

REFERENCES

[1] Poria, Soujanya, et al. "Context-dependent sentiment analysis in user-generated videos." *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*. 2017.

[2] Asokan, Ashish Ramayee, et al. "Interpretability for multimodal emotion recognition using concept activation vectors." *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022.