Three-Dimensional Lip Motion Network for Text-Independent Speaker Recognition

Jianrong Wang[1], Tong Wu[1], Shanyu Wang[1], Mei Yu[1], Qiang Fang[2], Ju Zhang[1,] and Li Liu[3*]
[1]College of Intelligence and Computing, Tianjin University, China.
[2]Institute of Linguistics, Chinese Academy of Social Sciences, China.
[3]Shenzhen Research Institute of Big Data, Shenzhen, China. E-mail: liuli@cuhk.edu.cn.

1. INTRODUCTION

Motivation

★ Lip motion can be used as a new kind of biometrics in speaker recognition.

★ Lots of works used 2D lip images to recognize speaker in a text-dependent context.

★ However, 2D lip easily suffers from face orientations.

Main Contribution

★ This is the first work that uses the sentence-level 3D lip motion (S3DLM) sequences in the speaker recognition task.

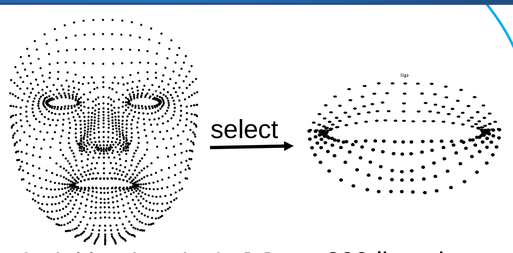
★ We present a novel 3D lip motion Network (3LMNet) based on S3DLM sequences by proposing the RFM and incorporating the priorknowledge of speaker's lip motion.

2. METHOD

S3DLM

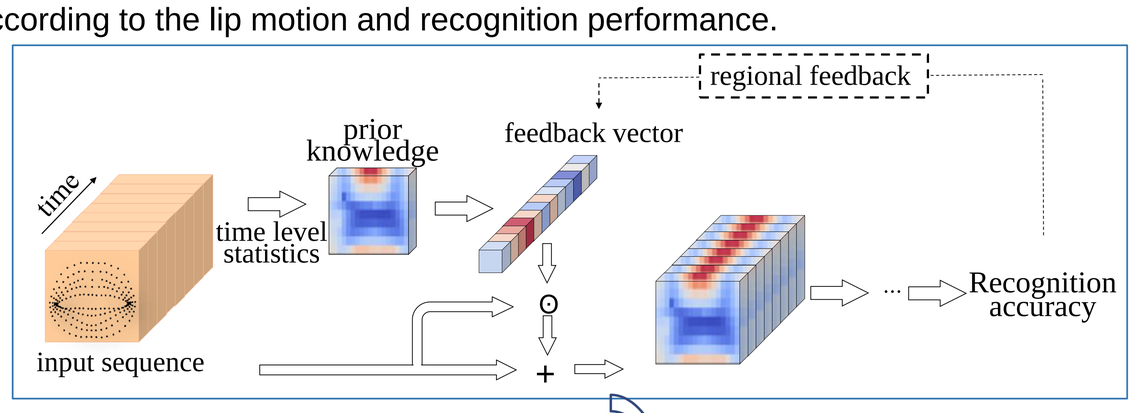
★ 200 lip points selected from the 1347 facial landmarks.

★ 28 frames in each sentence represent the motion of lip.



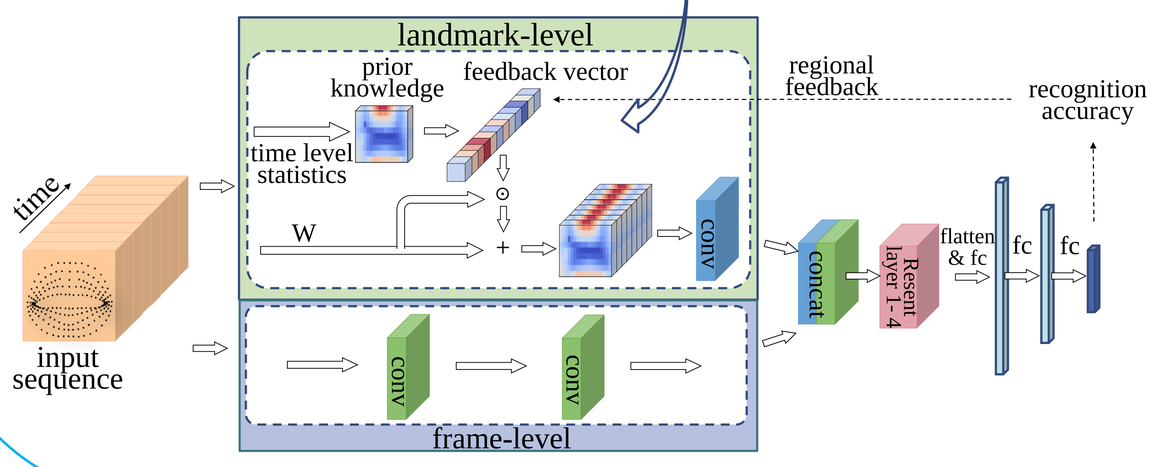
RFM & prior knowledge

★ RFM and prior knowledge of the lip motion is proposed to screen out key identifying information in lip motion. The RFM adjusts the feedback vector according to the lip motion and recognition performance.



3LMNet

★ The network learns landmark-level features, frame-level features of S3DLM sequences.



3. DATASET

Data

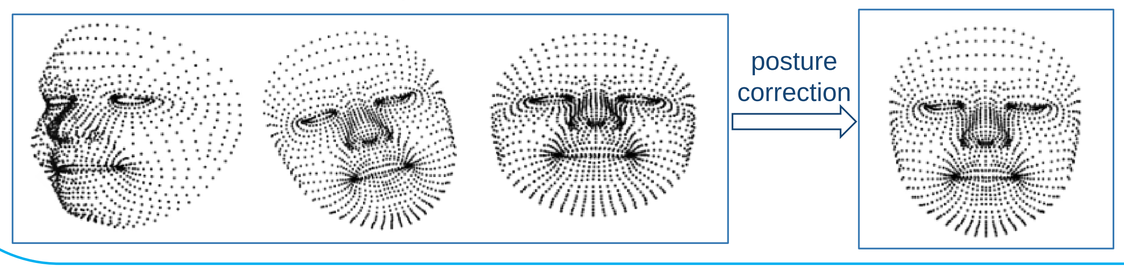
★ A large-scale depth-based multimodal audio-visual mandarin dataset, including 3D face point cloud composed of 1347 points, and 200 lip points were selected.

★ 69 speakers participated in the recording, and each speaker uttered 146 identical sentences.

★ 120 sentences of the corpus are used as the training set, and the remaining 26 sentences are used as the test set.

Data Preprocessing

★ Coordinate transformation; ★ Posture correction



3. EXPERIMENT

- **Discussion of the Relationship between Text and Lip Motion**

COMPARISON BETWEEN TEXT-BASED LIP MOTION IN THE LSD-AV DATASET								
Text-based sample	1-20	21-40	41-60	61-80	81-100	101-120	121-140	std
D_t	0.0067	-0.0048	0.0024	0.0038	-0.0017	0.0060	0.0004	**0.0046**

COMPARISON BETWEEN SPEAKER-BASED LIP MOTION IN THE LSD-AV DATASET								
Speaker-based sample	1-10	11-20	21-30	31-40	41-50	51-60	61-70	std
D_t	-0.0385	0.0293	-0.0096	0.0738	0.0187	-0.0465	-0.0272	**0.0431**

- **Performance of the S3DLM sequences**

The text-independent speaker recognition performance of using S3DLM sequences is better than that using 2D landmark sequences in three benchmark models.

Model	S3DLM	2D landmarks
LSTM	82.46%	76.23%
VGG-16	91.00%	87.10%
ResNet-34	93.50%	88.47%

Performance of the 3LMNet & Ablations study of the proposed 3LMNET

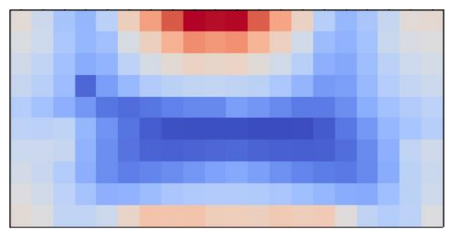
Model	Text-independent	Text-dependent
Lai et al.[2]	—	92.61%
Liao et al.[3]	—	97.11%
3LMNet-RFM-prior	93.91%	98.38%
3LMNet+RFM-prior	94.94%	98.73%
3LMNet+RFM+$prior_{opp}$	91.94%	97.73%
3LMNet+RFM+prior	95.22%	99.10%

The 3LMNet is better than these two text-dependent speaker recognition models using 2D lip images.

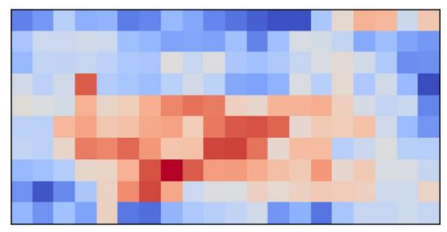
The proposed 3LMNet using both the RFM and the prior achieve the best performance.

Effect of RFM and the prior knowledge of lip motion

To further analyze the feedback effect of RFM and the prior knowledge of lip motion, we visualize the feedback information. We can see that by using RFM, large weights are assigned to the regions with small fluctuations. We conclude that strong fluctuations mainly related to the text, which interferes with personal characteristics. In contrast, slight fluctuations are less affected by the text, and thus shows more obvious personality characteristics that are beneficial for speaker recognition.



(a) Original lip points fluctuation



(b) Regional feedback visualization with prior knowledge of lip points

4. CONCLUSION

★ Three dimensional lip motion can be used for speaker recognition tasks, and achieve great performance.

★ The RFM can distinguish the contributions of different lip regions, and the prior knowledge derived from the known lip motion can further capture key lip regions.

5. REFERENCES

[1] J. Wang, L. Wang, J. Zhang, J. Wei, M. Yu, and R. Yu, "A largescale depth-based multimodal audio-visual corpus in mandarin," in 2018 IEEE 20th International Conference on High Performance Computing and Communications; 2018, pp. 881–885.

[2] J. Y. Lai, S. L. Wang, W. C. Liew, and X. J. Shi, "Visual speaker identification and authentication by joint spatiotemporal sparse coding and hierarchical pooling," Inf. Sci., vol. 373, pp. 219–232.

[3] J. Liao, S. Wang, X. Zhang, and G. Liu, "3d convolutional neural networks based speaker identification and authentication," in 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp.2042–2046.