A multi-channel/multi-speaker interactive 3D Audio-Visual Speech Corpus in Mandarin

Jun Yu[1,2,3], Rongfeng Su[1,2], Lan Wang[1,2] and Wenpeng Zhou[1,2]
[1]Key Laboratory of Human-Machine Intelligence-Synergy Systems,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]The Chinese University of Hong Kong, Hong Kong, China
[3]School of Information Science & Engineering, Lanzhou University, Lanzhou, China
18794853665@163.com, rf.su@siat.ac.cn, lan.wang@siat.ac.cn, wp.zhou@siat.ac.cn

Reporter: Rongfeng Su

CONTENT

Back Ground

Corpus Design

Recording

Data Processing

Conclusion

BACK GROUND (1/5)

- With the development of **Automatic Speech Recognition** (ASR) techniques, researchers have already made great progress in this field.

- Most of them mainly focus on either auditory or visual aspect.

- In face, combination of them can enhance the intelligibility of speech (noisy/hearing-impaired conditions)

- The first **Audio-Visual Speech Recognition** (AVSR) system: Petajan in 1984 [1]
  - Very simple task, single speaker, 100-words

- **Problem: lacking of suitable audio-visual corpora!**

BACK GROUND (2/5)

- Amounts of 2D audio-visual corpora have been designed for different tasks:
  - Digit recognition: Tulips1 [2] and CUAVE [3]
  - Isolated letter recognition: AVLetters [4]
  - Continuous speech recognition: Grid [5] and VidTIMIT [6]
  - Medium to large-vocabulary continuous speech recognition: AV-TIMIT [7], TCD-TIMIT [8] and IBM LVCSR [9] (largest, about 50 hours with 290 subjects)

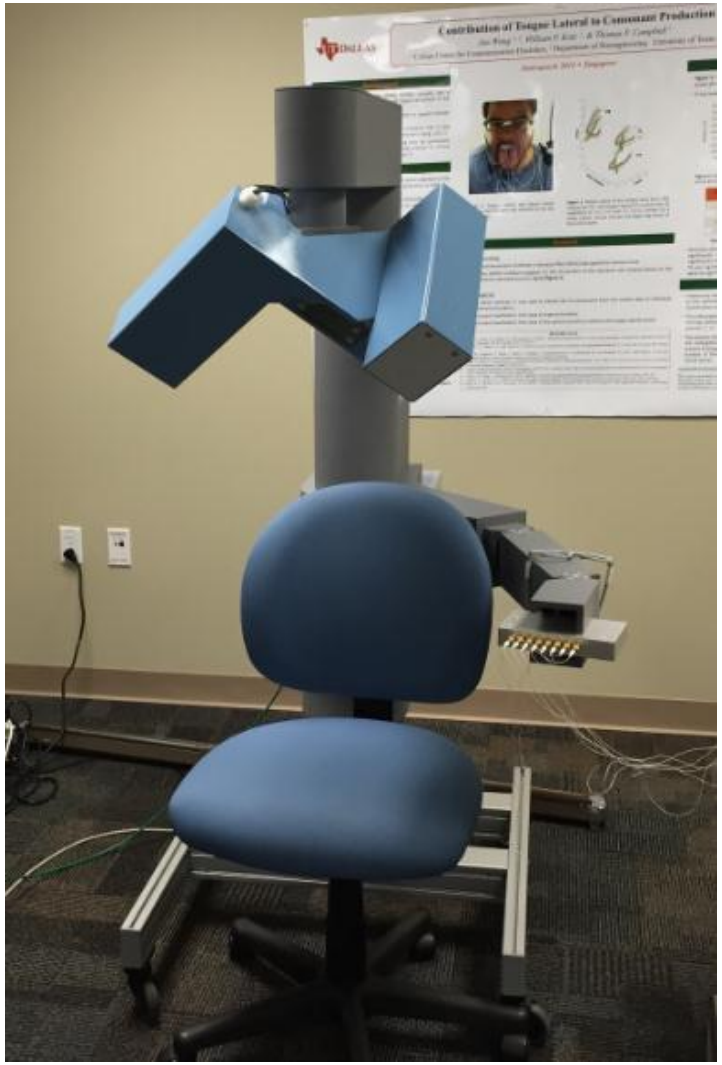
BACK GROUND (3/5)

- Comparison of 2D and 3D audio-visual corpora:
  - 3D facial features are immune to changing illuminations since they rely on the depth information embedded in them [10].
  - 3D facial motion representation is more accurate than 2D because it provides a higher discriminative power [11-12].

2D AV corpora	3D AV corpora
Low frame rate (normally 30-60 FPS)	High frame rate (up to 240 FPS)
Only using 2D visual information	Using 2D and 3D visual information
Many corpora existed	Few, especially **a gap in Mandarin**

BACK GROUND (4/5)

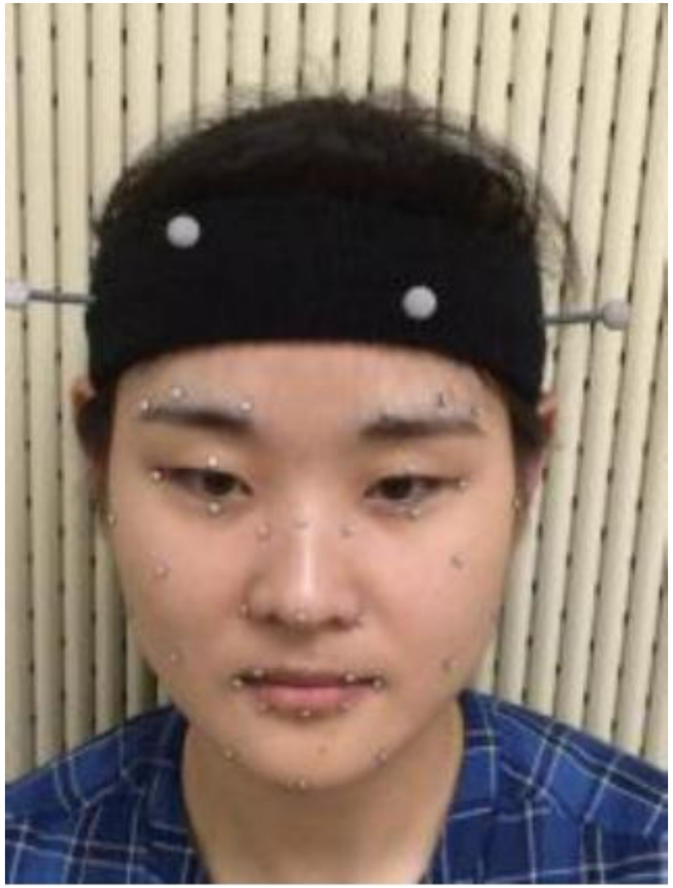
- Electromagnetic Articulography (EMA) [13] is one of the equipment to acquire high-precision 3D visual data:
  - Collect data by using sensors placed on tongue and other parts of the mouth
  - Hard to operate
  - Not allow for collecting large amounts of data



Back Ground (5/5)

- Solution: 3D facial motion capture (commercial system “OptiTrack”)
  - Very easy to operate
  - Collect data by using photogenic markers on the participants face

S250:E
832x832 | 250 FPS
100BASE-TX with PoE 802.3af
M003402
OptiTrack
© 2010 NaturalPoint Inc. Patents Pending.



CORPUS DESIGN

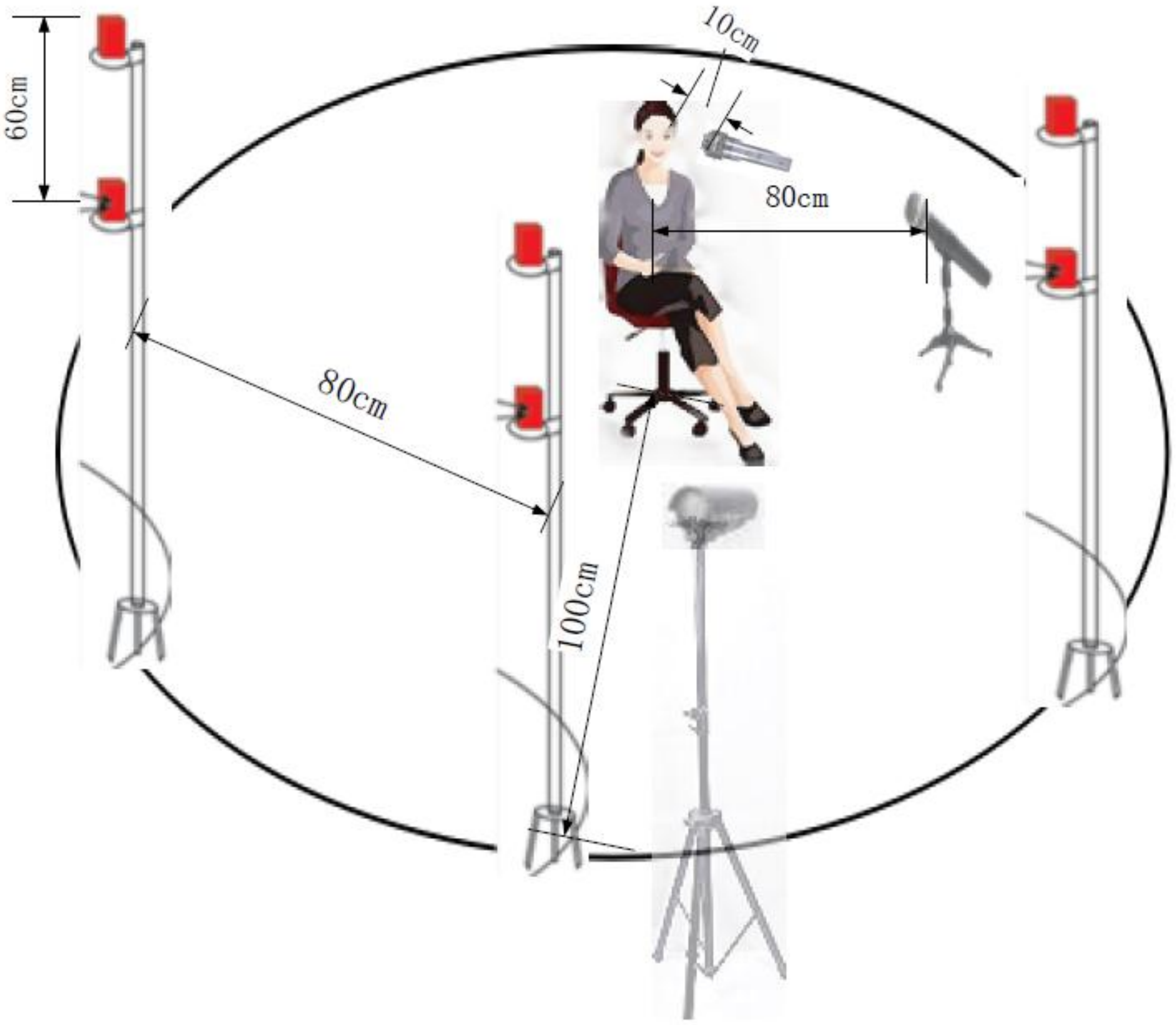
- Text corpus selection:
  - About 18k text samples from northern Chinese speech corpus (CASIA) [14]
  - Cover 116 Chinese phonemes, phonetically balanced
  - The whole set is divided into basic training set and test set (8:2) without overlap

- Subject selection:
  - 12 males and 12 females
  - All subjects from student body (age range 21-28) in northern China
  - Each subject should take a dialect-test, only typical north-accent speakers are considered
  - No speaker has no known history of speech, language or hearing impairments

RECORDING (1/4)

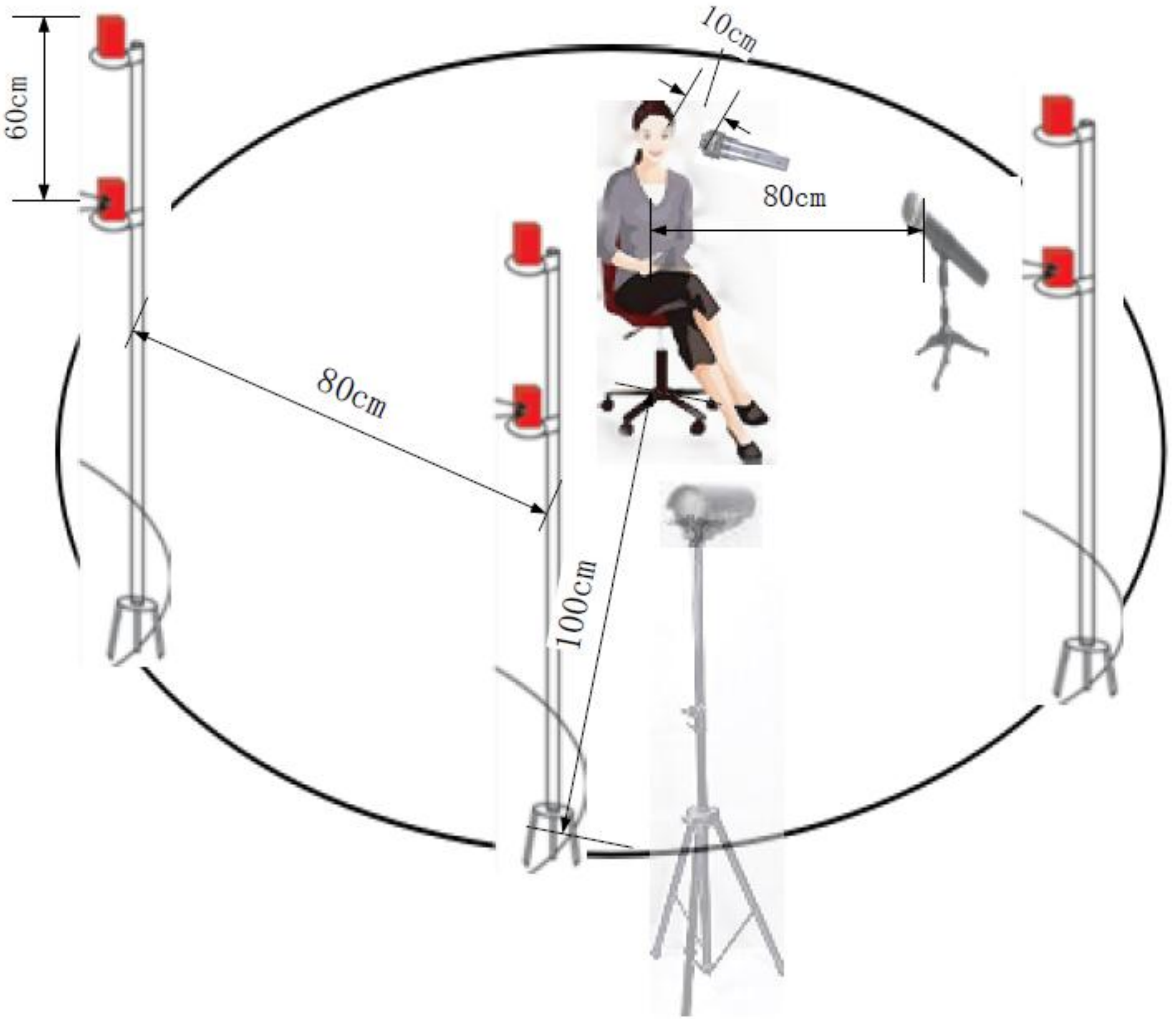
➢ Audio streams:

- Near-field and far-field
  - ◆ Near-filed: about 15cm
  - ◆ Far-filed: about 80cm
- Recording microphone:
  - ◆ Apple Blue Microphones Yeti Pro
- Sampling rate: 16kHz



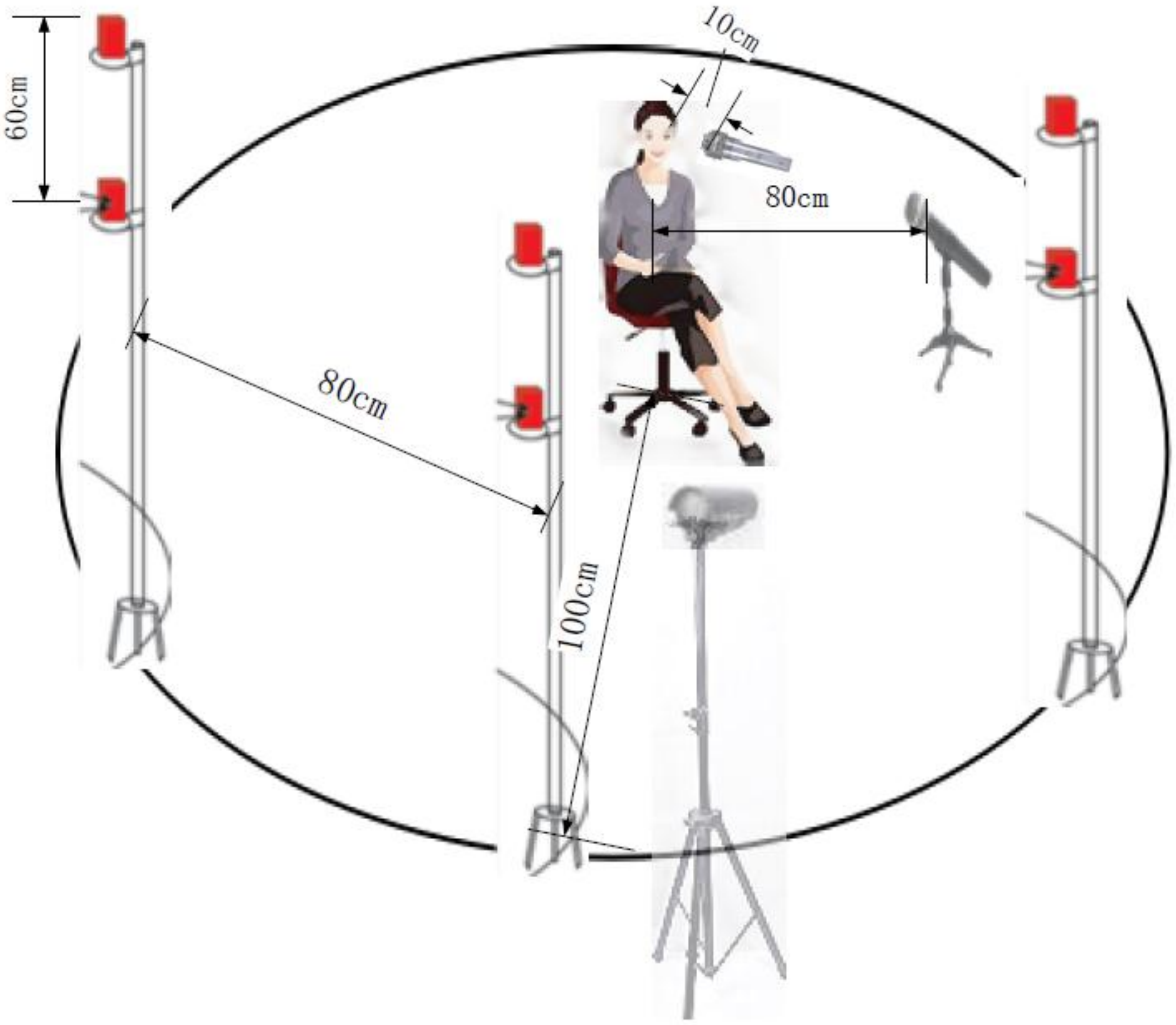
RECORDING (2/4)

- Audio streams (other requirements):
  - Speaker read each text sample loudly at a relatively slow speech rate
  - Each set (training/test set) is divided into several non-overlapping parts, and each speaker read only one part
  - Every 5 utterances are recorded in a file as one session
  - Guarantee 3-5 second silence at the beginning and at the end of each session, as well as between each utterance.



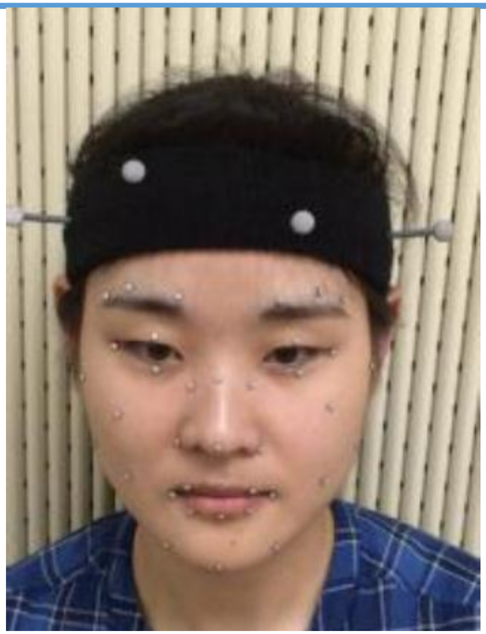
RECORDING (3/4)

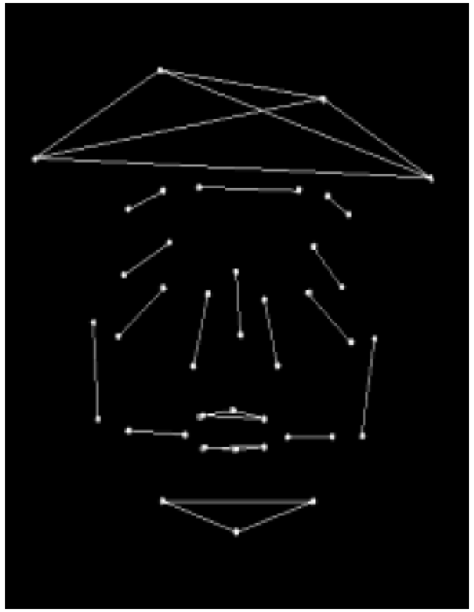
- Video streams:
  - OptiTrack: commercial 3D facial motion recording system
  - 6 infrared LED cameras
  - The distance between the sideway tripod and the middle tripod is about 80 cm
  - Two cameras are attached to each tripod by about 60 cm
  - The speaker sit in the front of the middle tripod, and the distance between them is about 1 meter
  - This guarantee the maximum overlap of the capture will be taking place



RECORDING (4/4)

- Video streams (other requirements):
  - 41 reflective markers, 4 hard headband markers and 37 facial markers
  - Recording frame rate: 100 Hz
  - Each session is started with a neutral pose, such as relaxed face, mouth closed, eyes open, looking straight ahead and so on
  - Avoid moving head violently to guarantee the facial movement no-offset
  - A simple warning tone is generated by the 3D video recording software at the beginning and at the end of each session for the synchronization between audio and video streams, the recording of 3D video data will only happen between the two signal



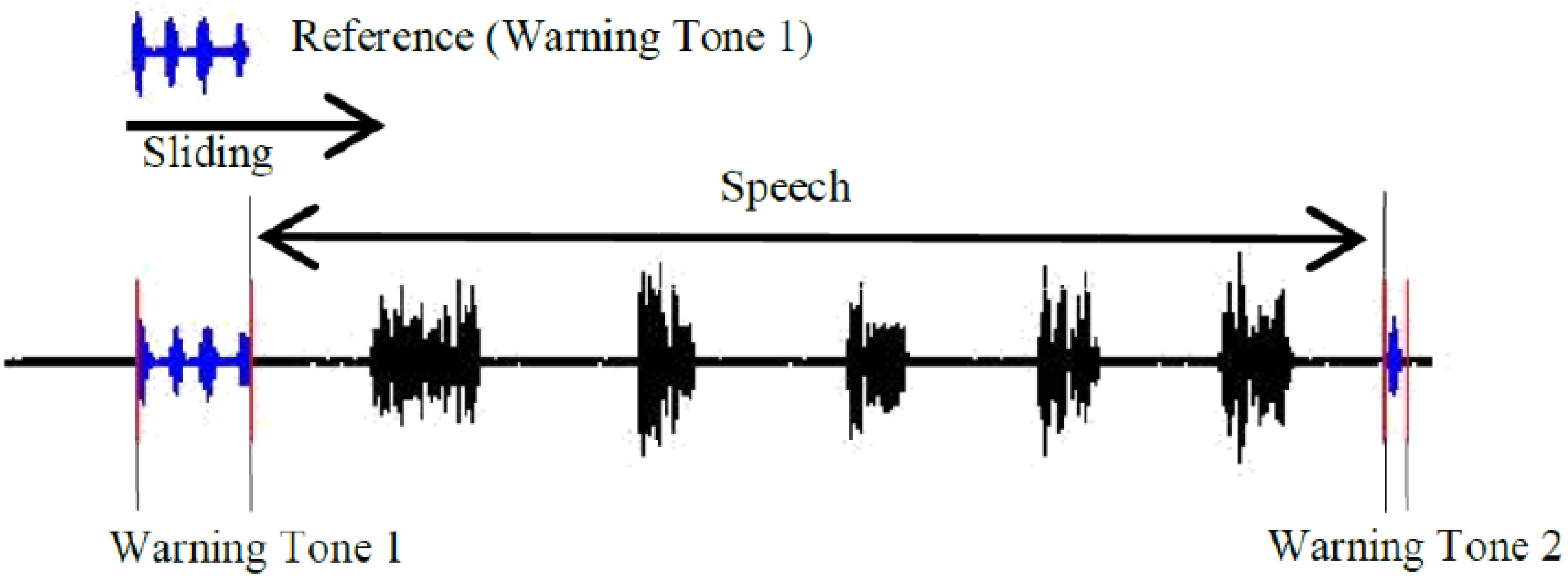


DATA PROCESSING (1/8)

- The data processing procedure contains 3 steps:
  - Semi-automatic synchronization of audio and 3D video streams
  - Detection and correction of outliers
  - Removing head motions

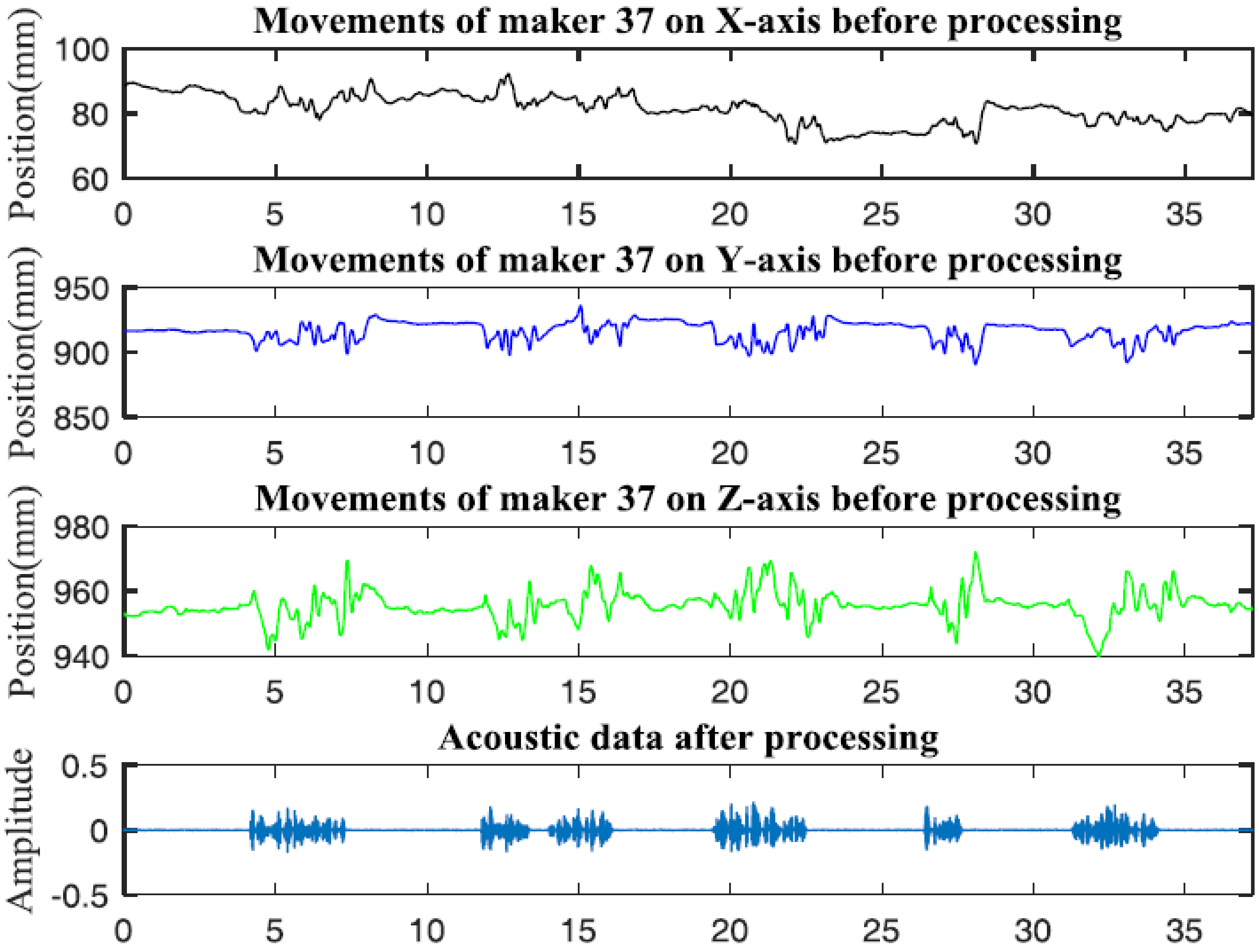
DATA PROCESSING (2/8)

- Semi-automatic synchronization of audio and 3D video streams
  1. Cut out manually the reference tone from the beginning and end of any session
  2. Slide the reference tone frame by frame
  3. Calculate the corresponding correlation coefficient value for each frame
  4. The position of the warning tone in a test audio stream could be automatically found by selecting the max correlation coefficient value



DATA PROCESSING (3/8)

- Semi-automatic synchronization of audio and 3D video streams (results)



Marker 37 is the lower lip

DATA PROCESSING (4/8)

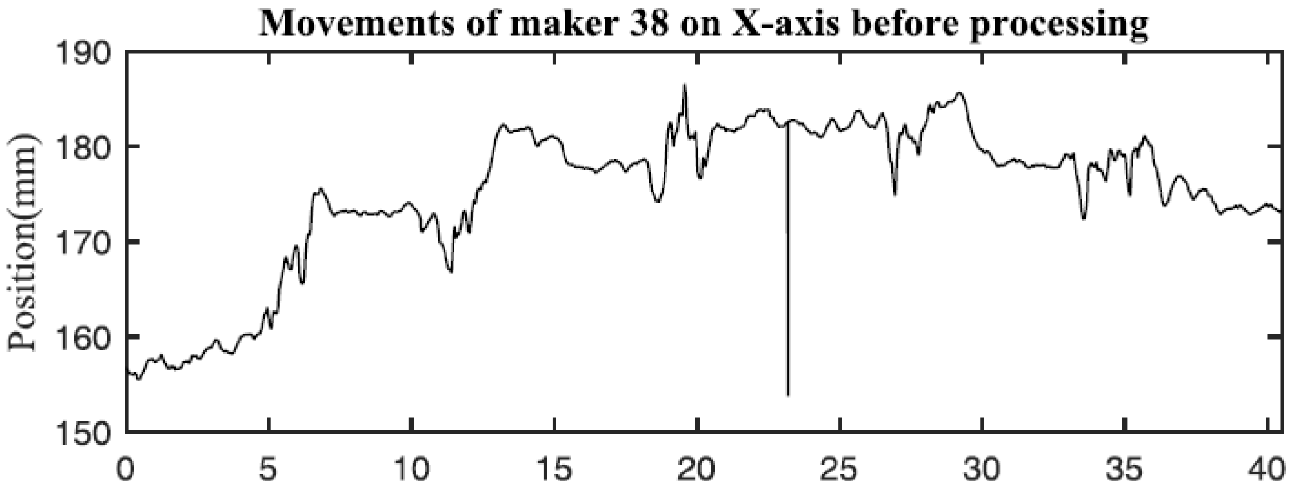
- Detection and correction of outliers
  - Based on k-nearest neighbor hood method [15], P_i is defined as a outlier, when

$$|P_i - \bar{P}| > \rho\delta$$

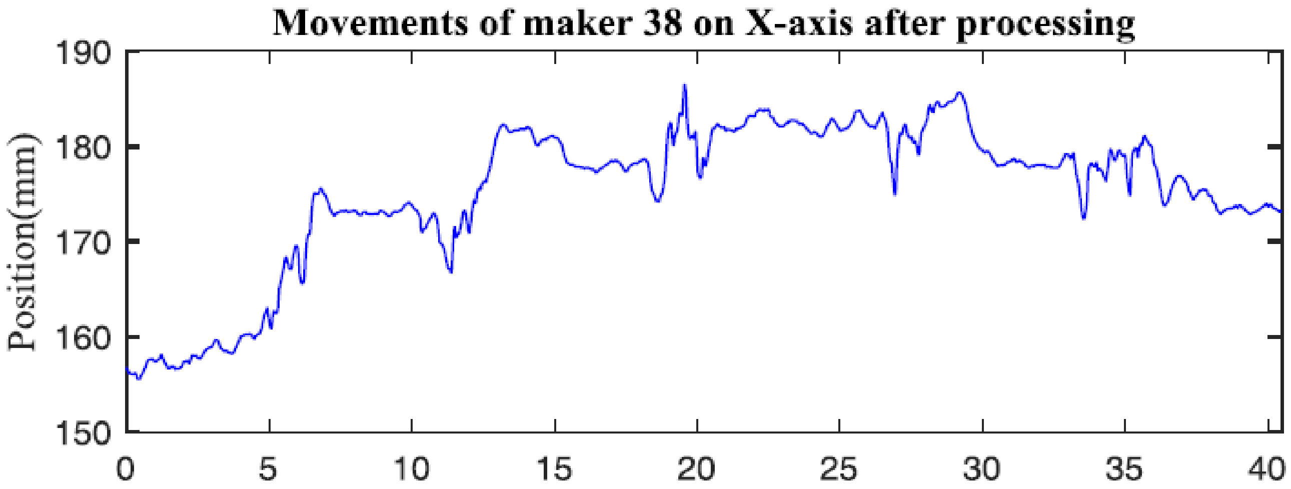
  - Once an outlier is detected, the original 3D position of the outlier will be replaced by the mean of k-nearest neighbor
  - Classes of outliers: repairable and unrepairable (massive error points occur in the time sequence, discarded in this paper)

DATA PROCESSING (5/8)

- Detection and correction of outliers (results)



Movements of maker 38 on X-axis before processing



Movements of maker 38 on X-axis after processing

Marker 38 is the left lower lip

DATA PROCESSING (6/8)

- Removing head motions:
  - Head displacements may cover some facial changes (main reason)
  - Idea: since the 4 headband markers are always separated by a fixed distances from each other, they can be used to extract global head motion from the other markers and thus make available head motion and facial deformation separately
  - The head motion can be removed by using the rotation matrix R_t and the translation matrix T_t

$$F_t = R_t F_t^{'} + T_t$$

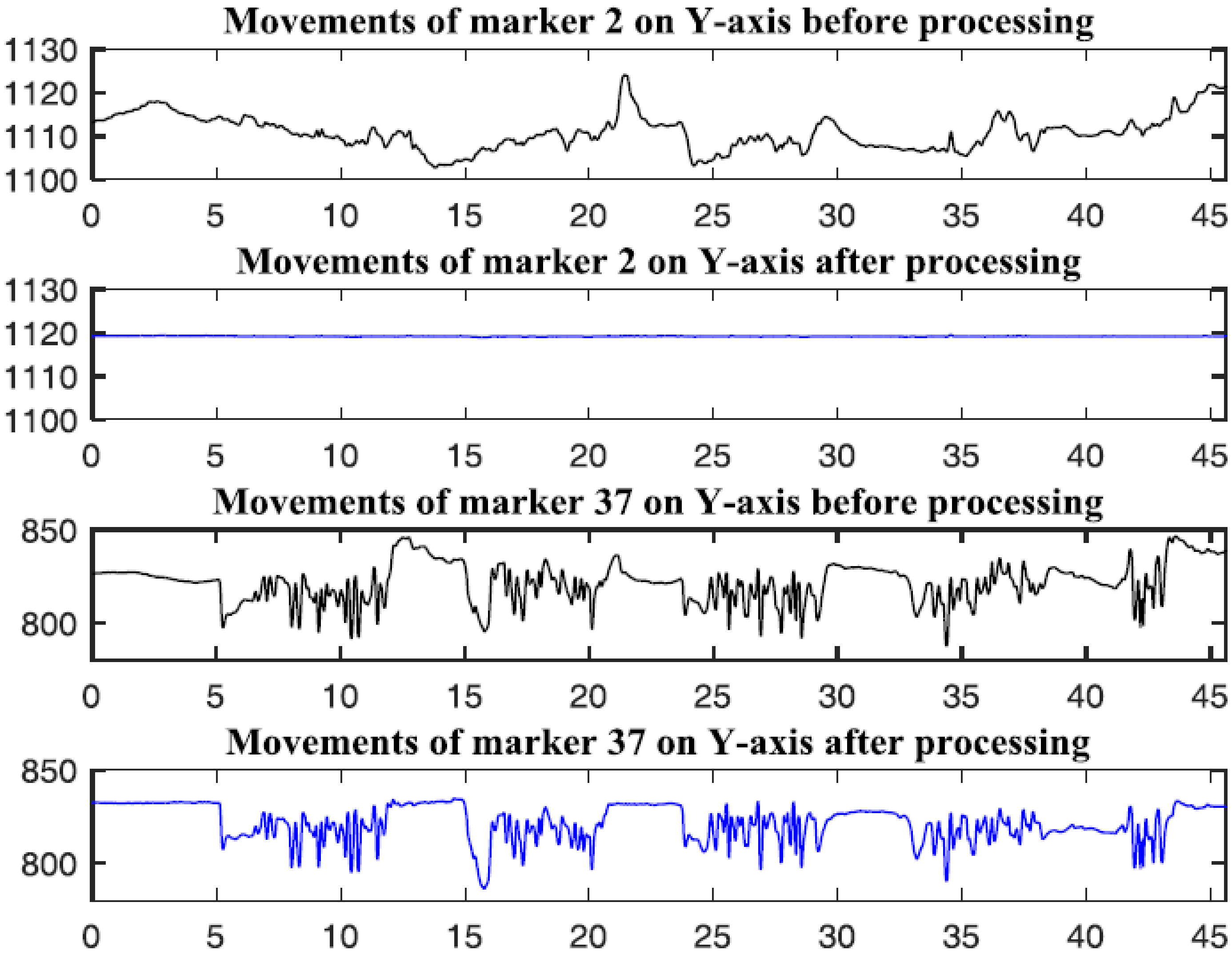
$$L_t = \sum_{i-1}^{4} (P_t^i - P_t^*)^\top (P_{ref}^i - P_{ref}^*)$$

$$L_t = (U_t V_t^*) \cdot (V_t \Sigma_t V_t^*) \qquad R_t = (U_t V_t^*)^\top$$

$$F_t^{'} = R_t^{-1}(F_t - T_t)$$

DATA PROCESSING (7/8)

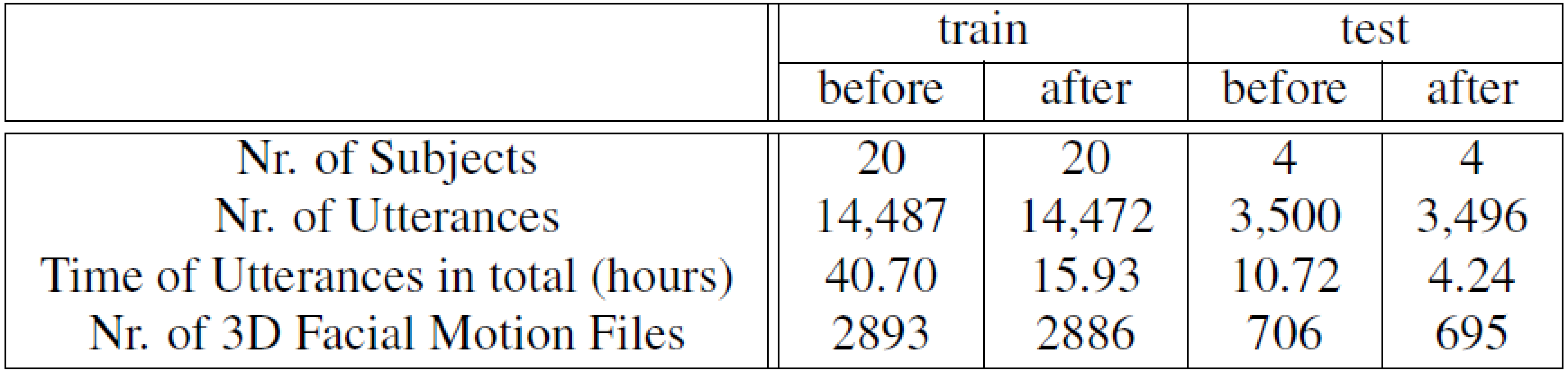
- Removing head motions (results)



Marker 37 is the lower lip

DATA PROCESSING (8/8)

➢ Summary after data processing



	train		test	
	before	after	before	after
Nr. of Subjects	20	20	4	4
Nr. of Utterances	14,487	14,472	3,500	3,496
Time of Utterances in total (hours)	40.70	15.93	10.72	4.24
Nr. of 3D Facial Motion Files	2893	2886	706	695

- The whole set contains about 18k utterances
- Totally 24 subjects are recruited. 20 subjects (10 males and 10 females) are selected to read the training set, while 4 subjects (2 male and 2 female) for the test set
- Roughly 48/65 minutes for each subject in the training/test set
- Total training set is about 16 hours, while the test set is about 4 hours
- Each frame of the 3D video data contains all 41 marks with the position (x; y; z), which consists of a column vector with 41*3=123 entries

CONCLUSION

- The present corpus provides a phonetically labeled, multi-channels and multi-speaker 3D facial motion corpus for large vocabulary Mandarin continuous speech recognition.

- It contains about 18k Mandarin utterances in total.

- We hope that the corpus could also be applied in other fields, such as speech visualization, speech synthesis, as well as rehabilitation training, audio-visual language animation in the further study.

- Future research will focus on facial expression corpus.

REFERENCE

[1] E. D. Petajan, “Automatic Lipreading to Enhance Speech Recognition (Speech Reading),” Ph. D dissertation, Univ. Illinois Urbana-Champaign, Champaign, IL, USA, 1984.

[2] J. Movellan, G. Tesauro, D. Touretzky, T. Leen, and S. Mateo, Eds., “Visual Speech Recognition with Stochastic Networks,” in Advances in Neural Information Processing Systems, Cambridge, MA, USA:MIT Press, vol. 7, pp. 851-858, 1995.

[3] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, “CUAVE: A new audio-visual corpus for multimodal human-computer interface research,” in Proc. ICASSP, pp. 2017-2020, 2002.

[4] I. Matthews, T. Cootes, J. Bangham, S. Cox, and R. Harvey, “Extraction of visual features for lip reading,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 2, pp. 198-213, 2002.

[5] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audiovisual corpus for speech perception and automatic speech recognition,” Journal of Acoustical Society of America, vol. 120, no. 1, pp. 2421-2424, 2006.

[6] C. Sanderson, “Biometric Person Recognition: Face, Speech and Fusion,” Saarbruecken, Germany: VDM, 2008.

[7] T. J. Hazen, K. Saenko, C.-H. La, and J. R. Glass, “A segment based audio-visual speech recognizer: Data collection, development, and initial experiments,” in Proc. 6th Int. Conf. Multimodal Interfaces, New York, NY, USA, pp. 235-242, 2004.

[8] N. Harte and E. Gillen, “TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech,” Transactions on Multimedia, pp. 603-615, 2015.

[9] I. Matthews, G. Potamianos, C. Neti, and J. Luettin, “A comparison of model and transform-based visual features for audio-visual LVCSR,” in Proc. Int. Conf. Multimedia Expo, 2001, pp. 22-25.

[10] C. Xu, Y. Wang, T. Tan and L. Quan, “Depth vs. intensity: Which is more important for face recognition?”, in: Proc. ICPR 2004, Vol.4, 342-354, 2004.

[11] A.F. Abate, M. Nappi, D. Riccio and G. Sabatino, “2D and 3D face recognition: A survey”, Pattern Recognition Letters, pp. 1885-1906, 2007.

[12] C. Sui , S. Haque, R. Togneri, et al. “A 3D Audio-Visual Corpus for Speech Recognition”, in SST. Proc, pp. 125-128, 2012.

[13] D. Zhang, X. Liu, N. Yan, L. Wang, Y. Zhu, and H. Chen, “A multi-channel/multi-speaker articulatory corpus in Mandarin for speech visualization,” in Proc. ISCSLP, pp. 299-303, 2014.

[14] H. Wang, L. Wang and X. Liu. “Multi-level adaptive network for accented Mandarin speech recognition”, IEEE International Conference on Information Science and Technology, pp. 602-605, 2014.

[15] P. Gogoi, B. Borah,and DK. Bhattacharya, “Outlier identification using symmetric neighborhoods,” Procedia Technology, pp. 239-246, 2012.

Thank You!