HIT-AVDB-II： A New Multi-view and Extreme Feature Cases Contained Audio-Visual Database for Biometrics

Xiaoxin Lin[1] Hongxun Yao[1] Xiaopeng Hong[1] Qian Wang[1]

[1]School of Computer and Science, Harbin Institute and Technology
E-Mail:{xxlin, yhx, xphong, qwang}@vilab.hit.edu.cn

Abstract

For research on the proper law of audio-visual speech and biometrics technology, and evaluation of algorithms and systems, we construct a multi-language and multi-view database HIT-AVDB-II with a corpus of various common and special sentences include Chinese and English poems, tongue twister, digits, Greek alphabet and music. The HIT-AVDB-II is ready to facilitate the investigation of multi-view biometrics technology and visual speech reading. HIT-AVDB-II contains formal and extreme feature cases for study. For fair comparison, we also establish related experiment protocols on view and intersection respectively. Further, we supplied one baseline speaker-identification recognition algorithm based on DCT to be compared.

Keywords: Biometric, visual speech, multi-view, database

1. Introduction

Audio-visual speech biometrics and speech reading are of great significance in pattern recognition, computer vision and security strategy areas. Lots of algorithms and systems have been proposed with the aim of meeting the demand of speaker recognition and visual speech [1]. In order to evaluate the properties of these algorithms and systems, various databases have been constructed [2,4,6,14], most of which are under constrained conditions [1]. However, in practical use, lip-reading and audio-visual speech biometrics are still challenging. This is due to many factors of practical use, such as huge vocabulary, illumination variations, blur caused by hand-held camera or head movement, rotation, etc. What is more, the law of human speech still needs to learn. Therefore, databases with large number of subjects, phonetically and visual word balanced with large corpus under unconstraint conditions are still expected [1, 2, 3].

As presented in [4], with profile visual information, the performance of audio-visual speech reading can be improved. In this paper we endorse this contribution of multi-view and endeavor to advance the research in multi-view visual speech reading and visual biometrics.

Besides the contribution of multi-view research on speech reading, we argue that personal habitual mouth movement is helpful for biometrics. The most difference lie in the mouth region between different people always occurs when the speakers exaggerate their mouth movements, which present his/her personal habit and features. In this paper, we regard all the special mouth movements as “exaggerate mouth”. This “exaggerate mouth” is defined as “extreme features”, short for “ext-feature” hereinafter.

For better explicitly taking advantage

Table 1 Parameter of Some Existing databases

Database	Fee	Sessions or recording duration	number of subjects	Corpus(languages and speech content)
CMU's PF AVSR Dataset(07)	___	1	10	150 words[1]
VALID(05)	Y	4 (in office) 1(in studio) (a month)	106 (29 F)	XM2VTS corpus, head rotation in studio [2] with illumination changes and acoustic noise
AVICAR(04)	Y	_____	100 (50 F)	isolated digits, letters, phone numbers, and TIMIT sentences, language backgrounds (60% native American English speakers) inside a car.[3,4]
BANCA(03)	N	12 (3 mouths)	208(104 F)	random 12-digit number, name, address and birth date (4 European languages)
MIT AV TIMIT(03)	Y	20 times	233	20 sentences
AV-ViaVoiceTM(03)	___	_____	290	10,500 single words
VidTIMIT(02)	Y	3 (1 week)	43 (19 F)	Ten sentences each person. Phonetically balanced sentences. Head rotation. NTIMIT corpus, appearance and mood change
CUAVE(02)	Y	1	36 individual, 20 pairs	Isolated and connected digits
XM2VTS(99)	N	4	295	Head rotation, 2 digits sentences and one sentence

Note: Blank: not mentioned in the published paper. 1: Frontal and profile; 2: four targets, placed above, below, left, and right of the camera. 3: An array of cameras mounted on the dashboard; 4: car idle, 35 and 55 mph windows rolled up, or just front windows rolled down, Car background

of the above factors, we construct a multi-view, multi-language, continuous speech database, which is named Harbin Institute of Technology Audio Visual Speech Database II (HIT-AVDB-II). HIT-AVDB-II is designed open and free. It is obvious that our databases can be used for research in: 1 visual speech; 2 biometrics; 3 lip detection/tracking and 4 multi-view etc.

In this paper, we review the relate work of databases in section 2. After expatiating the motivation of HIT-AVDB-II, we present our design principles, reasons and objectives in section 3. In section 4, practice protocols on HIT-AVDB-II is presented, followed with a baseline algorithm evaluation.

2. Relate work

After reviewing the growing course of visual speech database and corpus, we summarize them as follows. In early phase, databases always consist of isolated words. For example, Tulips1[17] consists of only 12 subjects with a corpus of 4 English digits. Other well-known databases includes AVletters[5], M2VTS[6], AMP/CMU[7] etc. Most of them are simple, ideal recording and with small vocabulary. Therefore, these databases cannot be used for practical lip reading research.

To overcome these shortcomings, larger and more complex databases have been developed gradually. Connected-digit strings, unconstraint background in some extent, visual and acoustic noise and large vocabulary are set to simulate the realistic daily life. XM2VTS[6], CUAVE[4], MIT's AV TIMIT[12] and BANCA [1] are well used.

Besides, AVICAR [13] is recorded insides a car with an array of cameras on the dashboard in 2004. VALID database[14] consists of 4 sessions in office and 1 session in studio. Face rotation images of five views, center, left, right, up and down, each approx 10 degrees from center are also recorded [15]. VALID is released in 2005. CMU profile-frontal AVSR dataset start in 2007 for profile investigation [8]. From the above mentioned, we find multi-view attracts attentions of the designers of these datasets.

More parameters of these databases are shown in Table.1. Note that most databases are not free or not open.

However, databases mentioned above are far from enough for visual speech and biometrics [1]. The ones consist of realistic variability are still expected. Due to this motivation, HIT-AVDB-II aims at filling the vacancy of simultaneous big-degree multi-view recorded and extreme feature added database with large-vocabulary, continuous speech, adequate number of subjects, realistic variability and some nonideal recording conditions like: time variations, visual and acoustic noises etc.

3. Design goals and record methods

As mentioned in Section 1, we emphasize the exaggerated pronunciations which ask for special mouth movements. We assume these feature includes some information that would improve the recognition rate.

We also assume that contributions of different views are different to both speech reading and biometrics tasks. Each view owns special information that others could not cover completely. This is more or less motivated by [4, 5, 9].

Following the above assumptions, HIT-AVDB-II is designed to maximize the speakers' speaking characteristics under natural conditions. The main design principles include: 1 multi-view recording simultaneously, 2 exhibiting as much extreme features as possible, 3 including multi-language and 4 Illumination variation and non-controlled surrounding sound recording.

According to Principle 1, HIT-AVDB-II is multi-view recorded. Unlike CUAVE [4], DAVID [9] and CMU databases [8], in which only the frontal and the profile views are used, HIT-AVDB-II are recorded in 4 different views simultaneously.

Based on Principle 2 to 3, the corpus of HIT-AVDB-II includes digits, Chinese poems, tongue twisters of Chinese and English, Greek alphabets, music notes, mandarin vowel.

Reasons about the above mentioned corpus are as follows:

1. The use of Chinese poems here is mainly because that reading poems leads the speakers to express her/his emotion and special tune, which could not be found in the four standard tunes of mandarin.

2. HIT-AVDB-II also selects some tongue twisters, containing some consonant pairs which are easily confused to each other. These expose the speakers' the habits and pronunciation "defects" in unconsciously fast mouth movements.

3. Greek alphabet letters are introduced for exhibit mouth movement sufficiently. Music notes and vowels of Mandarin are helpful for exposing some special tunes.

In HIT-AVDB-II, 30 speakers (15 female and 15 male) are asked to sit on the fixed chair to record. Considering the frame rate, the time length of corpus and the number of views, we believe the size of HIT-AVDB-II is adequate to some extent, which is similar to databases like CUAVE [4] and VidTIMIT[18].

HIT-AVDB-II is recorded in 3 sessions. Speakers are free except for the fixed chair and the corpus including 11

sentences. They speak each sentence twice naturally in each session. Intersession delay is 2 days. This delay allows the change of face, skin variation, hair style of subjects. We also encourage them to wear spectacles and hair ornaments to simulate realistic scenarios. The first session is recorded in the morning with the dark blue background, the second one in the afternoon with light green background and the last one in the evening with dark red background under fluorescent lamp.

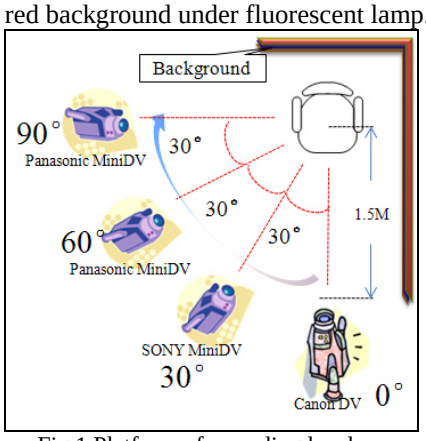

Fig.1 Platform of recording locale

Each Camera is tuned to view the above of chests. The locale and the cameras are set as the Fig.1. The videos are gathered through IEEE 1394 interface and finally compressed to MPEG2.

Sound waves are included in the video file. More configuration parameters and storage parameters are in Table2 and Table 3 respectively:

Table2. Configuration Parameter

Item	Value	Remark
Speaker	30	15 male and 15 female
Sentence	11	
Utterance	6	3sessions, 2 utterances each session
Background	3	Red, Green, Blue
Recording Time		Morning, afternoon and evening
View points	4	0°(facial), 30°, 60°, 90°(profile)

Note: Allow Angle deflection: 5°.
Gesture Requirement: Free.

Table3. Storage Parameter

Item	Value	Remark
Record Format	SP	Common Clear
Frequency	25fps	Noise like sinker knock on stones
Resolution	720×576	Pixels
Number of DVD	4	4.7G×4

For facilitating management and practice, the storage folder structure of HIT-AVDB-II is organized as following:

#view id// #person id// #sentence id// #Utterance ID

(a) 90° (b) 60° (c) 30° (d) 0°
Fig.2 Sample of Multi-view images from the left to right is 90°, 60°, 30°, 0° images

(a)dark red (b) Light green (c) dark blue
Fig. 3: Sample of multi-background

Fig.2 illustrates an example of the 4 views of a person. Fig.3 shows the influence on complexion from different backgrounds and illumination.

4. Practice protocol

From the database, we have 6 utterances of each sentence of each person in one view. Here, we denominate these six utterances U_1 to U_6.

As the Fig.2 shows, HIT-AVDB-II has 4 views. And as we have assumed, each view owns special information, which would also be helpful to the others in some extent. Besides the view factor, we also regard that algorithm and AVSR systems should be session-independent. Therefore, we propose two protocols that one view related and one intersection related for further understanding and fair comparison.

Here, we recommend to measure identification performance with the Rank-1 match rate [1] and verification performance with FAR and FRR [1, 18]. More details about the protocols of HIT-AVDB-II are as follows:

Protocol I: Multi-view protocol

Type A: Individual Test

As this paper has assumed, contributions of different views are different. They complement to each other mutually. Here, four views from frontal to profile are named as 0°, 30°, 60°and 90°. Individual Test is to do experience on each view respectively and compare the results.

Type B: Combination Test

For further understanding of the relationship among different views, Combination Test is to investigate the relationship of each views when put them together. View combination follows the rule of VIEW-TITLE and the view under this combination is VIEW-TITLE-View Id; For example: we combine the frontal and profile together, we get VIEW-Frontal_Profile consists of VIEW-Frontal_Profile-0° and VIEW-Frontal_Profile-90°. Totally, we could get 12 combinations and 4 individuals.

Protocol II: Intersection Protocol

Three sessions of HIT-AVDB-II are recorded at different time span of days. Each session includes two utterances. To evaluate the algorithms and find out the relationship of each session and utterance, Type A and Type B are proposed.

Type A: Session Test

In verification or identification with threshold, we propose one of these three sessions for training, one for evaluation set and one for test set. Any of these three could swap with each other.

Type B: Utterance Test

For identification application without an evaluation set, we propose each utterance as test set while the other five to be the training set. Each test client has 11 test sentences from 29 other unknown persons.

5. Experiences

5.1. Motivation

To offer a benchmark result for further comparison, we utilize Semi-Continuous Hidden Markov model (SCHMM) based on a common used feature extracting algorithms DCT [16].

5.2. Experience

Preprocessing, including sentence segmentation and landmark (four points of mouth rectangle) labeling have been completed manually. We use 2-D DCT computation with 64×64 mouth rectangle. We select the top-left 30 low frequency components from DCT coefficients in a zigzag-scanning manner. A SCHMM with 6 states and 8 mixtures is exploited for training and testing task. More experiences and results will be updated daily on the public available URL: http://vilab.hit.edu.cn/lipreading/avdbii.php.

6. Conclusion

In this paper, under the assumption of importance of the “extreme feature” represented by “exaggerate mouth” and multi-view information we constructed a Chinese face speaker database HIT-AVDB-II. We designed the corpus for HIT-AVDB-II with Chinese and English phrases, poems and tongue twisters, Greek alphabet letters and music notes and side-view recording. The HIT-AVDB-II consists of 6 utterances of each sentence of 30 subjects. To facilitate the further investigation, the research practice protocols are introduced.

7. Acknowledgement

We would like to acknowledge the support of Program for New Century Excellent Talents in University NCET-05-03 34 and National Natural Science Foundation of China under contract No.60775024. We are also extremely grateful to everyone that participates in this database, subjects, labelers etc.

8. References

[1] Aleksic, P.S.; Katsaggelos, A.K., "Audio-Visual Biometrics," *Proceedings of the IEEE*, Nov. 2006, vol.94, no.11, pp.2025-2044.

[2] Jankowski, C. Kalyanswamy, A. Basson, S. Spitz, J., "NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database," *Acoustics, Speech, and Signal Processing, 1990*, pp.109-112.

[3] C. Sanderson and K.K. Paliwal. "Identity Verification Using Speech and Face Information", *Digital Signal Processing*, 2004.14(5):449-480.

[4] Patterson, E.; Gurbuz, S.; Tufekci, Z.; Gowdy, J.N., "CUAVE: a new audio-visual database for multimodal human-computer interface research," *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002(ICASSP '02)*, 2002, Orlando, USA, pp.2017-2020.

[5] I. Matthews, "Features for Audio-Visual Speech Recognition", *Ph.D. thesis*, School of Information Systems, University of East Anglia, UK, 1998.

[6] K. Messer, J. Matas, J. Kittler, and J. Luettin, "Xm2vtsdb: The Extended M2VTS Database," *Proc. Second Int'l Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA '99)*, 1999, pp. 72-77.

[7] Kumar, K.; Tsuhan Chen; Stern, R.M., "Profile View Lip Reading," *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007(ICASSP 2007).* , 2007, Honolulu, Hawaii, pp.429-432.

[8] M. Faraj, J.Bigun, "Synergy of Lip-Motion and Acoustic Features in Biometric Speech and Speaker Recognition", *IEEE Transactions on Computers*, 2007, 55(9): pp1169-1175.

[9] C. C. Chibelushi, F. Deravi, and J. S. Mason, "BT DAVID Database " *Internal Rep., Speech and Image Processing Research Group,Dept. of Electrical and Electronic Engineering, Univ, les Swansea*, 1996.

[10] Gavat, I. Costache, G. Iancu, C., "Robust speech recognizer using multiclass SVM," *Neural Network Applications in Electrical Engineering, 2004*, pp. 63-66.

[11] T. Chen, "audiovisual speech processing," *IEEE Signal Processing Mag.*, vol. 18,pp. 9–21, Jan. 2001.

[12] T. Hazen, K. Saenko, C. La and J. Glass, "A segment-based audio-visual speech recognizer: Data collection, development and initial experiments," *Proc. ICMI*, State College, October 2004

[13] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "AVICAR: Audio-visual speech corpus in a car environment", *Proc. Conf. Spoken Language*, Jeju, Korea, 2004.

[14] N. A. Fox, B. A. O'Mullane, and R. B. Reilly, "VALID: A new practical audio-visual database, and comparative results", *Proc. of the 5th International Conference on Audio-and Video-Based Biometric Person Authentication (AVBPA-2005)*

[15] http://ee.ucd.ie/validdb/

[16] P. Duchnowski, U. Meier, and A. Waibel, "See me, hear me: Integrating automatic speech recognition and lip-reading" in Proc. *Int. Conf. Spoken Lang. Processing*, Yokohama, Japan, Sep. 18–22, 1994, pp. 547–550.

[17] J. R. Movellan, “Visual speech recognition with stochastic networks” in Advances in Neural Information Processing Systems, G. Tesauro, D. Toruetzky, and T. Leen, Eds. Cambridge, MA: MIT Press, 1995, vol. 7.
[18] Conrad Sanderson, **“**The VidTIMIT Database**”**, *IDIAP-Com 02-06, 2002.*