International Journal of Database Theory and Application
Vol.10, No.1 (2017), pp.1-10
http://dx.doi.org/10.14257/ijdta.2017.10.1.01

Construction of Bi-modal Database for Barrier-free Teaching System

Jiling Tang[1], Ping Feng[1] and Zhanlei Li [2]

[1]*College of Computer Science and Technology, Changchun University, Changchun 130022, China*
[2]*Business college, Dalian University of Technology, Panjin 124000,China*
934720633@qq.com, 17204696@qq.com, 752003002@qq.com

Abstract

This paper analyzes the application of Chinese speech recognition technology in the non-barrier education system, and studies the construction of bi-modal database for barrier-free teaching system. Based on the case study of the curriculum named "Foundation of Photoshop", the paper creates corpus to make acquisition of experimental data and annotation of corpora.Meanwhile we analyze and design the organization of data and build essential dictionary and grammar network in recognition system.

Keywords*: liberated learing;bi-moldal database;barrier-free teaching*

1. Introduction

According to second sampling survey of China disabled ,we estimate that the number of deaf people attain to just over 20,540,000,before 2010,which ranked second in all types of disables. So it is an important part that our country is suppose to carry out the hearing impaired education and teaching research in development of special education [1].

Computerized speech-to-text technology used in hearing-impaired education was firstly put forward by Stuckless in one of his papers in 1981 [2]. With IBM's strong technical support, Saint Mary's University Canada, set up the Barrier-Free Teaching Union to make speech recognition applied widely.

In China,the earliest study about lipreading was started by Yao HongXun from Harbin Institute of Technology.She took extraction method of lip color filter based on the features of lip to Identify five vowels.The lip-reading recognition for non-specific person in this method can reach above 90%.The Institute of Acoustics of the Chinese Academy of Sciences established the first Chinese speech dual-mode database by analyzing the structure of similar databases at home and abroad .The database is combined with the characteristics of Chinese pronunciation of which selected corpus over all the Chinese initials and finals . In a certain extent, it provided the help for the later researchers in acquisition and solution of the visual information.Moreover , Wuyi University, Dalian University of Technology, Zhejiang University, Southeast University and other institutions have also committed to lip reading study [4]. But there is no special study of the technology used in teaching, especially in teaching hearing-impaired students [3]. Which leads to the status quo that the recognition ability of speech recognition software used in teaching is good, but the recognition of professional vocabularies is more difficult and makes errors frequently.

In the Chinese language environment, the use of the dual mode speech recognition technology with audio and video combination for the teaching system (shown in Figure 1) can improve the recognition rate and robustness of the system.What we should understand is that the training process of speech recognition is a statistical process.In order to ensure

ISSN: 2005-4270 IJDTA

the accuracy of the model, we requires a basis of substantial audio and video data. In the field of pure speech recognition, a lot of standard speech corpus provides the baseline information, which plays a very important role in the development of speech recognition. However, in the field of audio visual ,due to its application in recording different requirements in different areas, there is no such is very comprehensive and system database. Therefore, this paper introduces construction process of audio and video database in the system.

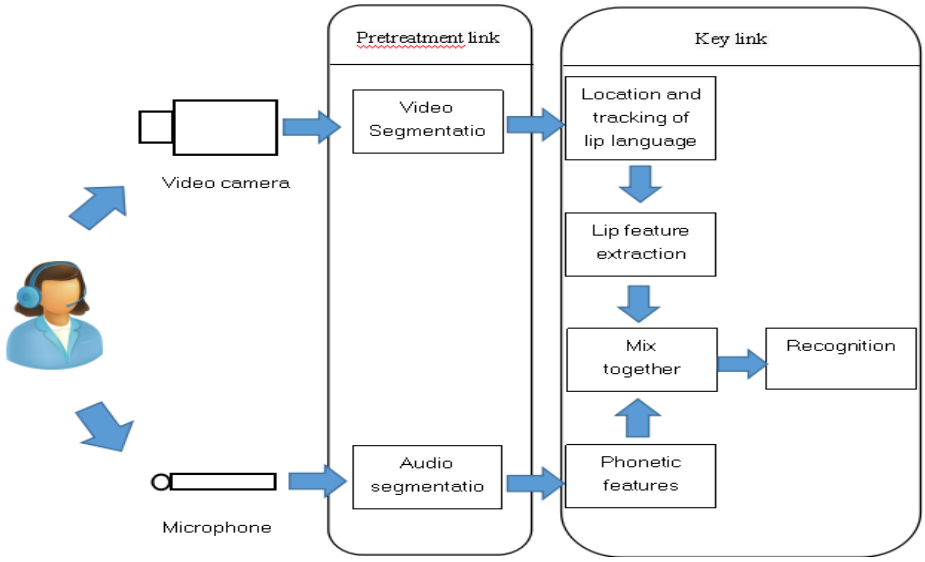

Figure 1. Structure of Barrier-Free Teaching System

2. Building a Database Based on Bi-Modal Database

Through the analysis and comparison of data at home and abroad, this paper summed up some of the basis of building a database, provided a reference for the identification system to build a database.

2.1. To Select Text Corpus Based on the Size of the Vocabulary

Pronunciation representation: large vocabulary system, asked to choose corpus to cover all the pronunciation, and the small vocabulary contains only system used pronunciation.

Apparent place representation: when choosing the corpus, bi-modal should not only consider the pronunciation, but also take into account the apparent position that is the shape of the lips. To have better representation, to be able to reflect more fully the Chinese auditory, visual law, the former includes all common syllables, the later to relate to all of the visual bits.

2.2. To Select a Recording and Imaging Conditions Depending on the Application Object and Purpose of Use

Determine the crews: non-specific system training data should include many people to record. If the user is female, the crews should contain a certain proportion of female members, crews should choose different ages, different cultures degree, and different regions of the accent etc based on the possible situations of the users.

Choose the noise: usually bi-modal databases need to add noise in the audio channel to test the robustness of the system.we should select the noise of different SNR and different scenarios. Depending on the different methods of anti-noise, noisy data used for training is different.Such as when using the anti-noise compensation method, you need to use

different noise environments to train the different models, when using the anti-noise characteristics method, you should have a training with noise-free voice.

Take the image angle: according to the different visual features selected, we should take images from different angles, usually based on the need, we should set up ranging from 1 to 3 sets of cameras to capture images from different angles, for instance, visual characteristics use the model approach, it is sometimes necessary to use to the side of the image, and therefore we need to take from the side image.

Select the light: when recording, light conditions include natural light,cold source light as well as the special lamps used in photography when the aim is to make the light more evenly.We used the cold source light indoor, natural light outdoor.

2.3. Selecting Data's Format Based on the System Requirements and Storage Requirements

The selection of the audio parameters: uncompressed PCM coding is selected and sampling rate, ranging from 8KHZ to 48KHZ, is not too high and not too low because the sampling rate too low will cause loss of information, too high will cause redundant data. If so, memory space will be wasted. In the course of training and testing 16KHZ is used while 44.1KHZ is used in recording because speech information plays a dominant role in bimodal speech recognition. Although some data seems wasteful, but if storage space is sufficient, the more the better.

The selection of the video parameters: the video features depend on lips features, so video parameters can be selected smaller. The video sampling rate are ranging from 20 frames / sec to 50 frames / sec. We chose a common 30 frames / sec.

The selection of the storage format: what kind of audio and video data selected to save depends on the hardware. Data formats are AVI, MEPG2, and MEPG4 formats, MPEG4 standard was introduced in recent years, which is based on the concept of encoding objects. For example, it is based on faces videos to define the facial animation parameters FAP (Facial Definition Parameter) and face-defined parameters FDP (Facial Animation Parameter) [9]. The best way to extract visual features is AAM (Active Appearance Model). AVI and the simulation system don't need compressing. So AVI format is used to save.

The data stored in the system includes the following:

First, the corpus selection includes a clear demand, clear content, collection of original text and removing redundancy.

The next step is the collection and primary process of voice and video data. Initial treatment also includes data preprocessing and data segmentation. Data preprocessing primary get rid of excess non-voice data, such as duplicate data, long silent segments, speech rate, volume, signal to noise ratio, contrast, lighting and other aspects of unqualified data; data segmentation is cut into the required length of voice.

The annotation of corpus voice segments is the final stage. It means increasing every segment of the international phonetic alphabet to represent the name of the segment and the Chinese Academy of Social Sciences Institute of Linguistics library can use the proposed prosodic transcription system C-To BI (Chinese-Tone Break Index). The standards are divided into eight levels, respectively Pinyin layer, label marked Accents, consonant / vowel layer, layer tone and intonation, pauses layer, layer accent, accent and topics layer conversion layer. Based on the needs of the database different levels can be used. And database in this system uses Pinyin layer.

3. Contribution of Dual-Mode Database

The work is heavy in the construction of corpus based on continuous speech recognition system. Because there is a corpus with practical application value, in other words, some elements of the corpus type, such as type, size and construction, must be on the basis of application requirements after carefully designed, rather than random combination. Only by doing on this way can we ensure that the work that we have put into is worth .Despite of the medium and small vocabulary, the corpus should be eliminated carefully in this system.

3.1. Corpus Selection Based on the Non - Barrier Teaching System

Because the course involves a large vocabulary, we build different corpus for different courses according to the practical needs of teaching, by choosing applicable common words in non - barrier teaching system .From the results of the survey of Jiang Mei [5], among the computer courses opened in China's colleges and universities, Flash, Photoshop are most widely accepted. Therefore, this paper chooses the "Foundation of Photoshop" course as the object to research the selection of the corpus.Because the course is a practical course, there are a lot of Practical Interactive demo links. Teachers should manipulate computer while they are trying to use sign language to explain, which makes the classroom teaching progress slow, impacts the teaching of fluency as well.There is much of textual information related to course content on the interface of photoshop software.So this paper only analyzes and organizes teaching vocabularies which teachers use to explain the operation to design "Foundation of Photoshop" course corpus with the total of 60 words, as shown in Table 1.

Table 1. Corpus Vocabularies of "Foundation of Photoshop" Course in Chinese

open-computer system	close-computer system	left-click the mouse	right-click the mouse	double-click the left mouse button
click on the icon	log in	log out	move the mouse	drag the mouse
open file	save file	close file	restore file	delete file
open window	open dialog box	open the shortcut	open toolbar	open sub menu
close window	close dialog box	close shortcut	close toolbar	close sub menu
open panel	close panel	drag the slider up and down	drag the slider left and right	set parameters
select data	confirm	sub-panel off	combine panel	input data
select tool	select options	adjust size	set height	set width
enlarge	narrow	direct	lock	erase
Rotate	wipe	align	copy	delete
Fill	return	merge	link	create
Cancel	modify	cut	paste	edit
Hide	convert	exit	add to	change

3.2. Acquisition of Audio and Video Data

We choose teachers of different accents, different ages and different gender as representative experimenters. By collecting equipment, we can get the the sample collection of the audio and video signals from subjects to generate the formation of a collection of documents as the basic data for the teaching system of the barrier.

In the experiment, we choose 30 teachers as test subjects (17 males and 13 females), between the ages of 25-60, and ask them to say the specified data with normal speed and intonation of Mandarin. Then we use notebook computer ， video capture software ， microphone and common camera as recording device. The format of data is avi format.The code of audio is the PCM coding. Sampling frequency is 41000Hz, single channel, 16 bit quantization;The rate of video frame is 30 frames/s, 12 bits quantization, 320*240 each frame of the image.

Recording are made in a relatively quiet laboratory environment and all of the data are collected for the system's training. After that, some of the data will be synthesized into the noise signal by the software to form the test data. The test data have the exact value of the signal noise ratio, so that it is easy to clarify the test standard.The results we obtioned respctivelly are -10dB, 0dB, 10dB, 15dB, 20dB.

Rules for naming the collecting files is: gender identity + age + code of origin+ "_" +number of subjects + "_ "+code of corpus.avi". For example, MOB_2_45.avi indicates No.45 audio and video file from No.2 test subject, an over 45 year-old man with a northern accent.

3.3. The Annotation of Corpus

The purpose of tagging the corpus is to obtain the data of each pronunciation more accurately. Label should include the information of pronunciation and the initial mark of the pronunciation data. The system uses the HTK training in audio processing. By considering the training video, we prefer to using audio frames for the mark instead of time parameters. Similarly, video processing is also marked by video frames. The labeling is performed by tagging software Wav2Label [6] as shown in Figure 2.

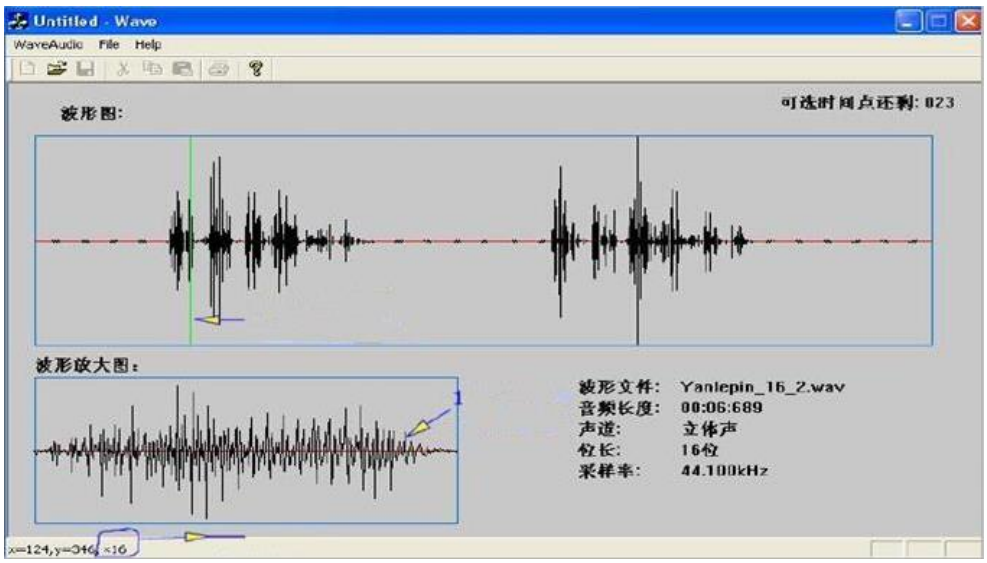

Figure 2. Wav2Label Tagging Software Interface

After being processed by Wav2Label, four different text contents of the document are generated: a TXT file including a voice sentence ,initial hits and end hits;a WRD file recording initial hits, end hits and pronunciation of every word in speech file; a VED file

recording the corresponding video frames of the starting time and ending time of every word's pronunciation in speech; a LAB file recording the starting time and the ending time (with MS as the time unit) and the pronunciation of each word in each sentence in order to be able to match the usage of HTK.

Table 2. Annotation Case

Corpus sample	*.txt	*.wrd	*.ved	*.lab
Open-computer system	0 135254	805 52038 da 53096 68090 kai 69501 80614 dian 80967 89964 nao 91022 101430 xi 102312 114660 tong	15 17 da 18 23 kai 23 27 dian 27 30 nao 30 34 xi 34 39 tong	1016 1180 da 1204 1544 kai 1576 1828 dian 1836 2040 nao 2064 2300 xi 2320 2600 tong

3.4. Organization of the Database

The data stored in the system includes the following:

(1)Personal information of test teachers includes name, gender, birth date, accent and the unique number assigned to each of tested person by database system, as shown in Table 3.In China, the working ages of practical teachers are from 25 to 60 years old. Because the characteristics of the sound will vary with age, the system divide the test objects into two groups, one is under 45 years of age and the other is above 45 years of age which are identified with the character Y and O, in order to distinguish between different audio and video data. Because the Chinese dialects includes seven language systems: Northern, Hakka, Hunan, Wu, Guangdong, Fujian, jiangxi. Teachers in China should also pass the National Proficiency Test of Putonghua when they obtain the Teacher's Professional Qualification Certificate.Therefore, the system only considers that what the teacher say in the class are universal words with local accent in the class .Respectively, the characters B, K, X, W, Y, M and G are used to describe the different regions of the accent.

Table 3. Personal Information of Test Teachers

variable name	type of data	Instructions of variable assignment
ID	int	Automatic generated by system;
sex	char	F(female),M(male);
birthday	data	1969-10-29
age	char	If ((sysday-birthday)>45,'O','Y'); O=old,Y=young.
Birthplace	char	B、K、X、W、Y、M、G

(2) collection of audio and video data

Each experimental teacher collects a corpus of AVI video files and WAV audio files, as well as the generation of the four annotation files.

(3)The relationship between the data: people of different gender and age differ in the voice frequency of speech, such as the frequency of the voice of the male made by vocal cords are generally lower than that of the female;People from different places have different tones and accents when they read the same Chinese characters because of the influence of dialects.All of these will affect the lip shape and tone when each Chinese character is pronounced, so the same corpus vocabularies will entail different AVI and WAV files and annotation file. As a consequence the relationship between the data is a pair of multi tree structure, as shown in Figure 3.

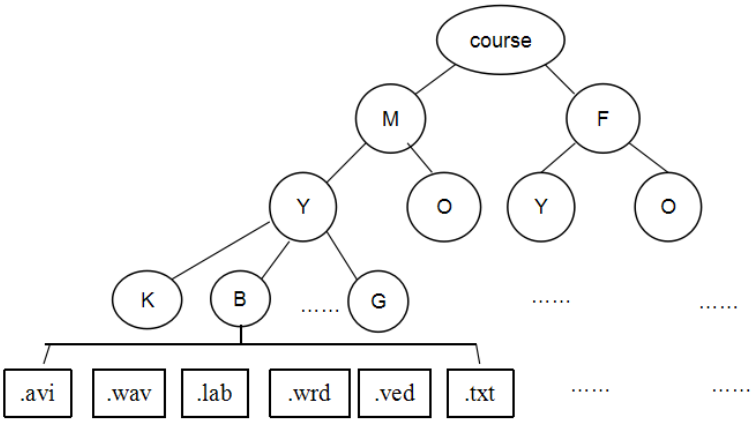

Figure 3. Data Structure of Curriculum Corpus

3.5. The Establishment of Dictionary and Grammar Rules

Continuous speech recognition is to house statistical knowledge of all speech in a unified framework of HMM. It can be divided into three levels: the first layer Acoustics, Speech layer, each syllable with a HMM model and the corresponding parameters shows that basic unit of HMM model is transferred arc (probability) among the states. The second layer is the word layer that rules a predetermined plurality of sub-tone composing entry. The third layer is the syntax layer, which defines the rules for entry constitutes the statement. The first layer is provided by the acoustic model, while the word layer and the syntax layer are provided by the language model. The language model in speech recognition system is very important.

Because people sometimes do not listen to each sound accurately.They often understand the speaker's intention by the front and back relations and a prior knowledge in the process of listening to the voice.The prior knowledge is a language model when it is used in machine recognition.

Portions of the modular system, identify module in particular, program in ATK (Application Toolkit for HTK) programming, and when the system HTK established to identify comparison, it needs to referring to two important documents ---- dictionary and word net. These two documents defined the number of search path for recognizing the result. These two documents are corresponding to the word layer and the grammar layer.

Wherein the content of the dictionary is the comparison table of the word and the acoustic model, the function is transferring the recognized word into a form of the model sequence that can be written with a text editor. By clarifying the corpus, it formed the dictionary, named Zidian. Words in the dictionary are divided into two categories: verbs and nouns.Therefore, we need to establish a dictionary and grammar rules.

The content of the dictionary is the contrast between the word and the acoustic model, and the function of it is to convert the recognized word into the form of the model sequence coded directly with a text editor. Classifying the vocabularies in the corpus forms a dictionary, as shown in Table 4.

Table 4. Dictionary

Verb				Noun	
word	acoustic model	Word	Acoustic model	word	Acoustic model
open	da3 kai1	Close	guan1 bi4	computer	dian4 niao3
click	dian3 ji1	doube click	shuang1 ji1	system	xi4 tong3
log in	deng1 lu4	log out	zhu4 xiao1	icon	tu2 biao1
move	yi2 dong4	drag	tuo1 dong4	data	shu4 ju4
save	cun2 chu3	restore	hui1 fu4	slider	hua2 kuai4
delete	shan1 chu2	select	xuan3 ze2	mouse	shu3 biao1
confirm	que4 ren4	combine	zu3 he2	file	wen1 jian4
set	she4 zhi4	adjust	tiao2 zheng3	panel	mian4 ban3
input	shu1 ru4	direct	zhi3 xiang4	parameters	can1 shu4
enlarge	fang4 da4	align	dui4 qi2	window	chuang1 kou3
rotate	xuan2 zhuan3	merge	he2 bin4	option	xuan3 xiang4
fill	tian1 cong1	cut	jian3 qie1	dialog box	dui4 hua4 kuang4
cancel	qu3 xiao1	exit	tui4 chu1	shortcut	kuai4 jie2 cai4 dan1
hide	yin3 cang2	lock	suo3 ding4	sub menu	zi3 cai4 dan1
narrow	suo1 xiao3	copy	fu4 zhi4	toolbar	gong1 ju4 lan2
wipe	ca1 chu2	link	lian2 jie1	size	da4 xiao3
return	fan3 hui2	paste	zhan1 tie1	height	gao1 du4
modify	xiu1 gai3	add to	tian1 jia1	width	kuan1 du4
convert	zhuan3 huan4	erase	qing1 chu1	up and down	shang4 xia4
edit	bian1 ji2	delete	shan1 chu1	right and left	zuo3 you4
change	gai3 bian4	create	jian4 li4	enter	hui2 che1

After the establishment of the dictionary word layer, then create a language model about syntax layer. Use the language model full and effective, which can significantly improve the recognition rate of recognition system. There are two ways to establish language model: one is based on the rules of grammar network, the other is based on the statistical language model. Both two methods in different situations have different effects. Rule-based approach is suitable for speech recognition of small vocabulary or in special occasions. The way based on statistics is suitable for continuous speech recognition of large vocabulary. This system has a small vocabulary, so it is based on the rules of grammar network.

Lexical network is to identify the system of recognition of the rules of grammar, this document records the possibility of the combination of each word. For example, if the teacher says the word "open", after it, possibly it will produce 14 sets of instructions including "computer system", "window", "sub menu", "dialog box "and so on. Below with "open" and "close" and "computer system" , "window" , "sub menu" , "dialog box ","panel ","toolbar ","shortcut menu ". The schematic diagram is as shown in Figure 4.

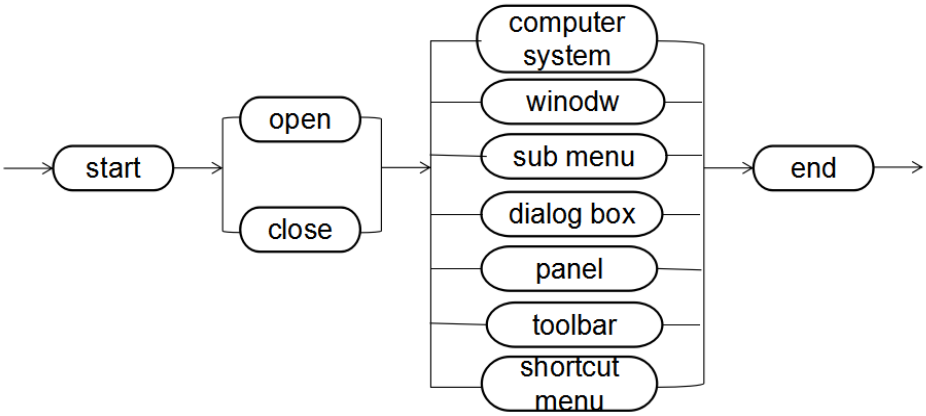

Figure 4. Schematic Diagram of Lexical Network

In the case of small vocabulary, lexical network can be written by text editor by hand. On the occasion that the vocabulary is larger and the rules are more complicated, HTK provides relatively understandable rules of writing which can be coded in accordance with the rules of writing and compiled into the format of the lexical network through the HParse command.

HTK writing rules of grammar:

| - Choose one from two or more;

[] - Option, said that the constants or variables in the [] is not essential

{} - Repeated zero time or many times;

<> - Repeated once or more;

Definitions of variables: Define variables started with "$"

Format:

$var= expression;

The lexical network in Figure 4 can be expressed as following scripts:

```
$Ming Ci=Window|Dialog box|Sub menu|Shortcut menu|
        computer system|panel|toolbar
$Dong Ci=Open|Close
(SEN_START $Dong Ci $Ming Ci SEN_END)
```

Through the statistics and summary of the corpus in the sentence, it can be divided into the following three categories

1) verb + noun. Such as "click" + "mouse", "input "+ "data"

2) verbs. Such as set, select, combine.

3) noun + noun. Such as "sub "+ "menu", "computer "+ "system"

On the basis of the rule of law, all the words in the dictionary will be edited into a script file gram, compiled to generate lexical network wdnet through using HParse.

4. Conclusion

This paper introduces the structure of the barrier-free teaching system for deaf students. Firstly, through the analysis of the existing bimoldal database at home and abroad, we summed up the basis and content of bimoldal database establishment. Then, the paper introduces the construction process of the dual mode data from the corpus selection, data collection, the construction of the database framework and the annotation of the text corpus. At last, it introduces the establishment process of the dictionary and grammar network which are necessary for the construction of the identification system, providing the sufficient data guarantee for the following model training and the research of the experimental evaluation.

References

[1] Y. Jiaosheng, "Research of English teaching by Interactive whiteboard", Master thesis of Yunnan Normal University, vol. 6, **(2014)**.
[2] S. Furui, "50 years of progress in speech and speaker recognition research", Ecti transactions on computer and information technology, vol. 1, no. 2, **(2005)**, pp. 64-74.
[3] X. Jing, "A Study on the Effect of Information Transmission and Its Influence Factors of Speech Recognition Technology on Classroom Teaching of Deaf Students of Advanced Learning Institutions", Master thesis of Chongqing Normal University, vol. 6, **(2007)**.
[4] Wudi, "Research on Lip Reading Algorithms", Master thesis of Beijing Jiao Tong University, vol. 4, **(2015)**.
[5] J. Mei, "How to train written communication ability in computer teaching of Deaf students", Journal of computer knowledge and technology, vol. 9, **(2013)**.
[6] Y. Lepin, "Design of speech control system on AV bimodal information and its Realization", Master thesis of South China University of Technology, vol. 5, **(2010)**.
[7] L. Xin, "Research on 2D and Bimodal Hybrid for face Recognition with a sigle Training Sample", Master thesis of Harbin Engineering university, vol. 3, **(2011)**.
[8] L. Gang, "Mandarin Chinese Visual-speech Database for Speech-impaired People", Journal of Chinese Biomedical Engineering, vol. 6, **(2007)**.
[9] H. Yuanlie and Y. Zilu, "Using K-D Tree to Implement Effective Query Bimodal Multimedia Database", Journal of Computer Engineering and Application, vol. 18, **(2003)**.
[10] C. Y. Xiang and L. Ming, "Research on Robustness of Audio-Visual Speaker Recognition Based on Articulatory Features", China Academic Journal Electronic, vol. 12, **(2010)**.
[11] Y. Jingjie, "Bimodal emotion recognition based on body gesture facial expression", Journal of image and graphics, vol. 9, **(2013)**.
[12] Z. Xin and D. Linin, "CVss1.0 Audio-Visual Database for Visual Speech Synthesis", Microcomputer Applications, vol. 3, **(2007)**.
[13] H. Yanfang, "Construction of Visual Speech Representation for Bi-modal Based Speech Recognition", Master thesis of Beijing university of technology, vol. 6, **(2013)**.
[14] F. Xiaohui, "Research on Noise Treatment of Speech Recognition with Lip-movement Information", Doctor thesis of of South China University of Technology, vol. 10, **(2010)**.
[15] H. Qiahua, Z. Zhengyu and F. Xiaohui, "Lip motion and voice consistency analysis algorithm based on shift-invariant dictionary", Journal of Huazhong University of Science and Technology, vol. 10, **(2015)**.

