A Telephone Speech Corpus of China's Minority languages for Automatic Language Identification

Xiuhua Zeng[1], Jian Yang[2], Libo Zuo[2], Yonghua Xu[2]
[1]School of Physics and Electronic Engineering
Qujing Normal University, Qujing, 655011, China
[2]School of Information Science and Engineering
Yunnan University, Kunming, Yunnan, 650091, China

***Abstract*—Research in language identification require corpus of multi-languages speech data to capture the distinguishable information within and across languages. In the past few decades, many statistical approaches to language identification have been developed based on two common and public-domain corpora which consist of telephone speech from about 26 languages and dialects. However, the China's minority languages have not been used as the target languages in the published papers up to now. In our work, we select 9 typical China's minority languages and Mandarin to construct our telephone speech corpus. These minority languages are composed of Naxi, Miao, Bai, Dai, Yi, Zhuang, Uygur language, Mongolian and Tibetan. Each minority language represents its minority nationality. The corpus can be used to study, develop, evaluate and compare minority languages identification algorithms. Moreover, it will promote the Linguistic researchers to pay more attention to the long history and splendid culture of our national minorities.**

Keywords-Language identification; Telephone speech; Corpus; Minority languages

I. INTRODUCTION

Automatic language identification (LID) is the process of identifying the language of a spoken utterance. With the rapid development of information technology, LID plays an important role in applications such as telephone-based multilingual conversation system, spoken language translation system, multilingual information retrieval system and security [1].

The language identification of Mandarin and China's minority languages is an important technology in these applications. Research in language identification require corpus of multi-languages speech data to capture the distinguishable information within and across languages. In the past few decades, many statistical approaches to LID have been developed based on two common and public-domain corpora which consist of telephone speech from about 26 languages and dialects [4]. However, the China's minority languages have not been used as the target languages in the published papers up to now. The Chinese phonemic loanwords are very popular in present-day China's minority languages. Because the China's minority language identification should contain not only language identification but also accent identification, it is crucial distinctive from the typical LID [8]. In order to study china's minority languages identification, we will construct a new multilingual telephone speech corpus consisted of speech taken from telephone conversations in 9 typical China's minority languages and Mandarin, respectively. We aim at developing new approaches to China's minority languages identification and designing training and recognition algorithms to treat them based on this speech corpus without phonetic transcription. Using this speech corpus, we will propose the LID methods for all China's minority languages with the Chinese loanwords. The work will promote the research and application of language identification technology in China.

In our work, we select 9 typical China's minority languages and Mandarin to construct our telephone speech corpus. These minority languages are composed of Naxi, Miao, Bai, Dai, Yi, Zhuang, Uygur language, Mongolian, Tibetan [7]. Each minority language represents its minority nationality. The corpus can be used to study, develop, evaluate and compare minority languages identification algorithms. Moreover, it will promote the Linguistic researchers to pay more attention to our national minority.

The rest of this paper is organized as follows. In section two, we briefly review the common and public-domain corpora and language recognition evaluation. Third part will describe the speech collection process. Then we describe our minority languages corpus in the fourth part. Part five introduces the post processing to our corpus. In the last we present conclusions and proposals for future work.

II. THE COMMON AND PUBLIC-DOMAIN CORPORA

Research in automatic language identification from speech has a history extending back to the 1970s. It was difficult to compare the performance of those language identification systems, as few of the algorithms had been evaluated on common corpora. To support and promote the research in these areas, which include language identification, multi-language speech recognition, word spotting and speech-to-speech automatic language translation, Muthusamy designed the Oregon Graduate Institute Multi-Language Telephone Speech (OGI) Corpus in 1992 [5][6]. The OGI corpus contains 100 different speakers in each of the following 11 languages: English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. From 1993-1996, the National Institute of Standards and Technology (NIST) has sponsored the evaluation of language identification systems using the OGI speech corpus at first, And then, Starting in 1996 the NIST evaluations have employ the Linguistic Data Consortium's CallFriend corpus which includes Arabic and the same languages as OGI corpus [11]. From then on, the language identification systems have become more complex. Its performance is more accurate. It is proved that Both OGI

and CallFriend corpora are standard and public-domain database in language identification area. Relying on the standard corpora, NIST has conducted a number of evaluations of language recognition technology, most recently in 2007. The 2007 evaluation is similar to those previous evaluations. The most signification differences are the increased languages and dialects. There are 26 languages and dialects which contain Cantonese, Mandarin (Mainland and Taiwan), Min and Wu dialects. The tasks of evaluation have some differences and researchers pay more attentions to the standard language corpora' dialects [4].

It is no doubt that China is a unified multinational country with 55 ethnic minorities and each nation has a long history. Most of minorities have its own language and writing system. We must know about its language and let others understand what minority it is, and then the minority's long history and brilliant culture will be known to the outside world. Research in multi-minority languages identification would be enhanced by the availability of minority languages corpus.

III. COLLECTION PROCESS

The speech is collected using DN081 Voice Board and the touch-tone phone. The speech was sampled at 8000 samples and 16 bit resolution. The device was programmed to answer the telephone. Moreover, We provided the toll-free telephone number. The software of recording works on the windows XP/NT operating system and program is run by Microsoft Visual C++6.0 and Microsoft Access2003.

The collection process was automated, once the recording protocol and the corresponding equipments were provided. The speech data could be recorded with native language speaker. Before we record the new minority language, we must contact standard native speaker to pre-record the instructions and promotes. Then, the sound recording of each minority language would be supervised by the native speaker. The sound recording process is showed in Fig1. Brief greeting was received once caller got into our collection system and heard the following prompts in each language to select a language which was represented with a digit from 0 through 9. All subsequent instructions and prompts which were recorded by the native speaker were given in the target language.

Fig.2 displays the working interface of our recording system. The recording system is composed of five sub-systems.

1) Caller management system: Input the caller's informance which contain name, minority and phone number, then the data report would be generated using Microsoft Access .

2) Sound recording system: to finish the tasks of automatic recording process, speech endpoint detection and speech storage.

3) Speech management system: speech Combining, copying, backuping, installation and so on.

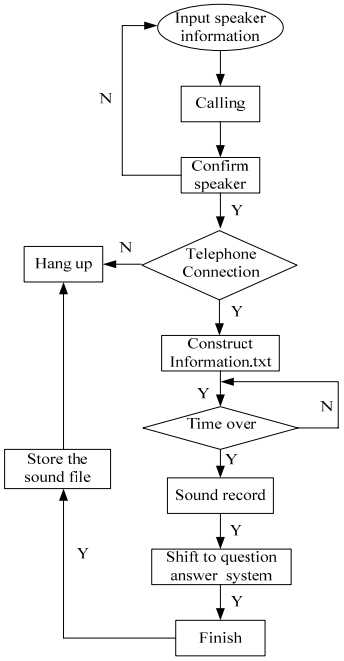

Figure 1. Sound recording process

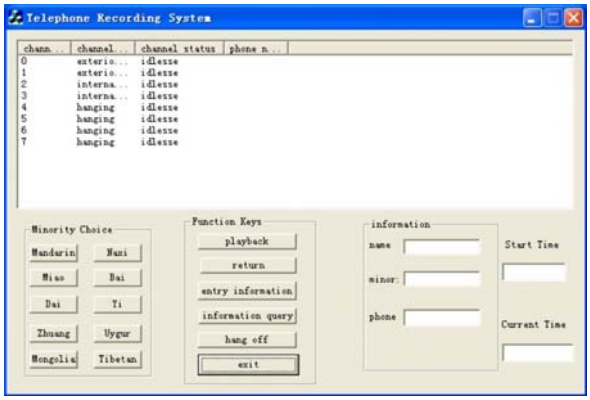

Figure 2. Working interface of recording system

4) Searching and analyzing system: searching each language speakers.

IV. MINORITY LANGUAGES TELEPHONE SPEECH CORPUS

In order to research the china's minority languages and Mandarin, we construct a new multilingual telephone speech corpus consisted of speech taken from telephone conversations in 9 typical China's minority languages and Mandarin. Each language utterances are collected from 50 different speakers. The initial phase of six languages data acquisition and processing were completed in May 2010. They are Naxi, Miao, Bai, Uygur language, Mongolian and Tibetan. The next four minority languages' recording is underway.

A. Choice of Minority Languages

Our telephone speech corpus chooses 9 typical China's minority languages and Mandarin. The minority languages are composed of Naxi, Miao, Bai, Dai, Yi, Zhuang, Uygur language, Mongolian and Tibetan. On one hand, these languages were selected based on a combination of

linguistic consideration and the availability of native speakers in china. On the other hand, the minority languages represent important geographic and political regions.

According to the fifth national census in 2000 statistics, these minority languages have much more population then other languages. Zhuang has a population of nearly 16.18 million. Naxi has the smallest population in 9 minorities about 310000. Data acquisition has been more convenient to obtain, because Yunnan province has 25 minorities and the minority population occupy the 1/3 of total population. Except Uygur language and Tibetan, other minority languages could be got in Yunnan province, moreover, Naxi is one of unique minority languages in Yunnan.

These languages represent a range of unrelated languages as well as languages from the same sub-family. For example, Mandarin is a Sino-Tibetan language family which includes Naxi, Miao, Bai, Dai, Yi, Zhuang and Tibetan. They belong to different branch of Sino-Tibetan language family. The Sino-Tibetan language family has many characters which could distinguish languages. Each syllable of Sino-Tibetan language family is a fixed tone. It retains part of consonant-consonant and has overlapping forms of the word. Both Uygur language and Mongolian belong to Ural-Altaic language. So, we can believe that it is less confusable to identify languages between Mandarin and Mongolian.

B. Data Acquisition

Most of Speaker who participate the sound recording are college students. Moreover, they are the native speakers of minority languages and they record the speech which is spoken in their native languages.

It is a good reference to construct the minority languages corpus as the OGI corpus. The recording protocol was designed to obtain fixed-vocabularies, short topic-specific descriptions and elicited continuous and free speech.

1) Fixed-vocabularies

- Language name: what is your native language?
- Speaker's gender.
- What is your Zodiac?
- The number: say the number zero through ten.
- Communicative languages: Please sit down; thank you very much; I'm sorry, you're welcome.
- Specific thing: the Sun; the moon; the star; the mountain.

2) Short topic-specific descriptions

- Tell something that you like about your hometown.
- Tell us about the climate in your hometown.
- Describe the room that you are calling from.
- Tell us your minority's festivals.
- Where is your hometown?
- What is the time and data now?
- What is the weather like today?
- What is your favorite sport?
- The members of your family.
- What is your most favorite food?

3) Elicited continuous and free speech

Elicited continuous and free speech was obtained by asking caller to speak on any topic that they want. You could talk anything you like, such as sport, shopping, swimming and so on. The callers were given sever seconds to think over before the sound recording to reduce the number of pauses and mistakes in the continuous and free speech. The duration of the free speech is on the average of 3 minutes.

4) The loanwords

The Chinese loanwords are very popular in present-day China's minority languages. With the development of society, there are more and more novel things happened to minorities, they could not describe using their native minority languages, so the loanwords appear. Loanword is the inevitable result of language contact. Written in the form of direct loanwords usually be regarded as Chinese foreign words, such as album, Story and so on. In our corpus, we also consider the Loanwords, they are

- Household appliances: computer, refrigerator.
- What day is today?
- What is your cell phone brand?
- What is kind of transportation vehicle for you?

In our corpus, each speaker contributed 21 utterances, a total of approximately 390 seconds of speech. Each target language has 50 speakers.

The OGI corpus provided orthographic and fine-phonetic transcriptions which were verified by native speakers. Orthographic transcriptions allow access to the database at the lexical and sentence levels. while our minority languages were difficult to gain. Most of the minority languages speakers just could speak their languages and seldom wrote. To solve this problem, we should develop new language identification algorithms to help study minority languages.

C. Speech post processing

The corpus contained a wide variety of speech due to different speakers, microphone, telephone handset, communication lines, background noise and the languages being spoken. It is enviable that there is some wrong with the utterances, so we should be processed again. Firstly, we should chop to remove the excess background noise at the beginning and the end of each utterance. Secondly, It is necessary to chop the native speaker's utterances, during the sound recording process, each target native language manager need connect to callers and their utterances would have effect on to the caller's speech. The last, the target native language manger must make judgment to the quality and content of speech. They noted the occurrences of the speech which contain breath noise, terrible environmental noise, caller didn't follow the instructions and wasn't native speaker and so on.

V. CONCLUSIONS

We construct a new multilingual telephone speech corpus consisted of speech taken from telephone conversations in 9 typical China's minority languages and Mandarin. The minority languages are composed of Naxi, Miao, Bai, Dai, Yi, Zhuang, Uygur language, Mongolian and Tibetan. Each language utterances are collected from their native speakers.

The sound recording process is divided into two phases. The initial phase of work is to finish six languages data acquisition and processing which were completed in May 2010. They are Naxi, Miao, Bai, Uygur language,

Mongolian and Tibetan. The next four minority languages' recording is underway and will be finished soon.

In this paper, we describe our corpus which aims at doing research to China's minority languages identification. It will promote more researchers to study the minority languages and their long history and splendid culture. Meanwhile the work will promote application of language identification technology in China.

The future works include further research on language identification. The algorithms on language identification will be transplanted to our minority languages corpus.

ACKNOWLEDGEMENTS

We gratefully acknowledge the supports from grants of the program of National Nature Science Fund (No. 60865002) and the school-level Programs of Qujing normal university (No. 2009QN006). Correspondence should be addressed to Jian Yang.

REFERENCES

[1] M. A. Zissman. "Comparison of four approaches to automatic language identification of telephone speech". IEEE Transactions on Speech and Audio Processing, vol. 4.1996.

[2] Y. K. Muthusamy, Etienne Barnard. Reviewing Automatic Language Identification [J]. IEEE signal processing magazine. Oct.1994.

[3] M. A. Zissman, Kay M. Berkling. automatic language identification [J]. Speech communication. 2001, pp.115-124.

[4] The 2007 NIST Language Recognition Evaluation Plan. www.nist.gov/speech/tests/lang.

[5] Y. K. Muthusamy."The OGI multi-language telephone speech corpus," in Proc. ICSLP'92, vol, 2, Oct.1992, pp.895-898.

[6] Y. K. Muthusamy."A Segmental Approach to Automatic Language Identification",PH. D. Thesis, OGI Technical Report, No.CSLU 93-002, Nov. 24, 1993.

[7] Yonghua XU, Jian Yang. " A Telephone Speech Corpus to Minority Languages Identification", NCMMSC'2009, August. 14. 2009.

[8] Yuanyuan Pu, Jian Yang. "A Mandarin Speech Database with Non-native Accent of Yunnan for Speech Recognition", Computer Engineering, pp.87-89. vol.10, 2003

[9] P. A. Torres-Carrasquillo, D. A. Reynolds, and J. R. Deller Jr. "Language identification using Gaussian Mixture Model Tokenization". In ICASSP, Orlando, Fl, USA. 2002.

[10] Pedro A. Torres-Carrasquillo. "Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features". In Proc. International Conference on Spoken Language Processing in Denver, CO, ISCA, pp. 33-36, 82-92 September. 2002.

[11] Alvin F. Martin and Mark A. Przybocki. "NIST 2003 language recognition evaluation". In Proceedings of Eurospeech, pp. 1341–1344. 2003.

[12] E. Singer, P. A. Torres-Carrasquillo, T. P. Gleason, W. M. Campbell, and D. A. Reynolds. "Acoustic, phonetic, and discriminative approaches to automatic language identification". In Proceedings of Eurospeech, pp. 1345–1348. 2003.

[13] William Campbell, Terry Gleason. "Advanced Language Recognition using Cepstral and Phonotactics:MITLL System Performance on the NIST 2005 Language Recognition Evaluation". In Proc. IEEE Odyssey. 2006.

[14] Xiuhua Zeng, Jian Yang, Dan Xu. Approaches to Language Identification Using Gaussian Mixture Model and Linear Discriminant Analysis[C]. IITA workshop 2008, pp.1109-1112

[15] Xiuhua Zeng, Jian Yang, Dan Xu. Approaches to Language Identification Using feature-level and decision-level fusion [C]. Advance in information and system sciences, Volume 3, Number 1−2, pp. 255−261.