Transcribing Southern Min Speech Corpora with a Web-Based Language Learning System†

Jun Cai[1,2], Jacques Feldmar[1], Yves Laprie[1], Dominique Fohr[1], Jean-Paul Haton[1]
[1] *Groupe Parole, LORIA-CNRS & INRIA, BP 239, 54600 Vandoeuvre-les-Nancy, France*
[2] *Dept. of Cognitive Science, Xiamen Univ., 361005 Xiamen, China*
{Jun.Cai, Jacques.Feldmar, Yves.Laprie, dominique.fohr, jph@loria.fr}

Abstract

The paper proposes a human-computation-based scheme for transcribing Southern Min speech corpora. The core idea is to implement a Web-based language learning system to collect orthographic and phonetic labels from a large amount of language learners and choose the commonly input labels as the transcriptions of the corpora. It is essentially a technology of distributed knowledge acquisition. Some computer-aided mechanisms are also used to verify the collected transcriptions. The benefit of the scheme is that it makes the transcribing task neither tedious nor costly. No significant budget should be made for transcribing large corpora. The design of a system for transcribing Min Nan speech corpora is described in detail. The application of a prototype version of the system shows that this transcribing scheme is an effective and economical way to generate orthographic and phonetic transcriptions.

1. Introduction

Southern Min language (a.k.a. Min Nan or Southern Fujian) refers to a family of Chinese dialects which are spoken mainly in southern Fujian, eastern and southwestern Guangdong, both of which are coastal provinces in Mainland China. The geographic distribution of Min Nan also includes Taiwan and some areas in Southeast Asia. It is usually called Taiwanese by residents of Taiwan. In 2005, the total number of its speakers is estimated at 49 million. In common parlance, Min Nan usually refers to Xiamen dialect (better known as the Amoy language) because Xiamen (Amoy) is the principal city of southern Fujian and Amoy accent is considered the most important, or even the standard accent in all variants of Min Nan.

Up to now no practical large vocabulary speech recognition system for Min Nan has been developed in Mainland China. In Taiwan, researches on speech recognition for Min Nan began at the end of the 1990s and some large vocabulary Min Nan speech recognition systems had been successfully developed since then [1, 2]. As the first attempt in Mainland China to develop a Min Nan LVCSR system, the Speech Group at Xiamen Univ. had collected a set of recordings of radio news broadcast in Min Nan for 150 hours and started on building acoustic models and language models for Min Nan speech recognition since Nov. 2007.

Transcribing the Min Nan speech recordings into both orthographic and phonetic forms is a resource- and labor-intensive procedure. Since there is no existing recognition system yet which could generate the transcription automatically, the only way to generate the transcription is manual annotating, which is such a tedious and time-consuming task for human annotators that the annotating procedure is usually error-prone. Until now, only a small portion of the recordings (for 20 hours) has been manually transcribed. Because transcribing speech corpora is intrinsically a task of adding annotations to a set of pieces of information and is similar to the task of labeling images, some basic ideas can be drawn from the Human-Computation-based games for labeling images [3] to deal with the difficulties in transcribing speech recordings. We propose combining the task of transcribing Min Nan speech with the pedagogical procedure of Min Nan language learning, via designing a Web system based on the concept of Human Computation. Other than the image-labeling games from which the players' benefit is only for fun, our Web system provides a platform for Min Nan learners to facilitate their training in listening comprehension and pronunciation.

2. Human computation and its application for labeling images

In computer science, Human Computation (or Human-based Computation, HC) is a technique when a computational process performs its function via outsourcing certain steps to humans [5]. The basic idea of HC is that there are a lot of problems that humans can easily solve but computer can not yet. Some of these problems can be solved by just making good use of human processing power [3, 6].

The technique of HC has been successfully adopted to design interactive systems either to perform tasks which appear to be difficult for computer programs to solve or to collect commonsense knowledge from the general public over the Web [3, 6-10]. Such interactive systems can be designed as computer games which people play only for fun. A typical example is the ESP game which addresses the problem of image labeling [3].

The ESP game combines people's desire to be entertained with the acquisition of meaningful labels for images. It is designed to be played by two randomly paired partners and usually played online by a large number of pairs simultaneously. For players, the goal of the game is to guess what the partner is typing for each image. Once both partners have input the same textual string for a certain image while the image is on their screens, the game moves on to the next image. This situation is described as they "agree on the image". Partners strive to agree on as many images in a fixed time period as they play the game. Every time two partners agree on an image, they get a certain number of points. By providing players with points for each image and bonus points for completing a set of images, the system can reinforce players' incremental success in the game and thus encourage them to continue playing. By this way, the system can collect a set of strings for every image.

Since the players can not communicate with each other, the easiest way for both partners to input the same string is to type something related to the content of the image. The agreement by a pair of independent players implies that the label is probably meaningful. From the perspective of the system, the textual string on which two players agree is typically a good label for the image. Furthermore, if a string has been agree by a lot of pairs of players, the probability that it could be a meaningful label would be high. So, a "good label threshold" (say, 40) can be used to further guarantee the quality of the labels.

This kind of game is usually referred to as the "game with a purpose" (GWP). Actually, such a game runs a distributed computation in people's brains instead of in silicon processors. Different mechanisms, such symmetric verification, asymmetric verification [6] can be used to implement a GWP. To make such GWP systems effective, two important issues should be considered in the design. The first is cheating prevention. If partners are able to communicate with each other, the agreement on images would not be related to their contents and thus meaningless. Similarly, players could cheat by being partnered with themselves or by agreeing on a unified strategy for every image. Several mechanisms, such as randomly pairing the players, pairing the players from different IP addresses, and utilizing pre-recorded game-play can make cheating impossible. The second issue concerns the popularity of the game. It is important to make the game interesting enough to attract a large number of players to participate. If there is not enough number of players, it would be impossible for the system to collect sufficient input strings so as to extract high-quality labels.

3. Design of the system

By drawing the essential idea of GWP, an HC-based Web system has been designed to deal with the problem of Min Nan speech transcription. This system combines the task of transcribing Min Nan speech with the pedagogical procedure of language learning. So the users can use the system to learn Min Nan pronunciation while performing the task of transcribing the speech recordings.

The framework of the system is depicted in Figure 1. The core part is a module for Web-based Min Nan language learning. It is this module that utilizes the technique of HC to realize the speech transcription. The input to the whole system is a set of speech sentences, while the output is the confident transcriptions of them.

3.1. Automatic segmentation

All the speech recordings are pre-partitioned into sentences of around 10s. In order to make the task less difficult, an automatic segmentation module has been developed to partition the input speech into short segments of about 2s. In Min Nan news broadcast, a 1s piece of speech contains on average 4 valid syllables. Like Mandarin, every syllable of Min Nan consists either of one syllable onset plus one syllable rime or of only one syllable rime, and every syllable corresponds to one and only one Chinese character. Therefore, the orthographic transcription of a speech segment of 2s

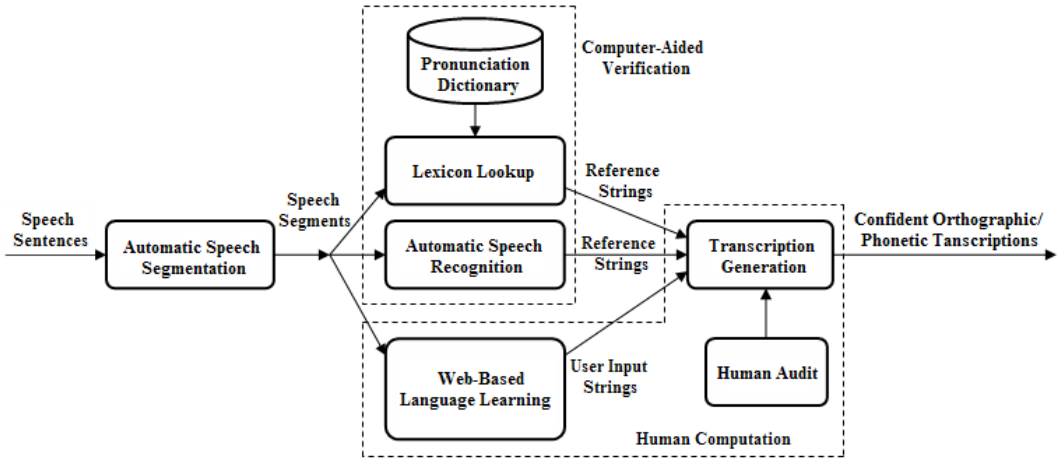

Figure 1. Framework of the human-computation-based transcription system

corresponds to 8 Chinese characters on average, and the corresponding phonetic transcription normally contains about 16 phonetic symbols. Labeling such segments would not be a difficult task for the users.

We use short-time-energy-based voice activity detection (VAD) technique [11] to perform the automatic segmentation. In Min Nan speech, there is normally no stop between the onset and the rime within a syllable. Continuous silence only happens between syllables. So, the end points decided with the VAD technique normally locate between the utterances of different Chinese characters. That implies there is no incomplete syllable in every speech segment generated by the automatic segmentation module.

3.2. User interface

Each speech segment is played to users, enabling them to practice training in listening comprehension and phonetics. In addition, the system collects up the user input labels and stores them in a set of XML files (that will be described later). Figure 2 depicts the Web page (http://59.77.21.117:8080/humanComputation/jsp/minnan.jsp) for Min Nan phonetic training. The users' task here is to listen to speech segments and to input corresponding phonetic symbols. There is an audio player for playing and re-playing each segment. To the right is a textbox which can display the corresponding text (if it has already been stored in the XML file) of the current speech sentence in order to give the user some clues to understand the speech. Under the audio player and the textbox, there is an input box to enable the user to input phonetic symbols. In this system, we adopt "Romanization of Taiwan Min Nan Language Phonetic Alphabet" [12] to label the phonemes of Min Nan. A special keyboard is set up on the Web page to facilitate the input of the phonetic symbols. The only way to input phonetic symbols is to use a mouse to click on this special keyboard. To help users master the pronunciation of each phoneme, we also introduce a "pronunciation" function into the keyboard. By double-clicking a key, the system can play the standard pronunciation of its corresponding phoneme. With the buttons under the keyboard, users can either choose to pass on a difficult segment, or to submit the input phonetic marks to the system. After the user presses the "Done" button, a popup window will appear to show the score of the current input.

Figure 2. Web page for phonetic learning

3.3. Transcription storage

In the system implementation, the speech data is stored as sentences. An XML file is attached to each sentence to store transcriptions and other information about the sentence. Since each sentence is partitioned into segments and what the users transcribe are exactly

these segments, the information in the XML file is organized according to the segmentation of the sentence. The XML schema is shown below.

```
<?xml version="1.0" encoding="UTF-8"?>
<ANNOTATION>
      <UTTERANCE> speech.file </UTTERANCE>
      <LENGTH> number of seconds </LENGTH>
      <TEXT> orthographic transcription of the sentence </TEXT>
      <SAMPLINGRATE> 16 <SAMPLINGRATE/>
      <WORDLENGTH> 16 </WORDLENGTH>
      <ENDIANNESS> little endian </ENDIANNESS>
      <NUMBER_SEGMENTS> n </NUMBER_SEGMENTS>
      <SEGMENT> segment01
         <FILENAME> segment.file </FILENAME>
         <START_TIME> start point </START_TIME>
         <SEG_LENGTH> number of seconds </SEG_LENGTH>
         <LABEL>
            <WORD_LABEL> word level annotation
                  </WORD_LABEL>
            <PHONE_LABEL> phone level annotation
                  </PHONE_LABEL>
         </LABEL>
         <ANNODATA> annotation01
            <WORD_LABEL> word transcription </WORD_LABEL>
            <WORD_CONFIDENCE> m </WORD_CONFIDENCE>
            <PHONE_LABEL> phonetic transcription
                  </PHONE_LABEL>
            <PHONE_CONFIDENCE> m </PHONE_CONFIDENCE>
         </ANNODATA>
                  …
         <ANNODATA> annotation20
            <WORD_LABEL> word transcription </WORD_LABEL>
            <WORD_CONFIDENCE> m </WORD_CONFIDENCE>
            <PHONE_LABEL> phonetic transcription
                  </PHONE_LABEL>
            <PHONE_CONFIDENCE> m </PHONE_CONFIDENCE>
         </ANNODATA>
      </SEGMENT>
            …
            …
      <SEGMENT> segment10
         <FILENAME> segment.file </FILENAME>
         <START_TIME> time of start point </START_TIME>
         <SEG_LENGTH> number of seconds </SEG_LENGTH>
            …

      </SEGMENT>
</ANNOTATION>
```

The segments are stored in at most 10 SEGMENT elements. Several child elements within each SEGMENT element store the file name, the start time, and the time duration of a segment. Transcriptions input by the users and their related confidence values are stored in the <ANNODATA> child elements. For each segment, 20 different transcriptions can be stored in the XML file.

3.4. Computer-Aided verification

For an input string, the GWP uses the number of players who agree on it to decide its quality. Though this mechanism is useful to collect common sense facts and knowledge, the Web-based game itself can not ensure the correctness of the collected information. Sometimes, a common sense could be a common error in spite of its popularity. To prevent the system from outputting totally wrong transcriptions, we have introduced automatic speech recognition (ASR) module and a lexicon lookup module to facilitate computer-aided verification of the input transcriptions.

The ASR module is built on the transcribed 20-hour subset of the speech data. The word level correctness of the recognition is about 80%. Every speech sentence is fed into the ASR module for being recognized automatically. For each sentence, the module outputs a Chinese character string. Afterwards, the lexicon lookup module generates a phonetic string according to the character string. The system uses these two strings as references to verify the quality of the user input strings of characters and phonemes. The Maximum Substring Matching algorithm [13] is used to compute the word error rate of the user input string. We define the consistency between the input string and its reference as follows:

$$\text{Consistency} = 1 - \text{word error rate} \quad (1)$$

The consistency of an input string can be used as a measure of its quality. In the current implementation, input strings with the consistency less than 50% are refused by the system. Thus a low-quality transcription has no chance to appear in the results of human computation, even though the transcription is a common sense.

3.5. Generating transcriptions

By collecting up a large amount of user inputs, the system utilizes HC techniques to generate orthographic and phonetic transcriptions of all speech segments. To describe the generation of transcriptions, we still take the procedure of phonetic transcribing as the example. After the user logins the system, the system plays a segment of a sentence to the user and waits for the user's response. If the user has input a phoneme string and pressed the "Done" button, the following algorithm is used to handle the input string.

Algorithm: HC-based Transcribing
INPUT: A speech segment and the corresponding XML file;
An input phoneme string ;
OUTPUT: A score of the input string.
PROCEDURE
BEGIN
(1) Compare the input string with its reference string and compute the consistency between them;
(2) **IF** (the consistency < 40%)
(2.1) Discard the input string ;
(2.2) **Return** a low score (say 1 point);
ELSE
(2.3) **IF** (the input string is new for the segment and there is a free slot in the XML file)
(2.3.1) input string → XML file;

(2.3.2) **Return** an adequate score;
ENDIF
(2.4) **IF** (the input string == a previously stored string)
(2.4.1) Increase the confidence measure of the stored string;
(2.4.2) **Return** a high score according to the confidence measure;
ENDIF
(2.5) **IF** (the consistency > minimum confidence value of the segment)
(2.5.1) Delete the string with the minimum confidence;
(2.5.2) input string → XML file;
(2.5.3) **Return** an adequate score;
ENDIF
(2.6) Discard the input string ;
(2.7) **Return** an adequate score.
ENDIF
END

For each input string, when it is stored in the XML file for the first time, its confidence measure is initialized to its consistency value computed with Eq. (1). Afterwards, every time the same string is input by a different user, its confidence measure is increased by 1. Therefore, the greater the number of users who agree on a transcription, the greater its confidence measure will be. Once every speech segment in the corpus has been repeatedly transcribed by a large amount of users, the best transcription can be decided based on the idea of HC principle.

However, the HC technique itself can not get rid of the situations that common errors survive in the final transcription. To prevent such situations, a human-audit module has been introduced into the system to facilitate human experts to selectively inspect the transcriptions. Combining human audit by experts with human computation by a large mass of people, we can ensure the high quality of the final transcriptions of the speech corpus.

4. Summary and future work

This paper has proposed an HC-based Web application system which is utilized to generate orthographic and phonetic transcriptions of Min Nan speech corpora. The system combines speech data transcription with language learning. It adopts human computation to collect transcriptions from learners of Min Nan language, and uses human audit to further guarantee the quality of the transcriptions. Though the formal version of the system is being under construction, the experimentation of a prototype version shows that the HC-based transcribing scheme is an effective and economical way to collect orthographic and phonetic labels. In this prototype version, 10 sentences are transcribed by 24 students. The accuracies of orthographic and phonetic transcriptions are 100% and 96%, respectively. We believe that if the system is used by a large amount of people, high-quality transcriptions can be generated based on the collected inputs.

Several prospective researches might be pursued in order to further improve the performance and the utility of the system. Firstly, the system for labeling Min Nan radio broadcast speech should be put into practical application oriented to a mass of Min Nan learners. Secondly, more language learning functions should be added into the system to make it more helpful to the learners.

5. References

[1] R-Y. Lyu, Y-J. Chiang, and W-P. Hsieh, "A large-Vocabulary Taiwanese (MIN-NAN) Multi-syllabic Word Recognition System Based Upon Right-context-dependent Phones with State Clustering by Acoustic Decision Tree," Proc. of the 5th International Conference on Spoken Language Processing, Sydney, Australia, Nov. 30–Dec. 4, 1998.

[2] R-Y. Lyu, Y-J. Chiang, and W-P. Hsieh, et al, "A Large-Vocabulary Speech Recognition System for Taiwanese (Min-nan)," Journal of the Chinese Institute of Electrical Engineering, Vol. 7, No. 2, pp. 123–136, May, 2000.

[3] L. von Ahn, and L. Dabbish, "Labeling Images with a Computer Game," Proc. of the ACM SIGCHI conference on Human factors in computing systems, Vienna, Austria, pp. 319–326, 2004.

[4] 16 A. Kosorukoff, "Human-based Genetic Algorithm," IEEE Transactions on Systems, Man, and Cybernetics, Vol. 5, pp. 3464–3469, Oct. 2001.

[5] L. Von Ahn, "Human Computation," http:// video.google.com/videoplay?docid=-82464639809766 35143. July 26, 2006.

[6] P. Singh, T. Lin, and E. T. Mueller, et al, "Open Mind Common Sense: Knowledge Acquisition from the General Public," Lecture Notes in Computer Science, Vol. 2519/2002, pp. 1223–1237, Feb. 2004.

[7] C. Gentry, Z. Ramzan, and S. Stubblebine, "Secure Distributed Human Computation," Proc. of ACM 9th International Conference on Financial Cryptograph and Data Security (FC2005), pp. 328–332, Feb 2005.

[8] L. von Ahn, M. Kedia, and M. Blum, "Verbosity: A Game for Collecting Common-Sense Facts," Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems, Montréal, Québec, Canada, pp. 75–78, 2006.

[9] Amazon.com, Inc., "Amazon Mechanical Turk," http://www.mturk.com, 2005-2007.

[10] E. Dong, G. Liu, and Y. Zhou，et al, "Voice Activity Detection Based on Short-time Energy And Noise Spectrum Adaptation," Proc. of the 6th International Conference on Signal Processing, Vol. 1, pp. 464 – 467, Aug. 2002.

[11] Ministry of Education (Republic of China), "Romanization of Taiwan Min Nan Language Phonetic Alphabet" http://www.ntcu.edu.tw/tailo, Oct. 2006.

[12] X. Huang, A. Acero, and H.-W. Hon, Spoken Language Processing: A Guide to Theory Algorithm and System Development, Prentice Hall PTR, 1st edition, pp. 421, Apr. 2001.