TAIWANESE CORPUS COLLECTION VIA CONTINUOUS SPEECH RECOGNITION TOOL

Yuang-chin Chiang [2], Zhi-siang Yang[1], Ren-yuan Lyu [1]
[1.] Dept. of Electrical Engineering, Chang Gung University, Taoyuan, Taiwan
[2.] Inst. of Statistics, National Tsing Hua University, Hsin-chu, Taiwan
Email: rylyu@mail.cgu.edu.tw, rylyu@ms1.hinet.net; Tel: 886-3-3283016ext5677

ABSTRACT

Corpora, in their different forms for different purposes, have been the bases for modern natural language processing technology. Taiwanese (MinNan), as other language members in the Sino-Tibet family, has been marginalized due to many reasons. One of the consequences of this marginalization is that no standard written script exists, and thus collecting corpus for these languages has been extremely difficult. By (almost) arbitrarily selecting the *hanlor* written script (mixture of *hanzi* and roman characters), we are still facing the problem that only few people are capable of phonetically transcribing a given Taiwanese text. On the other hand, reading a Taiwanese text is easier due to the existence of many commonly used *hanzi*. By recording a person's reading of Taiwanese text, we use a continuous speech recognizer for Taiwanese to automatically transcribe the text, and end up with two kinds of corpora, one in text, one in speech. The accuracy of the automatic phonetic transcription is about 96.05% in syllable count. For marginalized languages, this automatic transcription can be very useful for corpus collection if proper error spotting scheme is implemented.

1. INTRODUCTION.

Corpus has been the basis for modern natural language processing techniques. For different purposes, we need different kind of corpus. For marginalized languages such Taiwanese and Hakka, corpus collection can be formidable. The reasons include: no widely accepted written script, only few self-educated people can write in their native languages, and even less people is capable of the necessary background knowledge for a seemingly simple task such as phonetically transcribing a Taiwanese text.

For Taiwanese, there are at least three kinds of written scripts: all in *hanzi* (Chinese character), all in roman characters, and *hanlor* (mixture of *hanzi* and roman characters). Text written in *hanlor* has the advantages that it is easier to create, easier to read (both writer and reader do not have to learn more Taiwanese *hanzis*), and easier to process electronically (no user-defined characters to worry).

Given a Taiwanese *hanlor* text in electronic form, we have corpus in its raw form. Various language processing tool such as automatic segmentation and automatic phonetic transcription can then be readily applied if e-dictionary exists for Taiwanese. Using a 60K-dictioanary of Daiim input method [1], we are able to transcribe the phones of text with 85% accuracy in base-syllable count. The reasons for this far from perfect performance are that text-to-phone conversion technique is not perfect, that Taiwanese has a vast amount of *hanzis* with more than one pronunciation, and that the dictionary does not cover all the words and characters in the text. Manual correction seems inevitable.

But the lengthy manual correct proves to be costly, if not impossible. Due to the lack of Taiwanese education, it is difficult to find a person who is efficiently capable of this correction, especially for tones.

However, thanks to the easiness of *hanlor* written script, an educated native speaker can read Taiwanese text without much difficulty. By recording the reading of a speaker, we are able to use our speaker-dependent Taiwanese continuous recognizer to transcribe speech data. We thus have two kinds of corpus under our disposal: a text corpus with phonetic transcription, and a speech corpus for further study for recognition. Note that only speaker dependent speech recognizer is needed to serve our purpose of phonetic text corpus collection.

This paper is organized as follows. Section 2 describes our scheme of transcription using continuous speech recognizer. Section 3 discusses some error spotting schemes. Last section is for conclusion.

2. Automatic Phonetic Transcription of Text via Speech and Text Data.

In this section, we will describe our phonetic transcriber using known text and continuous speech recognizer. The performance of the transcriber will be reported.

A Taiwanese *hanlor* text consists of *hanzis* and words in roman characters with the latter representing the sounds. A Taiwanese *hanzi* can have several pronunciations, and even words in roman characters could have pronunciation variation, since Taiwanese is a tonal language and words in roman characters usually is in base-syllable form, that is, syllables without tone mark.

Suppose that we want phonetically transcribed text. An automatic phonetic transcriber could work as follows. Given a Taiwanese *hanlor* text, a speaker's reading of the text is recorded, and a search network of this text is constructed for use with continuous speech recognizer by connecting the possible pronunciations of each word or character. See Fig.1 in appendix for an example. The speech and the network are used for recognition, and then the phonetic transcription reported. This restrictive network can boost the correct rate of the recognition. We then end up with a corpus with phonetic transcription, which in turn can be used to improve the performance of a speech recognizer for more general purposes.

The setup of our recognizer is rather standard. See, for example, [2]. The speech is recorded at 16k sampling rate using a noise-canceling microphone in a relatively quite environment. Speech frames of 16 msec with frame shift 8 msec are used for short time analysis. Each frame is pre-emphasized, Hamming window applied, and mel-ceptrum and delta-ceptrum of dimension 12 each are computed. Together with energy and delta-energy, the feature vector is of 26 components. For the continuous hidden Markov model part, we use both inter-syllabic and intra-syllabic right-context-dependent phones as basic units. Each unit is modeled by a three-states and three-gaussian-mixtures, and trained as a speaker dependent model by a set of speech data on a Taiwanese phonetically balanced corpus. Note that the recognition units are those of base syllable, and thus recognition result cannot have information about tones. We use HTK toolkit [3] for the training and recognition process.

As a comparison, our previously reported Taiwanese large vocabulary recognizer achieve 93.2% syllable correct rate under the same setup [4][5]. And a continuous speech recognizer performs only 55.6% syllable correctness [6] due to the lack of ample corpus to train the language model in the recognizer.

The experiment is performed on the following materials. Three articles from the prose book "The Way of Youth" [7] are read by two speakers. The two speakers also record some speech for our phonetically balanced corpus [4][5] to train the speaker dependent recognizers. The experimental materials are summarized in Table 1.

	Syllable count	Speaker A	Speaker B
Article 1	1497	16.98 min	17.34 min
Article 2	1499	17.34 min	18.02 min
Article 3	2324	26.04 min	26.77 min

Table 1. Experimental Data. Each speaker recorded each article for three times

Those speech data are then recognized using the restrictive network built from the pronunciations variations. Table 2 shows the findings.

	Speaker A	Speaker B	Average
Syllable correct rate (%)	96.06	96.04	96.05
Sentence correct rate (%)	71.59	69.74	70.67

Table 2. Recognition correctness

Note that even with a much smaller network, it is not perfect in the phonetic transcription. A further error analysis shows that most of the errors come from the confusing set such as velar and pre-alveolar consonants. Other minor errors include accent differences, the reading of speaker being incorrect, the pronunciation dictionary not fully covering the correct pronunciation, and combination of two syllables into one due to strong co-articulation.

3. Error Spotting.

Whatever the reason the errors come from, we still need to find some way to relieve the error correction process from having to go through the whole text. Error spotting seems a reasonable direction to pursuit. If the transcriber can issue a warning on the syllable that is possibly incorrect based on certain criteria, and the warning percentage is reasonable, and the warnings include all that are errors, then manual inspection effort can be reduced significantly. Two methods of error spotting are considered in our study.

A. Error Spotting by Likelihood.

By inspecting the average log-likelihood scores (over all speech frames) generated by recognizer for each syllable, we might guess that one with low average log-likelihood might be an error and issue a warning. A simple criterion is to decide a threshold for each syllable, and issues a warning if a given recognized syllable has lower average log-likelihood.

However, to include all the errors in the recognition results, we have to set the thresholds so low that almost half of the syllables need to be inspected, that is, the false alarm rate can reach 50% to include all the true alarms.

Maybe this alarming criterion is too simple minded, but we did not pursuit further.

B. Error Spotting by duplicate speech recognition.

With two or more sets of speech data on the same text, we can compare their phonetic transcription, and issue a warning if there is a difference. By comparing the six sets of transcription results using the above speech data, we find that there are still 1.27% of the syllables that are recognition error but the transcriber gives the same result, that is, our recognizer consistently makes same mistakes not issuing warning. This result is somewhat disappointed.

4. Conclusions.

Given the non-perfect performance of the state of the art speech recognizer based on the hidden Markov model technique, it is expected that our phonetic transcriber will not be perfect. An additional reasonable error spotting mechanism will serve our corpus collection purpose. Although the two error spotting methods in this study are less than ideal, we hope that finer setup in the transcriber can improve the error spotting performance: higher speech sampling rate for better consonants discrimination, better alarm criteria based on the likelihood scheme. For marginalized languages such as Taiwanese, this automatic phonetic transcription scheme can be crucial for corpus collection.

Tone transcription is also an important part in the corpus, especially for Taiwanese. In addition to be a tonal language, Taiwanese is also rich in tone sandhi. Basically tone sandhi in Taiwanese consists of two problems: when and where-to. The question when tone sandhi happens is still not well understood, and a corpus with tone information will help. It is an issue needs to be studied.

Another technique useful for phonetic transcription is using the text-to-speech system. By generating a speech based on the phonetic transcription that might contain errors, and play the speech to ears of a native Taiwanese speaker, one can easily judge if possible errors exist according its naturalness. Note that the speech generation process can be as simple as concatenating monosyllabic speech files. This is a study currently undergoing.

Reference

[1] YuangChin Chiang. *Daiim Input Method for Windows 95/98/NT*. Version 4.1. National TsingHua University, Hsinchu, Taiwan. 1994.

[2] L. Rabiner and B.H. Jung. *Fundamental of Speech Recognition*. Prentice-Hall International. 1993.

[3] Entropic Research Laboratory, Inc. HTK: Hidden Markov Model Toolkit. 1993.

[4] Ren-yuan Lyu, YuangChin Chiang, Ren-jou Fang, Wen-ping Hsieh, "A Large-Vocabulary Taiwanese (Min-nan) Speech Recognition System Based on Inter-syllabic Initial-Final Modeling and Lexicon-Tree Search", *ROCLING XI Conference*, p.139~p.149, Hsinchu. Aug. 1998.

[5] Ren-yuan Lyu, YuangChin Chiang, Wen-ping Hsieh, Ren-zhou Fang, Zhi-xiang Yang, Zong-yi Lin. "A Large-Vocabulary Taiwanese (Min-nan)

Multi-syllabic Word Recognition System Based upon Right-Context-Dependent Phones with State Clustering by Acoustic Decision Tree," *International Conference on Spoken Language Processing*, Sydney, Australia. Nov. 1998.

[6] Zhi-xiang Yang. *An Initial Study On Continuous Taiwanese Speech Recognition*. Master Thesis, Dept. of Electrical Engineering, Chang Gung University, TauYuan, Taiwan. 1999

[7] CunHong Diunn. *The Way of Youth*. DaiLe Publishing, Taipei, Taiwan. (In Taiwanese.) 1994.

APPENDIX

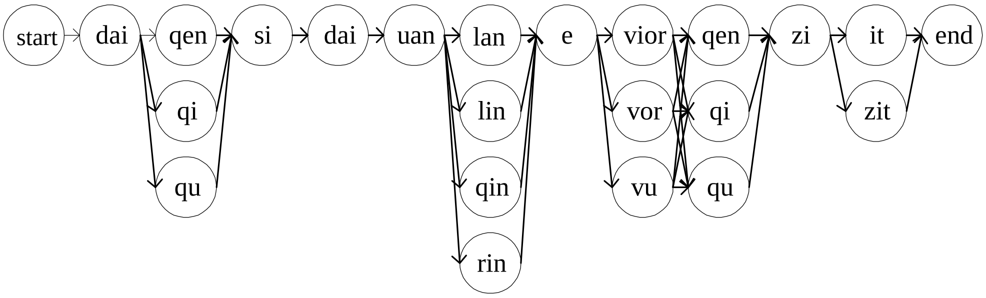

Fig.1. Example network built from a sentence. (Taiwanese is one of the mother tongues of Taiwanese. **台語是台灣人ｅ母語之一。**)