Taiwanese TV News-to-Document Index System

Dau-Cheng Lyu [1, 3], *Bo-Hou Yang*[1, 3], *Ren-Yuan Lyu* [2], *Chun-Nan Hsu*[3]

1. Dept. of Electrical Engineering, Chang Gung University, Taoyuan, Taiwan
2. Dept. of Computer Science and Information Engineering, Chang Gung University, Taoyuan, Taiwan
3. Institute of Information Science, Academia Sinica, Taipei, Taiwan
E-mail: rylyu@mail.cgu.edu.tw, TEL: 886-3-2218800ext5967

This paper describes an index system from Taiwanese TV speech news to World Wide Web Chinese text documents. This system is based on two main techniques: automatic speech recognition (ASR) and bi-lingual text alignment. For the former, we utilized the speech-to-text approach to recognize the utterance of anchors in the TV news as Taiwanese tonal syllable sequences. Then we translated the Chinese text documents which obtained from the corresponding news website to the Taiwanese tonal syllables by a bi-lingual pronunciation lexicon. Afterward, a dynamic programming algorithm is used in the syllable-level alignment for linking the TV news and the documents. A corpus of speech data about 100 speakers and the text data with 840k Chinese characters were used to train the acoustic and language models in ASR. A bi-lingual lexicon contains 70k vocabularies is used as the resource of the pronunciation model for ASR and the statistical translation model for bi-lingual text alignment. Finally, the experiment of the TV news with 40 stories was evaluated for the document index system, and the accuracy rate of index is over 82% on average.

1. Introduction

A. Automatic collecting and transcription speech data:

The corpus is one of the most important materials for the ASR, but most of the released corpora for Chinese languages around the world or even in Taiwan is Mandarin. It is very time and money consuming if we want to collect Taiwanese or Hakka corpora on our own. Therefore, a straightforward approach to this problem consists of collecting audio data and then generating the transcription automatically using a large vocabulary speech recognition system. We only need to process the post-transcription for the news spontaneous speech if we utilize the relative news documents.

B. Bi-lingual text alignment:

The lexicon is also an important resource for speech recognition. In the lexicon, it contains the pronunciation, syntactic and semantic tags, or the translation between colloquial speech and classic literature. We can find more related or new phrase by video-to-text aligning, and the feedback can improve our lexicon for robustness.

C. Video search service:

Massive quantities of audio and multimedia content, such as TV radio and television programs, are becoming available on the internet in the global information infrastructure nowadays [1] [2]. We can easily provide the users with the spoken document retrieval in the website by the video-to-text index system. A searchable index that provides the ability to play the segments of the user's interest within the audio file of these shows would make these archives much more accessible for listeners interested in a particular topic.

The overall architecture of the TV news-to-document index system is shown in Fig. 1.The whole system can be separated into three parts. The upper dotted square of Fig. 1 is the signal processing and automatic speech recognition (ASR). The lower dotted square is the text alignment. The middle dotted square contains the statistic models which are trained from the speech and text corpora. The detail part will be described separately in the following sections.

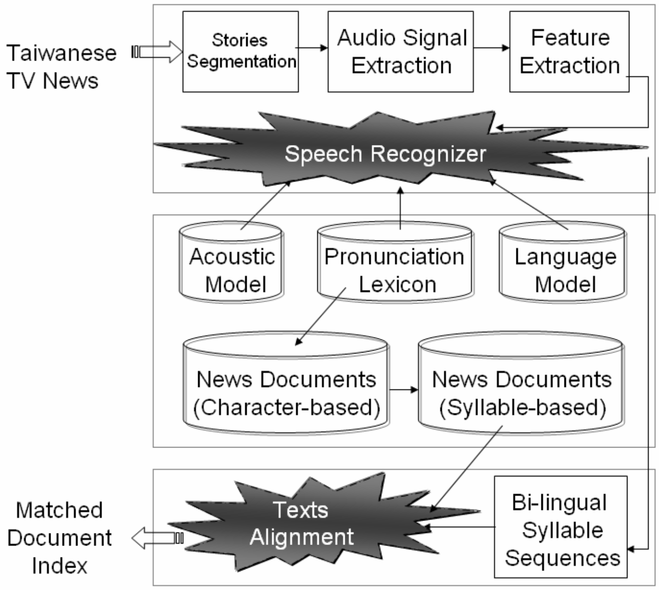

Figure 1. The overall architecture combining ASR and texts alignment techniques for Taiwanese TV news-to-document index system. The input is the Taiwanese video TV news, and the output is the matched document which is obtained from the Internet.

This paper is organized as follows. Section 2 introduces three kinds of data for both Taiwanese and Mandarin. The two main techniques: automatic speech recognition and bi-lingual text alignment will be described in section 3 and 4. After that, the experimental results and analysis are presented in section 5; conclusion and future work are made in the final section.

2. Multimedia Resources

Three types of resources were collected for this work: a bi-lingual speech corpus containing 11.5 hours; the manual segmented Taiwanese TV news data, and a corpus of Chinese news documents automatically obtained from the Internet.

2.1 Taiwanese Speech Corpus

A Taiwanese read speech database produced by 55 male and 55 female speakers over 16k 16 bits microphone was provided in Multi-media Signal Processing Laboratory at Chang Gung University in Taiwan. The statistics of the corpus considered here are listed in <table.1>. The duration of training data is 11.2 hours, and is uttered by 100 speakers for a total of 46086 utterances. For testing, we chose another 10 speakers, with 17 minutes of speech of 1000 words out of a vocabulary of 40 thousands words.

	Training data	Testing data
No. of Speakers	100	10
No. of Utterances	46086	1000
No. of Hours	11.2	0.28
No. of Syllable per Utterance	1.9	2.6

Table 1. The statistics of the Taiwanese speech corpus for acoustic model training and baseline testing.

2.2 TV News Data

The video data of TV news from Formosa Television News (FTVN) were collected during a period of one week in November 2002 for the indexing purposes. The original video data record 1 hour morning news everyday, and the total amount of the data was 5 hours. A section of TV news generally contained four main parts which are: anchors, reporters, interviewees and weather report. An anchor and the following reporter parts are a pair and this pair is also defined as a story. A story almost can find the corresponding document in the FTVN website. Our research goal is automatic finding the linking of the story and the related document. Therefore, we manually segmented the anchor parts at this stage. Totally two anchors with 40 stories were segmented and the whole duration is about 13 minutes. Then, the video stories were extracted the audio signal(right channel) is stored as 16kHz with a resolution of 16 bits for speech recognition. We also manually transcribed this signal to Taiwanese tonal syllables with the help of a language expert. Totally about 2567 syllables are labeled in the transcripts. Those transcriptions were just prepared as the standard answer for the speech recognition in evaluation.

2.3 Text Documents Data

The text resources consist of 840k Chinese characters from FTVN news website [3], and the most of the data are automatic mined twice a day from the November ’02 –December ’02 by a web agent [4]. We collected 7 classes of news, including sport, economics, art, life, social headline, and politics. There are a total of 3079 stories and about 100 stories everyday on average. In addition, the texts were preprocessed to remove undesirable material (table, lists, punctuation marks, etc) and categorized by the date. The texts were then further processed for language model training. First the texts were segmented into sentences and then normalized in order to better approximate a spoken form by the texts analysis module [5]. Because there is no natural boundary between 2 successive words in Chinese characters, we have to transcribe the Chinese characters to phonetic representation by word segmentation. In the cause to align the texts to the video by continuous speech recognition, we used the bi-lingual pronunciation lexicon (will mention in section 3.2) for automatic translating the literature form of Chinese characters to spoken form of Taiwanese tonal syllables. All the statistics of the TV news and text document data are listed in <table.2>.

	TV news		Document
No. of Stories (Duration)	13 (3 min)	27 (10 min)	3079 100(Doc./day)
No. of Sentences	64	182	74k
No. of Characters	928	1969	840k

Table 2. The statistics of the TV news (extracted 2 anchors) and the FTVN text data (Nov. ’02-Dec. ’02)

3. Automatic Speech Recognition Overview

In this index system, ASR plays an important role, and it can be divided into three parts: acoustic, pronunciation and language model. The general formula in speech recognition can be expressed as following:

$$\hat{w} = \arg\max_{w}\{p(w)\cdot\max_{i}[P(v_i \mid w)p(x \mid v_i)]\} \qquad (1)$$

where v_i is the i^{th} multiple pronunciation of the word *w*. $P(x \mid v_i)$ is the acoustic likelihood of pronunciation v_i . The unigram probability distribution of the pronunciations of *w* is given by $P(v_i \mid w)$, subjected to the normalization constraint:

$$\sum_{i}^{N} P(v_i \mid W) = 1 \qquad (2)$$

where N is the total number of pronunciations of word W.

3.1 Acoustic Processing

In this part, we use a single phonetic set to cover all the sounds of Taiwanese and Mandarin speech. The phonetic transcription system called Formosa Phonetic Alphabet (ForPA) [6] was designed to transcribe three major languages (Mandarin, Taiwanese, Hakka) in Taiwan. Sounds which are represented by the same ForPA symbol share one common phoneme category [7]. We have a set of 2878 tonal syllables for Taiwanese.

3.2 Pronunciation Modeling

In the pronunciation modeling, we want to integrate the multiple pronunciations between bi-lingual, and calculate the pronunciation variants with monolingual.

There are two approaches to dealing with pronunciation variations, i.e., the knowledge-based approach and the data-driven approach.[8] The former consists of generating variants by using phonological rules, and the later consists of performing phone recognition to obtain information on the pronunciation variations in the data. We adopt the rule-based approach to the issues of Taiwanese tone sandhi, and the pronunciation variation between spontaneous and read speech. On the other hand, we use a data-driven approach based on confusion matrix for finding syllable mapping between the real pronunciations and canonical pronunciations. In the knowledge-based approach, we use a statistical technique to build a mapping from each character to its multiple pronunciations using the Formosa Lexicon. Then every character has a reliable probability mapping to possible pronunciations. In the data-driven approach, the syllable confusion matrix is constructed by a dynamic-programming technique to align the recognition results of an evaluation data set. We chose the most variational pronunciations by the confidence measure which is the occurrence possibility, and eliminate the relatively small counts. The sum of all the occurrence possibility of each character is then normalized to unity for fair competition in the Viterbi search.

4. Bi-lingual Text Alignment

Bi-lingual text alignment is one of the most important research issues in machine translation for past few decades [9], and an alignment is a segmentation of the two texts such that the *n*th segment of one text is the translation of the *n*th segment of the other [10]. It also assumes that translational equivalence is a relation that can be learned from data [11]. The best translation models are those whose parameters correspond best with the sources of variance in the data. Therefore, we used the bi-lingual pronunciation lexicon as the data for training a probabilistic translation model to translate the Chinese character documents to the Taiwanese tonal syllables.

We use a text analysis model to fit the speaking style of the anchors when translating the written form text documents to the colloquial text. This module contains two main stages: language translation and digital sequence representation. In language translation, we used the bi-lingual pronunciation lexicon with about 70k words as the knowledge source and then adopt a word segmentation algorithm based on the sequentially maximal-length matching in the lexicon. For each segmented word, there may not be only one pronunciation. To deal with the multi-pronunciation problem, a network with word frequencies as nodes and word transitional frequencies as arcs was constructed for each sentence and a Viterbi search for the best pronunciation is then conducted. In the digital sequence analysis, each of almost all Taiwanese single-syllabic words has 2 distinct manners of pronunciation: one for classic literature such as the Chinese traditional poems, and the other for colloquial expression in daily lives. However, for digits, these 2 manners of pronunciation exist in daily lives. Consider the following phrase as an example. "1221 公斤" is pronounced as "1(*zit7*) 千(cing1) 2(*nng2*) 百(bah4) 2(*ri2*) 拾 (zap3) 1(*it7*) 公斤 (gong1-gin1)", where the first "1" and the last "1" are pronounced differently as "zit" (oral) and "it" (classic) respectively, and similarly, "2"'s are pronounced as "nng2" (oral) and "ri2" (classic) respectively. The manner of pronunciation depends on the position of the digit in a sequence, which can be summarized in rules. On the other hands, if a digit sequence does not represent a quantity, it is pronounced digit by digit as the classic pronunciation. For examples, "西元1221 年" is pronounced as "西(se1) 元(quan1) 1(it7) 2(ri2) 2(ri2) 1(it7) 年(ni2) ", where all "1" and "2" are pronounced as their classic pronunciations "it" and "ri", respectively.

After the translation, we used the dynamic time-warping approach to align the ASR results of the anchors' utterances into the FTVN news documents. A story is one-to-one alignment and found the best match if the document and the result has the most hit points in syllable level.

5. Experimental Results

5.1 Experimental Setup and Baseline Syllable Recognition

First of all, the signal processing for speech recognition in feature extraction is applied to a 20 ms frame of speech waveform every 10 ms. For each speech frame, 12 mel-frequency cepstral coefficients (MFCCs), the logarithmic energy and the pitch were extracted, and these coefficients along with their first and second time derivatives were combined to from a 42-dimensional feature vectors [12]. In addition, to compensate for the mismatch of the training and testing data in channel effects, utterance-based cepstral mean subtraction (CMS) is applied. The HTK toolkit was used to train context-dependent Initial and tonal Final models. The HMM topology was three-state, left-to-right without skips. The number of Gaussian mixtures at each state is variable, which depends on the availability of training data for each state. Finally, an average of 8 mixtures per state was obtained. The language model was trained as syllable-level bi-gram by the 840k text data, and the perplexity is 98. The baseline used <Table 1> testing data which contain totally 10 speakers, and the syllable accuracy rate is 68%.

5.2 The Results and Analysis

The results are shown in <table 3>. The candidates of the document in the same day of the TV news for anchor 1 were 104, and 121 for anchor 2. That means the system chose one of the document from 104 candidates for anchor 1 by using the dynamic programming algorithm. For evaluation our speech recognition in spontaneous speech, we used the manual transcriptions as the collect syllable answers for anchor speech, and the syllable accuracy rate are 42%, 40% for both anchors. The results are lower than the baseline, the main reason we thought is that the anchor's speech is spontaneous, and the baseline speech is read. The spontaneous speech is affected by the speakers accent, speaking style. Therefore, the results are always worse than the read speech. In the third row, the correct number of index in anchor 1 is 11 and 22 for anchor 2 when we used the FTVN documents as the candidates. However, the news documents in the FTVN may not be exactly the same as the anchor's speech in TV news. Therefore, we used the manual transcriptions as the candidate documents for another evaluation; the results showed that 100% index accuracy rate are achieved. Obviously, the main error of the index is not caused by the in the ASR error or bi-lingual texts alignment, but is caused by the inconsistency between the anchor's speech and the FTVN documents. We analyzed two examples of the manual and FTVN documents, and the results are shown in Fig. 2. The left side picture (average correlation 0.99) shows the example of the syllable points between the manual and FTVN documents in correct index, and right side picture (average correlation 0.4) shows the example of an error index.

	Anchor 1	Anchor 2
Syllable Accuracy Rate (bi-gram)	42%	40%
Index Accuracy Rate (web documents from FTVN)	85% (11/13)	81% (22/27)
Index Accuracy Rate (manual transcribed document)	100%	100%

Table 3. The results of the syllable accuracy rate and index accuracy rate in different type documents for two anchors.

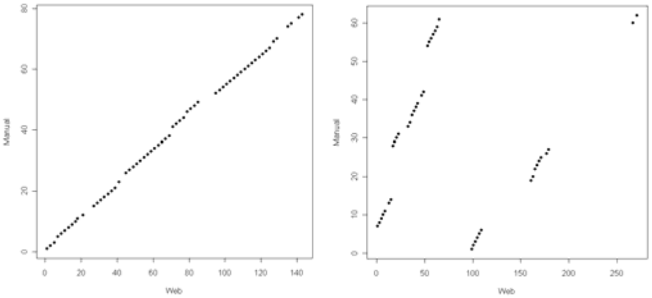

Figure 2. The index examples for the corresponding syllable points between the manual and web documents.

6. Concluding Remarks and Future Work

This paper presented the initial results of the Taiwanese TV news that was aligned to the FTVN news documents index system. Due to the popularity of the internet and multimedia, this research area has become very important. We are successful in integrating the automatic speech recognition and bi-lingual text alignment in the index system, and the results also showed that the index accuracy rate can achieve over 80%, and the most error is due to the mismatch between the anchor's speech and the FTVN news documents.

In the future work, firstly we will develop the automatic segmentation technique for the TV news, and collect more data for the Taiwanese speech recognition. The bi-lingual translation model must be improved for strengthening the text alignment, and reduce the mismatch of the anchor's speech and the web news documents.

7. Reference

[1] Hsin-min Wang, "Experiments in Syllable-based Retrieval of TV News Speech in Mandarin Chinese," Speech Communication, 32(1-2), pp. 49-60, Sept. 2000

[2] Jean-Manuel Van Thong, et al, "SpeechBot: a Speech Recognition based Audio Indexing System for the Web," In Proc. International Conference on Computer-Assisted Information Retrieval (RIAO), 2000

[3] http://www.ftvn.com.tw/

[4] Chia-Hui Chang, Harianto Siek, Jiann-Jyh Lu, Jen-Jie Chiou and Chun-Nan Hsu, "Reconfigurable Web wrapper agents" IEEE Intelligent Systems, 18(5):34-40, Special Issue on Web Information Integration, September/October 2003

[5] Ren-yuan Lyu, et al., "A Taiwanese (Min-nan) Text-to-Speech (TTS) System Based on Automatically Generated Synthetic Units", ICSLP2000, Oct. 2000

[6] Min-siong Liang, et al., "An Efficient Algorithm to Select Phonetically Balanced Scripts for Constructing a Speech Corpus," 2003. Beijin, China

[7] Dau-Cheng Lyu, et al., "Speaker Independent Acoustic Modeling for Large Vocabulary Bi-lingual Twaiwanse/Mandrain Continuous Speech Recognition," In Proc. SST 02, Melbourne, December 2002

[8] Judith M. Kessens, Catia Cucchiarini, Helmer Strik, "A data-driven method for modeling pronunciation variation" Speech Communication, Vol 40, pp. 517 – 534**,** June 2003

[9] Jean Veronis, "Parallel Text Processing Alignment and Use of Translation Corpora," Kluwer Academic, 2000

[10] I. Dan Melamed, "Empirical Methods for Exploiting Parallel Texts" 2001

[11] Cheung, L. Y. L., et al., "Some Considerations on Guidelines for Bilingual Alignment and Terminology Extraction", The 1st SIGHAN Workshop on Chinese Language Processing, COLING, Taipei, August 2002

[12] Dau-Cheng Lyu, et al, "Large Vocabulary Taiwanese (Min-nan) Speech Recognition Using Tone Features and Statistical Pronunciation Modeling" In Proc. EuroSpeech, Switzerland, 2003.