A Bi-lingual TV News-to-Document Indexing System

Dau-Cheng Lyu [1, 3], Ren-Yuan Lyu [1], Yuang-chin Chiang[2], Chun-Nan Hsu[3]
[1] Chang Gung University, Taiwan
[2] National Tsing Hua University, Taiwan
[3] Academia Sinica, Taiwan
{daucheng chunnan}@iis.sinica.edu.tw, rylyu@mail.cgu.edu.tw

ABSTRACT

Based on a bi-lingual speech recognition engine and a parallel text alignment technique, we initially developed an audio search engine for indexing bilingual broadcast news to documents found on the World Wide Web (WWW). First, the automatic audio segmentation technique is used for extracting the anchor streams, and then the speech recognizer transformed the bilingual audio data to tonal syllable sequences. Secondly we utilized the parallel text and text-to-phone processing to translate the Chinese characters to the language dependent spoken style tonal syllable sequences. Afterward, the dynamic programming algorithm scored both sequences, and output the index number as the answer. Our system finally indexed both 22 Mandarin and 40 Taiwanese news on a wide range of topics, and the performance is achieved 82% accuracy rate.

Keyword：bi-lingual, ASR, index system, text alignment

1. INTRODUCTION

While more and more multimedia data, including audio, video, text and image are captured and stored on the standalone computers and the Internet, effective automatic indexing and retrieval systems are needed in order to fully utilize those information [8][11]. The techniques of audio classification and segmentation are the pre-requisites to automatic indexing and retrieval.

A number of academic groups have built indexing systems based on speech recognition technique recently [1][12]. In this paper, we deal with two languages (Mandarin and Taiwanese) simultaneously, and do the parallel text alignment between the Taiwanese spoken documents and the Chinese written documents. The overall architecture of the proposed index system was shown in Figure 1. The procedures are as follows: the spoken signals of both Taiwanese and Mandarin TV news, pass the video-to-audio extraction and automatic audio segmentation processing; the anchor's speech of the signal will be extracted from the audio stream. After processed by Automatic Speech Recognition (ASR) engine, the anchor's speech will become syllable sequences. In another part, the written documents were obtained from the Internet by a robot agent [4]; using the text-to-phone processing to transform [21] the read-style documents to the syllable sequences. Afterwards, the parallel text processor [9] will align the two syllable sequences; finally the output is the matched index document, which is the most relative to the input spoken and reports in the TV news. On the other hand, in order to evaluate the performance of our system and error analysis, all the TV content are manually annotated to the syllable sequences by the human transcription (the middle part of Figure 1).

The three purposes of this project are: 1. Automatic collecting spontaneous speech data form Internet, because in recent years, amount of audio and video data can be easily acquired from the Internet. Therefore, we can rapidly build our multimedia data by way of the ASR technique to transcribe the spoken documents. 2. We can renovate the bilingual lexicon for the parallel text alignment by the results of the transcription of the TV news and the WWW documents, therefore the relationship between spoken style pronunciation for Taiwanese and read style Chinese phrases can be discovered. 3. As the growing of the data which is obtained from the Internet, we can do the TV news retrieval.

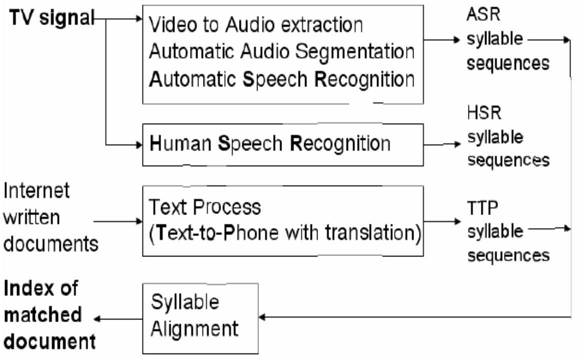
TV signal
Video to Audio extraction
Automatic Audio Segmentation
Automatic Speech Recognition
ASR syllable sequences
Human Speech Recognition
HSR syllable sequences
Internet written documents
Text Process
(Text-to-Phone with translation)
TTP syllable sequences
Index of matched document
Syllable Alignment

Figure 1. The overall architecture of the index system containing three main techniques: automatic audio segmentation, automatic speech recognition, and parallel text alignment. The input can be video signal of Taiwanese or Mandarin TV news, and the output is documents found in the World Wide Web

This paper is organized as follows. Section 2 introduces three kinds of data for both Taiwanese and Mandarin. The three main techniques: automatic audio segmentation, automatic speech recognition and bi-lingual text alignment will be described in section 3, 4 and 5. After that, the experimental results and analysis are presented in section 6; conclusion is made in the final section.

2. BI-LINGUAL RESOURCES

In order to connect the video news with the corresponding document news, three types of resources were collected for this work: bi-lingual TV news video; a bi-lingual speech corpus for automatic speech recognition and a corpus of Chinese news documents automatically mined from the Internet.

2.1 TV News Data

First of all, the video data of bi-lingual TV news are collected from the Formosa Television News (FTVN)[5] and the Public Television Service (PTVS) were collected [19]. In general, a section of the news in Taiwan has half hours, and contains four main parts which are: anchor's report, interviewers, weather report and advertisements/ music, which is shown in Figure 2. The distributions of these five parts are 20%, 38%, 7%, 25% and 10% for average. In other words, in the half hour's TV news is formed on a series of stories, and a story is composed of the first two parts –anchor's report and interviewers. The interviewer is the more detailed content or extension of the anchor's report. For each story, we can almost find the corresponding document in the websites. In order to recognize the anchor's speech to text using automatic speech recognition, the diction of each anchor's parts of a story in one section of TV news is necessary. The automatic diction technique

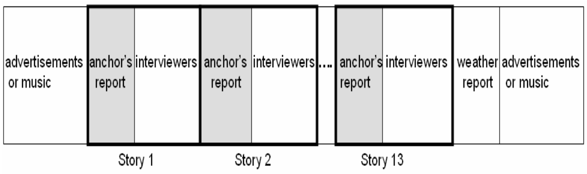

Figure 2. The progress of TV news in half hour, which includes anchor's speech, interviewers, weather report and advertisements/music.

The data is described in the following. In one section of FTVN, two anchors with 40 stories were extracted and the whole duration is about 13 minutes, and in the PTVS, we have one anchor and 22 stories with 6 minute. The video data is recorded as MEPG1 format. However, for the automatic speech recognition, the video data is extracted the audio signal and stored as 16k Hz 16 bits. In addition, in order to have a good performance in speech recognition, the anchor's speech is selected. The reason is anchor's speech is much clear and with little background music or noise. Furthermore, the automatic detecting anchor's speech in one section of T V news is necessary, and the detail is addressed in Section 3. Another, for evaluating the performance of the automatic audio segment, we manually segment the video data, and transcribe the anchor's speech to syllables by a language expert. Totally, there are about 4733 syllables labeled in the transcripts, and those transcriptions were prepared for the standard answer for the speech recognition in evaluation and index error analysis.

2.2 Bi-lingual Corpus

In order to generate a bi-lingual speech recognizer, the language of Mandarin and Taiwanese read speech database is collected. The corpus is totally contained 23 hours with 120 speakers, among which, 100 speakers for training the acoustic model and others for evaluating the performance of speech recognition. In the training set, every speaker records both languages, and in the evaluation set for each speaker, only single language is required. All of the data is recorded with 16k/16 bits microphone in an office environment. The all statistics of the corpus are listed in Table.1.

Table 1. The statistics of the bi-lingual speech corpus for acoustic model training and baseline testing. M: Mandarin, T: Taiwanese

	Training Set		Evaluation Set	
Language	M	T	M	T
No. of Speakers	100	100	10	10
No. of Utterances	43078	46086	1000	1000
No. of Hours	11.3	11.2	0.28	0.28

2.3 Text Documents Data

The written documents consisting of 850k Chinese characters were obtained from FTVN and PTVS news website. Most of the data are automatically mined twice a day during one month in using web agent [2]. The documents is correspond to the TV news. They totally contain 3079 stories, i.e. about 100 stories for one day in average.

There are two main purposes for collecting these documents, training language model of the speech recognition and for aligning bi-lingual texts. The procedures are addressed in the following: First the texts were segmented into sentences and then normalized in order to better approximate a spoken form by the texts analysis module [21]. Second, each sentence have to segment as word or phrase level for transcribing that written in Chinese characters to phonetic representation in both languages. However, there is no natural boundary between 2 successive words in Chinese characters, the word segmentation is used the method of maximum sequence match to find the best segmentation. Third, in the cause to align the texts to the video by continuous speech recognition, we used the bi-lingual pronunciation lexicon for automatic translated the literature form of Chinese characters to spoken form of Taiwanese syllables. All the statistics of the TV news and text document data are listed in Table.2.

Table 2. The statistics of the TV news (extracted only 3 anchors) and the WWW text documents

	Anchor's Utterances		WWW Documents
Language	T	M	M
No. of Stories (total duration)	40 (13min)	22 (6min)	3097 (one month)
No. of Sentences	246	167	75k
No. of Characters	2897	1836	850k

3. AUDIO SEGMENTATION

Although, the TV news can be divided into four parts, the anchor's report is the most important of all. The reasons are because the part is the abstract and essence of each story and the anchor's speech can be almost found in the topic sentence of corresponding document in the WWW. Therefore, in this section is addressed the method that how to automatically detect the duration of anchor's speech in one section of TV news.

The topics of using different features to discriminate the speech, non-speech, music, or environment sound and silence have been studied in [13] [14]. Several approaches have been suggested in the literatures to classify speech and non-speech audio segments [6][16]. In 1990, Prinsloo proposed a method for segmenting syllables and syllable classes in continuous speech. The implementation involves a combination of phonological knowledge, speaker adaptable features, vector quantization and a hidden Markov modeling technique.[15] In 2004, Prasad employed an algorithm, based on group delay processing of the magnitude spectrum to determine segment boundaries in the speech signal. His experiments have been carried out on TIMIT and TIDIGITS databases. The performance in segment boundary is that 70% of the syllable boundaries are within 20% of syllable duration. Furthermore, overall 5% insertions and deletions have also been observed [22].

A good performance of automatic segmentation in speech signal is based on reasonable features and scheme.[17] Therefore, in the feature selection, we used many discriminative features both in temporal and spectral domains of the speech signal, such as, spectrum flux, spectrum centroid and spectrum variance. Besides, 12 mel-frequency cepstral coefficients (MFCCs), normalized log energy, pitch with their first and second order derivatives are also used.

In the segmentation scheme, as shown in Figure 3, we used top-down strategy to extract the element of the news.

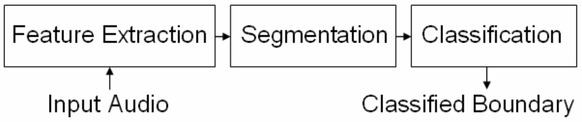

Figure 3. The Segmentation scheme for TV audio news.

The input is the TV audio news stream, using the above features to segment the audio data, then, we have some boundaries. As mentioned before, a section of TV news can be identified four parts, and that also be classified as speech and non-speech parts by the audio signal. In the non-speech parts, these could be music, silence or environment noise. According the research reports [10][25][15], some temporal and spectrum related features are used to get high performance separating speech and non-speech section. The formulas are addressed in the following:

(I). Waveform related Features

Let $x[n]$ be the original speech signal of length L, $x_t[m]$ is the t th frame of length N in $x[n]$

i. Absolute Magnitude,

$$A_t = \sum_{m=0}^{N-1} |x_t[m]|$$

ii. Zero Crossing Rate,

$$Z_t = \sum_{m=0}^{N-1} f\left(x_t[m]x_t[m+1]\right),$$

$$\text{where} \quad f(v) = \begin{cases} 1, \; \mathit{for} \;\; v < 0 \\ 0, \mathit{otherwise} \end{cases}$$

iii. Fundamental frequency (namely, Pitch), P_t, analyzed by the autocorrelation method. [21]

iv. Low Short Time Energy Ratio

$$LSTER = \frac{1}{2}\sum_{m=0}^{N-1}\left[\operatorname{sgn}\left(0.5avEng - STE(m) + 1\right)\right]$$

(II). Spectrum related Features:

Let $P_t[k]$ be the Discrete Fourier Transform for $x_t[m]$, where $0 \leqq k < N$

$$p_{K_t}[k] = \frac{|X_t[k]|}{\sum_{k=0}^{N-1} |X_t[k]|},$$

is the normalized absolute spectrum at t th frame, could be looked upon as the p.d.f. of the random variable K_t, which represents the frequency index, and the energy is located.

v. Spectrum Mean,

$$M_t = \frac{1}{N}\sum_{k=0}^{N-1} |X_t[k]|$$

vi. Spectrum Variance,

$$V_t = \frac{1}{N}\sum_{k=0}^{N-1} |X_t[k]|^2 - M_t^{\,2}$$

vii. Spectrum Centroid,

$$C_t = \sum_{k=0}^{N-1} k \cdot p_{K_t}[k]$$

viii. Spectrum Centroid Moment, (namely, Diversity of spectrum),

$$D_t = \sum_{k=0}^{N-1} k^2 \cdot p_{K_t}[k] - \left(C_t\right)^2$$

ix. Spectrum Entropy,

$$E_t = \sum_{k=0}^{N-1} p_{K_t}[k] \cdot \log\left(\frac{1}{p_{K_t}[k]}\right)$$

x. Spectrum Flux,

$$F_t = \frac{1}{T}\sum_{\tau=t}^{\tau=t+T-1} f_\tau ,$$

where $f_t = \frac{1}{N}\sum_{k=0}^{N-1}\log(|X_t[k]|+\delta) - \log(|X_{t-1}[k]|+\delta),$

and δ is an arbitrary small positive number

There are two examples in Figure 4 and Figure 5 to represent the concept that the features mentioned above are obvious to segment the boundary of speech, music or environment noise.

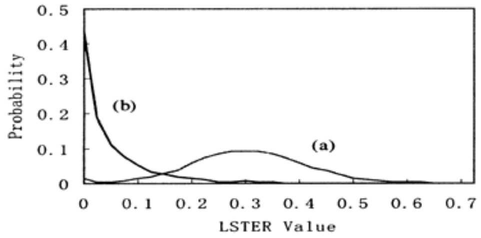

Figure 4. The distribution of LSTER (Low Short Time Energy Ratio) on (a): speech signal and (b): music signal.

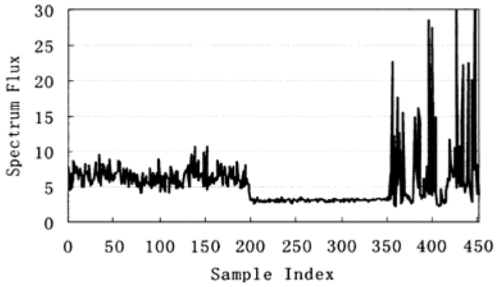

Figure 5. The spectrum flux value on three types of signals. The durations of sample index during 0~200; 200~35; 350~450 are signal of the speech; music and environment noise, respectively.

Therefore, we keep the speech part, and those parts are cut into many fragments by the above features. In order to detect the anchor's speech from those fragments, the speaker-based hidden Morkov model (HMM) is also trained using MFCC features. The duration of one frame is 5 seconds, and shifts one second. Then, some fragments are identified as anchor reports. In general, an anchor's section is continued 10 seconds at least. We merge or discard the fragments that are identified as anchor's speech which has more or less 10 seconds duration.

To compare the results of two different segmentation approaches, we look upon one as reference boundaries and the other as testing boundaries as shown in Figure 6. For each boundary in the reference boundaries, the nearest one in the testing boundaries are picked out and marked as the matching ones. The matching difference (in sec.) is also recorded simultaneously. Those who in the testing boundaries could not be matched to any one in the reference boundaries were marked as the insertion ones. Those who in the reference boundaries could not find an appropriate matching one in the testing boundaries were marked as the deletion ones. It can be easily shown that

$N_{matching} + N_{insertion} = N_{Testing}$,
$N_{matching} + N_{deletion} = N_{Reference}$,

where $N_{matching}$,$N_{insertion}$, and $N_{deletion}$ are the number of the matching, insertion, and deletion boundaries respectively between the reference and testing set, while $N_{Reference}$ and $N_{Testing}$ are the number of the boundaries in the reference set and testing set, respectively.

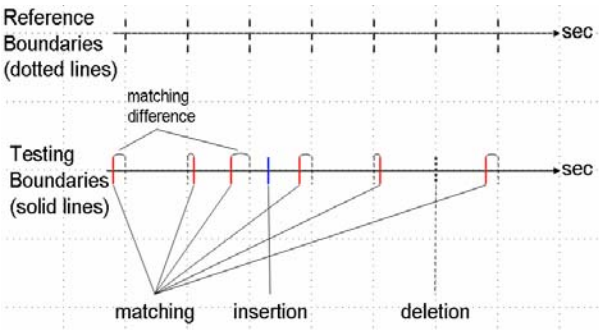

Figure 6. The performance measurement between 2 segmentation approaches by matching, insertion and deletion rate

The insertion rate (I) , deletion rate (D), were thus defined as

$I=N_{insertion}/N_{Testing}*100\%$;
$D=N_{deletion}/N_{Reference}*100\%$;

In order to evaluate the performance of the automatic audio TV news segmentation on our method; we used manual segmentation as the reference. The result is shown in Table 3 and Table 4.

Table 3 Segmentation consistency between manual and automatic where I, D are the boundary insertion and deletion rates

	0.1 sec.	0.2 sec.	0.3 sec.
I (%)	5.40	11.1	9.67
D (%)	2.77	11.1	22.22

Table 4. The Classification results for four types, where A, B, C, and D represent anchors report, interviewers, weather report and advertisements/ music respectively. The first row is reference part, and the first column is testing part.

	A	B	C	D
A	100%	0%	0%	0%
B	5%	90%	5%	0%
C	0%	0%	100%	0%
D	0%	20%	0%	80%

In Table 3 is addressed the insertion and deletion rate of segmentation performance on three different time scale measures, and the latter is represented the confusion matrix for the reference and testing types on the four types of scene in a section of TV news. In the Table 4, the first row is reference part, and the first column is the testing part. The segmentation accuracy rate of the anchor/non-anchor part is achieved over 92.5%, and the average of the boundary variance is about 50 micro seconds (ms).

4. ASR OVERVIEW

The goal of the ASR in Chinese language is to translate acoustic features to character sequences. In this index system, ASR plays an important role, because it continues the video TV news and succeeds to the next stage, text alignment. The core of ASR can be divided into three parts: acoustic, pronunciation and language model as shown in Figure 7.

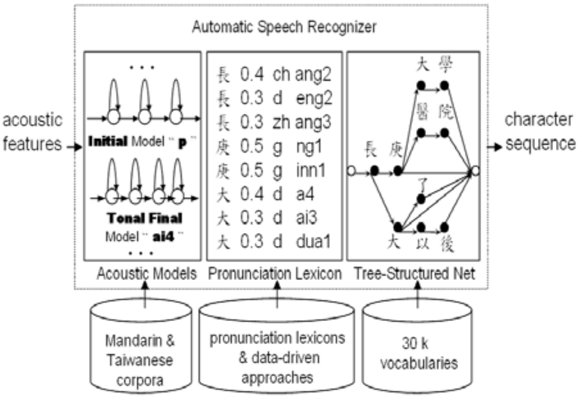

Figure 7. A diagram of automatic speech recognition, including acoustic, pronunciation and language models.

The general formula in speech recognition can be expressed as following:

$$\hat{w} = \arg\max_{w}\{ p(w) \bullet \max[p(v_i \mid w) p(x \mid v_i)]\} \qquad (1)$$

where v_i is the i^{th} multiple pronunciation of the word w. $p(x|v_i)$ is the acoustic likelihood of pronunciation v_i. The unigram probability distribution of the pronunciations of w is given by $p(v_i|w)$, 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.

In this paper, the acoustic is trained using the HHM with the context dependent Initials and tonal Finals on both Mandarin and Taiwanese corpora, which is mentioned in section 2.2. It has been shown that the performance of acoustic models trained by combined speech database from multiple languages is better than that trained by speech data from a single language [23][3]. For this reason, we use ForPA (Formosa Phonetic Alphabet), which is an inventory of phoneme symbols, to transcribe the corpus in these two languages discussed here. Table 5 shows the statistical information of the phonemic inventory in different phonetic levels. Sounds in different languages that are transcribed using the same phonemic symbols in ForPA share the same speech material. Combining two languages in this manner reduces the number of syllables by 21%. In order to easily integrate tone information, we used the context-dependent Initial and tonal Final as acoustic units, and trained these models by sharing the data which belonged to the same acoustic unit. Then, a divisive clustering algorithm was used to create context querying decision trees using four question sets, including an Initial set, a tonal Final set, the set of language properties, and a tonal information set. The above clustering approach could achieve significant improvement compared to previous results [3].

Table 5. The statistic information of all Mandarin (M) and Taiwanese (T) linguistic units in four levels: the numbers of Tonal Syllables (N_{TS}), Initials (N_I), Tonal Finals (N_{TF}), and context-dependent Initial/tonal Finals (N_{CDIF}). ∩and∪ mean intersection and union respectively.

	M	T	M∪T	M∩T
N_{TS}	1288	2878	3519	647
N_I	17	19	22	14
N_{TF}	295	225	416	104
N_{CDIF}	1656	3496	4374	778

Furthermore, in order to more efficiently merge the similar part of the sound for one phoneme or triphone model in both languages, we used a tying algorithm based on a decision tree to cluster the HMM models by using the maximum likelihood criterion [18]. For the question sets, we used phonetic knowledge to design a total of 63 questions, including 10 language-dependent questions, 11 common questions, 28 Initial questions, and 14 Final questions. Then, the tree grew and split as we chose the optimal one among all the questions to maximize the increase in the likelihood scores or the decrease in uncertainty. Finally, the convergence condition was set to halt the growth of the decision tree. The acoustic model used in the experiment depended on the different splitting and convergence criteria adopted

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 deal with pronunciation variations, i.e., the knowledge-based approach and the data-driven approach [20]. The former consists of generating variants by using phonological rules, and the later consists of performing phone recognition to obtain information on the pronunciation variation in the data. We chose the most various pronunciations by the confidence measure which is the occurrence possibility, and eliminate the relatively small counts. The sum of all the occurred possibility of each character is then normalized to unity for fair competition in the Viterbi search.

In the searching net, we use a large-vocabulary tree structured word net, because the perplexity can be reduced in the tree-structured searching net against the linear searching net. This part is trained using the web news documents.

5. BI-LINGUAL TEXT ALIGNMENT

Bi-lingual text alignment is one of the research issues in machine learning for past few decades [9]. It also pointed out that translational equivalence is a relation that can be learned from data. 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 trained a probabilistic translation model to translate the Chinese character documents to the Taiwanese tonal syllables.

In order to get more performance between the spoken style text and document, the text analysis for text-to-speech technique most is included. The analysis module contains two main stages: language translation and digital sequence representation.

In language translation, we used the bi-lingual pronunciation lexicons, ForLex (Formosa Lexiocn) and Gang's Taiwanese lexicon, with about 70k words as the knowledge source and then used a word segmentation algorithm based on the sequentially maximal-length matching in the lexicon. The ForLex is derived from Mandarin lexicon, and thus many commonly used Taiwanese terms are missing. The Gang's Taiwanese lexicon contains Taiwanese expressions from a sampling of radio talk show. A Chinese character in Taiwanese commonly can have a classic (Chinese) literature pronunciation (CLP) and a daily life pronunciation (DLP) [24]. Many of the characters have even more pronunciations. Those lexicons can provide optionally pronunciations for each segmented word. Some statistics of the two sub-lexicons are summarized in Table 6 and Table 7.

In digital sequence representation, 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

For each segmented word, there may be existed not only one pronunciation. To deal with the multi-pronunciation problem, a network with word frequencies as node information and word transitional frequencies as arc information has been 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 oral expression in daily lives. However, for digits, these 2 manners of pronunciation exist in daily lives.

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

Table 6. The number of pronunciation of Formosa bi-lingual Lexicons, including classic literature pronunciation (CLP) and daily life pronunciation (DLP).

	CLP Taiwanese	DLP Taiwanese	Total
1-Syllable	2319	8040	10359
2-Syllable	21337	49222	70559
3-Syllable	7163	11367	18530
4-Syllable	55	15525	15580
5-Syllable	1	711	712
6-Syllable	0	497	497
7-Syllable	0	478	478
8-Syllable	0	195	195
9-Syllable	0	3	3
10-Syllable	0	20	20
Total	30875	86060	116935

Table 7. The distribution of Gang's Taiwanese lexicon.

	Taiwanese
1-Syllable	8153
2- Syllable	46587
3- Syllable	13241
4- Syllable	2106
5- Syllable	175
Total	70262

6. EXPERIMENT RESULTS

6.1 Experimental Setup and Baseline Syllable Recognition

The feature extraction, acoustic modeling and pronunciation modeling were used our pervious setup

[16]. 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 was used to train context-dependent Initial and tonal Final models. The language model was trained as syllable-level bi-gram by the 840k text data, and the perplexity is 664. The baseline used Table 1 testing data was totally 20 speakers for Mandarin and Taiwanese, and the syllable accuracy rate was 68%.

6.2 The Results and Analysis

The results are shown in Table 8. It is shown us that the syllable accuracy rate is about 45% for both languages, and those results are much lower than the baseline. The main reason is that the spoken style of the anchors in the TV news is spontaneous speech; however our training corpus is read speech. This kind of mismatch is caused by the speech rate, the co-articulation or accent where has influenced the results of indexing accuracy rate in the Mandarin TV news. The index accuracy rate for Taiwanese is 82% and 86% for Mandarin and the documents is collected form the Internet, and the syllable transcription is automatic transcribed by the TTP. Besides, the candidates (or perplexity) of the documents for the Taiwanese are 89, and 96 for Mandarin.

We initial thought that if we have sophisticated ASR system, that we can improve the performance of the index results. Therefore, we used the transcription of HSR against TTP which is the transcription of the Internet documents. Consequently, we assumed the results of the ASR are totally correct, and those syllable sequences were aligned to the match document from the Internet. However, the results are not improved too much in Taiwanese.

We analyzed the error parts of the documents in Taiwanese, we found that the content of anchor’s speaking is not exactly the same with the Internet documents, because the document in the Internet is represented as the read form of Chinese characters, whereas, the spoken form of Taiwanese is different from Mandarin read form in phrase order or slang. We draw two examples between the syllable level of the Internet document and the anchor’s speaking in the Figure 3. The left picture is correctly indexed if the correlation of the Internet document and anchor’s speaking is high, but the right picture is wrongly indexed if the order of the WWW documents and the anchor’s content are mismatch. Therefore we have replaced some of the WWW documents to the manual transcription, as the result, the index accuracy rate achieved 100% for Taiwanese.

However, the index accuracy rate did not increase as we changed the syllable sequences from the results of ASR to HSR. Therefore the indexing error must be the acoustic signal mismatch, for this reason, we used the MLLR technique to adapt read speech to spontaneous speech by another anchor’s speech as the training samples. Finally, the results encouraged us; the Mandarin index accuracy rate achieved 100%.

Table 8. The results of the index and syllable accuracy rate for two anchors. SAR: Syllable Accuracy Rate; IAR: Index Accuracy Rate. ASR vs HSR: the syllable sequences from ASR against syllable sequences from HSR.

Taiwanese	Mandarin		Perplexity
SAR (ASR vs. HSR)	42%	45%	664
IAR (ASR vs. TTP)	82% (33/40)	86% (19/22)	89/96
IAR (HSR vs. TTP)	85% (34/40)	100% (22/22)	89/96
IAR (ASR vs. HSR)	100%	86%	89/96
IAR (+MLLR, ASR vs. HSR)	100%	100%	89/96

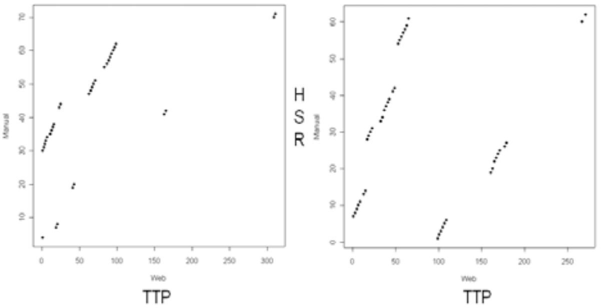

Figure 8. The examples of the corresponding syllable points between the HSR and TTP documents.

7. CONCLUSION

In this paper we have described an initial experimental result of speech recognition and parallel text alignment based audio indexing system which enables all the documents are found on the World Wide Web. We also have integrated both automatic audio segmentation and text-to-speech processing to deal with the index system in advance.

The major part of this paper has successfully included the bilingual text alignment to translate the Taiwanese spoken document to the Chinese character text document. The bilingual pronunciation lexicons have played an important role in the parallel text alignment, and so is the TTS technique to translate the read style document to the spoken style.

The results are shown that the index accuracy rate can achieve over 80%, however, the most error index is the mismatch between the Internet documents and the spoken documents, even though the ASR accuracy rate is 45% in syllable level.

In the future we plan to use this automatic index technique to build a video news retrieval system as the service for the people who has interesting in the video news. Based on the automatic speech recognition technique, our system can be queried by either text or speech input. Therefore, the next obvious direction is automatic collecting large and representative datasets in order to automatically evaluate and build the system using web agent techniques. Finally in order to evaluate our system extensive user experiments are required. User experiments for evaluating our service are planned for the future.

8. REFERENCES

[1] B. Logan, Pedro Moreno, Jean-Manuel Van Thong, Ed Whittaker, "An Example Study of An Audio Indexing System for The Web," 2000, *Proc. of the International Conference on Spoken Language Processing*, 2000, Beijing.

[2] C.H. Chang, C.N. Hsu and J.J. Lui, "Automatic information extraction from semi-structured Web pages by Pattern Discovery" Decision Support Systems, 35(1):129-147, *Special Issue on Web Retrieval and Mining*, 2003.

[3] D.C. Lyu, B.H. Yang, M.S. Liang, R.Y. Lyu, C.N. Hsu, "Speaker Independent Acoustic Modeling for Large Vocabulary Bi-lingual Taiwanese/Mandarin Continuous Speech Recognition," Proc. *of the 9th Australian International Conference on Speech Science & Technology*, 2002, Melbourne, Australia.

[4] E. Jeong and C.N. Hsu, "Integration and Reuse of Heterogeneous XML DTDs for Information Agents," *Decision Support Systems*, 2001, *IAT*, Japan.

[5] Formosa Television News, http://www.ftvn.com.tw/

[6] G. Lu, and T. Hankinson, "An Investigation of AutomaticAudio Classification and Segmentation," *Proc, of the International Conference on Spoken Language Processing*, 2000, Beijing, pp.776-781.

[7] G. J. Prinsloo and M. W. Coetzer, "Automatic syllabification and phoneme class labelling with a phonologically based hidden Markov model and adaptive acoustical features," *International Journal of Computer Speech and Language*, 4(3), July 1990, Pages 247-262.

[8] H.M. Wang, "Experiments in Syllable-based Retrieval of Broadcast News Speech in Mandarin Chinese," *International Journal of Speech Communication*, 32(1-2), pp. 49-60, Sept. 2000.

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

[10] J. Huang, Z. Liu, Y. Wang, "Joint scene classification and segmentation based on hidden Markov model," *IEEE Transactions on Multimedia*, 7(3), 2005, pp.538 – 550.

[11] J.M. Van Thong, et al, "SpeechBot: a Speech Recognition based Audio Indexing System for the Web," *Proc. International Conference on Computer-Assisted Information Retrieval*, 2000.

[12] J. Garfolo, E. Vorhees, C. Auzanne, V. Stanford, and B. Lund. "Spoken document retrieval track overview and results," *Proc. of the 7th Text Retrieval Conference*, 1998.

[13] L. Lu, H. Jiang and H.J. Zhang "A robust audio classification and segmentation method," *Proc. of Multimedia Conference*, 2001, Ottawa, Canada.

[14] L. Lu, H.J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Trans. on Speech and Audio Processing*, 10, October 2002, pp.504-516.

[15] L. Chaisorn and T.S. Chua, "The Segmentation and Classification of Story Boundaries in News Video," *Proc. of 6th IFIP working conference on Visual Database Systems VDB6*, 2002, Australia.

[16] M. Judith, Kessens, C. Cucchiarini, and H. Strik, "A data-driven method for modeling pronunciation variation," *International Journal of Speech Communication*, 40, 2002, pp. 517-534.

[17] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *International Journal of Acoustic Society American*, 58, 1975, pp. 880-883.

[18] P.Y. Liang, J. L. Shen, L. S. Lee, "Decision Tree Clustering for Acoustic Modeling in Speaker-Independent Mandarin Telephone Speech Recognition," *Proc. of the International Symposium on Chinese Spoken Language Processing*, 1998, Singapore, pp. 207-211.

[19] Public Television Service, http://www.pts.org.tw/

[20] R.Y. Lyu, et al, "A Unified Framework for Large Vocabulary Speech Recognition of Mutually Unintelligible Chinese Regionalects," *Proc, of the International Conference on Spoken Language Processing*, 2004, Jeju Island, Korea.

[21] R.Y. Lyu, et al, "A Taiwanese (Min-nan) Text-to-Speech (TTS) System Based on Automatically Generated Synthetic Units," *Proc, of the International Conference on Spoken Language Processing*, 2000, Beijing.

[22] V. Kamakshi Prasad, et al., "Automatic segmentation of continuous speech using minimum phase group delay functions," *International Journal of Speech Communication*, 42(3-4), 2004, pp. 429-446.

[23] Y. Liu, and P. Fung, "Partial change accent models for accented Mandarin speech recognition," *Proc. of the IEEE Workshop on ASRU*, 2003, St. Thomas, U.S., Virgin Islands.

[24] Y.C. Chiang, M.S. Liang, H.Y. Lin, and R.Y. Lyu, "The Multiple Pronunciations in Taiwanese and the Automatic Transcription of Buddhist Sutra with Augmented Read Speech," *Proc. of the European Conference on Speech Communication and Technology*, 2005, Lisbon, Portugal.

[25] Z. Liu, Y. Wang, T. Chen, "Audio feature extraction and analysis for scene segmentation and classification." *International Journal of VLSI Signal Processing Systems*, 1998, 20, pp. 61-79.