Cross-Lingual Audio-to-Text Alignment for Multimedia Content Management[*]

Dau-Cheng Lyu [2,3], Ren-Yuan Lyu [1], Yuang-Chin Chiang [4], Chun-Nan Hsu [3†]

[1] Dept. of Computer Science and Information Engineering, Chang Gung University, Taiwan

[2] Dept. of Electrical Engineering, Chang Gung University, Taiwan

[3] Institute of Information Science, Academia Sinica, Taiwan

[4] Institute of statistics, National Tsing Hua University, Taiwan

Abstract

This paper addresses a content management problem in situations where we have a collection of spoken documents in audio stream format in one language and a collection of related text documents in another. In our case, we have a huge digital archive of audio broadcast news in Taiwanese, but we do not have transcriptions for it. Meanwhile, we have a collection of related text-based news stories, but they are written in Chinese characters. Due to the lack of a standard written form for Taiwanese, manual transcription of spoken documents is prohibitively expensive, and automatic transcription by speech recognition is infeasible because of its poor performance for Taiwanese spontaneous speech. We present an approximate solution by aligning Taiwanese spoken documents with related text documents in Mandarin. The idea is to take advantage of the abundance of Mandarin text documents available in our application to compensate for the limitations of speech recognition systems. Experimental results show that even though our speech recognizer for spontaneous Taiwanese performs poorly, we still achieve a high (82.5%) alignment accuracy.

Keywords: Speech recognition; Audio document retrieval; Cross-language information retrieval; Parallel document alignment

1. Introduction

This paper addresses a content management problem in situations where we have a collection of spoken documents in audio stream format in one language and a collection of related text documents

[*] This paper is dedicated to the memory of Ms. Chin-Lien Grace Wu (1935—2006).

[†] Corresponding author. Phone: +886-2-27883799x1801, Fax: +886-2-27824814, chunnan@iis.sinica.edu.tw.

in another. For example, suppose that we have a large digital archive of Taiwanese spoken documents that are recorded from TV or radio, but their corresponding transcripts are not available. To enable users to access such a digital archive, we can search for related Mandarin text documents from large corpora on the Web and associate spoken documents to text documents with similar contents by parallel text alignment and cross language translation [1-2] to establish the index. Based on this idea, we have developed an automatic approach that aligns spoken documents in Taiwanese and text documents in Mandarin so that we can index and manage the digital archive. This paper describes our approach and its empirical evaluation.

In Taiwan, a rapidly growing volume of digital spoken documents are in Taiwanese. Taiwanese (or Min-Nan) is a "regionalect" [3] in the large family of Han-Chinese languages. Though not an official language, Taiwanese is one of the most popular languages in Taiwan with approximately 82% of the population fluent in it. Even so, there is no standard written form for Taiwanese, so existing text documents in the language are written in many different ways. Although the most popular written form of Taiwanese is based on Chinese characters, not all Taiwanese phrases can be written in Chinese characters. Furthermore, without a commonly agreed standard, many Taiwanese phrases are written in several different ways in Chinese characters.

We need to be creative in managing the growing volume of Taiwanese spoken documents, because existing content management technologies, most of which are based on information retrieval (IR) methods for text-based content, cannot be applied directly to the current problem. Manual transcription using software packages, such as Transcriber [4] and Praat [5], is prohibitively expensive due to the lack of a standard written form for Taiwanese and the lack of skillful human transcribers. Automatic transcription by speech recognition is a possible solution [6-7]. Over the last decade enormous progress has been made in the field of speech recognition such that word-level error rates of speech recognizers can be as low as 10 percent under certain conditions [8]. Therefore, automatic transcription has been widely applied to the content management of spoken documents [9-12]. For example, Cook et al. [13] developed an automatic speech recognition system to manage audio broadcast news documents.

However, applying speech recognition techniques to the automatic transcription of spoken documents is not straightforward, even for English and Chinese, the world's most popular languages. Most spoken documents contain spontaneous speech. Large vocabulary continuous speech recognition currently only achieves high accuracy for read speech, as in dictation systems. Speech recognition for spontaneous and conversational speech is still a challenge, unless the application domain is limited [14]. Since speech recognition for Taiwanese is not as mature as that for English or Mandarin, directly applying speech recognition to automatic transcription may not be a viable solution in our case.

In this paper, we present an approximate, yet effective, solution to the content management problem by aligning Taiwanese spoken documents with related text documents in Mandarin. The idea is to take advantage of the abundance of Mandarin text documents available in our application. Content management systems can index Taiwanese spoken documents by indexing the aligned text document in Mandarin. In other words, with our approach, we can use text documents in Mandarin to represent the content of our spoken documents. Text documents can also compensate for the limitations of speech recognition systems. Although our speech recognizer's performs for spontaneous Taiwanese is not satisfactory (barely above 55%), we still achieve a high (82.5%) alignment accuracy in our experiment.

Aligning cross-lingual multi-media documents (spoken and text) poses a number of challenges. These challenges are summarized as follows:

1. How can we segment a continuous broadcast news session into news stories so that we can align the stories with text documents? How can we determine the boundary between one news story and another in an audio stream?
2. Once we have segmented an audio stream into news stories, how can we determine which text document contains the same news story as its audio counterpart? Since the reports are in different languages, should we define a common language and translate the text documents into the common language? How can we perform such translation? This is a complicated issue because there is no one-to-one mapping between Chinese characters and Taiwanese syllables.
3. How can spoken documents be transcribed into the common language? If speech recognition techniques are applied, how can they deal with spontaneous speech, especially when a broadcast news story may cover a number of domains?

To address these challenges, we integrate various intelligent multi-media content management technologies, including audio classification, segmentation, speech recognition, cross-lingual translation, and parallel texts alignment. We provide a solution to cross-lingual multi-media content management by integrating our pervious research results in the technologies mentioned above [3, 15-19]. Moreover, with our cross-lingual audio-to-text alignment approach, IR-based content management techniques can be applied to manage audio content so that users can access digital audio Taiwanese news archives by sending conventional keyword search queries in Chinese characters.

The remainder of this paper is organized as follows. Section 2 gives an overview of our approach. Section 3 describes how we segment a broadcast news session into a sequence of news stories. Section 4 reviews our speech recognition method. Section 5 explains how to translate a text document in Mandarin into a sequence of Taiwanese tonal syllables. Section 6 explains how to align two tonal syllable sequences. Section 7 reports our experimental evaluation, and Section 8 contains our conclusions and suggestions for future work.

2. System Overview

Our approach to automatic alignment of cross-lingual spoken document and text document consists of four components. Figure 1 illustrates the data flow between these four components.

a. **Segmenter**: The segmenter takes an audio broadcast news session in Taiwanese as input and segments it into *p* individual Taiwanese spoken documents, $s_T^1, s_T^2, \ldots, s_T^p$. Each segmented spoken document corresponds to a news story in Taiwanese.

b. **Speech recognizer**: Next, we apply the speech recognizer to transcribe the content of each spoken document into a written form of Taiwanese tonal syllables. Let the output be $r_T^1, r_T^2, \ldots, r_T^p$ for $s_T^1, s_T^2, \ldots, s_T^p$, respectively.

c. **Language translator**: This is a shallow language translator that provides a word-by-word translation of a Mandarin text document into a sequence of Taiwanese tonal syllables, denoted by $l_T^1, l_T^2, \ldots, l_T^q$.

d. **Alignment system**: The alignment system then aligns two sequences, $r_T^1, r_T^2, \ldots, r_T^p$ and $l_T^1, l_T^2, \ldots, l_T^q$ if their contents match.

In the end, a collection of q text documents in Mandarin, $t_M^1, t_M^2, \ldots, t_M^q$, that report a similar set of news stories as the audio news broadcast session in Taiwanese is identified and aligned with $s_T^1, s_T^2, \ldots, s_T^p$. The final output is produced by tracing back to the pair of original spoken and text documents (s_T^i, t_M^j), based on the alignment results. With these components, we can automatically associate contents in different languages and in different multi-media formats.

Figure 1 also shows the corpora required to train the components. Since the algorithms used in our components are data-driven and model-based, corpora and training algorithms are needed to train the models to achieve the desired performance. For example, the language and acoustic models in our speech recognizer require speech and text corpora, respectively. The speech corpus contains 22 hours of recordings of 100 male and female speakers. In order to process bi-lingual spoken documents, the speech corpus contains bi-lingual speech data with phonetic transcriptions. The text corpus contains news documents in various categories, including sports, politics, business, and technology. The text corpus contains 3,000 documents with more than 850,000 Chinese characters. We use the text documents to estimate the probabilities of two syllables between the contexts so that we can obtain a bi-gram-based language model for the speech recognizer. The corpora for the other components are described in the following sections.

3. Automatic Audio Segmentation

The purpose of audio segmentation is to divide a session of broadcast news into a sequence of individual news stories. In this section, we analyze the structure of a broadcast news session and define our segmentation problem as a sequence labeling problem. We solve the problem by training a Hidden Markov Model (HMM) that identifies homogeneous regions according to their background conditions and the characteristics of their audio signals. We report our empirical evaluation the performance of our audio segmenter.

3.1 Structure of a news session

Most broadcast news sessions follow a similar well-defined structure. Figure 2 illustrates the structure of a typical session of Formosa Television News (FTVN) in Taiwan. According to this structure, we define five types of scenes:

A. **Highlight**: In this type of scenes, which lasts 30 to 60 seconds, the news anchor introduces the top stories that will be covered in this session. At the end of this segment, the anchor typically introduces him/herself and the date of the broadcast.

B. **Anchor's report**: When the anchor is speaking, he/she reports a news story that may last 15 to 60 seconds. In this type of scenes, only the anchor's voice is heard, i.e., there is no background music or noise.

C. **Interview**: The anchor may transfer the session to an on-site reporter or an expert on the topic. A graphic animation and a sound track will often accompany the voice-over of the reporter. At the end of the interview, the reporter transfers the session back to the anchor after naming him/herself and mentioning the location where he/she has been reporting.

D. **Weather report**: In this type of scenes, a weather expert forecasts the weather for the next few days in Taiwan and worldwide. Usually, music is played in the background.

E. **Advertisement**: An advertising scene usually lasts 30-240 seconds and consists of a series of 15, 30, or 60 second TV commercials.

The statistics of the five scenes in a four-hour broadcast news corpus are shown in Table 1. The top two largest scenes are *anchor's report* and *interview*. We define a pair of consecutive *anchor's report* and *interview* as a *story*, as shown in Figure 2 (a). The goal of segmentation is to identify and extract *stories* from the audio stream of a broadcast news session. Table 1 also categorizes the signal texture of different types of scenes as "speech", "music", "background sounds", "silence", "speech with music", or "speech with background sounds."

Since both *interview* and *advertisement* may contain all kinds of signals, it is difficult to distinguish them. But we can identify *anchor's report* easily because it only contains speech signal. Nevertheless, we have to use a set of discriminating features to distinguish these scene types. In the next section, we describe the features that we use to discriminate the above scenes.

3.2 Features Selection and Extraction

To accurately segment an audio stream, we select a set of features that capture both the temporal and spectral characteristics of the audio signal and the characteristics of vocal track. These features are as follows. Let N be the total number of frames in a one-second audio window.

i. **High Zero Crossing Rate Ratio** (HZCRR):

HZCRR has been proved to be useful in characterizing a variety of audio signals and has been widely applied in speech/music classification algorithms [20]:

$$HZCRR = \frac{1}{2N}\sum_{n=0}^{N-1}[\mathrm{sgn}(ZCR(n) - 1.5avZCR) + 1] \qquad (1)$$

$$where \qquad avZCR = \frac{1}{N}\sum_{n=0}^{N-1} ZCR(n), \qquad (1.1)$$

ZCR is the zero-crossing rate in a one-second window, and *sgn(.)* is the sign function.

ii. **Low Short-Time Energy Ratio** (LSTER):

This feature represents variations in short-time energy (STE) and is effective for distinguishing speech and music signals. The idea is that, since there are more silent frames in speech, the LSTER measure will be much higher for speech than that for music [21].

$$LSTER = \frac{1}{2N}\sum_{m=0}^{N-1}[0.5avSTE - sgn(STE(m)) + 1] \qquad (2)$$

$$where \qquad avSTE = \frac{1}{N}\sum_{n=0}^{N-1} STE(n). \qquad (2.2)$$

iii. **Spectrum Flux** (SF):

SF is the average variation value of the spectrum between a pair of adjacent frames in a one-second window. In general, the SF value of speech is higher than that of music [22].

$$SF = \frac{1}{(N-1)(K-1)}\sum_{n=1}^{N-1}\sum_{k=1}^{K-1}[\log(A(n,k) + S) - \log(A(n-1,k) + S)]^2 \qquad (3)$$

where *A(n,k)* is the discrete Fourier transform (DFT) of the *n*-th frame of the input signal, *K* is the order of DFT, and *S* is a small value to avoid numeric overflow.

iv. **Mel-Frequency Cepstrum Coefficients** (MFCCs):

MFCCs are widely used in speech recognition [23] to evaluate the human auditory system and represent speech amplitude in a compact form. They also capture acoustic information about phonemes. In general, a set of MFCCs contains 12 dimensions with a 26-bank mel-filter. A complete description of the process is given in [24].

v. **Mel-Frequency Spectrum Coefficients** (MFSCs):

Music and speech sounds are also different in spectrum domain. Therefore, we designed a method to extract mel-frequency spectrum coefficients from speech signals. This method is similar to the method for SF except for the omission of DFT that translates spectrum signals to cepstrum signals. MFSCs reveal the spectrum characteristics by amplifying perceptually meaningful frequencies that compress 1024 spectrum components into 12 bins.

Consequently, we have a total of 41 dimensions as follows: 12 dimensions for MFCCs, one dimension for normalized energy, plus their first order derivatives. In addition we use 12 dimensions for MFSCs, and 1 dimension each for HZCRR, LSTER, and SF.

3.3 Story/Scene Segmentation

The segmentation process is performed as follows. The audio stream of the input broadcast news session is sliced into overlapping audio frames with a constant window size. We then extract the features of each audio frame, as described in Section 3.2, and send the extracted feature vectors to a trained HMM. Each state in the HMM model represents a scene type and applies a Gaussian Mixture Model (GMM) to estimate the probability that the input audio frame belongs to the scene type given the input feature vector. Based on the estimated probabilities, the HMM model will output the most probable scene sequence. An example is given in Figure 2 (b). Adjacent audio frames in the scene sequence with different scene labels mark the boundaries of the scene types. We can then segment the audio stream at those boundaries as the final result.

There is a strong point to be made for using HMM on top of GMM as the audio segmenter instead of using GMM only. Because GMM does not take the order of the scenes into account, GMM alone may produce an unreasonable scene type sequence where some frames might have a label different from their neighbor frames. If that happens, those frames must be misclassified and should be labeled with the same scene type as its neighbors because an audio stream should be continuous and it is highly unlikely that scene types would change abruptly or frequently. Transition probabilities in HMM can naturally fix this type of errors. Another advantage of HMM is that there is no need to change the topology of a fully-connected HMM model even if the scene type order is changed. For example, suppose that in another TV station, *anchor's report* may be followed immediately by *weather report*, and several *advertisement* sessions may be inserted between *interview*. In this case, we can update the transition probabilities between new scene type pairs to have non-zero values to adapt to the new scene type order.

3.4 Evaluation of Story/Scene Segmentation

We used a four-hour corpus of Taiwanese audio broadcast news to empirically evaluate our audio segmenter. We used three hours of data for training and the rest for testing. Both training and test sets were manually labeled with five scene types. We recorded the audio data in a 16 KHz, 16-bit form and sliced the audio stream into audio frames with a three-second-window duration and a half duration overlap. We then extracted 41 feature dimensions from the audio frames, as described in Section 3.2. We applied the Expectation Maximization (EM) algorithm to train the HMM model, which contains five fully-connected states. The output distributions tied to all states in the HMM model are modeled with a mixture of 64 diagonal-covariance Gaussian densities. With the trained HMM model, we can perform the segmentation by applying the log-space Viterbi decoding algorithm.

Table 2 shows the accuracy of our HMM audio segmenter. The diagonal entries show the number of scenes that were correctly classified into their corresponding scene type, while the off-diagonal entries show the number of misclassified scenes. Overall, our segmenter achieves a classification rate above 95 percent. More importantly, our segmenter can successfully detect the boundaries of news *stories* --- consecutive pairs of *anchor's report* and *interview* scenes.

4. Automatic Speech Recognition

In this section, we describe the algorithm for building a HMM-based speech recognizer comprised of a multi-lingual acoustic model, a bi-tonal-syllable-based language mode and a pronunciation variation model. The contributions of these techniques have been published in our previous works [2, 17, 19]. We also present the baseline result of speech recognition for a bi-lingual corpus.

4.1 Acoustic Modeling

Since we need to handle multi-lingual speech in the acoustic modeling, we use the International Phonetic Alphabet (IPA) scheme to transcribe the results of speech recognition. IPA is an inventory of common phoneme symbols for describing the phonetic pronunciation of almost all languages in the world, and is widely used in multilingual speech recognition systems [25-27]. With IPA, speech sounds perceived as the same in different languages are transcribed with the same phonemic symbols. Table 3 shows the statistics of IPA in different phonetic levels for Mandarin and Taiwanese. According to Table 3, combining two languages in this manner reduces the number of syllables by 21%, implying that, with the same training sample size, the amount of training data available for each parameter increases as the number of parameters decreasing, which improves the robustness of the trained model.

Tone modeling is an inevitable and challenging issue in speech recognition for tonal languages. In order to handle Taiwanese spoken documents, we must identify tone classes. In addition to the MFCC features described in Section 3.2, we also use pitch as a feature. We apply an auto-correlation-based method [28] as our pitch extraction algorithm and use the following 42 features for speech recognition:

- 12 MFCC coefficients
- 1 normalized energy
- 1 pitch feature
- 14 delta coefficients (the first derivatives) of all of the above
- 14 acceleration coefficients (the second derivatives) of the above

To simplify the integration of tonal information, we use the context-dependent *Initial* and tonal *Final* as acoustic units, and train their models by sharing data belonging to the same acoustic unit. According to the 3rd column of Table 3, tonal syllables (N_{TS}) are about 3.5 times more frequent than those without tonal syllables (N_S), implying that with the same training sample size, N_S will get more data than N_{TS}, but using N_S alone to classify the acoustic signals cannot discriminate the difference among tones.

Besides the IPA scheme, we also used decision tree-based states to overcome the problem of data insufficiency. More precisely, we used a tying algorithm based on a decision tree to cluster the HMM models by the maximum likelihood criterion [17, 29]. The decision tree in our algorithm is a binary tree that classifies target objects by asking binary questions in a hierarchical manner. The questions are based on phonetic knowledge and constitute a total of 63 questions, including 10 language-dependent questions, 11 tonal 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, a convergence condition is specified to halt the growth of the decision tree. The acoustic model used in the experiment depends on different splitting and convergence criteria adopted. This approach has the advantage that every possible context acoustic model state can be classified by the tree, so any backing-off models can be avoided. This is crucial to the design of a statistical speech recognizer for multilingual tasks [3].

4.2 Language Modeling

The success of a large vocabulary speech recognition system depends on the effectiveness of its language model. The bi-gram model is a state-of-the-art language model that assumes the probability of a word given the preceding sequence can be approximated by the probability of the immediately preceding word. [30-31].

Usually, a text corpus is used to improve the performance of spontaneous speech recognition in broadcast news because we can use the corpus to estimate the bi-gram probabilities in the language model [32]. Unfortunately, such a text corpus is not available in our case. Therefore, to generate a language model suitable for Taiwanese spontaneous speech, we use the results of our language translator component (see Section 2). The language translator uses a bi-lingual pronunciation lexicon to translate text documents written in read-based Mandarin Chinese characters into Taiwanese spontaneous-based text form, which serves as the text corpus to construct the language model. We describe details of the language translator in Section 5.

Another issue is that, in spontaneous speech, pronunciation variations caused by co-articulation occur frequently and degrade the performance if they are not properly accounted for [33]. We have developed an approach to the issue of pronunciation variations in spontaneous speech based on a statistical lexicon of pronunciation taken from both dictionaries and real speech data [19]. The lexicon contains the pronunciation variations of each word and their probability distributions. The number of pronunciation variations in the lexicon is optimized to balance the tradeoff between coverage and search space.

4.3 Evaluation of Speech Recognition

We trained the HMM-based acoustic models with a 22.5-hour bi-lingual corpus (Mandarin and Taiwanese). The training set of the corpus contains speech from 100 speakers (50 male, 50 female). Each of them recorded the speech in both languages. The evaluation set contains voice of 20 speakers (10 male, 10 female). Each of them recorded the speech in only one language. The corpus was recorded with a 16K/16bits microphone in a lab environment, and the standard HTK toolkit [23] was used as the experimental platform. The statistics of the corpus are listed in Table 4. The final test results of speech recognition in terms of syllable accuracy rate are 70.7% and 66.3% for Mandarin and Taiwanese, respectively. These results were achieved by integrating the tonal acoustic model, the bi-gram language model, and the pronunciation variation model.

5. Cross-lingual Translation

The cross-lingual translator is an important component for aligning Mandarin text documents with Taiwanese spoken documents. The goal of the translator is to translate the Mandarin text into Taiwanese tonal syllables, which can then be aligned with the tonal syllable sequences obtained by applying speech recognition to the spoken documents.

One of the best approaches to effective language translation for human readers is to translate at the sentence level [34]. However, since our goal is automatic alignment, we only perform word-level, or word-by-word translation. Word-by-word translation is also simpler and achieves a higher accuracy than sentence level translation.

5.1 Bi-lingual Lexicons

Our word-by-word translation approach is based on two bi-lingual lexicons.

1. **Formosa Lexicon**, which contains approximately seventy thousand words derived from four major Mandarin newspapers in Taiwan, and was translated into Taiwanese by a linguist.
2. **Gang's Taiwanese Lexicon**, which contains approximately thirty thousand words extracted from a Taiwanese radio talk show. The vocabulary is the same as that used by native Taiwanese speakers in daily conversation. The author, a professional linguist, translated the Taiwanese vocabulary into Mandarin.

In both lexicons, each entry contains word and phrase level Chinese characters accompanied by their tonal syllabic pronunciation in both Mandarin and Taiwanese.

A Chinese character, which corresponds to a single-syllable word in Taiwanese, usually has two distinct types of pronunciation: one for classic, formal literature, such as poems, and the other for oral, colloquial expressions used in daily conversation. However, the distinction between the two types is much more complex than formal versus colloquial and can be context dependent. For example, the first “one” and the second “one” in “one hundred and one” are pronounced differently in Taiwanese. *Formosa Lexicon* provides a set of rules for different types of pronunciation.

Also, care must be taken when translating a text document written in the reading style of Mandarin to the spoken style of Taiwanese. *Gang's Taiwanese lexicon* contains both styles in Chinese characters. We give some examples in Table 5. For example, in the first row, the first column shows the reading style of Mandarin for the phrase “very long time” that may appear in a text document. Our goal is to translate it to the spoken style Taiwanese, appearing in the second column, though both columns are written in Chinese characters.

5.2 Segmentation

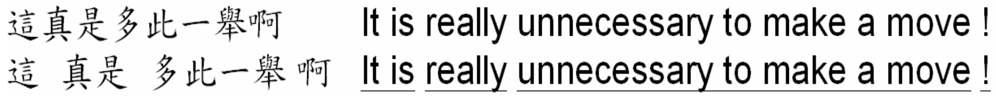
Since Chinese does not have explicit word delimiters (such as spaces used in English), we must use an explicit word segmentation process for Chinese text documents. However, the definition of a “word” in Chinese is ambiguous, as the examples of word segmentation ambiguity given in Table 6, where the same character sequence can be segmented into three different word sequences with valid syntax and semantics.

In our previous work, we developed a word segmentation algorithm called the *sequentially maximal-length matching algorithm* [15], which resolves most ambiguity problems. In an experiment on 28 Taiwanese articles, our algorithm achieved 87.2% accuracy on average. To objectively evaluate our segmentation algorithm, the articles were manually segmented by two Taiwanese linguistic experts.

5.3 Translation

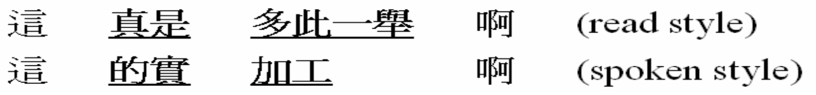
The complete translation procedure is as follows:

1. **Segmentation**: Given a text document in Mandarin, each sentence in the document is segmented into words by the sequentially maximal-length matching algorithm based on two bi-lingual lexicons. For example:



這真是多此一舉啊 It is really unnecessary to make a move !
這 真是 多此一舉 啊 It is really unnecessary to make a move !

2. **Monolingual translation**: In this step, the reading style words are mapped to the spoken style words. Both the input and output are in Chinese characters. For example:



這 真是 多此一舉 啊 (read style)
這 的實 加工 啊 (spoken style)

3. **Cross-lingual translation**: Finally, based on the lexicons, the spoken style words in Chinese characters are translated into a sequence of Taiwanese tonal syllables.

As mentioned earlier, there may be more than one pronunciation for a segmented word. To search for the best pronunciation, during the last step, our translator constructs a network in which the nodes are syllables with their word frequencies and the links are transitions between the syllables weighted by the word transitional frequencies. We then apply the Viterbi search algorithm, as in HMM, to search for the most likely pronunciation sequence.

6. Alignment

After the introduction of the serial of techniques to translate the Taiwanese spoken documents to Chinese character-based text documents, we are ready to build connections between these documents to text news documents extracted from the Web.

Document alignment associates document pairs that cover similar stories, events, etc. In our case, the purpose of document alignment is to match pairs of sequences in Taiwanese tonal syllables. One of them is derived from the spoken document of a news anchor’s speech and is the output of the speech recognizer. The other is taken from a text document translated from Mandarin to Taiwanese.

In order to align these sequence pairs, we apply the *Dynamic Time Warping (DTW)* algorithm [36], which searches for an optimal match between two sequences that may contain stretched and compressed sections. DTW is also commonly used as a distance measure between sequences and widely applied in the fields of speech recognition, image processing, and bioinformatics [37-38]. The DTW algorithm creates an *m*-by-*n* matrix, where *m* and *n* are the lengths of the input sequences. The warping between the two sequences can then be used to find matched regions and determine the similarity between them. The shortest path through this matrix is proved to be an optimal alignment of the input sequence pair. Intuitively, the distance of this path is a measurement of the similarity of the two sequences; the shorter this distance, the more similar they are; thus, for identical sentences, the distance will be zero. To compute the shortest path, dynamic programming is used. Dynamic programming can efficiently compute the distance of the matched regions in the sequence pair recursively in advance and cache the cumulative results. The shortest path from a syllable in one of the sequence to the other sequence is simply from the position of the given syllable to its closest neighboring matched syllable in the other sequence. In order to match syllables, indicators, such as proper nouns, numbers, dates, etc, are vital clues. Besides, term similarity as in latent semantic analysis can also be applied. Therefore, if two sequences refer to the same news story, they will have a minimum DTW distance measure and will be deemed related. In this case, the corresponding documents of the two syllable sequences, i.e., the spoken and text documents will be aligned together.

Figure 3 shows two alignment results, where the horizontal axis represents the syllable positions translated from a text document and the vertical axis represents those from a spoken document. A spot at <*i*,*j*> on the chart means that the syllable in position *i* of the translated text document matches the syllable in position *j* of the spoken document. In fact, both examples are alignment results of spoken and text documents that refer to the same news story. As we can see, in some cases, such as the right chart, the order of syllables is quite different in spoken Taiwanese and reading Mandarin even if they refer to the same news story.

7. Experimental Evaluation

We divided the experimental evaluation of our approach into two stages:

1. Evaluation of how well our components prepare sequences of Taiwanese tonal syllables for alignment. That is, we evaluated the performance of the speech recognition component and the language translation component.
2. Evaluation of the accuracy of the alignment, given imperfect sequences.

The purpose is to evaluate the feasibility of our cross-lingual alignment approach for managing multi-media cross-lingual contents.

7.1 Experiment Setup

We designed the experiment procedure as shown in Figure 4. To perform stage 1 (S1), we used a collection of 40 spoken documents (audio streams) containing a news anchor’s speech. The collection of spoken documents contains three thousands words according to manual transcription. The outputs S1 and S2 are Taiwanese tonal syllable sequences. S1 was extracted by the speech recognizer, and S2 was derived from the manual transcription by Taiwanese linguistic experts.

Next, we applied our Web information extraction algorithm [39] to crawl four hundred text documents from FTVN's web site for news stories in the same week as when the news stories of the spoken documents occurred. The purpose was to evaluate whether the alignment component can correctly match spoken documents with a large number of text documents. Among the 4,000 text documents, 40 report exactly the same news stories as the 40 spoken documents. A bi-lingual expert translated these 40 text documents and the results are denoted S4, while all 4,000 text documents were sent to our language translators and the output is denoted as S3.

As mentioned in Section 4.2, we need a text corpus to train the language model in order to improve automatic speech recognition of spontaneous speech. We used S3 as the text corpus because it is in spoken Taiwanese tonal syllables and the contents were extracted from related news documents. The language model is a syllable-based bi-gram model that contains about 480,000 Taiwanese tonal syllables.

7.2 Experimental Results of Speech Recognition and Cross-lingual Translation

The purpose of this experiment is to evaluate whether we can automatically and effectively prepare Taiwanese tonal syllable sequences for alignment. The results of speech recognition for broadcast news and that of language translation are shown in Table 7(a). The tonal syllable accuracy rate is defined as the ratio of the number of hypothesized syllables to the number of correct syllables minus the number of inserted syllables. We assume that the human results (i.e., S2 and S4) are correct and compare them with the automatic outputs (i.e., S1 and S3) to obtain the experimental results.

From Table 7(a), we observe that the speech recognizer's performance is poor. This is due to mismatch of the training and test speech. The former is read speech, but the latter is spontaneous speech. The result of the cross-lingual translation is satisfactory.

7.3 Experimental Results of Spoken and Text Documents Alignment

We applied our alignment component to align different combinations of human results and automatic results to evaluate the performance of our approach. The results are shown in Table 7(b). The goal of the experiment was to align both sets of automatic results (i.e., S1 and S3). Thirty-three out of forty spoken documents were aligned with text documents, which amount to an 82.5% accuracy rate. As noted by [1-2], it is possible to obtain good alignment despite poor accuracy rates for speech recognition. Note that our speech recognition performance is only 57.7%. We thought that this low accuracy was undesirable since it impacted on alignment performance.

On the other hand, when we tried to align both sets of human results (i.e., S2 and S4), we were surprised to find that the accuracy rates were not perfect either. Comparing A(S2,S3) and A(S2, S4), we only obtained a slight improvement, even when human transcriptions of the spoken documents were used. We believe the errors in alignment of S2 and S4 are due to the inherent ambiguity of words in cross-lingual translations. Although the topics of both reports are the same, their syllable sequences may not be exactly the same (they have a 13.3% syllable disagreement rate). In our experiment, the upper bound of the alignment accuracy rate is 95% (both documents used manual transcription and translation).

It is noteworthy that the alignment performance of S1 and S4 is worse than that of S1 and S3. The reason may be that S1 is the result of speech recognition, whose language model is trained from S3. S1 is somehow derived from S3, while S4 is translated by humans and the translation style may be more diversified than machine translation based on bi-lingual lexicons. It is also interesting that the alignment performance of S2 and S3 is 85%, better than that of S1 and S4. This result shows that speech recognition still dominates the alignment result assuming that cross-lingual translation by human is perfect.

8. Conclusion and Future Work

We have described an approach to automatic cross-lingual alignment of spoken documents and text documents. Its performance is comparable with manual alignment, which is known to be very tedious and time consuming. Our approach can be summarized as follows. First, based on the structure of television news programs we extract the anchor's reports and interviews, using a HMM-based audio segmenter. Then, we apply speech recognition to transcribe the news anchor's speech as a tonal syllable sequence in Taiwanese. Meanwhile, a cross-lingual translator translates text documents in Mandarin to tonal syllable sequences in Taiwanese, using bi-lingual statistical lexicons. Once we have pairs of tonal syllable sequences, we apply the DTW algorithm to align them and select the best matches.

The experimental results show that our approach is sufficiently accurate for use by search engines in content management applications. This would allow users to browse and search for audio spoken documents by associating them with well-organized text documents. We believe that our approach can be applied to similar problems in other languages.

In our future work, we will try to improve our speech recognition system by adapting the characteristics of news anchors' voices to a read-based acoustic model, and use the model to recognize the anchors' speech. This may improve speech recognition of spontaneous speech. We will also try to implement a bootstrapping approach whereby correctly aligned results can be used as training data to further improve the performance of all components in the system and reduce the need for human-labeled corpora.

Reference

[1] J.Y. Nie, P. Isabelle, M. Simard, and R. Durand, Cross-language information retrieval based on parallel texts automatic mining of parallel texts from the Web, In Proceedings of ACM Conference on Research and. Development in Information Retrieval, (Berkeley, CA, USA, 1999).

[2] P.-J. Cheng, J.-W. Teng, R.-C. Chen, J.-H. Wang, W.-H. Lu, and L.-F. Chien, Translating Unknown Queries with Web Corpora for Cross-Language Information Retrieval, In Proceedings of ACM Conference on Research and. Development in Information Retrieval, (Sheffield, UK, 2004).

[3] R.-Y. Lyu, D.-C. Lyu, M.-S. Liang, M.-H. Wang, Y.-C. Chiang, and C.-N. Hsu, A Unified Framework For Large Vocabulary Speech Recognition of Mutually Unintelligible Chinese Regionalects, In Proceedings of International Conference on Spoken Language Processing, (Jeju Island, Korea, 2004).

[4] Praat, http://www.fon.hum.uva.nl/praat/

[5] Transcriber, http://www.etca.rr/cta/gip/projets/transcriber/

[6] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava, Speech and Language Technologies for Audio Indexing and Retrieval, Proceedings of The IEEE 88(8) (2000).

[7] J.-M. van Thong, P.J. Moreno, B. Logan, B. Fidler, K. Maffey, and M. Moores, Speechbot: An Experimental Speech-Based Search Engine For Multimedia Content on The Web, IEEE Transactions on Multimedia 4(1) (2002).

[8] J. L. Gauvain and L. Lamel, Large Vocabulary Continuous Speech Recognition: Advances Applications, Proceedings of IEEE 88(8) (2000).

[9] H. D. Wactlar, A. G. Hauptmann, and M. J. Witbrock, Informedia: News-On-Demand Experiments in Speech Recognition, In Proceedings of DARPA Speech Recognition Workshop, (Harriman, NY, USA, 1996).

[10] J. Nouza, D. Nejedlová, J. Žďánský, and J. Kolorenč, Very Large Vocabulary Speech Recognition System For Automatic Transcription of Czech Broadcast Programs, In Proceedings of International Conference on Spoken Language Processing, (Jeju Island, Korea, 2004).

[11] P. Beyerlein, X. Aubert, R. Haeb-Umbach, and D. Klakow, Automatic Transcription of English Broadcast News, In Proceedings of the European Conference on Speech Communication Technology, (Budapest, Hungary, 1999).

[12] A. G. Hauptmann, R. E. Jones, K. Seymore, S. T. Slattery, M. J. Witbrock, and M. A. Siegler, Experiments in Information Retrieval from Spoken Documents, In Proceedings of DARPA Broadcast News Transcription and Understanding Workshop, (Lansdowne, VA, USA, 1998).

[13] G. Cook, J. Christie, D. Ellis, E. Fosler-Lussier, Y. Gotoh, B. Kingsbury, N. Morgan, S. Renals, T. Robinson, and G. Williams, An Overview of the SPRACH System for the Transcription of Broadcast News, In Proceedings of DARPA Broadcast News Workshop (Herdon, VA, USA, 1999).

[14] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and W.-J. Zhu, Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives, IEEE Transactions on Speech Audio Processing, 12(4) (2004).

[15] R.-Y. Lyu, Z.-H. Fu, Y.-C. Chiang, and H.-M. Liu, A Taiwanese (Min-Nan) Text-To-Speech (TTS) System Based on Automatically Generated Synthetic Units, In Proceedings of the International Conference on Spoken Language Processing, (Beijing, China, 2000).

[16] R.-Y. Lyu, C.-Y. Chen, Y.-C. Chiang, and M.-S. Liang, A Bi-Lingual Mandarin/Taiwanese (Min-Nan), Large Vocabulary, Continuous Speech Recognition System Based on the Tong-yong Phonetic Alphabet (TYPA), In Proceedings of the International Conference on Spoken Language Processing, (Beijing, China, 2000).

[17] D.-C. Lyu, B.-H. Yang, M.-S. Liang, R.-Y. Lyu, and C.-N. Hsu, Speaker Independent Acoustic Modeling for Large Vocabulary Bi-Lingual Taiwanese/Mandarin Continuous Speech Recognition, In Proceedings of Australian International Conference on Speech Science & Technology, (Melbourne, Australia. 2002).

[18] D.-C. Lyu, M.-S. Liang, Y.-C. Chiang, C.-N. Hsu, and R.-Y. Lyu, Large Vocabulary Taiwanese (Min-Nan) Speech Recognition Using Tone Features Statistical Pronunciation Modeling, In Proceedings of the European Conference on Speech Communication Technology, (Geneva, Switzerland, 2003).

[19] D.-C. Lyu, R.-Y. Lyu, Y.-C. Chiang, and C.-N. Hsu, Modeling Pronunciation Variation for Bi-Lingual Mandarin/Taiwanese Speech Recognition, Computational Linguistics & Chinese Language Processing, 10(3) (2005).

[20] L. Lu, H. Jiang, and H.-J. Zhang, A Robust Audio Classification Segmentation Method, In Proceedings of ACM International Conference on Multimedia, (2001).

[21] L. Chaisorn and T.-S Chua, The Segmentation Classification of Story Boundaries in News Video, In Proceedings of Working Conference on Visual Database Systems, (Australia, 2002).

[22] Z. Liu, Y. Wang, and T. Chen, Audio Feature Extraction Analysis For Scene Segmentation Classification, Journal of VLSI Signal Processing Systems, 20(1-2) (1998).

[23] S. J. Young et al., HTK: Hidden Markov Model Toolkit Version 3.1, Engineering Department, Cambridge University (2001).

[24] L. Rabiner and B. Juang, Fundamentals of Speech recognition (Prentice-Hall, Upper Saddle River, NJ, 1993).

[25] T. Schultz and A. Waibel, Multilingual Cross-lingual Speech Recognition, In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, (Lansdowne, VA, USA 1998).

[26] V. Digalakis, P. Monaco, and H. Murveit, Genones: Generalized mixture tying in continuous hidden Markov model-based speech recognizers, IEEE Transactions on Speech and Audio Processing, 4(44) (1996).

[27] F. Palou, P. Bravetti, O. Emem, V. Fischer, and E. Janke, Towards a Common Phone Alphabet for Multilingual Speech Recognition, In Proceedings of the International Conference on Spoken Language Processing, (Beijing, China, 2000).

[28] B. Paul, Accurate Short-Term Analysis of the Fundamental Frequency the Harmonics-To-Noise Ratio of a Sampled Sound, Proceedings of the Institute of Phonetic Sciences 17 (1993).

[29] P.-Y. Liang, J.-L. Shen, and L.-S. Lee, Decision Tree Clustering For Acoustic Modeling in Speaker-Independent Mandarin Telephone Speech Recognition, In Proceedings of the International Symposium on Chinese Spoken Language Processing, (Singapore, 1998).

[30] S. Furui, K. Maekawa, H. Isahara, T. Shinozaki and T. Ohdaira, Toward the Realization of Spontaneous Speech Recognition - Introduction of a Japanese Priority Program and Preliminary Results, In Proceedings of The International Conference on Spoken Language Processing, (Beijing, China, 2000).

[31] T. Yokoyama, T. Shinozaki, K. Iwano, and S. Furui, Unsupervised Class-Based Language Model Adaptation For Spontaneous Speech Recognition, In Proceedings of International Conference on Acoustics, Speech, Signal Processing, (Hong Kong, 2003).

[32] X. Huang, A. Acero, and H. Hon, Spoken language processing, (Prentice Hall, Upper Saddle River, NJ, 2001).

[33] M. Wester, J. M. Kessens, and H. Strik, Pronunciation Variation in ASR: Which Variation to Model? In Proceedings of the International Conference on Spoken Language Processing, (Beijing, China, 2000).

[34] I. D. Melamed, Empirical Methods for Exploiting Parallel Texts, (MIT, Cambridge, MA, USA, 2001).

[35] W.-K. Lo, H. Meng, and P. C. Ching, Cross-Language Spoken Document Retrieval Using Hmm-Based Retrieval Model With Multi-Scale Fusion, ACM Transactions on Asian Language Information Processing, 2(1) (2003).

[36] C.S. Myers and L.R. Rabiner, A Comparative Study of Several Dynamic Time-Warping Algorithms for Connected Word Recognition, The Bell System Technical Journal, 60(7) (1981).

[37] R. Manmatha and N. Srimal: Scale Space Technique for Word Segmentation in Handwritten Manuscripts. In Proceedings of International Conference on Scale-Space Theories in Computer Vision, (1999).

[38] W. Abdulla, D. Chow, and G. Sin, Cross-Words Reference Template for DTW-Based Speech Recognition Systems, In Proceedings of Convergent Technologies for Asia-Pacific Region Conference, (2003).

[39] C.-H. Chang, C.-N. Hsu, and S.-C. Lui, Automatic Information Extraction from Semi-Structured Web Pages by Pattern Discovery, Decision Support Systems, 35(1) (2003).

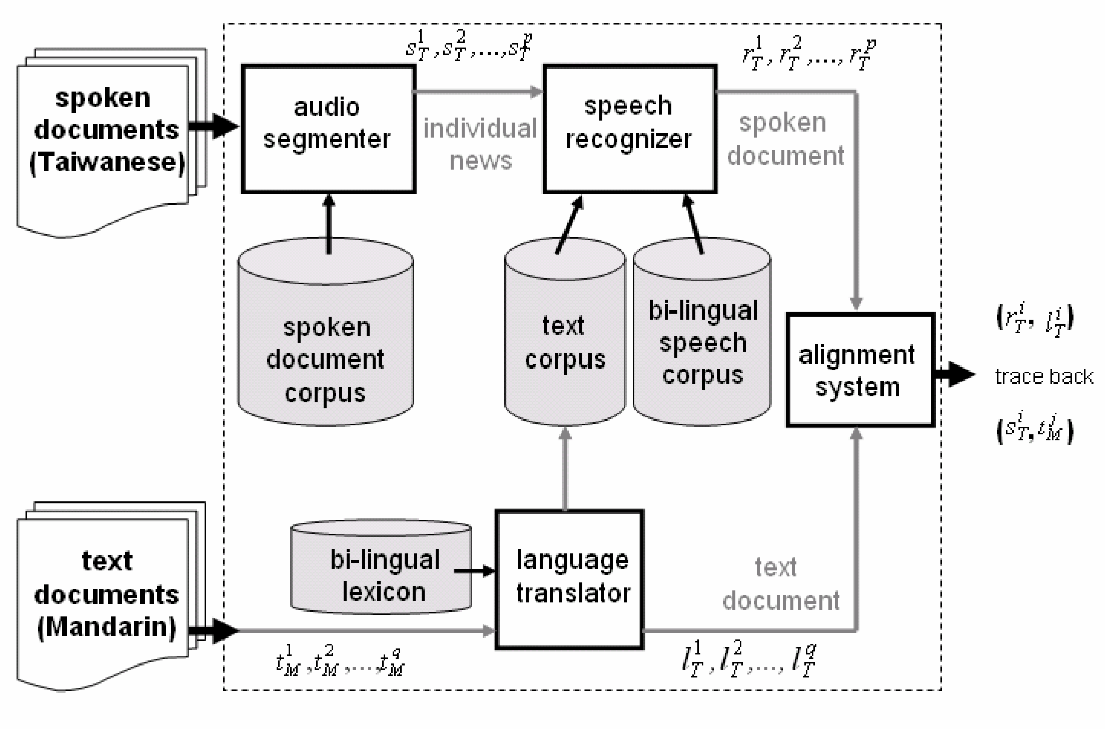
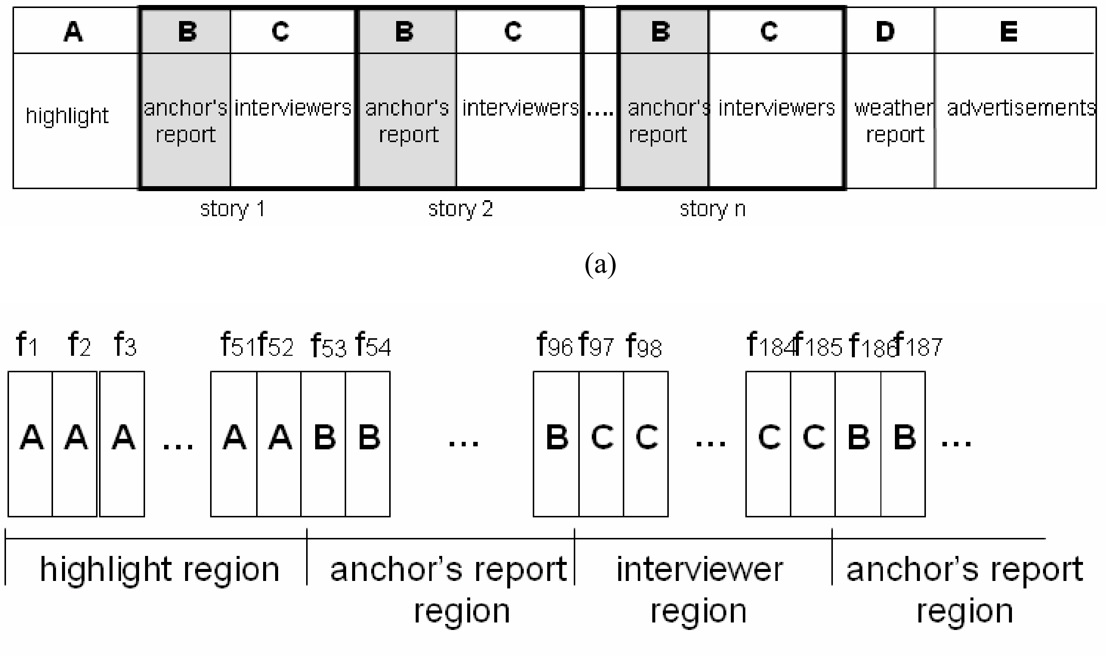


Figure 1. Components and data flow orders of the multi-media, cross-lingual document alignment system.



(a)

(b)

Figure 2: (a) Structure of a typical television broadcast news session on FTVN. (b) An example of the output sequence of the HMM scene type segmenter.

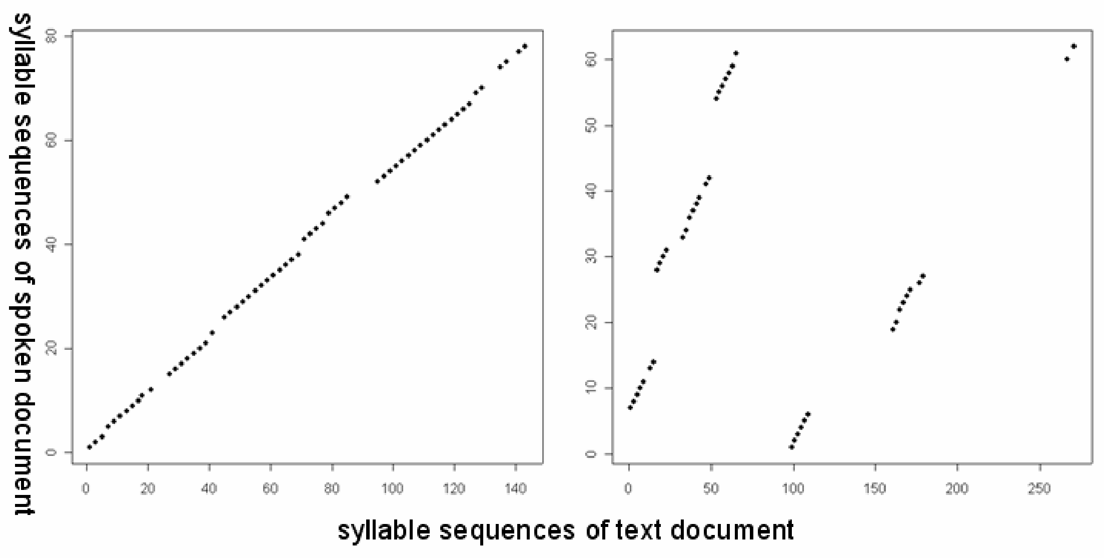


Figure 3. Two examples of syllable sequence alignment between spoken and text documents; the left side scores 0.9, and the right side scores 0.4.

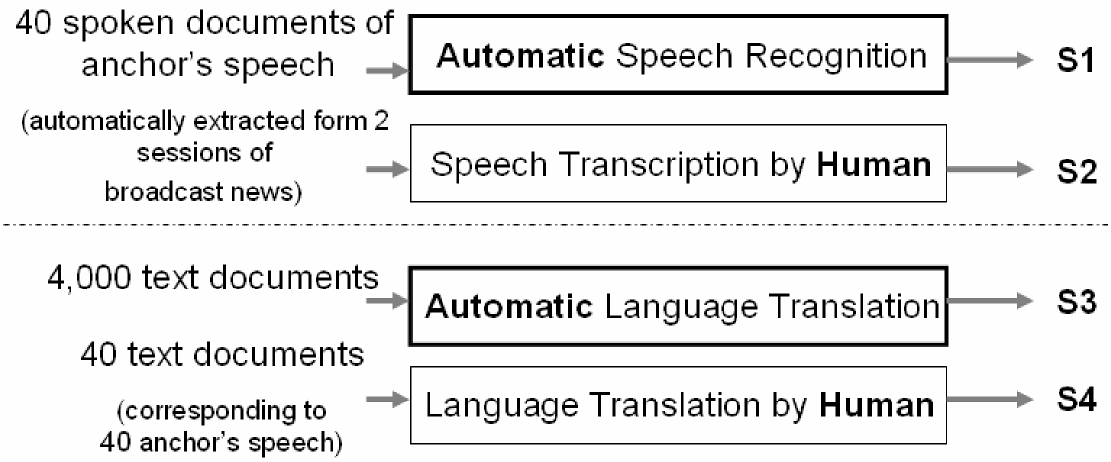

40 spoken documents of anchor's speech
(automatically extracted form 2 sessions of broadcast news)
Automatic Speech Recognition
S1
Speech Transcription by Human
S2
4,000 text documents
Automatic Language Translation
S3
40 text documents
(corresponding to 40 anchor's speech)
Language Translation by Human
S4

Figure 4. Procedures for preparing sequences for evaluating the alignment of spoken documents and text documents.

Scene types	Percentage (in time)	Signal texture
Highlight	4.53%	Speech with music
Anchor’s report	16.65%	Speech only
Interview	57.20%	All
Weather report	7.60%	Speech only, Speech with music
Advertisement	14.02 %	All

Table 1. Statistics of the proportions and signal texture of five scene types in a four-hour broadcast news corpus.

Classified as→	A	B	C	D	E
A) highlight	2				
B) anchors report		40			
C) interview		2	37	1	
D) weather report				2	
E) advertisement					2

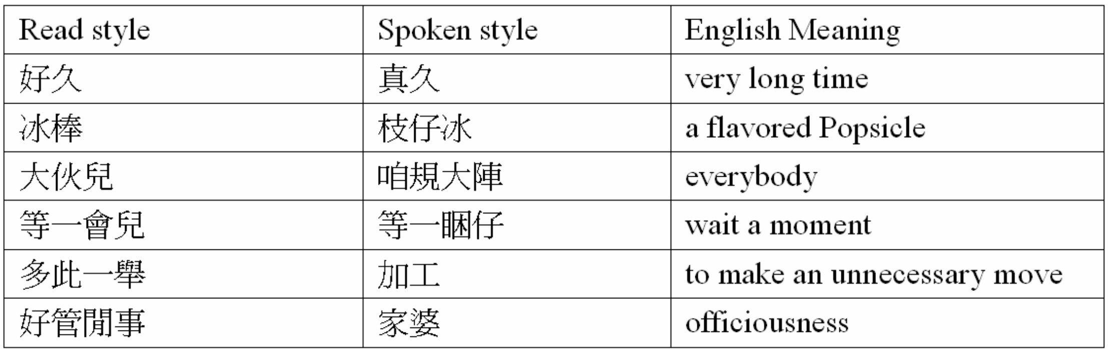
Table 2. Confusion table of the classification results for five scene types.

	M	T	M∪T	M∩T
N_S	408	709	925	192(21%)
Tone	5	7	9	3
N_{TS}	1288	2878	3519	647(18%)
N_I	17	19	22	14(63%)
N_{TF}	295	225	416	104(25%)
N_{CDIF}	1656	3496	4374	778(18%)

Table 3: Statistics of the number of syllables (N_S) and tones (Tone) of Mandarin (M) and Taiwanese (T) and their linguistic units: the number of Tonal Syllables (N_{TS}), Initials (N_I), Tonal Finals (N_{TF}), and context-dependent Initial/Tonal Finals (N_{CDIF}). ∩and∪ denote the intersection and union of the two languages, respectively.

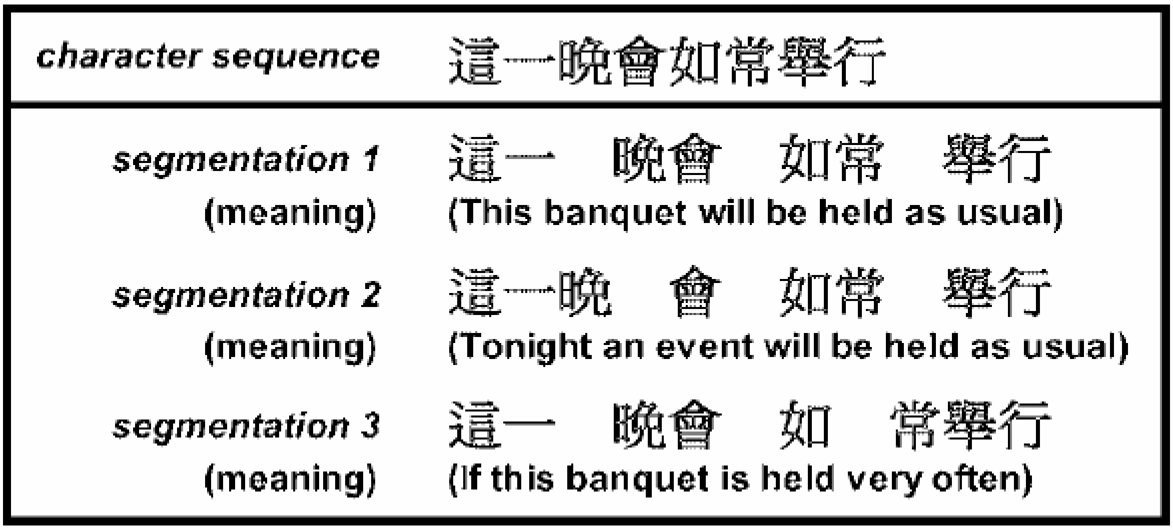
	Training Set		Evaluation Set	
Language	M	T	M	T
No. of Speakers	100	100	10	10
No. of Utterances	43078	46086	1000	1000
No. of Hours	11.3	11.2	0.28	0.28

Table 4. Statistics of the bi-lingual speech corpus for acoustic model training and baseline testing (M: Mandarin, T: Taiwanese)



Read style	Spoken style	English Meaning
好久	真久	very long time
冰棒	枝仔冰	a flavored Popsicle
大伙兒	咱規大陣	everybody
等一會兒	等一睏仔	wait a moment
多此一舉	加工	to make an unnecessary move
好管閒事	家婆	officiousness

Table 5. Examples of terms in Taiwanese with the same meaning used in the reading and spoken styles as written in Chinese characters and their meanings in English.



character sequence	這一晚會如常舉行
segmentation 1	這一 晚會 如常 舉行
(meaning)	(This banquet will be held as usual)
segmentation 2	這一晚 會 如常 舉行
(meaning)	(Tonight an event will be held as usual)
segmentation 3	這一 晚會 如 常舉行
(meaning)	(If this banquet is held very often)

Table 6. An example of ambiguity in Chinese word segmentation, from [35].

Terms	(auto, manual)	Tonal Syllable Accuracy Rate (%)
Speech Recognition	S(S1, S2)	57.7%
Cross-lingual Translation	S(S3, S4)	89.1%

(a)

	Alignment Accuracy Rate (%)
For evaluation	
A(S1,S3)	82.5 (33/40)
For analysis	
A(S1,S4)	77.5 (31/40)
A(S2,S3)	85.0 (34/40)
A(S2,S4)	95.0 (38/40)

(b)

Table 7. (a) Results of syllable accuracy rates for speech recognition of spoken documents and cross-lingual translation of text documents. (b) Comparison of alignment accuracy rate results of aligning spoken documents and text document by either automatic or manual methods.