Computational Linguistics and Chinese Language Processing
Vol. 9, No. 2 , August 2004, pp. 13-28

Multiple-Translation Spotting for Mandarin-Taiwanese Speech-to-Speech Translation

Jhing-Fa Wang[*], Shun-Chieh Lin[*], Hsueh-Wei Yang[*], and Fan-Min Li[*]

Abstract

The critical issues involved in speech-to-speech translation are obtaining proper source segments and synthesizing accurate target speech. Therefore, this article develops a novel multiple-translation spotting method to deal with these issues efficiently. Term multiple-translation spotting refers to the task of extracting target-language synthesis patterns that correspond to a given set of source-language spotted patterns in conditional multiple pairs of speech patterns known to be translation patterns. According to the extracted synthesis patterns, the target speech can be properly synthesized by using a waveform segment concatenation-based synthesis method. Experiments were conducted with the languages of Mandarin and Taiwanese. The results reveal that the proposed approach can achieve translation understanding rates of 80% and 76% on average for Mandarin/Taiwanese translation and Taiwanese/Mandarin translation, respectively.

Keywords: Multiple-Translation Spotting, Speech-to-Speech Translation

1. Introduction

Automatic speech-to-speech translation is a prospective application of speech and language technology [See JANUS III [Lavie *et al*. 1997], Verbmobil [W. Wahlster 2000], EUTRANS [Casacuberta *et al*. 2001] and ATR-MATRIX [Sugaya *et al*. 1999]]. However, the unsolved problems in speech-to-speech translation are how to obtain proper source segments and how to generate accurate target sequences while the system performance is degraded by speech input. With the rising importance of parallel texts (bitexts) in language translation, an approach called translation spotting has been applied for proposing appropriate translations, referring to the TransSearch system [Macklovitch *et al*., 2000] and sub-sentential translation memory systems [M. Simard, 2003]. Previous works in this area have suggested that manual review or

[*] **Corresponding author:**
Prof. Jhing-Fa Wang, Department of Electrical Engineering, National Cheng Kung University, No.1, Dasyue Rd., East District, Tainan City 70101, Taiwan, R.O.C.
Email: wangjf@csie.ncku.edu.tw Tel: 886-6-2757575 ext. 62341 Fax: 886-6-2746867

crafting is required to obtain example bases of sufficient coverage and accuracy to be truly useful.

Translation spotting (TS) is a term coined by Vḗronis and Langlais [2000] and refers to the task of identifying word tokens in a target-language (TL) translation that correspond to some given word-patterns in a source-language (SL) text. This process takes as input a couple, i.e., a pair of SL and TL text segments known to be translation patterns, and an SL query, i.e., a subset of the patterns of the SL segment, on which the TS will focus its attention. In more formal terms:

Ÿ The input to the TS process is a pair of SL and TL text segments $\langle S, T \rangle$ and a contiguous, non-empty input sequence of word-tokens in SL, $q = s_1 \mathbf{L} s_n$.

Ÿ The output is a pair of sets of translation patterns $\langle r_q(S), r_q(T) \rangle$: the SL answer and TL answer, respectively.

Table 1 shows some examples of TS, where the words in italics represent the SL input, and the words in bold are the SL and TL answers. As can be seen in these examples, the patterns in the input q and answers $r_q(S)$ and $r_q(T)$ may or may not be contiguous (examples 2 and 3), and the TL answer may possibly be empty (example 4) when there is no satisfactory way of linking TL patterns to the input. By varying the identification criteria, the translation spotting method can help evaluate units over various dimensions, such as frequency ranges, parts of speech and even speech features of spoken language.

Table 1. Translation spotting examples.

Query		Sentence Pair	
		SL (Mandarin)	TL (Taiwanese)
1.	*q*:*待 幾 天*	你 預計 要 待 幾 天	lie phahsngx bueq doax kuie jit
		$r_q(S)$={待,幾,天}	$r_q(T)$={doax,kuie,jit}
2.	*q*:*我 要 訂 兩 間 單人房*	我 明天 要 訂 兩 間 有 淋浴 設備 的 單人房	minafzaix goar bueq dexng lerng kefng u sea sengqw e danjiin paang
		$r_q(S)$={我,要,訂,兩,間,單人房}	$r_q(T)$={goar,bueq,dexng,lerng,kefng,danjiin paang}
3.	*q*:*今晚 有 […] 雙人房 嗎*	請 問 你們 今晚 有 一 間 雙人房 嗎	chviar bun lirn ehngf u cit kefng sianglaang paang but
		$r_q(S)$={今晚,有,雙人房,嗎}	$r_q(T)$={ehngf,u,sianglaang,paang,but}
4.	*q*:*包括 … 在 內*	有 包括 早餐 在 內?	u zafdngx but
		$r_q(S)$={包括,在,內}	$r_q(T)$={ f }

However, translation spotting can only draw out the TL answer from the best translation; it can not handle an SL query whose word-tokens are distributed in different translations. Consequently, we propose conducting multiple-translation spotting of a speech input using multiple pairs of translation patterns. Figure 1 shows an example of multiple-translation spotting of a speech input. When a speaker inputs an SL speech query ”今晚會有三間單人房嗎 ”, the proposed system can obtain a TL speech pattern set that includes five elements, ”ehngf”, ”kvaru”, ”svaf”, ”kefng”, and ”danjiinpaang”, according to the spotted SL speech patterns ”今晚”, “會有”, “間”, “嗎”, ”三”, and ”單人房”. The rest of this article is organized as follows. Section 2 presents the framework of the proposed system. Section 3 presents system data training for Mandarin and Taiwanese. Section 4 describes the proposed translation method for speech-to-speech translation. Section 5 presents experimental results. Finally, Section 6 draws conclusions.

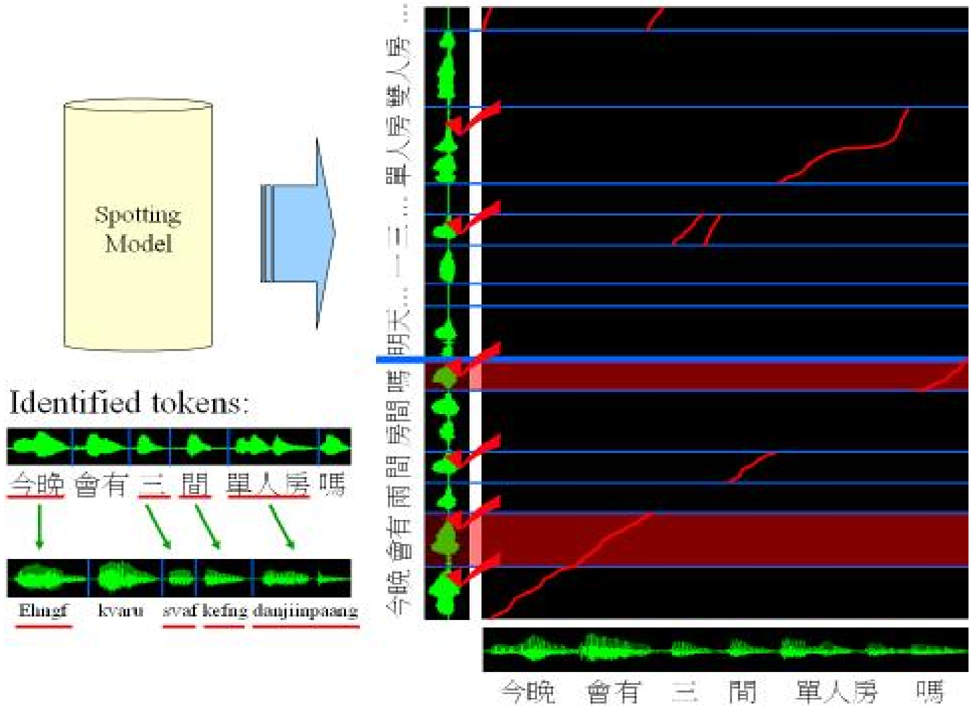
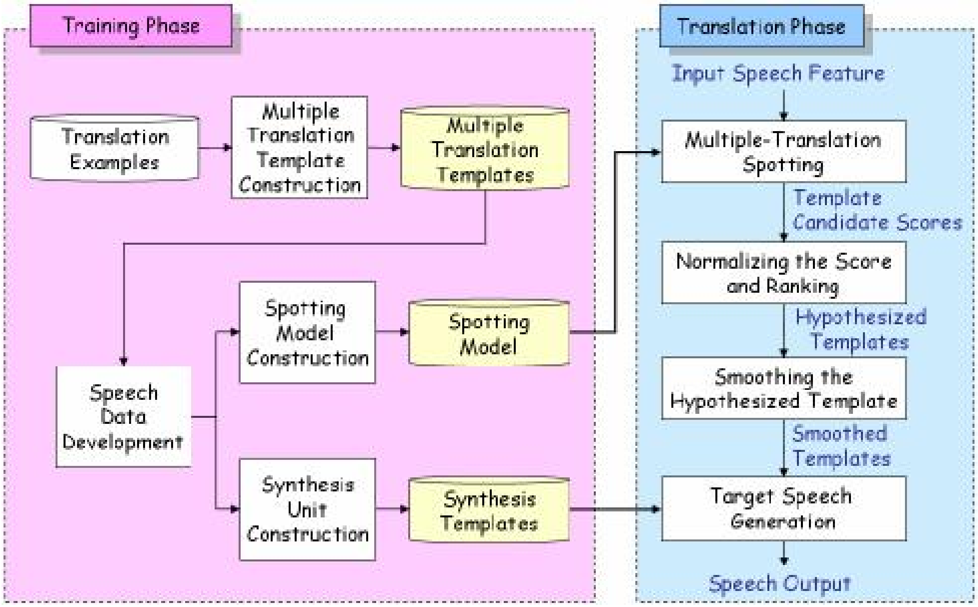

Spotting Model
Identified tokens:
今晚 會有 三 間 單人房 嗎
Ehngf kvaru svaf kefng danjiinpaang
今晚 會有 兩 間 房間 嗎 明天... 一 二 ... 單人房 雙人房 ...
今晚 會有 三 間 單人房 嗎

Figure 1. An example of multiple-translation spotting.

2. Framework of the Proposed System

The proposed speech-to-speech translation system is divided into two phases – a training phase and a translation phase. In the training phase, the developed translation examples are imported to derive multiple-translation templates and develop speech data. In the following step, the developed speech data are applied to construct multiple-translation spotting models and synthesis templates. Figure 2(a) shows a block diagram of the training phase.


Training Phase
Translation Examples
Multiple Translation Template Construction
Multiple Translation Templates
Speech Data Development
Spotting Model Construction
Spotting Model
Synthesis Unit Construction
Synthesis Templates
Translation Phase
Input Speech Feature
Multiple-Translation Spotting
Template Candidate Scores
Normalizing the Score and Ranking
Hypothesized Templates
Smoothing the Hypothesized Template
Smoothed Templates
Target Speech Generation
Speech Output

(a) (b)

Figure 2. Framework of the proposed system: (a) a training phase; (b) a translation phase.

Figure 2(b) shows a block diagram of the translation phase. A one-stage based spotting method is adopted to identify input spoken phrases for each spotting template, and the template candidates are assigned in the following score normalization and ranking process. However, the hypothesized word sequence generally includes noise-like segments. Accordingly, the segments are adjusted by smoothing the hypothesized word sequences. After the hypothesized word sequences of all template candidates have been smoothed, the hypothesized target sequences are generated using the translation template with the maximum number of spotting tokens of speech input. The obtained target speech segments are used to produce target speech by means of the corresponding synthesis template in the final target generation process.

3. Data Training Phase

As for the task of translating Mandarin and Taiwanese language pairs, although these languages both belong to the family of Chinese languages, their language usages still have various development by language families and their origins, Mandarin belongs to Altaic language family, and Taiwanese belongs to Sinitic language family [Sher *et al*., 1999]. Therefore, in the following section, we will consider their language usages for three template construction.

3.1 Multiple Translation Template Construction

While translation templates can be fully constructed, one major issue in translation pattern exploitation, called “divergence,” makes straightforward transfer mapping extraction impractical. Dorr (1993) describes divergence in the following way: “translation divergence arises when the natural translation of one language into another result in a very different form than that of the original.” Therefore, we choose translations with no divergence to practice constructing templates. An example of a simple translation template derived from a practicable translated example is shown below.

Translated Example: SL: “我 朋友 要 訂 房間”

↔ TL: ”goarn pengiuo bueq dexng pangkefng”

Intention Translation: M_{p1} 要 訂 M_{p2} ↔ T_{p1} bueq dexng T_{p2}

alignments

Variable Translation: If M_{p1}↔ T_{p1}, 我 朋友 ↔ goarn pengiuo

If M_{p2}↔ T_{p2}, 房間 ↔ pangkefng

The translation template is composed of a translated example, an intention translation, and two variable translations. The example shows how a sentence in Mandarin (SL) that contains an intention “要 訂” with two variables, M_{p1} (我 朋友) and M_{p2} (房間), can be translated into a sentence in Taiwanese (TL) with an intention “bueq dexng” and two variables, T_{p1} (goarn pengiuo) and T_{p2} (pangkefng). According to the template, the number of variable translations should be expanded to improve the capability for spotting the speech input. From the preceding example, variable translation expansion can be illustrated as follows:

Variable Translation Expansion:

If M_{p1} ↔ T_{p1},

我 ↔ goar

我 朋友↔ goarn pengiuo

If M_{p2} ↔ T_{p2}, 房間 ↔ pangkefng

票 ↔ phiaux

Therefore, we can obtain corpus-specific multiple translations in a template constructed from three translation patterns, which are “我 朋 友 要 訂 房間↔goarn pengiuo bueq dexng pangkefng”, “我↔goar”, and “票↔phiaux”.

3.2 Spotting Model Construction

Taiwanese is a typical oral language and still has no uniform system of writing. In the literature, there are two ways to represent Taiwanese words: Chinese characters and alphabetic writing. [Sher *et al*., 1999]. Chinese characters have huge hieroglyph character sets; therefore, it is difficult to systematize developed examples. Although alphabetic writing would be an appropriate representation form, a universal phonemic transcription system is still not available.

Therefore, for the purpose of practical system construction, a collection of speech data is developed from derived text-form templates not only to obtain spotting models but also to transcribe text data as waveform-based representations. For one of the translating languages, the speech data, including intention speech and related variable speech, are used in chorus to construct spotting reference models for use in multiple-translation spotting. Such spotting reference models are embedded with latent grammars from the constructed templates. When dealing with Mandarin-Taiwanese speech feature models, we build the database by extracting LPCC features from recorded template speeches. Hence, when speech recognition is performed, the LPCC features are extracted from the recorded template speeches, and the LPCC features of speech input are used in combination to compute the degree of dissimilarity. After language pairs of both Taiwanese and Mandarin speech data are developed, the transfer mapping information for a pair of Taiwanese and Mandarin speech segments known to be similar in terms of text-form word alignment is constructed.

3.3 Synthesis Template Construction

Both Mandarin and Taiwanese are tonal languages, and it is difficult to determine whether a morpheme will take its inherent tone or the derived tone when every word in a sentence is synthesized. [Wang *et al*., 1999; Sher *et al*., 1999]. Therefore, we utilize the obtained intention speech and variable speech as synthesis templates that include intention synthesis units and variable synthesis units. These synthesis units can be used to generate a speech output to be processed using a waveform segment concatenation-based synthesis method [Wang *et al*., 1999]. For each synthesis unit in the obtained speech data, the following features are stored:

- the waveform and its length,
- the code of the synthesis unit.

4. Translation Phase

4.1 Multiple-Translation Spotting Method

To deal with the problem of spotting between a speech input X_1^L and a translation pattern set $\left\{\left\langle s_j^{(v)}, t_j^{(v)} \right\rangle\right\}_{j=1}^{J}$ in the *v*-th translation template (r_v), we use the standard notation *l* to represent the frame index of $X_1^L, 1 \le l \le L$, j to represent the spotting pair ($\left\langle s_j^{(v)}, t_j^{(v)} \right\rangle$) index of r_v , $1 \le j \le J$, and *k* to represent the frame index of *j*-th spotting pattern $s_j^{(v)}$, $1 \le k \le K_j$. Then for each input frame, the accumulated distance $d_A(l,k,j)$ is computed by

$$d_A(l,k,j) = d(l,k,j) + \min_{k-2 \le m \le k}\left(d_A(l-1,m,j)\right) \cdot \tag{1}$$

For $2 \le k \le K_j$, $1 \le j \le J$, where $d(l,k,j)$ is the local distance between the *l*-th frame of X_1^T and the *k*-th frame of source pattern $s_j^{(v)}$. The recursion of (1) is carried out of for all internal frames (i.e., $k \ge 2$) of each source pattern. At the speech pattern boundary, i.e., $k = 1$, the recursion can be calculated as follows:

$$d_A(l,1,j) = d(l,1,j) + \min\left\lfloor \min_{1 \le m \le J}\left(d_A(l-1,K_m,m)\right), d_A(l-1,1,j) \right\rfloor \cdot \tag{2}$$

The final solution for the best path is

$$d_G^{(v)} = \min_{1 \le j \le J}\left[d_A\left(L, K_j, j\right)\right] \tag{3}$$

The details of the multiple-translation spotting algorithm are given below:

```
/*  Parameter descriptions
      $\{t_j^{(v)}\}_{j=1}^{J}$ : the spotting results of $\{s_j^{(v)}\}_{j=1}^{J}$, where
      $t_j^{(v)} = \begin{cases} 1, & \text{if SL speech pattern } s_j^{(v)} \text{ is spotted by } X_1^L \cdot; \\ 0, & \text{otherwise.} \end{cases}$
      $w_v \leftarrow \{t_j^{(v)} \mid t_j^{(v)} = 1,\ 1 \le j \le J\}$: the hypothesized TL synthesis patterns;    */

/*  Initialization    */
      $l \leftarrow k \leftarrow j \leftarrow 1$;
      $t_j^{(v)} \leftarrow 0, 1 \le j \le J$;
      $w_v \leftarrow \{\ f\ \}$;
/*  v-th template spotting    */
    while ( $l \le L$ )
        for each spotting pattern $s_j^{(v)}$
            while ( $k \le K_j$ )
                if (k = 1)
                    $d_A(l,k,j) \leftarrow d(l,k,j) + \min[\min_{1 \le j \le J}[d_A(l-1,K_j,j)], d_A(l-1,k,j)]$
```

```
          $p(l,k,j) \leftarrow \arg\min[\min_{1\le m\le J}[d_A(l-1,K_m,m)], d_A(l-1,k,j)]$
        else if ($k > 1$)
          $d_A(l,k,j) \leftarrow d(l,k,j) + \min_{k-2\le m\le k}(d_A(l-1,m,j))$
          $p(l,k,j) \leftarrow \arg\min_{k-2\le m\le k}(d_A(l-1,m,j))$
        else if ($k = K_j$)
          $k \leftarrow 1$;
        else
          $k$++;
        end if
      end while
    end for
    $l$++;
  end while
  $d_G^{(v)} \leftarrow \min_{1\le j\le J}[d_A(L,K_j,j)]$;
  $\hat{j} \leftarrow \arg\min_{1\le j\le J}[d_A(L,K_j,j)]$;
/*    Trace back and TL synthesis pattern extraction    */
  $\{\tau_j^{(v)}\}_{j=1}^{J} \leftarrow$ trace back$[(L,K_{\hat{j}},\hat{j})]$; /*, $t_j^{(v)}$ is assigned as 1 or 0*/
  for each $t_j^{(v)}$, $j$=1,2,…,$J$
    If ($t_j^{(v)} = 1$)
      $w_v \leftarrow w_v \cup \{t_j^{(v)}\}$;
    end if
  end for
  return $w_v$ and $\{\tau_j^{(v)}\}_{j=1}^{J}$;
```

4.2 Normalizing the Score and Ranking

The length of the matching sequence can severely impact the cumulative dissimilarity measurement, so a length-conditioning weight is applied to overcome this defect. Scoring methods that involve the length measurement $\Delta\left(X_1^L, s^{(v)}\right)$($s^{(v)} = \bigcup_{j=1}^{J} s_j^{(v)}$) [J.N.K. Liu and L. Zhou, 1998] can be defined in a number of similar ways:

$$\Delta(X_1^L, s^{(v)}) = \max(\|X_1^L\|, \|s^{(v)}\|)(\text{or } \min(\|X_1^L\|, \|s^{(v)}\|), \tag{4}$$

$$\Delta(X_1^L, s^{(v)}) = \|X_1^L\| * \|s^{(v)}\|, \tag{5}$$

$$\Delta(X_1^L, s^{(v)}) = N(L, \sum_{j=1}^{J} K_j) + F(L, \sum_{j=1}^{J} K_j)/3 \text{,} \tag{6}$$

where $\left\|X_1^L\right\|$ is the number of frames in speech input *x*; $\left\|s^{(v)}\right\|$ is the total number of search frames in $\left\{s_j^{(v)}\right\}_{j=1}^{J}$; $N(L, \sum_{j=1}^{J} K_j)$ is the number of frames compared; and $F(L, \sum_{j=1}^{J} K_j)$ is the number of frames that fail to be matched. To improve the flexibility and reliability of the dissimilarity measurement, an exponential $\Delta\left(X_1^L, s^{(v)}\right)$ is exponentially defined as follows:

$$\Delta (X_1^L , s^{(v)}) = \partial^{w_{X,s^{(v)}}} , \tag{7}$$

where $\partial^{w_{X,s^{(v)}}}$ is a weighting factor and $w_{X,s^{(v)}} = \left(\left\|X_1^L\right\| - \left\|s^{(v)}\right\|\right) \cdot \left\|s^{(v)}\right\|^{-1}$. The weighting factor of Eq. (7) has two features: one is length correlation normalization, and the other is exponential score normalization. For length correlation normalization, the tendency to choose a template $\left\|s^{(v)}\right\|$ with the same length difference of $\left\|X_1^L\right\|$ but smaller length multiplication is eliminated. With exponential score normalization, when the difference between the speech input and each template is larger, a higher dissimilarity score is obtained and spotting discrimination improves. Finally, the normalized measured dissimilarity is determined as follows:

$$d_G^{(v)} = d_G^{(v)} \cdot \partial^{w_{X,s^{(v)}}} . \tag{8}$$

The experimental analysis shown in Fig. 3 indicates that the interval ∂ that yields the most accurate dissimilarity measurement is $[1.2-d, 1.2+d]$. Therefore, the value of ∂ chosen here is 1.2. The weighting factor is determined using the feature models of the first speaker for inside training. The feature models are different from the test data; thus, ∂ is a test-independent weighting factor. After all the templates are ranked, the retrieval accuracy is estimated using the criterion that the intention of the source speech is located in the set of the best N retrieved translation templates.

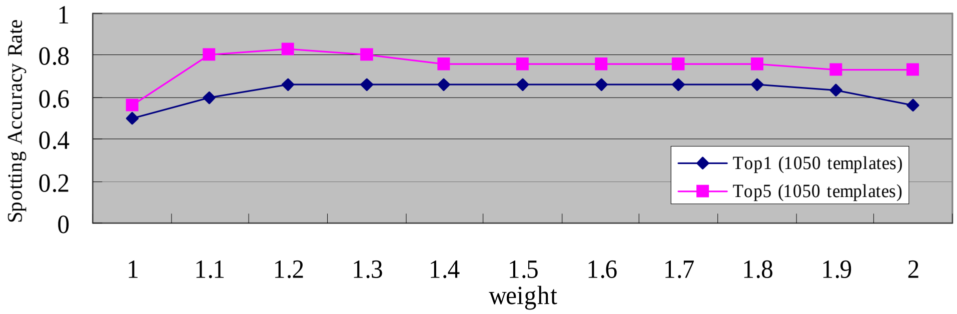


Figure 3. Time-conditioned weight convergence for dissimilarity measurement

4.3 Smoothing the Hypothesized Template

The main weakness with the one-stage algorithm for multiple-translation spotting is that it provides no mechanism for controlling the resulting sequence length, that is, for determining the optimal token sequence of arbitrary length. The algorithm finds a single best path whose sequence length is arbitrary. Therefore, the hypothesized token sequence generally includes noise-like components. The components should be in the form of duplications, and their durations should be below a threshold. Based on this assumption, hypothesized token outputs with segmented durations below the threshold are considered for further smoothing. With Mandarin and Taiwanese, the duration of a syllable is 0.3 sec on average [Sher *et al*., 1999], and this value is set as the relevant threshold to sift out noise-like components whose durations are less than 0.3 sec. These are the preliminary speaker-dependent results of our experiments. This system is able to adjust the threshold when a speaker speaks at different rates. Additionally, this system is corpus-specific, and out of vocabulary (OOV) words are rejected based on their high dissimilarity scores. After the token sequences of all the TopN templates have been smoothed, the hypothesized target sequences is generated using the translation template with the maximum number of spotting tokens of speech input.

4.4 Target Speech Generation

Once the hypothesized target sequences have been determined, the target speech generation process is straightforward, similar to the waveform segment concatenation-based synthesis method. In this method, waveform segments are extracted beforehand from the recorded intention synthesis units and variable synthesis units of the synthesis template, and they are rearranged with adequate overlapping portions to generate speech with the desired energy and duration. The merits of the method are the small computational cost in the synthesis process and the high level of intelligibility of the synthesized speech. The generation process includes complete matching, waveform replacement, and waveform deletion; thus, it is similar to the example-base translation method [J. Liu and L. Zhou, 1998].

5. Experimental Results

5.1 The Task and the Corpus

We built a collection of Mandarin sentences and their Taiwanese translations that usually appear in phrasebooks for foreign tourists. Because the translations were made sentence by sentence, the corpus was sentence-aligned at birth. *Table 2* shows the basic characteristics of the collected corpus.

Table 2. Basic characteristics of the collected translated examples.

	Mandarin	Taiwanese
Number of sentences	2,084	2,084
Total number of words	14,219	14,317
Number of word entries	6,278	6,291
Average number of words per sentence	6.82	6.87

In this work, the content of the high divergent example sentence pairs needed to be collated or sieved out to improve the accuracy and effectiveness of alignment exploration between word sequences and the derivation of multiple translation templates. *Table 3* shows the basic characteristics of the derived multiple translation templates. The derived templates were used to develop the speech corpus, which was used to construct spotting models and synthesis templates.

Table 3. Basic characteristics of the derived translation templates.

Number of templates	1,050
Number of intentions	1,050
Total number of translation patterns	5,542
Number of translation entries	1,260
Average number of translations per template	5.28

In order to evaluate the system performance, a collection of 1,050 utterances were speaker-dependent trained, and 30 additional utterances of each language were collected by using one male speaker (Sp1) for inside testing and by using two bilingual male speakers (Sp2 and Sp3) for outside testing. All the utterances were sampled at an 8 kHz sampling rate with 16-bit precision on a Pentium® IV 1.8GHz, 1GB RAM, Windows® XP PC.

5.2 Translation Evaluations

For the speech translation system, we found that the recognition performance of 39-dimension MFCCs and 10-dimension LPCCs was close. Therefore, we adopted 10-dimension LPCCs due to their advantages of faster operation and simpler hardware design. Speech feature analysis of recognition was performed using 10 linear prediction coefficient cepstrums (LPCCs) on a 32ms frame that overlapped every 8ms.

For estimating the computational load of the proposed MTS algorithm, a complexity analysis is shown in *Table 4*. Parts of the overall computation of the local frame distance depend on the feature dimension, so we used O(LPCC_add) and O(LPCC_mul) to represent the complexity of additions and multiplications, respectively. We applied Itakura type in each internal dynamic programming path selection employed 3 additions to decide the last node and 1 addition to accumulate the node distance, and 3 multiplications for slope weighting. In *Table 4*, the second row, *Distance computation*, presents the computational complexity of computing the local distance, and the third row, *Path selection*, presents the computational complexity of selecting the best path, that is, the computational overload of MTS for each template.

Table 4. Complexity analysis of the MTS algorithm.

	Computational Load	
	Addition	Multiplication
Distance computation	$L\cdot\sum_{j=1}^{J}K_j^{(v)}\cdot O(LPCC\_add)$	$L\cdot\sum_{j=1}^{J}K_j^{(v)}\cdot O(LPCC\_mul)$
Path selection	$5\cdot L\cdot\sum_{j=1}^{J}K_j^{(v)}$	$3\cdot L\cdot\sum_{j=1}^{J}K_j^{(v)}$
Total for each template	$L\cdot\sum_{j=1}^{J}K_j^{(v)}\cdot(5+O(LPCC\_add))$	$L\cdot\sum_{j=1}^{J}K_j^{(v)}\cdot(3+O(LPCC\_mul))$
Total for all templates	$\sum_{v}\left(L\cdot\sum_{j=1}^{J}K_j^{(v)}\cdot(5+O(LPCC\_add))\right)$	$\sum_{v}\left(L\cdot\sum_{j=1}^{J}K_j^{(v)}\cdot(3+O(LPCC\_mul))\right)$

When input speech is being spotted, a major sub-problem in speech processing is determining the presence or absence of a voice component in a given signal, especially the beginnings and endings of voice segments. Therefore, the energy-based approach, which is a classic one and works well under high SNR conditions, was applied to eliminate unvoiced components in this research. The measurement results were divided into four parts: the dissimilarity measurement of linear prediction coefficient cepstrum (LPCC)-based (baseline), the baseline with unvoiced elimination (unVE), the baseline with the time-conditioned weight (TcW), and the combination of unVE and TcW considerations with the baseline. A given translation template is called a *match* when it contained the same intention as the speech input. The reason for adopting this strategy was that variables could be confirmed again while a dialogue was being processed, while wrong intentions could cause endless iterations of dialogue. The experimental results for proper template spotting are shown in *Table 5 and Table 6*.

Based on the constructed translation templates, when the template or vocabulary size increases, more templates would possibly lead to more feature models and more similarities in speech recognition, thus causing false recognition results and lower spotting accuracy. Additionally, multiple speaker dependent results were obtained using three speakers. The first speaker's feature models (spotting models) were used to perform tests on the other two speakers, and the results are shown in Table 7. The experimental results show that although the feature models were trained by Sp1, the spotting accuracy of Sp2 and Sp3 was only reduced by 10 to 15 percent.

A bilingual evaluator was used to classify the target generation results into three categories [Yamabana *et al*., 2003]: *Good*, *Understandable*, and *Bad*. A *Good* generation needed to have no syntactic errors, and its meaning had to be correctly understood. *Understandable* generations could have some variable translation errors, but the main intention of the source speech had to be conveyed without misunderstanding. Otherwise, the translations were classified as *Bad*. With this subjective measure, the percentage of *Good* or *Understandable* generations for the Top 5 was 80% for Mandarin to Taiwanese (M/T) translation and 76% for Taiwanese to Mandarin (T/M) translation. The percentage of *Good* generations for the Top 1 was 63% for M/T translation, and it was 60% for T/M translation. We examined the translation templates in a specific domain and found that 100% translation accuracy could be achieved. In other words, translation errors occurred only as a result of speech recognition errors, such as word recognition errors and segmentation errors. The results show that T/M had poorer performance than M/T. This is perhaps because spoken Taiwanese has more tones than Mandarin; thus, it is harder for T/M translation spotting to find an appropriate translation template.

Table 5. Average accuracy of baseline spotting and the improvement in Mandarin-to-Taiwanese Translation.

	1		2		3		4	
Template Size	Baseline		Baseline + unVE		Baseline + TcW		Baseline + unVE +TcW	
	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5
150	0.5	0.63	0.6	0.83	0.63	0.83	0.76	1
250	0.5	0.63	0.6	0.83	0.63	0.83	0.76	1
350	0.46	0.6	0.56	0.8	0.6	0.8	0.73	0.96
450	0.46	0.6	0.56	0.8	0.6	0.8	0.73	0.96
550	0.43	0.6	0.56	0.76	0.6	0.76	0.7	0.93
650	0.43	0.56	0.53	0.73	0.56	0.76	0.7	0.93
750	0.43	0.5	0.53	0.73	0.56	0.73	0.7	0.9
850	0.4	0.5	0.5	0.7	0.53	0.73	0.66	0.86
950	0.4	0.46	0.5	0.7	0.5	0.66	0.66	0.83
1050	0.4	0.43	0.46	0.66	0.46	0.66	0.63	0.8

Table 6. Average accuracy of baseline spotting and the improvement in Taiwanese-to-Mandarin Translation.

Template Size	1		2		3		4	
	Baseline		Baseline + unVE		Baseline + TcW		Baseline + unVE +TcW	
	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5
150	0.46	0.6	0.6	0.83	0.6	0.76	0.76	1
250	0.46	0.6	0.6	0.83	0.6	0.7	0.73	0.96
350	0.46	0.56	0.56	0.8	0.56	0.7	0.7	0.96
450	0.43	0.56	0.56	0.76	0.56	0.66	0.7	0.93
550	0.43	0.53	0.53	0.76	0.56	0.66	0.66	0.86
650	0.43	0.53	0.53	0.73	0.53	0.6	0.66	0.86
750	0.4	0.5	0.5	0.7	0.5	0.6	0.63	0.83
850	0.4	0.5	0.5	0.7	0.5	0.56	0.6	0.8
950	0.4	0.46	0.46	0.66	0.46	0.56	0.6	0.76
1050	0.36	0.43	0.43	0.66	0.46	0.56	0.6	0.76

Table 7. Average accuracy of spotting in multiple speaker testing.

			Template Size (Sp1 model)									
		(Top5)	150	250	350	450	550	650	750	850	950	1050
Baseline +unVE +TcW	Sp1	M2T	1	1	0.96	0.96	0.93	0.93	0.9	0.86	0.83	0.8
		T2M	1	0.96	0.96	0.93	0.86	0.86	0.83	0.8	0.76	0.76
	Sp2	M2T	0.9	0.86	0.83	0.8	0.76	0.73	0.73	0.7	0.66	0.66
		T2M	0.86	0.83	0.8	0.76	0.76	0.73	0.7	0.7	0.7	0.66
	Sp3	M2T	0.86	0.83	0.8	0.76	0.73	0.73	0.7	0.66	0.66	0.63
		T2M	0.83	0.8	0.76	0.76	0.73	0.7	0.7	0.66	0.63	0.63

6. Conclusion

In this work, we have proposed an approach that retrieves identified target speech segments by carrying out multiple-translation spotting on a source input. According to the retrieved speech segments, the target speech can be further generated by using the waveform segment concatenation-based synthesis method. Experiments using Mandarin and Taiwanese were performed on Pentium® PCs. The experimental results reveal that our system can achieve an average translation understanding rate of about 78%.

Acknowledgements

This article is a partial result of Project NSC 90-2215-E-006-009, sponsored by the National Science Council, Taiwan, R.O.C.

References

Lavie, A., A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavalda, T. Zeppenfeld and P. Zahn, "JANUS III: Speech-to-Speech Translation in Multiple Languages," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 22(I) 1997, pp. 99–102.

Wahlster, W., "Verbmobil: Foundations of Speech-to-Speech Translation," New York: Springer-Verlag Press, 2000.

Casacuberta, F., D. Llorens, C. Martinez, S. Molau, F. Nevado, H. Ney, M. Pastor, D. Pico, A. Sanchis, E. Vidal and J. M. Vilar, "Speech-to-Speech Translation Based on Finite-State Transducers," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 26(I) 2001, pp. 613–616.

Sugaya, F., T. Takezawa, A. Yokoo and S. Yamamoto, "End-to-End Evaluation in ATR-MATRIX: Speech Translation System between English and Japanese," *Proceedings of European Conference on Speech Communication and Technology*, 6(I) 1999, pp. 2431–2434.

Macklovitch, E., M. Simard and P. Langlais, "TransSearch: A Free Translation Memory on the World Wide Web," *Proceedings of International Conference on Language Resources & Evaluation*, 3(I) 2000, pp. 1201–1208.

Michel, S., "Translation Spotting for Translation Memories," *Proceedings of HLT-NAACL Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, 2003, pp. 65–72.

Véronis, J. and P. Langlais, "Evaluation of Parallel Text Alignment Systems – The ARCADE Project," in J. Véronis (ed.): *Parallel Text Processing*. Dordrecht: Kluwer Academic, 2000, pp. 369–388.

Dorr, B. J., "Machine Translation: A View from the Lexicon," The MIT press, 1993.

Wang, J. F., B. Z. Houg and S. C. Lin, "A Study for Mandarin Text to Taiwanese Speech System," *Proceedings of the 12th Research on Computational Linguistics Conference*, 1999, pp. 37–53.

Sher, Y. J., K. C. Chung and C. H. Wu, "Establish Taiwanese 7-Tones Syllable–based Synthesis Units Database for the Prototype Development of Text-to-Speech System," *Proceedings of the 12th Research on Computational Linguistics Conference*, 1999, pp. 15–35.

Liu, J. and L. Zhou, "A Hybrid Model for Chinese-English Machine Translation," *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, 2(I) 1998, pp.1201–1206.

Yamabana, K., K. Hanazawa, R. Isotani, S. Osada, A. Okumura and T. Watanabe, "A Speech Translation System with Mobile Wireless Clients," *Proceedings of the Student Research Workshop at the 41st Annual Meeting of the Association for Computational Linguistics*, 41(II) 2003, pp. 119–122.