Behavioral synchronization between speech and finger tapping provides a novel approach to the improvement of speech recognition accuracy. We combine a sequence of finger tapping timings recorded alongside an utterance using two distinct methods: in the first method, HMM state transition probabilities at the word boundaries are controlled by the timing of the finger tapping; in the second, the probability (relative frequency) of the finger tapping is used as a ‘feature’ and combined with MFCC in a HMM recognition system. We evaluate these methods through connected digit recognition under different noise conditions (AURORA-2J) and LVCSR tasks. Leveraging the synchrony between speech and finger tapping provides a 46 % relative improvement and a 1 % absolute improvement in connected digit recognition experiments and LVCSR experiments, respectively.

1. INTRODUCTION

Our movements consist of various behaviors, and these behaviors often become synchronized with one another. Many studies have provided analysis and models of synchronizations between different human behaviors, including finger tappings. In [1], for example, a coupled oscillator has been proposed as a mathematical model. Cummins et al. studied the synchronization between speech utterances based on extensive experiments on synchronized speech [2],[3]. Through the experiments, self-entrainment which is the same as that in other types of behavioral synchronization is confirmed.

Following on this, we utilize the synchronization between finger tapping and speech utterance in order to improve speech recognition accuracy in noisy environments, because finger tappings provide information which can help to disambiguate word phrasing and boundaries. Since finger tapping can be generated eyes free and can be recorded through various input devices, for example, the click of a mouse button, a key hit on a keyboard or a tap on a touch pad, it provides a simple, easily obtained modality which nonetheless offers the potential to compensate for degradation in speech information.

Push-to-talk systems are those in which the user manually signals the start of an utterance. The method proposed in this paper is more powerful than push-to-talk due to its specific capability of robustness against tapping errors. As discussed in Section 5, the definition of a word as a single utterance unit is often unclear in agglutinative languages such as Japanese, in which tapping at each word fails for approximately 20% of words. The modeling of the synchronization and coordination between finger tappings and utterances is the fundamental issue addressed in this paper.

The rest of this paper consists of the following sections. In Section 2, the relationships between word boundaries and tapping timing are analyzed. In Section 3, we propose two different approaches to modeling the synchronization between an utterance and a finger tapping sequence. We describe our experiments and give evaluations for connected digit utterances and LVCSR in Sections 4 and 5, respectively.

2. FINGER TAPPING SEQUENCE

Figure 1 shows an example of a speech waveform (a) and the associated finger tapping timings (b), recorded for a connected digit utterance (4-5-1, /yoNgo:ichi/). The subject is instructed to tap for every word during recording. In this example, finger tappings are consistently located at the beginning of every digit uttered. However, in general, the relative position of a finger tapping to the word boundary is not fixed, as shown in Figure 2, where the histogram is calculated using 220 subjects’ connected digit utterances.

The figure shows that the distribution of finger tapping timings is centered at the word boundary position but has an 88 ms standard deviation, which is approximately 1/4 of the averaged duration (360 ms) of the digit utterances. (Word boundaries are estimated by forced alignment by applying matched acoustic models to clean speech.) The synchronous finger tapping sequence provides relevant timing information for the given utterance, however, the integration
of speech and finger tapping must deal with the observed fluctuations.

3. MODELING SYNCHRONIZATION

In order to utilize the finger tapping timing information for speech recognition, two different strategies, namely, making inter-word transitions synchronous with tapping, and feeding finger tapping timing to the HMM as an element of the feature vector, are exploited.

3.1. Synchronous State Transition (SST)

In the first strategy, HMM state transitions between words are controlled so that the inter-word state transition occurs only around the finger tapping. This method can be implemented by giving a penalty score $\gamma[n]$ to each sentence hypothesis when the inter-word transition occurs at the $n$th frame. The simplest example of the penalty is a binary function which has the penalty in the regions distant from tapping, as shown in Figure 1 (c). The function $\gamma[n]$ is designed empirically.

3.2. Tapping Probability Feature (TPF)

The second strategy is to generate a feature stream which is associated with the alignment information from the given finger tapping sequence. In the implementation, the below vector stream $X[n]$ is used for the feature vector of HMMs, i.e.,

$$X[n] = [x_1[n], \ldots, x_N[n], p[n]]^T.$$  

(1)

Here, $x_i[n]$ is the $i$th element of the speech spectral feature, for example, MFCC, and $p[n]$ is the feature associated with the tapping probability at frame $n$. In the simplest case, a binary point sequence:

$$t[n] = \sum_{i=1}^{N} \delta[n - n_i],$$  

(2)

where $N$ is the total number of tappings and $n_i$ is the location of the $i$th tapping, can be used for $p[n]$.

However, as shown in Figure 2, the location of the tapping has a distribution, and a smoothed version of the tapping sequence, i.e.

$$p[n] \equiv w[n] * t[n],$$  

(3)

is used. For the filter $w[n]$, we used a Gaussian window,

$$w[n] = \frac{1}{\sqrt{2\pi}\alpha} \exp\left\{-\frac{1}{2}\left(\frac{n - \beta}{\alpha}\right)^2\right\},$$  

(4)

with parameters $\alpha$ and $\beta$. $\alpha$ controls the width of the distribution and $\beta$ controls the bias and is calculated from the histogram shown in Figure 2. An example of the tapping probability feature is shown in Figure 1 (d).

4. CONNECTED DIGIT RECOGNITION EXPERIMENT

In agglutinative languages like Japanese, a word is not a clearly defined unit in an utterance. Therefore, in order to evaluate the potential performance of the above proposed method, we use connected digit utterances where the boundary between digits is clear.
Table 1. Baseline results: WER under the same evaluation framework with AURORA2. The performance is obtained for the utterances with finger tapping.

<table>
<thead>
<tr>
<th>Reference Word Error Rate</th>
<th>Set A</th>
<th>Set B</th>
<th>Set C</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi</td>
<td>6.62%</td>
<td>8.81%</td>
<td>11.78%</td>
<td>8.53%</td>
</tr>
<tr>
<td>Clean</td>
<td>47.95%</td>
<td>47.09%</td>
<td>48.47%</td>
<td>47.71%</td>
</tr>
<tr>
<td>Average</td>
<td>27.29%</td>
<td>27.95%</td>
<td>30.13%</td>
<td>28.12%</td>
</tr>
</tbody>
</table>

Fig. 3. Relative improvements obtained by the proposed methods. N: Naive method (predetermines the digit length according to the number of tappings), S: Synchronous State Transition, F: Tapping Probability Feature, SS: Spectral Subtraction.

4.1. Experimental Setup

For the experimental evaluation, we have recorded 40886 utterances of 1-7 digit strings from 220 (110 male and 110 female) speakers. At the recording, each speaker was instructed to make a finger tapping at each word. In this recording, only 0.0318% of digit utterances do not have a corresponding finger tap.

For the digit strings, Japanese translations of TI-DIGIT strings in the AURORA2 [4] standard evaluation corpus were used so that the same evaluation framework with AURORA2 (the noise robustness is evaluated for different noise and SNR conditions under matched and mismatched conditions) can be applied[5]. The overall performance is measured by the average relative improvement of the word accuracy from the baseline result, which is listed in Table 1.

The recognition system is also the same as the standard AURORA setup. Each of the 11 digit utterances was modeled by a strictly left-to-right word HMM with 16 states. Each state has 20 Gaussian mixture PDF. Silence is modeled by a 3-state HMM with 36 Gaussians per state, whereas short pauses are modeled by a single-state HMM tied to the second state of the silence HMM.

4.2. Results

The experimental results in Figure 3 show the average relative improvements from the baseline. In addition to the proposed methods, the results of a naive method for using the number of finger tappings in order to limit the length of the digit string are also given as (N). Obviously, any error in tapping produces a recognition error in this naive method, and thus it is not applicable to general utterances in agglutinative languages. Both of proposed methods (S and F) provide more than 15% improvement. A better result was obtained when we combined the two methods (S+F). The combined result (S+F) outperformed the naive method and spectral subtraction result (SS). Since the finger tapping contains different information from a speech waveform, combining the proposed method and spectral subtraction can further improve the result to be 46% (SS+D+F).

5. APPLICATION TO LVCSR

5.1. Effects of missing tap

For agglutinative languages, there is no distinct definition of a word as an utterance unit, and therefore the positions of finger tappings at each word are expected to be inconsistent and frequently missed. In order to evaluate the robustness of the proposed method against the missing of finger tappings, we performed the same experiment as before, but with finger tapping records removed from the corpus at a given probability.

As shown in Figure 4, as the ratio of missing tappings increases, the improvement decreases. However, more than a 10% relative improvement is still obtained when 20% of finger tappings are missed. One should note that when using the naive method, that is, when assuming the number of tappings is equal to the number of words in the sentence, as many errors as there are missing taps will occur.
5.2. LVCSR Experiment

Finally, we have conducted a preliminary experiment in applying the proposed method to LVCSR. For the test corpus, 10 male speakers each read 100 sentences extracted from newspaper articles with finger tapping; thus, in total 1000 sentences were recorded. Since the definition of a word is not clear in Japanese, for the data collection, subjects were instructed to tap at each bunsetsu. (A bunsetsu in Japanese corresponds to a phrasal unit such as a noun phrase or a prepositional phrase in English.)

For 5584 bunsetsu in the test corpus, 5789 finger tappings were recorded as the result of 446 (8%) deletions and 651 (11%) insertions from/to the bunsetsu boundaries. The number of errors is incredibly high compared to that of the connected digit utterances. The insertion and deletion numbers are quite different for each speaker, that is, they range from 5 to 177 and from 3 to 165, respectively.

The results of the recognition experiments are plotted in Figure 5. Here, car-noise is added to the clean recorded speech at 20 and 15 dB SNR in order to determine the degree of improvement under noisy conditions. Unlike the connected digit case, 2k state triphone acoustic models and a 20k vocabulary trigram language model trained by 40k read sentences and newspaper articles of 75 months, respectively, are used for the basic recognition system.

As shown in Figure 5, even though 20 % of finger tappings do not precisely correspond to the bunsetsu boundaries, combining the tapping timing information improved the recognition accuracy in the noisy environment. However, this improvement is small compared with that for the connected digit cases.

5.3. Discussions

There are several reasons for the small improvement in LVCSR compared to that in the simulated error condition in Section 5.1. First, the tapping timing in the newspaper readings has much greater fluctuation than the connected digit utterances. The standard deviation of the distribution of the tapping timing relative to the bunsetsu boundaries was 2.5 times greater (222 ms) than that in the connected digit case. Second, insertion errors of tapping, which are not discussed in Section 5.1, were contained in the utterances. Furthermore, the baseline recognition accuracy of LVCSR is heavily dependent on the language model, and therefore the balancing between the language probability weight and the parameters in the proposed methods may need further optimization.

6. CONCLUSIONS

We proposed a novel approach to the enhancement of speech interfaces by using simultaneous finger tapping. Two different strategies for combining the finger tapping information and speech spectral information, that is, Synchronous State Transition (SST) and Tapping Probability Feature (TPF), are proposed. We found that the basic effectiveness of using synchrony between speech and finger tapping provides a 46 % relative improvement and a 1 % absolute improvement in the connected digit recognition experiments and LVCSR experiments, respectively.

In this paper, we used words or bunsetsu in Japanese as the basic unit for synchronization with finger tapping, because these seem to be natural units of utterance. Further investigations on aspects of microscopic synchronization such as finger tapping and syllable onset may improve the results for read or natural speech.

7. REFERENCES