SPEECH RECOGNITION WITH A RE-SPEAK METHOD FOR SUBTITLING LIVE BROADCASTS

Toru Imai, Atsushi Matsui, Shinichi Homma, Takeshi Kobayakawa, Kazuo Onoe, Shoei Sato, and Akio Ando

NHK (Japan Broadcasting Corp.)
Science and Technical Research Laboratories, Tokyo
{imai.t-mq, matsui.a-hk, homma.s-fc, kobayakawa.t-ko, onoe.k-ec, satou.s-gu, andou.a-io}@nhk.or.jp

ABSTRACT
This paper describes a "re-speak" method for subtitling live TV broadcasts using a speech recognition system. Original on-location speech in live sport or music programs contains background noise, spontaneous or emotional speech, and the voices of speakers unknown to the recognition system, all of which cause recognition performance to deteriorate. However, if a different individual, to which the system has been adapted, carefully rephrases the original utterances in a studio, these problems can be largely overcome. Recognition experiments showed that rephrasing the commentary was effective in reducing perplexities and word error rates compared with simply repeating it. Speech recognition using the re-speak method was applied in practice to a music-based variety show and the 2002 Winter Olympic Games in order automatically to produce simultaneous subtitles for hearing-impaired viewers. A word error rate below 5% and a subtitle display delay time below three seconds were achieved.

1. INTRODUCTION
There is a great need for simultaneous subtitling of live broadcast programs for the hearing impaired and elderly. Since March 2000, NHK (Nippon Hosho Kyokai; Japan Broadcasting Corp.) has been subtitling broadcast news in Japanese every day by using a speech recognition system developed at NHK [1]. A reason to use speech recognition is that it is rather difficult to recognize all of such commentary with practically satisfied accuracy by a current speech recognizer, its output results are not accurate enough for subtitles. The reasons are background noise, unspecified speakers, and speaking styles that may not match acoustic models and language models. It is difficult to collect enough training data of audio and texts in the same domain as the target program. Therefore we adopt the re-speak method to eliminate such problems.

In the method a different speaker from the original speakers of the target program carefully rephrases what he or she hears. We call him or her a "re-speaker". The re-speaker wears headphones, listens to live TV programs, and speaks what he or she hears in different words if needed so that its meaning is clearer or more acceptable and the expression is advantageous to speech recognition. The method has the following advantages for speech recognition:

- The re-speaker is in a silent studio with headphones. The utterances have no background noises even if the target live program has high levels of noises.
- The re-speaker is known and acoustic models can be adapted prior to the program with relatively many adaptation data.
- The re-speaker arranges the phrases by substitution, deletion, or insertion of words so that the speech recognition becomes easier, the recognized words are more accurate, and the subtitles are easily understood.
- Filled pauses, exclamatory phrases, or sounds indicating hesitation are not repeated.
• Incomplete sentences without a subject word, which often appear in Japanese conversational speech, are supplemented with a subject.
• Complicated construction of a sentence like inversion is changed into a simple sentence.
• Conversational speech is rephrased into a kind of prepared read speech.
• Knowing the ability of the speech recognizer, the re-speaker immediately selects words that can be easily recognized from among the original words.
• When replacing words, the re-speaker takes synonyms that can be easily recognized by experience.

Apparently the way of re-speaking affects the speech recognition performance. It is necessary to develop the skill of the re-speakers so that the final subtitles become better.

3. EXPERIMENT

3.1. Evaluation data
Experiments were performed to examine how the re-speak method was effective to speech recognition. Target programs were speed skating and ski jumping of the 2002 Olympic Winter Games. By using recorded videotapes of those games, four female re-speakers simulated the commentary in two different ways. One was precise “repeating” just as they hear. The other way was careful “rephrasing” under some instructions above. Table 1 shows the number of words and sentences of the evaluated speech in both speaking strategies. The rephrased sentences are less than repeated ones due to deletion of words or sentences.

<table>
<thead>
<tr>
<th>program</th>
<th>speed skating</th>
<th>ski jumping</th>
</tr>
</thead>
<tbody>
<tr>
<td>method</td>
<td>repeat</td>
<td>rephrase</td>
</tr>
<tr>
<td>#sentence</td>
<td>977</td>
<td>11,289</td>
</tr>
<tr>
<td>#word</td>
<td>11,289</td>
<td>8,545</td>
</tr>
<tr>
<td>weight</td>
<td>50K</td>
<td>1,616</td>
</tr>
<tr>
<td>#sentence</td>
<td>1,616</td>
<td>20,302</td>
</tr>
<tr>
<td>#word</td>
<td>1,367</td>
<td>15,419</td>
</tr>
</tbody>
</table>

Table 1: Evaluation data.

3.2. Language Model
N-gram language models of bigrams and trigrams were trained [3] by using some kinds of weighted texts after Japanese morphological analysis [4]. The base texts were manuscripts of NHK’s general news. Then we added some kinds of related texts to the target programs with higher mixture weights in a count level, as shown in Table 2. The weighting values were sub-optimally determined by using a similar test set in a preliminary experiment. The N-gram language models were smoothed by backing-off and Good-Turing discounting where cutoffs were set to 2 and 3 for bigrams and trigrams, respectively. The vocabulary size was initially 18K from the general news and 33K finally after the addition of all of the training texts.

3.3. Acoustic Model
The evaluated speech data were acoustically analyzed into 39 parameters (12 MFCCs with log-power and their first- and second-order regression coefficients) every 10msec after digitization at 16 kHz and 16 bits with the Hamming window of 25msec width. Acoustic models of gender-dependent and speaker-independent state-clustered 16-mixture triphone HMMs were initially trained from NHK’s news data of 200 hours. Then they were adapted to each re-speaker by MAP [5] and MLLR [6] with 3-hour speech data recorded when the re-speakers trained themselves by using different programs with the evaluated ones. The number of Japanese phonemes was 42. The number of triphone HMMs was 3,126 with 1,973 states.

<table>
<thead>
<tr>
<th>source</th>
<th>#sentence</th>
<th>#word</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcriptions of preliminary games in this Olympic</td>
<td>4.2K</td>
<td>50K</td>
<td>400</td>
</tr>
<tr>
<td>Expected word list (athletes, rules, and etc.)</td>
<td>300</td>
<td>1K</td>
<td>100</td>
</tr>
<tr>
<td>Transcriptions of past Olympic games</td>
<td>11K (skate)</td>
<td>130K</td>
<td>230K</td>
</tr>
<tr>
<td>Transcriptions of past Olympic games</td>
<td>18K (jump)</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Sports news on the web</td>
<td>17K (skate)</td>
<td>330K</td>
<td>10</td>
</tr>
<tr>
<td>Transcriptions of sports news</td>
<td>19K (jump)</td>
<td>360K</td>
<td>10</td>
</tr>
<tr>
<td>Manuscripts and transcriptions of sports news</td>
<td>250K</td>
<td>4M</td>
<td>10</td>
</tr>
<tr>
<td>Manuscripts of general news</td>
<td>2.8M</td>
<td>107M</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2: Training texts for the language models.

3.4. Decoder
In order to output recognition results for subtitles as quickly as possible, the decoder makes an early decision without waiting for the end of input utterances [7]. The low latency decoder is based on 2-pass searches. During the first pass the decoder periodically executes the second pass that rescores partial N-best word sequences up to that time. If the rescored best word sequence has words in common with the previous one, that part is regarded as likely to be correct and is decided to be a part of the final result. This method is not theoretically optimal but makes a quick response with a negligible increase in word errors.

In the first pass, the word-dependent N-best algorithm [8] is carried out with a modified Viterbi beam search. During the search top-\(n\) \((n<<N)\) paths that have different previous words are saved in each HMM state. If multiple paths reach the same state with the same previous word, only the best path is saved. At a word-end phoneme, its word-id, the word-end frame, scores and back-pointers to the previous word-ends for the \(n\) paths are saved as a word lattice, and only the best path propagates to following words. Pruning in the beam search is based on a global beam width of log-likelihood and a narrower word-end beam width.

The first pass procedure runs on a tree-structured phoneme network [9] where initial phoneme sequences are shared by some words. Instead of copying the trees dynamically, we use a single and static tree. To apply a language model earlier, the maximum bigram probability among words sharing a phoneme node is used for factoring and pruning. The maximum bigram probabilities are computed quickly on the fly by comparing the
pre-processed maximum values from existing bigram entries and unigram entries (i.e. backing-off cases) at each node.

Every $\Delta t$ frames during the first pass the uncompleted word lattice is traced back from the word-end pointed by the node with the best score to get $N$-best word sequences up to that frame. The $N$-best word sequences do not include the premature word to which the best node belongs if the node is not a word-end. If there are words previously decided, word-end branches in that portion need not be expanded for the $N$-best. Then the $N$-best word sequences are rescoring as the periodical second pass using a trigram language model to get the 1-best word sequence up to that frame. The current 1-best word sequence is compared with the previous 1-best word sequence obtained $\Delta t$ frames before. If some words are repeated after the last decision, those words are regarded as likely to be correct and are decided as a part of the final result. If there are no repeated words, the 1-best word sequence is regarded as unstable and no decision is made. Since words close to the current frame are unstable, the latest $M$ words are excluded from the target to be decided. Even if the best word in a pre-determined portion has been changed, it is ignored.

To examine only the effect of the re-speak method, the early decision was not made in this experiment, but it was done in the next section for real broadcasting with the values of $n=4$, $N=200$, $\Delta t=30$ and $M=1$. In the broadcasting, to allow re-speakers more than one in rotation at ten- or twenty-minute shifts and use adapted acoustic models for each re-speaker, the phoneme network in the decoder has multiple links to every adapted acoustic model and accepts manual selection of the re-speaker without stopping the decoder.

### 3.5. Result

Table 3 shows the performance of two language models for speed skating and ski jumping with different speaking strategies of "repeat" and "rephrase". Their references were made from evaluation data for each method individually. The table indicates that rephrasing greatly reduces the perplexities and rates out of vocabulary (OOV). Rephrasing was also effective to increase the rates that trigram entries hit the evaluation data.

The recognition results are shown in Table 4. The word error rates were also reduced by means of rephrasing. The reduction rates of word errors were 37% and 27% relative for speed skating and ski jumping, respectively.

Table 3: Performance of the language models.

<table>
<thead>
<tr>
<th>program</th>
<th>speed skating</th>
<th>ski jumping</th>
</tr>
</thead>
<tbody>
<tr>
<td>method</td>
<td>repeat</td>
<td>rephrase</td>
</tr>
<tr>
<td>perplexity</td>
<td>142.2</td>
<td>99.6</td>
</tr>
<tr>
<td>trigram hit</td>
<td>61.5%</td>
<td>65.9%</td>
</tr>
<tr>
<td>OOV</td>
<td>0.28%</td>
<td>0.19%</td>
</tr>
</tbody>
</table>

Table 4: Recognition results for off-line experiments.

<table>
<thead>
<tr>
<th>program</th>
<th>speed skating</th>
<th>ski jumping</th>
</tr>
</thead>
<tbody>
<tr>
<td>method</td>
<td>repeat</td>
<td>rephrase</td>
</tr>
<tr>
<td>word error rate</td>
<td>7.6%</td>
<td>4.8%</td>
</tr>
</tbody>
</table>

As shown in Table 3, rephrasing is effective to decrease perplexities and entropy. We classified the rephrased evaluation utterances for speed skating into five patterns dependent on the difference from the repeated utterances: insertion of words, replacement of words, deletion of words, deletion of sentences, and their mixed cases. Each pattern decreases entropy from the case of repeating as is, and their ratios are shown in Table 5 with the numbers of occurrences. Despite the highest occurrence, the pattern of deleting sentences showed the lowest ratio. The most effective pattern was to insert words for the supplement of meanings or grammar in spite of the least frequent occurrence.

Table 5: Classification of rephrased utterances.

<table>
<thead>
<tr>
<th>pattern</th>
<th>number of occurrences</th>
<th>contribution to reducing entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insertion of words</td>
<td>118</td>
<td>26.1%</td>
</tr>
<tr>
<td>Replacement of words</td>
<td>123</td>
<td>19.6%</td>
</tr>
<tr>
<td>Deletion of words</td>
<td>114</td>
<td>19.6%</td>
</tr>
<tr>
<td>Deletion of sentences</td>
<td>310</td>
<td>13.0%</td>
</tr>
<tr>
<td>Mixed cases</td>
<td>128</td>
<td>21.7%</td>
</tr>
</tbody>
</table>

Figure 1: Perplexities for speed skating with various mixture weights.

Figure 2: Perplexities for ski jumping with various mixture weights.
The effectiveness to give higher weights to relevant texts in
the training of the language models is shown in Figure 1 and 2
for speed skating and ski jumping, respectively. They show
perplexities when the mixture weights in a count level for
transcriptions of preliminary games in this Olympic were
changed from 1 to 2000. Perplexities for rephrased speech were
less than those for repeated speech at any weights by
approximately 30%. Although the weighted texts were
precisely transcribed from the commentary of very related
programs, the perplexities for the repeated speech did not
become less than those for the rephrased speech even if the
mixture weights became larger.

4. BROADCASTING WITH SUBTITLES
The speech recognition system with the re-speak method was
practically applied to NHK’s live broadcasts of Japanese top-
rated music-based variety show and the 2002 Olympic Winter
Games in order to automatically produce simultaneous subtitles
for hearing impaired. The utterances spoken by re-speakers
were fed into the speech recognizer (COMPAQ AlphaStation
DS20E 833MHz), and its results were directly sent to viewers
as the teletext without editing so that subtitles are immediately
on the screen and stay relevant to the scene. The language
models were trained by using relevant texts in a similar way as
the experiments in the previous section.

As shown in Table 6, the word error rates for the music-
based variety show and the opening ceremony of the Olympic
were equal to or less than 5% and acceptable for subtitles. The
delay of subtitles from the original speech was less than three
seconds. Despite the fact that the speech recognition results
with some errors were broadcast as is, we got many positive
responses from viewers. Hearing-impaired viewers expressed
delight at finally being able to enjoy the programs with their
families. Figure 3 is the captured image from the live broadcast
of the opening ceremony of the 2002 Olympic Winter Games
with simultaneous subtitles produced by the speech recognition.

<table>
<thead>
<tr>
<th>program</th>
<th>music-based variety show</th>
<th>opening ceremony of the Olympic</th>
</tr>
</thead>
<tbody>
<tr>
<td>perplexity</td>
<td>35.5</td>
<td>39.6</td>
</tr>
<tr>
<td>word error rate</td>
<td>5.0%</td>
<td>2.8%</td>
</tr>
</tbody>
</table>

Table 6: Recognition results for real subtitling.

Figure 3: Live broadcast of the opening ceremony of the 2002 Olympic Winter Games with simultaneous subtitles produced by the speech recognition.

5. CONCLUSIONS
This paper presented speech recognition with the re-speak
method for subtitling live broadcasts. In the experiments, it was
shown that careful rephrasing reduced the perplexities and
word error rates compared with repeating just as heard. Above
all rephrasing ways, inserting words complementarity was most
effective. The application to live broadcasts of the music-based
variety show and the 2002 Olympic Winter Games was
realized for simultaneous subtitles with word error rates equal
to or less than 5% and was positively welcomed by viewers of
hearing impaired.

6. REFERENCES