VOLUNTEER-BASED IPA JAPANESE DICTATION FREE SOFTWARE PROJECT

Kiyohiro Shikano(Nara Institute of Science and Technology), Tatsuya Kawahara(Kyoto University),
Katsunobu Itou(Electrotechnical Laboratory), and Kazuya Takeda(Nagoya University)

ABSTRACT

The IPA Japanese dictation project is aiming at building a Japanese LVCSR free software platform, with the collaboration of many speech and language researchers. Our LVCSR platform will not only encourage the development in the speech recognition applications but also raise the potential in speech recognition research in Japan. As this project is partially supported by IPA, the actual activities rely on the volunteer efforts, which include (1) LVCSR speech and text database IPSJ WG efforts, and (2) IPA Japanese dictation project efforts.

1. INTRODUCTION

LVCSR (Large Vocabulary Continuous Speech Recognition) is an important basis of application developments of speech technologies in the coming decade, which include a voice-activated word processors, voice dialog systems such as car navigation, voice-activated games, and CAI. Dictation free software will be useful for LVCSR application development as shown in Figure 1.

In USA and Europe, research projects such as ARPA and SQALE have resulted in a substantial technology progress in LVCSR[1,2]. In Japan, LVCSR research has not been supported with the government these ten years. We had to construct LVCSR speech database and develop Japanese LVCSR algorithms by the volunteer-based or independent efforts. We have been involved in the following two volunteer-based activities. (1) IPSJ LVCSR speech database WG, and (2) IPA Japanese dictation free software project. The duration and web sites for these activities are shown in Table 1.

2. IPSJ LVCSR SPEECH DATABASE WG [3,4]

In Japan, the ASJ continuous speech database (ASJ-PB) which contains about 10,000 phonetically balanced sentence utterances has been widely used as a public continuous speech database. This database is not quite enough for LVCSR research. LVCSR speech database working group was initiated by young researchers in summer, 1995 to catch up with the USA and Europe LVCSR speech database construction movements. IPSJ Special Interest Group on Spoken Language Processing mentally supported this WG. The WG officially began in November, 1995. The WG original targets were summarized as follows; (1) large size of text corpus, (2) speech database for dictation, and (3) basic models and tools for dictation. The WG members are shown in Table 2.

We had not been able to use a large text database such as newspaper articles, because Japanese texts are written without spacing between words. Moreover, pronunciation ambiguity of Kanji (Chinese characters) makes the use of large texts difficult.

Table 3 shows an example of Japanese newspaper articles and their morphological analysis outputs. The WG intended even to develop a morphological analysis system and a pronunciation annotation system. These system implementation efforts were succeeded to the IPA dictation project.

As for text corpus for dictation, Mainichi newspaper articles between 1991 and 1994 were chosen. The Mainichi
Newspaper is one of major nation-wide general newspapers in Japan. Our first target was a sentence selection for LVCSR speech database. There was no public morphological analysis system, which was accurate enough to construct a language model. We decided to use the morphological tagged corpus of the Mainichi newspaper, which is distributed as a part of the RWCP-Text-Corpus.

First, we extracted all of the article paragraphs from the original CD-ROM with RWCP-text-Corpus. Next, paragraphs without a period were removed for readability filtering. These removed paragraphs are poems, recipes, tables, lists, and so on. As another readability filter, sequences of morphological units (words) between special symbols such as round brackets, which were automatically estimated as unread expression, were removed. Finally, the paragraphs were divided into sentences according to periods or equivalent symbols. It was necessary to divide the text corpus into a language model training section and a sentence selection section. The most recent three months' data were kept for the sentence selection. The rest of 45 months was used for language modeling. The first step to construct a language model is to make a word-frequency list from the data we kept for the sentence selection. The rest of 45 months was used for language modeling. The first step to construct a language model is to make a word-frequency list from the training texts using their morphological tags. Next a word bigram language model was generated using the CMU SLM Toolkit [9], where vocabulary sizes are 5k and 20k words.

Table 4: Sentence selection criteria

<table>
<thead>
<tr>
<th>Vocabulary Size</th>
<th>MID size (5k words)</th>
<th>LARGE size (20k words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of Sentence</td>
<td>NORMAL</td>
<td>15 – 19 for MID size, 5 – 29 for LARGE size</td>
</tr>
<tr>
<td>Word Perplexity</td>
<td>Low</td>
<td>0 – 10 for MID, 0 – 70 for LARGE</td>
</tr>
<tr>
<td>Number of Out-of-Vocabulary words</td>
<td>+</td>
<td>one OOV word included</td>
</tr>
</tbody>
</table>

Table 5: Design of sentence set

<table>
<thead>
<tr>
<th>Perplexity</th>
<th>MID</th>
<th>Mid</th>
<th>High</th>
<th>MID</th>
<th>Mid</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>MID</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>LARGE</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>LARGE+</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>LARGE++</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6: Members of IPA dictation project

| Leader: K. Shikano (NAIST) |
| Phoneme Model G: K. Takeda (Nagoya Univ.) |
| M. Minamisato (Toyokashi Univ.) |
| Recognition G: T. Kawahara (Kyoto Univ.) |
| T. Kobayashi (Waseda Univ.) |
| Language Model G: K. Itou (ETL) |
| K. Shikano, M. Yasumoto (Tokyo Univ.) |
| Morphological Analysis G: A. Yamada (Astem) |
| T. Ueno (NAIST) |

Figure 2: R&D background and plan

**1997**

5K System by EWS

**1998**

20K System by EWS

**1999**

20K, 40K System by PC

Basic System 20K Vocab. System Improvement

**Figure 3: IPA project goals**
3.1. Acoustic Models

The acoustic models are based on Gaussian mixture HMM. They are available in the HTK[8] format. We have trained several kinds of Japanese acoustic models from context-independent phone (monophone) to triphone models. In the triphone modeling, the HTK decision tree-based clustering is carried out to make physical triphones that group similar contexts and can be trained with reasonable data. By changing the threshold of clustering, we set up a variety of models, whose numbers of the HMM states are about 1000, 2000, and 3000. Numbers of Gaussian mixtures in an HMM state are 4, 8, 16, and 32 according to their accuracy. Every HMM phone model consists of three states. These HMMs also include gender-dependent and gender-independent HMMs.

The set of 43 Japanese phones defined by ASJ speech database committee is adopted. These HMM models are trained with ASJ speech database with ASJ-PB and ASJ-JNAS, as shown in Figure 2. The speech data were sampled at 16 kHz and 16 bits. Twelfth-order mel-frequency cepstrum coefficients (MFCC) are calculated every 10 ms. The cepstrum difference coefficients (delta-MFCC) and delta-power are also used. Cepstrum mean normalization (CMN) is performed based on the whole utterance average.

3.2. Language Models

The lexicon is also provided in the HTK format. The Mainichi newspaper articles from Jan. 1991 to Sep. 1994 (45 months) are used to make language models as shown in Figure 2. These articles are divided into words (morphological units) with a morphological analysis program (ChaSen) and a reading annotation program (ChaWan). These programs, ChaSen and ChaWan, are also developed or improved, and are available to the public.

The lexicon consists of the most frequent words. The lexical coverages for 5k and 20k lexicons are 85.8% and 95.7%, respectively. Word bigram and trigram language models are constructed using the CMU-Cambridge SLM toolkit[9] for the predefined 5k and 20k lexicons. They are available in the CMU-Cambridge SLM toolkit format. We also developed an algorithm for reducing the size of back-off trigram models based on the cross entropy criterion. We can reduce the size of trigram parameters into 1/3 ~ 1/10 without the recognition rate degradation[7].

3.3. Decoder

The recognition engine named JULIUS[6] is developed to interface the acoustic models and the language models. JULIUS is composed of two decoding passes. The first pass uses the word bigram, and the second pass uses the backward trigram.

In the first pass, a tree-structured dynamic lexicon with bigram probabilities is adopted with the frame-synchronous beam search algorithm. In order to reduce the computation amount and memory size, one-best approximation is adopted, rather than word-pair approximation. The degradation by one-best approximation in the first pass is successfully recovered by the tree-trellis search in the second pass.

3.4. Japanese Dictation System

The configuration overview of the decoder, the acoustic models and language models is illustrated in Figure 4. In the first pass of JULIUS, the word bigram is applied and the HMM phone model deals with only intra-word phonetic context. The one-best word candidates are stored in the form of word trellis. In the second pass, the word trigram and inter-word phonetic context by the HMM triphone model, which are precise but computationally expensive, are carried out efficiently on the word trellis.

The components, such as the HMM phone models, the language models, and decoder, developed independently at the different sites, are successfully integrated for 5k vocabulary and 20k vocabulary dictation systems.

Since there are several choices of the HMM phone models and the language models, we decide to develop two kinds of dictation systems, the fast version and the accurate version. The fast version uses the HMM monophone model and introduces several approximations such as a tree-structured lexicon fixed with unigram probabilities and the reduced word trigram. The accurate version, of course, uses the HMM triphone model.

3.5. System Evaluation and Target

Performance evaluation by word recognition rate and processing speed was carried out using the fast version and the accurate version for 200 sentence utterances from 46 speakers in February, 1999.

First, the gender-dependent HMM models were used for evaluation. The fast version with the 16 Gaussian mixture monophone model and the 1/10 reduced 20k language model works almost realtime at a standard workstation and PC. The fast version word recognition is 85.4%. The accurate version with the 2000 HMM state and 16 Gaussian mixtures shows 8.4 times realtime and 93.7% word recognition rate. These performance results are summarized in Table 7. These recognition rates are evaluated using an automatic scoring tool, which is an extension of the NIST scoring tool. This scoring tool can evaluate the system performance from various viewpoints of words, word pronunciation, characters, and character pronunciation. This scoring tool shows almost same word recognition rates as manually scored ones.

Second, gender-independent 2000-state 16 Gaussian mixture HMM phone model was evaluated. Evaluation of the
gender-independent HMM phone model was carried out using the 1/10 reduced language model. The word recognition rate is 91.7%, which is comparable with the word recognition rate of 93.3% for the gender-dependent HMM models.

Lastly, these HMM phone models and the decoders were applied to 5k vocabulary dictation in order to evaluate the last one year progress. The recognition speed for the first version and the accurate version was improved drastically more than two times during the last one year. These progress results are illustrated in Figure 8, together with 20k vocabulary dictation performance. The 20k vocabulary performance is almost same as the 5k one for the accurate version. For the fast version, some significant degradation is observed. We have to introduce more accurate phone models instead of the monophone models for the fast version improvement.

4. Conclusion and Future Plan
The baseline Japanese dictation platform we have been developing is proven reasonably well. The baseline platform now works in Unix workstation and Linux PC. We are planing to distribute CD-ROMs and hold a summer school again.

The future plan of the IPA project is to improve each module and its speech application program interface as follows: (1) vocabulary size and language model will be extended to deal with 40k vocabulary size, (2) the morphological analysis program (ChaSen) and the pronunciation annotation program (ChaWan) will be officially released to the public, (3) tied-mixture HMM will be trained and released in order to improve the fast version word accuracy, (4) the accuracy and speed of the decoders will be improved, (5) Windows / Linux PC versions will be released, (6) speech application program interface will be improved, and (7) adaptation programs to a task, a speaker, and environment will be released. Moreover, all kinds of software and tools will be released as a LVCSR workbench with their detailed manuals and supports through networks.

Our volunteer-based efforts to construct LVCSR speech database and to implement free software for Japanese dictation have proved that this kind of a virtual laboratory works successfully and efficiently.

Acknowledgement: First of all, we are grateful to other IPA dictation project members and its advisory members for their contributions and cooperation. We are grateful to IPSJ LVCSR WG members for their various contributions and a lot of efforts. We are also grateful to the ASJ speech database committee for their database collection collaboration and distribution efforts. Lastly, we deeply thank IPA for the understanding and financial support.

This research is partially supported by the IPA (Information Technology Promotion Agency).

REFERENCES

Table 7: JULIUS evaluation(20k vocabulary)

<table>
<thead>
<tr>
<th>Phone as HMM</th>
<th>Monophone</th>
<th>Triphone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language Model</td>
<td>Reduced Trigram (1/10)</td>
<td>Trigram</td>
</tr>
<tr>
<td>Word Rec. Rate</td>
<td>85.4%</td>
<td>93.3%</td>
</tr>
<tr>
<td>Proc. Time</td>
<td>2×RT</td>
<td>4×RT</td>
</tr>
</tbody>
</table>

Figure 8: Year 98 evaluation results and Year 99 targets