SHARABLE SOFTWARE REPOSITORY FOR JAPANESE LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION [*]

Tatsuya Kawahara (Kyoto Univ.)[†], *Tetsunori Kobayashi* (Waseda Univ.),
Kazuya Takeda (Nagoya Univ.), *Nobuaki Minematsu* (Toyohashi Univ. of Tech.),
Katsunobu Itou (ETL), *Mikio Yamamoto* (Tsukuba Univ.),
Atsushi Yamada (ASTEM), *Takehito Utsuro* (Nara Inst. of Sci. & Tech.),
Kiyohiro Shikano (Nara Inst. of Sci. & Tech.)

ABSTRACT

The project of Japanese LVCSR (Large Vocabulary Continuous Speech Recognition) platform is introduced. [1] It is a collaboration of researchers of different academic institutes and intended to develop a sharable software repository of not only databases but also models and programs. The platform consists of a standard recognition engine, Japanese phone models and Japanese statistical language models. A set of Japanese phone HMMs are trained with ASJ (Acoustic Society of Japan) databases of 20K sentence utterances per each gender. Japanese word N-gram (2-gram and 3-gram) models are constructed with a corpus of Mainichi newspaper of four years. The recognition engine **JULIUS** is developed for assessment of both acoustic and language models. The modules are integrated as a Japanese LVCSR system and evaluated on 5000-word dictation task. The software repository is available to the public. [2]

1. INTRODUCTION

In order to build a Large Vocabulary Continuous Speech Recognition (LVCSR) system, high-accuracy acoustic models, large-scale language models and an efficient recognition program (decoder) are essential[1]. Integration of these components and adaptation techniques for real-world environment are also needed. In order to promote both research of various component technologies and development of such complex systems, we have recognized the necessity of a common platform, that is a sharable software repository of not only databases but also models and programs. Sharing a state-of-the-art baseline system enables researchers to concentrate on specific issues and to evaluate new methods under a reasonable environment.

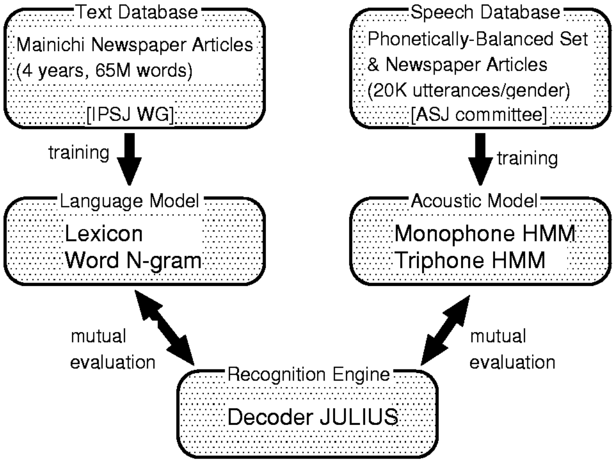

Figure 1: Platform of LVCSR

Japanese LVCSR systems have not been developed mainly because word segmentation has to be done on Japanese texts that are not written without spacing between words. But recent progress of morphological analyzers enables automatic segmentation, thus training of language models with large corpora[2].

We have adopted Mainichi Newspaper, one of the nation-wide general newspapers in Japan, for the sharable corpus of both text and speech[3], and organized a project to develop a standard software repository that includes acoustic and language models and recognition programs. The three-year project (1997-2000), funded by the IPA (Information-technology Promotion Agency), Japan, is a collaboration of researchers of different academic institutes. The software repository as the product of the project is being available to the public. The overview of the corpus and software mentioned here is depicted in Figure 1.

[*] WORK IS SPONSORED BY THE IPA (INFORMATION-TECHNOLOGY PROMOTION AGENCY), JAPAN.

[†] School of Informatics, Kyoto University 606-8501, Japan.

[1] `http://www.itakura.nuee.nagoya-u.ac.jp/~takeda/IPA/`

[2] `mailto: dictation-tk-request@astem.or.jp`

The specifications of acoustic models, language models and recognition engine are described in this paper. We also report assessment of each module under 5000-word Japanese dictation task.

2. SPECIFICATION OF MODELS AND PROGRAMS

2.1. Acoustic Model

Acoustic models are based on continuous density HMM. They are available in the HTK format[4].

We have trained several kinds of Japanese acoustic models from context-independent phone model to triphone models, as listed in Table 1. They are all gender-dependent, namely we set up both male models and female models.

The set of 43 Japanese phones are listed in Table 2. The phone notation is defined by Acoustical Society of Japan (ASJ) committee on speech database. Here, the symbols `a:`~`o:` stand for long vowels and the symbol `q` for a double consonant. Three pause models, `silB`, `silE` and `sp`, are introduced for pauses at the beginning, at the end of utterances and between words, respectively.

Table 1: List of Acoustic Models

model	#states	#mixtures
monophone	129	4, 8, 16
triphone 1000	1000	4, 8, 16
triphone 2000	2000	4, 8, 16
triphone 3000	3000	4, 8, 16

Table 2: List of Japanese Phones

```
a i u e o a:  i:  u:  e:  o:  N w y
p py t k ky b by d dy g gy ts ch
m my n ny h hy f s sh z j r ry
q sp silB silE
```

The acoustic models are trained with ASJ speech databases of phonetically balanced sentences (ASJ-PB) and newspaper article texts (ASJ-JNAS). In total, around 20K sentences uttered by 132 speakers were available for each gender.

The speech data were sampled at 16kHz and 16bit. Twelfth-order mel-frequency cepstral coefficients (MFCC) are computed every 10ms. The difference of the coefficients (Δ MFCC) and power (Δ LogPow) are also incorporated. So the pattern vector at each frame consists of 25 (=12+12+1) variables. Cepstral mean normalization (CMN) is performed on whole utterances to offset the channel mis-match.

Every phone model consists of three states. The decision tree-based clustering is performed to build physical triphones that group similar contexts and can be trained with reasonable data. By changing the threshold of clustering, we set up variety of models whose number of the states is 1000, 2000 and 3000, respectively.

2.2. Lexicon

A lexicon is also provided in the HTK format[4].

It is consistent with both the acoustic model and the language model. The phone symbols are covered with the acoustic model. The lexical entries match the language model.

The vocabulary set consists of the most frequent words (=morphs) in Mainichi newspaper articles from Jan. 1991 to Sep. 1994 (45 months) [3]. In Japanese, lexicon is mainly defined by a morphological analyzer that segments undelimited texts. In order to improve language model, we distinguish lexical entries by not only their notations but also their morphological attributes. The lexical coverage of various vocabulary sizes is listed in Table 3.

Not a few morphs have multiple baseform entries because Japanese Kanji usually has multiple pronunciations. Several entries of notational symbols are rewritten as pauses in pronunciation.

Currently, a lexicon of 5K vocabulary size is available. A 20K lexicon will also be released soon.

Table 3: Lexical Coverage

vocabulary size	coverage
5000	85.8%
8129	90.0%
20047	95.7%
27634	97.0%

2.3. Language Model

N-gram language models are constructed based on the lexicon. Specifically, word 2-gram and 3-gram models are trained using back-off smoothing. They are available in the CMU-Cambridge SLM toolkit format[5].

Notational symbols are also included in the statistical language models. As a result, the occurrence of short pauses between words is estimated by the probabilities of symbols that correspond to pauses.

The training corpus (Mainichi newspaper '91/01-'94/09) after pre-processing has 2.4M sentences and 65M words (=morphs). The cut-off thresholds for the baseline N-gram entries are 1 for 2-gram and 2 for 3-gram.

Specification of the resultant model for the 5K lexicon is shown in Table 4. For the decoder that performs forward-backward search, the backward 3-gram model is trained. More compact models are also prepared for memory efficiency by setting higher cut-off thresholds (4 and 8). Perplexity of the models is computed using test-set sentences in a different period ('94/10-'94/12).

Table 4: Specification of 5K N-gram

model	cut-off	#entries	perplexity
2-gram	1	578,653	107
3-gram	2	1,978,931	70

2.4. Decoder

The recognition engine named JULIUS[6] is developed to interface the acoustic model and the language model. It can deal with various types of the models, thus can be used for their evaluation.

JULIUS performs two-pass (forward-backward) search using word 2-gram and 3-gram on the respective passes.

In the first pass, a tree-structured lexicon dynamically assigned with 2-gram probabilities is applied with the frame-synchronous beam search algorithm. The 2-gram probabilities are factored into tree nodes according to the best word history.

Here, we assume one-best approximation rather than word-pair approximation. The degradation by the rough approximation in the first pass is recovered by the tree-trellis search in the second pass. Here, the word-trellis index form is adopted to realize efficient stack decoding. It turned out to achieve the same accuracy with much less computation and storage, compared with the word-graph search using word-pair approximation. Especially, necessary memory size for the search space is drastically reduced so that the program can be loaded at standard PCs.

In the second pass, inter-word context-dependency is also handled for accurate recognition. The second pass based on the stack decoder outputs correct N-best sentence candidates.

Overview of the decoder is summarized in Table 5.

Table 5: Overview of Decoder JULIUS

	acoustic model	language model	search approx.
1st pass	intra-word CD	2-gram	1-best
2nd pass	inter-word CD	3-gram	N-best

3. ASSESSMENT OF MODULES BY JAPANESE DICTATION TASK

As integration of the modules specified in the previous section, a Japanese dictation system is designed and implemented.

The acoustic model and language model are integrated based on the decoder specification. In the first pass, word 2-gram is applied and only intra-word phonetic context dependency is handled. Word 3-gram and inter-word context dependent model, which are more precise and computationally expensive, are incorporated in the second pass to re-score and search on reduced candidates.

As the first step, a baseline 5000-word dictation system is developed. The components independently developed at different sites are successfully integrated.

The integrated system can be used to assess the component modules, in turn. By changing the modules, we can evaluate their effects with respect to the recognition accuracy and efficiency.

For the evaluation, we have used portion of ASJ-JNAS speech database that are not used for training of the acoustic model. We picked up 10 speakers [3] and 10 utterances per speaker. [4] The uttered sentences are text-open to language model training.

Word accuracy is used as the evaluation criterion. It is computed with results of the first pass (2-gram) and with results of the second pass (3-gram), respectively. The first pass (2-gram) involves several search errors due to the one-best approximation, but the final results (3-gram) are reliable. The experiments for assessment are done by real-time factor of 5 to 10. The accuracy was almost saturated, but could be increased a bit by a larger beam width. [5]

3.1. Acoustic Model Assessment

At first, we present assessment of variety of acoustic models. Here, the baseline language model (cut-off 1-2) and the final tuned decoder are used.

The word accuracy is listed in Table 6 for male and Table 7 for female speakers, respectively.

It is observed that the monophone model needs many mixture components to achieve high accuracy, while increase of model complexity of the triphone does not improve so much. It suggests that much more data is needed to train the triphone model to the full extent. There is not much performance difference between male and female results.

[3] Speaker IDs are 006,014,017,021,026,089,102,109,115,122 for both male and female.

[4] Sentence IDs are No.01-10 (NORMAL MID LPP-HPP). Texts for each speaker are totally different.

[5] The results in this paper are as of April 1998.

Table 6: Evaluation of Acoustic Models (male)

model	mix.4	mix.8	mix.16
monophone	78.1 (67.2)	86.1 (74.7)	87.4 (80.4)
triphone 1000	88.2 (77.7)	91.4 (80.4)	91.7 (82.3)
2000	90.0 (78.0)	91.9 (80.1)	92.8 (82.6)
3000	90.2 (76.9)	92.4 (80.4)	92.7 (80.3)

word accuracy (%): with 3-gram (with 2-gram)

Table 7: Evaluation of Acoustic Models (female)

model	mix.4	mix.8	mix.16
monophone	79.1 (66.4)	84.7 (75.7)	88.6 (80.0)
triphone 1000	91.0 (79.1)	90.6 (82.9)	90.0 (82.3)
2000	91.3 (78.7)	91.8 (81.1)	93.2 (82.9)
3000	91.1 (79.3)	92.9 (80.6)	92.6 (81.7)

word accuracy (%): with 3-gram (with 2-gram)

3.2. Language Model Assessment

Next, we present assessment of language models. As the acoustic model, the male triphone 2000x16 is used.

The result is shown in Table 8. Slight degradation of the accuracy is observed with the coarse model of higher cut-off thresholds. It is memory efficient though it does not affect recognition time.

Table 8: Evaluation of Language Models

model	
baseline cutoff 1-2	92.8 (82.6)
cutoff 4-8	91.2 (82.8)

word accuracy (%): with 3-gram (with 2-gram)

3.3. Decoder Assessment

The decoding algorithm and techniques are evaluated by using the acoustic model of male triphone 2000x16 and the baseline language model (cut-off 1-2).

In Table 9, effect of several adopted techniques in our decoder is figured out. The accuracy with 3-gram (2nd pass) is much better than that with 2-gram (1st pass). Tuning the parameters of LM weight and insertion penalty has a little effect. Incorporation of inter-word context dependent model brings about large improvement. In our system, the lexical entries are defined by morphs which are smaller than ordinary 'words'. Thus, handling inter-word articulation is significant.

Table 9: Breakdown of Decoder Improvement

incorporated techniques	
2-gram only	(80.2)
3-gram	86.0 (80.2)
LM weight tuned for 2-pass	86.9 (80.2)
insertion penalty used	87.5 (82.6)
inter-word CD handled (=final)	92.8 (82.6)

word accuracy (%): with 3-gram (with 2-gram)

4. CONCLUSION AND ONGOING WORK

The baseline platform we are developing is standard and portable. As the formats and interfaces of the modules are general, any modules can be easily replaced. Thus, the toolkit is suitable for research on individual component techniques as well as development of specific systems. Moreover, it is possible to replace or integrate modules that are developed at different sites and evaluate them.

It is proven that our platform demonstrates reasonable performance when adequately integrated. The current version of the software (decoder) works under standard Unix platform. It needs about 64MB memory including space for the language model.

Ongoing work of the project is to improve the modules so that (1) they can be applied to 20K vocabulary task, and (2) they can be ported to standard PC platform.

Acknowledgement: The authors are grateful to advisory members of the project for their comments and cooperation. We are also debt to Mr. Akinobu Lee, who has been engaged in development and assessment.

References

[1] S.J.Young. A review of large-vocabulary continuous-speech recognition. *IEEE Signal Processing magazine*, 13(5):45–57, 1996.

[2] T.Matsuoka, K.Ohtsuki, T.Mori, S.Furui, and K.Shirai. Japanese large-vocabulary continuous-speech recognition using a business-newspaper corpus. In *Proc. ICSLP*, pages 22–25, 1996.

[3] K.Itou, M.Yamamoto, K.Takeda, T.Takezawa, T.Matsuoka, T.Kobayashi, K.Shikano, and S.Itahashi. The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus. In *Proc. ICSLP*, 1998.

[4] S.Young, J.Jansen, and J.Odell D.Ollason P.Woodland. *The HTK BOOK*, 1995.

[5] *The CMU-Cambridge Statistical Language Modeling Toolkit v2*, 1997.

[6] A.Lee, T.Kawahara, and S.Doshita. An efficient two-pass search algorithm using word trellis index. In *Proc. ICSLP*, 1998.