JAPANESE DICTATION TOOLKIT
– PLUG-AND-PLAY FRAMEWORK FOR SPEECH RECOGNITION R&D – *

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.)

http://winnie.kuis.kyoto-u.ac.jp/dictation/
E-mail: dictation-tk-request@astem.or.jp

ABSTRACT

A sharable software repository for Japanese LVCSR (Large Vocabulary Continuous Speech Recognition) is introduced. It is designed as a baseline platform for research and developed by researchers of different academic institutes under the governmental support. The repository consists of a recognition engine, variety of acoustic models and language models as well as Japanese morphological analysis tools. These modules can be easily integrated and replaced under a plug-and-play framework, which makes it possible to fairly evaluate components and to develop specific application systems. In this paper, specifications of the current version is described and assessment in 20000-word dictation task, which was also set up in our project, is reported. The software repository is freely available to the public.

1. INTRODUCTION

A Large Vocabulary Continuous Speech Recognition (LVCSR) system is a complex of high-accuracy acoustic models, large-scale language models and an efficient recognition program[1]. On the other hand, most of researchers are interested in specific components and try to demonstrate the effectiveness of new methods by integrating with other components. This background motivated us to develop a free sharable platform that can be used as a baseline and reference. Such a platform will also suffice for a baseline in developing application systems and for an open entry to those of other research fields or foreign countries.

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

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

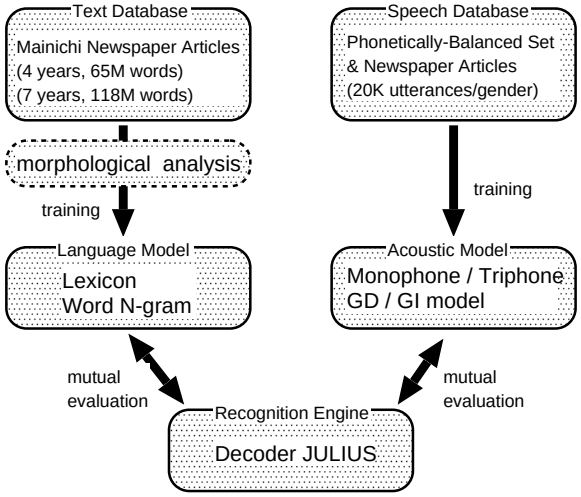

Figure 1: Platform of LVCSR

It is rather easy to have agreement of a common interface and format in the LVCSR system. It realizes a plug-and-play framework for research and development. Namely, researchers can put and test a new component and system developers can replace and tune components for specific applications.

We adopted Mainichi Newspaper, one of the nationwide general newspapers in Japan, for the sharable corpus of both text and speech[2], and organized a project to develop a standard software repository that includes a recognition program together with acoustic and language models under the support of the IPA (Information-technology Promotion Agency), Japan[3]. An overview of the corpus and software is depicted in Figure 1.

Specifications of the acoustic models, language models and recognition engine as well as Japanese morphological analysis tools are described in this paper. We also report evaluation of these modules under 20000-word Japanese dictation task.

2. SPECIFICATION OF MODELS AND PROGRAMS

2.1. Acoustic Model

Acoustic models are based on continuous density HMM. We basically adopt the HTK format as it is an ASCII file.

We have trained several kinds of Japanese acoustic models from a context independent phone model to triphone models, as listed in Table 1. We set up both gender dependent and gender independent models. Users can choose an adequate model according to the purpose.

Table 1: List of Acoustic Models

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

GD: Gender Dependent, GI: Gender Independent

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 are 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. Cepstral mean normalization (CMN) is performed on every utterance. 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 a variety of models whose number of the states is 1000, 2000 and 3000, respectively.

2.2. Morphological Analysis and Lexicon

A lexicon is a set of lexical entries specified with their notations and baseforms. It is also in the HTK format. The lexicon is consistent with both the acoustic model and language model.

The vocabulary consists of the most frequent words (=morphs) in Mainichi newspaper articles from January 1991 to September 1994 (45 months)[2].

In Japanese, definition of vocabulary depends on morphological analysis system that segments undelimited texts. We adopt a morphological analyzer ChaSen that has been developed at Nara Institute of Science and Technology. Major modification has been made for speech recognition purpose.

In Japanese, there are many morph entries that have different part-of-speech tags and also a lot of Kanji (Chinese character) entries that have multiple pronunciations. Generally, words of different part-of-speech tags have different tendency of possible adjacent words, even if they are same in notation. Pronunciation of some words is also dependent on adjacent words. In order to improve language modeling, we distinguish lexical entries by not only their notations but also their part-of-speech tags and phonetic transcriptions. If word pronunciation is not uniquely identified by the morphological analysis, one entry is allotted with multiple baseforms.

The lexical coverage of various vocabulary sizes is listed in Table 2.

Table 2: Lexical Coverage

vocabulary size	coverage
5000	88.2%
6135	90.0%
20000	96.5%
22959	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. Witten-Bell discounting method is used to compute back-off coefficients. We adopt the CMU-Cambridge SLM toolkit format as it is also an ASCII file.

We have compared two training sets, which are of Mainichi newspaper articles and different in their sizes. One is of 45-month articles ('91/01-'94/09; 65M words) and the other is of 75-month articles ('91/01-'94/09, '95/01-'97/06; 118M words).

The cut-off threshold for the baseline N-gram entries is 1 for both 2-gram and 3-gram [cutoff-1-1]. Then, elimination of N-gram entries is explored for memory efficiency. Conventionally, it has been done by setting a higher cut-off threshold. Here, we prepare a model with the cut-off threshold of 4 [cutoff-4-4]. In addition, we have introduced a new method based on the

model entropy, not word occurrences[4]. The method incrementally picks out 3-gram entries so that ML estimation of the reduced model gives the smallest increase of entropy. As a result, 3-gram entries are reduced to 1/10 [compress-10%].

Since we have applied the above methods to the two training sets, the list of language models gets as in Table 3.

Table 3: List of 20K Language Models

	2-gram entries	3-gram entries
45-month cutoff-1-1	1,238,929	4,733,916
45-month cutoff-4-4	657,759	1,593,020
45-month compress-10%	1,238,929	473,176
75-month cutoff-1-1	1,675,803	7,445,209
75-month cutoff-4-4	901,475	2,629,605
75-month compress-10%	1,675,803	744,438

2.4. Decoder

A recognition engine named Julius has been developed to interface the acoustic and language models. It can deal with various types of the models, thus can be used for their evaluation.

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

In the first pass, a tree-structured lexicon assigned with language model probabilities is applied with the frame-synchronous beam search algorithm. As the baseline, 2-gram probabilities are dynamically factored into all tree nodes according to the best word history. An efficient version assigns pre-computed 1-gram factoring values to the intermediate nodes, and applies 2-gram probabilities at the word-end nodes [1-gram factoring]. 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. The word-trellis index form is adopted to efficiently look up predicted word candidates and their scores[5].

In the second pass, we compute ten candidates by the stack decoder and sort them for the final output. An efficient version terminates the search by the first candidate [1-best candidate].

Cross-word context dependency is basically handled only in the second pass. However, we have also enhanced cross-word handling of the first pass with approximation which applies best model for the best history for an accurate version [cross-word enhanced].

An overview of the decoder is given in Table 4.

Table 4: 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

CD: Context Dependent model

3. EVALUATION OF MODULES

By integrating the modules specified in the previous section, a Japanese dictation system is realized. The integrated system can be used to evaluate the component modules, in turn. By changing the modules under the plug-and-play framework, we can evaluate their effects with respect to the recognition accuracy and efficiency.

As the IPA-98-TestSet, [1] we have used a portion of the ASJ-JNAS speech database that were not used for training of the acoustic model. It consists of 100 samples by 23 speakers for each gender. The sample sentences are open to the language model training. Word accuracy is computed using our tool that processes compound words.

3.1. Evaluation of Acoustic Models

At first, we present evaluation of a variety of acoustic models. Here, the baseline language model [75-month cutoff-1-1] and the baseline decoder are used. The word accuracy is listed in Table 5 for male and Table 6 for female speakers, respectively. It is observed that the monophone model needs many mixture components to achieve high accuracy, while the performance of the triphone model is almost saturated at the complexity of 2000 states. Gender independent models increases the error rates by around 2%.

3.2. Evaluation of Language Models

Next, we present evaluation of language models. As the acoustic model, the male triphone 2000x16 is used. The memory size and the word accuracy are shown in Table 7 for each language model. The models trained with 75-month articles consistently achieve better accuracy than those with 45-month data. As for memory-efficient models, the entropy-based compression method [compress-10%] is more effective than the simple cut-off method [cutoff-4-4].

[1] www.milab.is.tsukuba.ac.jp/jnas/test-set/male/male1_LARGE.txt
www.milab.is.tsukuba.ac.jp/jnas/test-set/female/female1_LARGE.txt

Table 5: Evaluation of Acoustic Model (male; accuracy)

	mix.4	mix.8	mix.16
monophone	75.3	79.6	83.9
triphone 1000	86.9	90.0	90.5
triphone 2000	90.5	90.9	92.0
triphone 3000	89.2	90.5	90.5
GI monophone	68.3	78.0	81.7
GI triphone 2000	87.4	88.7	90.0

Table 6: Evaluation of Acoustic Model (female; accuracy)

	mix.4	mix.8	mix.16
monophone	75.5	80.7	88.9
triphone 1000	90.3	91.6	92.6
triphone 2000	91.0	92.2	93.2
triphone 3000	90.5	90.4	91.3
GI monophone	76.0	80.8	84.7
GI triphone 2000	89.9	91.8	90.5

3.3. Evaluation of Decoder

The decoding algorithms are evaluated by using the acoustic model of male triphone 2000x16 and the baseline language model [75-month cutoff-1-1]. The word accuracy for tested methods is listed in Table 8. First, it is confirmed that enhancement of the cross-word handling of the first pass reduces the error rate by 25% at the expense of computation increase. In the lower portion of the Table, effect of several techniques for time efficiency is investigated. Use of 1-gram factoing instead of 2-gram in the tree-structured lexicon leads to great degradation of the accuracy in the first pass (2-gram). But it is almost recovered by the tree-trellis search in the second pass (3-gram). When the search is terminated by the first (1-best) candidate, the accuracy is slightly lowered. Each technique leads to increase of recognition speed by around 10%.

Table 7: Evaluation of Language Model

	accuracy	LM size
45-month cutoff-1-1	89.8	54MB
45-month cutoff-4-4	89.3	23MB
45-month compress-10%	89.3	28MB
75-month cutoff-1-1	92.0	79MB
75-month cutoff-4-4	90.9	34MB
75-month compress-10%	91.8	38MB

Table 8: Evaluation of Decoding Algorithms

	word accuracy 3-gram (2-gram)
baseline	92.0 (78.9)
+ cross-word enhanced	94.0 (85.1)
+ 1st-pass 1-gram factoring	91.2 (73.9)
+ 2nd-pass 1-best candidate	90.8 (78.9)

4. CONCLUSION

Key property of the software toolkit is generality and portability. As the formats and interfaces of the modules are widely acceptable, any modules can be easily replaced. Thus, the toolkit is suitable for research on individual component techniques as well as development of specific systems. Actually, the experiments in this paper are done by integrating and replacing modules that are developed at different sites. The results prove that the plug-and-play framework actually works and our platform demonstrates reasonable performance when adequately integrated.

The software repository is freely available to public. Though it is not complete yet especially in documentation, the repository is being used as a baseline in research community and industrial sides including countries other than Japan. The project is still ongoing to further improve the modules both in accuracy and efficiency and to enhance the portability.

Acknowledgement: The authors are grateful to advisory members of the project for their comments and cooperation.

References

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

[2] K.Itou et al. The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus. In *Proc. ICSLP*, pages 3261–3264, 1998.

[3] T.Kawahara et al. Sharable software repository for Japanese large vocabulary continuous speech recognition. In *Proc. ICSLP*, pages 3257–3260, 1998.

[4] N.Yodo, K.Shikano, and S.Nakamura. Compression algorithm of trigram language models based on maximum likelihood estimation. In *Proc. ICSLP*, pages 1683–1686, 1998.

[5] A.Lee, T.Kawahara, and S.Doshita. An efficient two-pass search algorithm using word trellis index. In *Proc. ICSLP*, pages 1831–1834, 1998.