J. Acoust. Soc. Jpn. (*E*) **20**, 3 (1999)

Japanese Dictation Toolkit — 1997 version —

Tatsuya Kawahara,[*1] Akinobu Lee,[*1] Tetsunori Kobayashi,[*2]
Kazuya Takeda,[*3] Nobuaki Minematsu,[*4] Katsunobu Itou,[*5]
Akinori Ito,[*6] Mikio Yamamoto,[*7] Atsushi Yamada,[*8]
Takehito Utsuro[*9] and Kiyohiro Shikano[*9]

[*1] *Kyoto University,*
Yoshida-Hon-machi, Sakyo-ku, Kyoto, 606-8501 Japan
[*2] *Waseda University,*
3-4-1, Ohkubo, Shinjuku-ku, Tokyo, 169-8555 Japan
[*3] *Nagoya University,*
Furo-cho 1, Chikusa-ku, Nagoya, 464-8603 Japan
[*4] *Toyohashi University of Technology,*
1-1, Hibarigaoka, Tenpaku-cho, Toyohashi, 441-8580 Japan
[*5] *Electrotechnical Laboratory,*
1-1-4, Umezono, Tsukuba, 305-8568 Japan
[*6] *Yamagata University,*
4-3-16, Jyonan, Yonezawa, 992-8510 Japan
[*7] *University of Tsukuba,*
1-1-1, Tennodai, Tsukuba, 305-8573 Japan
[*8] *ASTEM RI,*
17, Chudoji-Minami, Shimogyo-ku, Kyoto, 600-8813 Japan
[*9] *Nara Institute of Science and Technology,*
8916-5, Takayama-cho, Ikoma, 630-0101 Japan

(*Received 5 January 1999*)

【English translation of the same article in J. Acoust. Soc. Jpn. (J) **55**, 175-180 (1999)】

The Japanese Dictation Toolkit has been designed and developed as a baseline platform for Japanese LVCSR (Large Vocabulary Continuous Speech Recognition). The platform consists of a standard recognition engine, Japanese phone models and Japanese statistical language models. We set up a variety of Japanese phone HMMs from a context-independent monophone to a triphone model of thousands of states. They are trained with ASJ (The Acoustical Society of Japan) databases. A lexicon and word N-gram (2-gram and 3-gram) models are constructed with a corpus of Mainichi newspaper. The recognition engine JULIUS is developed for evaluation of both acoustic and language models. As an integrated system of these modules, we have implemented a baseline 5,000-word dictation system and evaluated various components. The software repository is available to the public.[†1]

Keywords: Large vocabulary continuous speech recognition, Software

PACS number: 43. 72. Ne

[†1] For further information:
http: //www.lang.astem.or.jp/dictation-tk/
mailto: dictation-tk-request@astem.or.jp

1. INTRODUCTION

Large Vocabulary Continuous Speech Recogni-

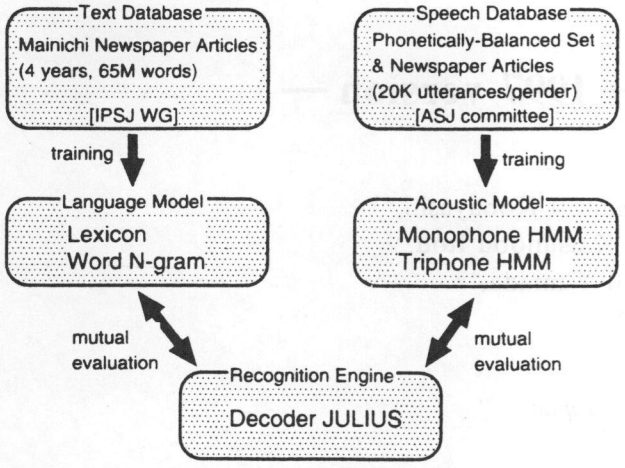

Fig. 1 Platform of LVCSR.

tion (LVCSR) is a basis of speech technology applications in the next generation including a voice-input word processor and dictation of broadcast programs or personal audio tapes. Its component technologies can also be used in various applications such as spoken dialogue interfaces.

In order to build an LVCSR system, high-accuracy acoustic models, large-scale language models and an efficient recognition program (decoder) are essential.[1-4] 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.

We have adopted Mainichi Newspaper, one of the nation-wide general newspapers in Japan, for the sharable corpus of both text and speech,[5] and organized a project to develop a standard software repository that includes acoustic and language models and recognition programs.[6] 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 available to the public. The overview of the corpus and software mentioned here is depicted in Fig. 1.

The specifications of acoustic models, language models and recognition engine are described in this paper. We also report evaluation of each module under 5,000-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.[7]

We have trained several kinds of Japanese acoustic models from a 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. Users can choose an adequate model according to the purpose. A simpler model realizes faster recognition at the expense of accuracy degradation.

The set of 43 Japanese phones are listed in Table 2. The phone notation is defined by the 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.

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

The speech data were sampled at 16 kHz and 16 bit. Twelfth-order mel-frequency cepstral coefficients (MFCC) are computed every 10 ms. 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

Table 1 List of acoustic models.

Model	#States	#Mixtures
Monophone	129	4, 8, 16
Triphone 1000	1,000	4, 8, 16
Triphone 2000	2,000	4, 8, 16
Triphone 3000	3,000	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 channel mis-match.

Each phone model consists of three states excluding the initial and final states that have no distributions. The state transitions are all left-to-right, and the path from the initial state and that to the final state are limited to one.

When a triphone model is applied to CSR, it has to cover all possible combinations of the phones. Thus, an extra file is used to define the mapping from the possible tuples (logical triphone) to prepared models (physical triphone). In actual, there are not sufficient data for all logical triphones. So 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 1,000, 2,000 and 3,000, respectively.

2.2 Lexicon

A lexicon is a set of lexical entries specified with their notations and baseforms. It is provided in the HTK format.[7]

The lexicon is consistent with both the acoustic model and the language model. The phone symbols used in the baseforms are covered with the acoustic model. For each lexical entry, its probability is given by the language model (at least 1-gram).

The vocabulary consists of the most frequent words (=morphs) in Mainichi newspaper articles from Jan. 1991 to Sep. 1994 (45 months).[5] In Japanese, lexical entries are mainly defined by a morphological analyzer that segments undelimited texts. In the 1997 version, we adopt the morphological analysis given by the RWCP text database. Generally, the morphs of different parts of speech have different tendency of possible adjacent words, even if they are same in notation. In order to improve language modeling, we distinguish lexical entries by not only their notations but also their morphological attributes (part-of-speech tags). The lexical coverage of various vocabulary sizes is listed in Table 3.

Table 3 Lexical coverage.

Vocabulary size	Coverage
5,000	85.8%
8,129	90.0%
20,047	95.7%
27,634	97.0%

Not a few lexical entries have multiple baseform entries because Japanese Kanji usually has multiple pronunciations. The lexicon also includes entries of comma, period and question marks that are rewritten as a pause in pronunciation.

In the 1997 version, a lexicon of 5K vocabulary size is available. A 20K lexicon will be released in the 1998 version.

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. The models are available in the CMU-Cambridge SLM toolkit format.[8]

The comma, period and question marks are also included in the statistical language models. As a result, the occurrence of short pauses between words is estimated by the probabilities of these 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. More compact models are also prepared for memory efficiency by setting higher cut-off thresholds (4 and 8).

Specification of the resultant model for the 5K lexicon is shown in Table 4. The perplexity of the models is computed using test-set sentences in a

Table 4 Specification of 5K N-gram.

Model		Cutoff	#Entries	Perplex.
Cutoff 1-2 (baseline)	2-gram	1	577,765	90.6
	3-gram	2	1,974,061	57.3
Cutoff 4-8	2-gram	4	327,787	93.9
	3-gram	8	711,598	64.2

different period ('94/10-'94/12). Each entry occupies 18 bytes for 2-gram and 6 bytes for 3-gram in our decoder. For the decoder that performs forward-backward search, the backward 3-gram model is trained.

2.4 Decoder

The recognition engine named JULIUS[9] has been 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. The word-trellis index is a set of survived word-end nodes in the first pass, their scores and their corresponding starting frames, thus enables the second pass to efficiently look up predicted word candidates and their scores. The search algorithm achieves 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.

The parameters of language model weight and insertion penalty as well as the beam width can be adjusted for the respective passes.

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

CD : Context-Dependent model.

3. JAPANESE DICTATION SYSTEM

By integrating the modules specified in the previous section, a Japanese dictation system has been designed and implemented.

The block digram of the system is illustrated in Fig. 2. 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 (CD) 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 the reduced

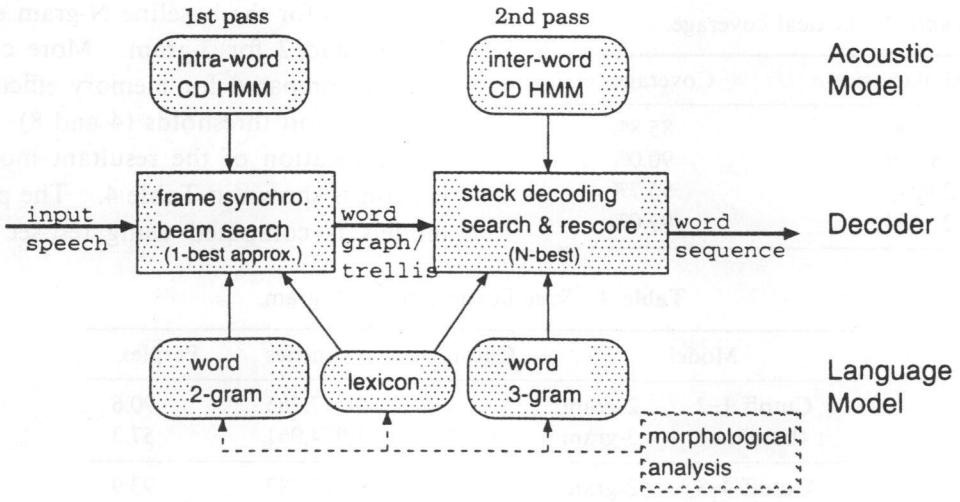

Fig. 2 Block diagram of Japanese dictation system.

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).

candidates.

Since there are several variations in both acoustic model and language model, we can design different systems accordingly. Typically, use of monophone model instead of context dependent model makes an efficiency-oriented system. Setting of decoding parameters such as the beam width may also yield variations of the system.

As the first step, a baseline 5,000-word dictation system was developed. The components independently developed at different sites were successfully integrated.

4. EVALUATION OF MODULES AND SYSTEMS

The integrated system can be used to evaluate 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 a portion of the ASJ-JNAS speech database that were not used for training of the acoustic model. We picked up 10 speakers[†2] and 10 utterances per speaker.[†3] The uttered sentences are text-open to the language model training. There are no unknown words in the test-set samples.

Word accuracy is used as the evaluation criterion. Definition of the word accuracy in Japanese involves several issues. Here, notational symbols are ignored, and recognized words are matched by Katakana transcription basis. The accuracy 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 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.[†4]

4.1 Evaluation of Acoustic Models

At first, we present evaluation of a 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.

4.2 Evaluation of Language Models

Next, we present evaluation of language models. As the acoustic model, the male triphone 2000×16 is used.

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

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

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

The test-set perplexity and the word accuracy are shown in Table 8. Slight degradation of the accuracy is observed with the model of higher cut-off thresholds. The coarse model is memory-efficient, but does not affect recognition time.

4.3 Evaluation of Decoder

The decoding algorithm and techniques are evaluated by using the acoustic model of male triphone 2000×16 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 (=language model) 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'.

Table 8 Evaluation of language models.

Model	Test-set perplex.	Word accuracy
Baseline cutoff 1-2	80.5	92.8 (82.6)
cutoff 4-8	90.6	91.2 (82.8)

Word accuracy (%): with 3-gram (with 2-gram).

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).

Table 10 Specification of typical systems.

Acoustic model	Monophone 129×16	Triphone 3000×8	Triphone 2000×16
Decoding	candidates reduced	small beam	large beam
CPU time	3×RT	6×RT	12×RT
Accuracy (male)	85.2	91.3	92.8
Accuracy (female)	87.3	91.2	93.2

CPU: Ultra SPARC 300 MHz
RT (Real Time): 4.1 s/utterance

Thus, handling inter-word articulation is significant.

4.4 System Assessment

Finally, performance of the total system is summarized in Table 10, where three typical system configurations are listed.

The efficiency-oriented system with the monophone model performs recognition within 3 times of the real time at a standard workstation. The triphone 3000×8 model realizes the word error rate of 9% with the speed of 6 times of the real time using a small beam width in the decoding process. The accuracy-oriented system with the precise acoustic model (2000×16) and sufficient decoding resources achieved the word error rate of 7%.

5. CONCLUSION AND ONGOING WORK

The 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. 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 64 MB memory including space for the language model.

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

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 Process. Mag. **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," Proc. ICSLP 96, 22-25 (1996).
3) P. S. Gopalakrishnan, L. R. Bahl, and R. L. Mercer,

"A tree search strategy for large-vocabulary continuous speech recognition," Proc. IEEE-ICASSP 95, 572-575 (1995).

4) L. R. Bahl, S. V. de Gennaro, P. S. Gopalakrishman, and R. L. Mercer, "A fast approximate acoustic match for large vocabulary speech recognition," IEEE Trans. Speech Audio Process. **1**(1), 59-67 (1993).
5) 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," Proc. ICSLP 98, 3261-3264 (1998).
6) T. Kawahara, T. Kobayashi, K. Takeda, N. Minematsu, K. Itou, M. Yamamoto, T. Utsuro, and K. Shikano, "Sharable software repository for Japanese large vocabulary continuous speech recognition," Proc. ICSLP 98, 3257-3260 (1998).
7) S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland, *The HTK BOOK* (1995).
8) *The CMU-Cambridge Statistical Language Modeling Toolkit v2 manual* (1997).
9) A. Lee, T. Kawahara, and S. Doshita, "An efficient two-pass search algorithm using word trellis index," Proc. ICSLP 98, 1831-1834 (1998).