IPA JAPANESE DICTATION FREE SOFTWARE PROJECT

Katsunobu Itou (ETL)***, Kiyohiro Shikano** (NAIST)****, Tatsuya Kawahara** (Kyoto Univ.)*****, kazuya Takeda** (Nagoya Univ.)******, Atsushi Yamada** (ASTEM)*******, Akinori Itou** (Yamagata Univ.)**, Takehito Utsuro** (NAIST)**, Tetsunori Kobayashi** (Waseda Univ.)**, Nobuaki Minematsu** (Toyohashi Univ.)**, Mikio Yamamoto** (Tsukuba Univ.)**, Shigeki Sagayama** (JAIST)**, Akinobu Lee** (Kyoto Univ.)

*1-1-4 Umezono, Tsukuba, Ibaraki, 305-8568, Japan, **1-6-10 Takayama, Ikoma, Nara, 630-0101, Japan,
*kito@etl.go.jp, **shikano@is.aist-nara.ac.jp, ***kawahara@kuis.kyoto-u.ac.jp, ****takeda@nuee.nagoya-u.ac.jp,
*****yamada@astem.or.jp

Abstract

Large vocabulary continuous speech recognition (LVCSR) is an important basis for the application development of speech recognition technology. We had constructed Japanese common LVCSR speech database and have been developing sharable Japanese LVCSR programs/models by the volunteer-based efforts. We have been engaged in the following two volunteer-based activities.

a) IPSJ (Information Processing Society of Japan) LVCSR speech database working group.
b) IPA (Information Technology Promotion Agency) Japanese dictation free software project.

IPA Japanese dictation free software project (April 1997 to March 2000) is aiming at building Japanese LVCSR free software/models based on the IPSJ LVCSR speech database (JNAS) and Mainichi newspaper article text corpus. The software repository as the product of the IPA project is available to the public. More than 500 CD-ROMs have been distributed. The performance evaluation was carried out for the simple version, the fast version, and the accurate version in February 2000. The evaluation uses 200 sentence utterances from 46 speakers. The gender-independent HMM models and 20k/60k language models are used for evaluation. The accurate version with the 2000 HMM states and 16 Gaussian mixtures shows 95.9 % word correct rate. The fast version with the phonetic tied mixture HMM and the 1/10 reduced language model shows 92.2 % word correct rate and realtime speed.

The CD-ROM with the IPA Japanese dictation free software and its developing workbench will be distributed by the registration to http://www.lang.astem.or.jp/dictation-tk/ or by sending e-mail to dictation-tk-request@astem.or.jp.

1. INTRODUCTION

Large vocabulary continuous speech recognition (LVCSR) is an important basis for the application developments of speech technology in the coming decade, which include a voice-activated word processors, a voice dialog system such as a car navigation device, a voice-activated game, voice input for PDA, and CAI.

Dictation systems, which are almost the same meaning as LVCSR, have been actively researched and developed for English in USA. English dictation systems have been also used in various kinds of application tasks such as in various medical fields and lawyer's dictation fields.

Moreover, dictation technology is promising and inevitable for word/sentence input for near future PDA including cellular phones. IBM has been selling famous dictation software, ViaVoice for several different languages. Dragon Systems, NaturallySpeaking, has been used in the same kinds of application fields. Microsoft has been developing dictation software, mainly in order to design and specify better speech SAPI (speech application program interface). These USA dictation Research efforts have been supported with the US government, DARPA. Especially, database construction and performance evaluation by the DARPA project has pushed forward with English dictation (LVCSR) technology.

Thus, dictation technology is not only useful for the dictation purpose, but also useful for the future PDA word/sentence input. The recognition accuracy improvement and computation amount reduction must be crucial for the future PDA and mobile applications.

1.1. Japanese Dictation Systems

Japanese dictation systems have been independently developed, without any strategic governmental supports. Japan IBM is selling Japanese dictation software, ViaVoice, and NEC begins to sell the same kind of dictation software, SmartVoice.

In Japan, word-processor including Kana-Kanji conversion is so popular that Japanese dictation have a great potential even for PC applications. One of the direct future dictation applications is broadcast news automatic caption generation. These efforts are also conducted with NHK, NTT, IBM, Toyohhashi University, and Electrical Communication University. Speech data indexing is studied in Ryukoku University and Tokyo Institute of Technology.

To construct Japanese dictation system, there are

several Japanese language dependent problems as follows,

(1) Definition of words is ambiguous, because Japanese texts are written without spacing between words. We need an accurate sharable morphological analysis program.

(2) Chinese character (Kanji) has usually several ways of reading/pronunciation. Reading is highly dependent on the context. We further need a sharable reading annotation program.

(3) Accurate speaker-independent phoneme models are not supplied to the public, because we have only small size of public available and distribution-free LVCSR speech database.

(4) We have to develop a dictation algorithm including Kana-Kanji conversion.

We had to construct common LVCSR speech database and develop sharable Japanese LVCSR programs/models by the volunteer-based or independent efforts.

We have been involved in the following two volunteer-based activities.

a) IPSJ (Information Processing Society of Japan) LVCSR speech working group.
b) IPA (Information Technology Promotion Agency) Japanese dictation free software project.

The web sites and related information for these volunteer-based activities are shown in Table 1. These activities and free dictation software/workbench will promote speech recognition application development.

1.2. Coming Application of Dictation Systems

Speech recognition technology has been introduced in various kinds of application fields these several years. Speech recognition technology can be roughly classified into the following three areas from the viewpoints of the application fields.

(1) Speech recognition chips or middle-ware

Speech recognition LSIs and speech recognition oriented DSPs are widely developed, which have several hundred word recognition capability. The most popular speech recognition chips are used with cellular phones. Other applications are found in the application of portable games and car navigation devices. In the coming several years, much more speech recognition chips will be installed in PDA and home electronics. These chips and middle-ware are getting very popular especially in Japan.

(2) Telephony service oriented speech recognition system

Information inquiry through telephone network (call center) is getting more popular, but operators have to stand by for inquiry. To reduce the cost of operators and extend the service hour, speech recognition and speech synthesis systems are getting to be installed in the various areas of telephone information inquiry. These telephony service oriented speech recognition systems have been introduced especially in USA. The telephone speech quality adaptation to the changes of cellular phone systems must be dealt with by speech data collection and acoustic model re-training at least.

(3) Dictation software for PC

According to the PC performance progress and dictation technology improvement, PC-oriented dictation software is getting popular and is widely sold. These kinds of dictation software are applicable to command input and sentence input. Further applications in the dictation software will be broadcast news summary, speech data indexing, and general-purpose dictation.

Another software implementation of speech recognition is seen for home game devices. Speech recognition LSI chips and middle-ware are most used among the above three speech recognition application areas. Especially in Japan, merging of cellular phone and PDA will be realized rapidly according to widely spread various services with cellular phones. The LSI chips and DSPs are only dealt with several hundreds of words so far, but dictation capability will be installed even in these chips and DSPs in the coming several years. Another possible application field of speech recognition is public services such as ticket vending or ATM banking.

2. IPSJ LVCSR SPEECH DATABASE WG

ASJ [1] continuous speech database (ASJ-PB) which contains about 10,000 phonetically balanced sentence utterances has been widely used as a public continuous speech database in Japan. 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 word segmentation outputs by the morphological analysis. 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 free software 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 sentence selection for LVSCR 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[2]-Text-Corpus.

[1] ASJ: The Acoustical Society of Japan
[2] RWCP: Real World Computing Partnership

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 also 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 training texts with their RWCP morphological tags. Next a word bigram language model was generated using the CMU SLM Toolkit (CMU-Cambridge,1997), where vocabulary sizes are 5k and 20k words.

The sentence selection was carried out considering the following factors, (1) vocabulary size, (2) length of sentences, (3) morphological unit (word) perplexity, and (4) number of out-of-vocabulary words. These sentence selection criteria are shown in Table 4. According to these criteria, design of sentence set was made as summarized in Table 5. This statistically controlled set consists of 90 sentences.

In addition to the designed set, a few paragraphs and 50 ATR phonetically balanced sentences are included in each design set. Each set was visually checked by four WG members from viewpoints of correctness in morphological analysis and reading annotation, and political correctness (privacy, political bias, etc.). Finally 150 sets were obtained.

The speech database committee of the ASJ asked to record several sentence sets for committee members. The speech data were recorded in collaboration with 39 sites. The utterances were recorded with two micro phones, a close-talk microphone and a desktop microphone. The speech data was compressed into 16 CD-ROMs and titled JNAS (Japanese Newspaper Article Sentences) (Itou,et al.,1998(Cocosda), Itou, et al.,1998(ICSLP)). The JNAS CD-ROMs have been released to the public.

3. IPA DICTATION FREE SOFTWARE PROJECT

Free dictation software will be useful for LVCSR application development as described in Section 1.2. Our project proposal to IPA is illustrated in Figure 1.

To build an LVCSR system, a high-accuracy acoustic model, a large-scale language model and an efficient recognition program are at least necessary. The original third target of the IPSJ LVCSR WG activities supported financially by IPA in April, 1997. We reorganized a project to develop a standard software repository that included acoustic and language models, recognition programs, and a morphological analysis program. The IPA project uses the JNAS LVCSR speech database and Mainichi newspaper article text corpus to develop LVCSR programs and tools, as shown in Figure 2. The software repository as the product of the IPA project should be available to the public. Members of the IPA project are limited in university and governmental researchers due to free software oriented activities. However, company researchers are supporting us as advisory members. Members of the IPA project are listed in Table 6. The original annual targets of the IPA project are shown in Figure 3.

3.1. Acoustic Models

The acoustic models are based on Gaussian mixture HMM. They are available in the HTK(Young, et al.,1996) format. We have trained several kinds of Japanese acoustic models, which include context-independent phoneme (monophone) models and triphone models. In the triphone modelling, 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 variety of models, whose numbers of the HMM states are about 1000, 2000, and 3000. Numbers of Gaussian mixtures a HMM state are 4, 8, 16, and 32 according to their accuracy. Every HMM phoneme model consists of three states. These HMMs also include gender-dependent and gender-independent HMMs.

The set of 43 Japanese phonemes defined by ASJ speech database committee is adopted. These HMM models are trained with ASJ speech database, ASJ-PB and ASJ-JNAS, as shown in Figure 2. The speech data are 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.

In 1999, the final year of the IPA project, phonetic tied mixture (PTM) HMM models for efficient and accurate decoding are successfully trained and implemented(Lee,2000). The PTM models are synthesized from context-independent phoneme (monophone) models with 64 mixture components per state by assigning different mixture weights according to the shared states of triphones. Gaussian parameters and component weights are then re-estimated by HTK for the further optimization. The PTM training process is explained in Figure 4.

Basically, we have three types of HMMs as follows,

(a) Monophone models for simple decoding,
(b) PTM models for fast and accurate decoding,
(c) Triphone models for accurate decoding.

The performance for these models is summarized in Table 7.

3.2. Language Models

The lexicon is also provided in the HTK format. The Mainichi newspaper articles from 1991 to Sep. 1997 (81 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 and 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(CMU-Cambridge,1997) 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 (Yodo, et al.,1998).

In 1999, 60k language models are developed using the Mainichi newspaper articles from 1991 to 1997 (81 months). The lexical coverage for 60k is 99.2%. There is no significant recognition rate degradation for decoding. The recognition time increase for the 60k language models is also small. The trigram parameter reduction program is also successfully applied.

3.3. Decoder

The recognition engine named *JULIUS* (Kawahara,1998) 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 as shown in Figure 5.

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 the one-best approximation in the first pass is successfully recovered by the tree-trellis search in the second pass.

In 1999, the first pass improvement is carried out by dealing with the phoneme context between words for PTM and triphone HMMs. The recognition accuracy improvements are attained as shown in Table 7. The recognition time increases in the factor of two.

The JULIUS decoder for PTM is developed to attain the fast and accurate decoding by introducing the Gaussian pruning techniques. Several Gaussian pruning techniques attain to reduce the computation ammount to about 20%.

3.4. Japanese Dictation Configuration

The configuration overview of the decoder, the acoustic model and the language model is illustrated in Figure 5. In the first pass of JULIUS, the word bigram is applied and the HMM triphone models are applied to only the intra-word phonetic context in JULIUS 98 version. In JULIUS 99 version, the rough inter-word phonetic context is dealt with based on the maximum likelihood for succeeding phoneme contexts. The one-best word candidates are stored in the form of word trellis. In the second pass, the word trigram and the precise inter-word phonetic context by the HMM triphone model, which are precise but computationally expensive, are used efficiently on the word trellis.

The components, such as the HMM phoneme models, the language models, and the decoder, were developed independently at the different sites, and successfully integrated for 5k, 20k and 60k vocabulary dictation systems.

Since there are several choices of the HMM phoneme models and the language models, we decide to develop three kinds of dictation systems, the simple version, the fast version and the accurate version. The simple 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 fast version uses PTM HMM triphone model, and the Gaussian pruning techniques are adopted. The accurate version, of course, uses the Gaussian mixture HMM triphone model.

3.5. Dictation System Evaluation

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

First, the gender-dependent HMM models were used for evaluation. As for the 20k language models, the cutoff 1-1 original language model (78.5MB) is adopted for the accurate version. The 1/10 compressed language model (30MB) is adopted for the fast version and the simple version. The simple version with the 16 Gaussian mixture monophone model and the 1/10 reduced 20k language model works realtime at a standard workstation and PC. The simple version word recognition is 85.3%. The fast version with PTM and the Gaussian pruning decoder attains 93.1% word recognition rate, and 2.8 times realtime recognition time. The processing time is measured with the UltraSPARC 300MHz. The high speed PC such as Pentium III can attain almost realtime processing speed. The accurate version with the 2000 HMM states and 16 Gaussian mixtures shows 8.4 times realtime and 93.7% word recognition rate for JULIUS 98 version, and 12.9 times realtime and 95.8% word recognition rate for JULIUS 99 version. These performance results are also included 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 recognition 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 mixture HMM phoneme models are evaluated. The word recognition rate for the accurate JULIUS 99 version is 94.7%, which is

comparable with the word recognition rate of 95.8% for the gender-dependent HMM models. The evaluation results are shown in Table 8. More evaluation results for the fast version and the simple version are also included in Table 8.

Lastly, the 60k language models are used for the accurate version and the fast version. The accurate version adopts the cutoff 1-1 original 60k language model (99.7MB), and the fast version adopts the 1/10 compressed 60k language model (54.5MB). The word recognition rates are 94.8% and 91.8% respectively. The processing time is almost same as the 20k language models. The evaluation results for 60k lexicon are also summarized in Table 9.

3.6. LVCSR Workbench

We have been developing various kinds of programs and tools as well as the decoding programs and models. These programs and tools are released as a LVCSR workbench with their detailed manuals and supports through networks. The workbench is shown in Figure 6.

Table 7: JULIUS evaluation (as of Jan. 2000)
(20k language model, UltraSPARC 300MHz,
200 sentences from 46 speakers)

HMM type		Number of Gaussians	Word Correct	Proc. Time
Tri-phone	Accurate Julius99	32,000 (2000 × 16)	95.8%	12.8 ×RT
	Accurate Julius98		93.7%	8.4 ×RT
PTM(Phonetic tied mixture)	Fast Julius99	8,256 (129×64)	93.1%	2.8 ×RT
	Fast Julius99		92.1%	2.3 ×RT
Mono-phone	Simple Julius98	2,064 (129×16)	85.3%	1.1 ×RT

Table 8: JULIUS evaluation for gender-independent HMM

JULIUS version		Gender-independent	Gender-dependent
Accurate version	Julius 98	91.7%	93.7%
	Julius 99	94.7%	95.8%
Fast version	Julius 99	91.1%	92.1%
Simple version	Julius 98	84.0%	85.3%

Table 9: JULIUS evaluation for 60k language model

	Word correct(%)		Proc. time	
Language model	60k	20k	60k	20k
Accurate version	94.8	95.8	16.9xRT	12.8xRT
Fast version	91.8	92.1	2.9xRT	2.3xRT

CONCLUSION and FUTURE PLAN

The baseline Japanese dictation platform we have been developing is proven to work reasonably well. The baseline platform now works in Unix workstation and Linux PC. We are distributing CD-ROMs, which include the dictation programs, the models, the workbench and the manuals. The CD-ROM with the IPA Japanese dictation free software and its developing workbench will be distributed by registration to http://www.lang.astem.or.jp/dictation-tk/ or by sending e-mail to dictation-tk-request@astem.or.jp. We are also planning to have a summer school, where our developed dictation programs and workbench are used.

We are planning to start a consortium to develop and maintain our programs, models, and workbench. The future plan of the consortium is to improve each module and its speech application program interface as follows: (1) Network grammar based continuous speech recognition program, (2) the morphological analysis program (*ChaSen*) and the pronunciation annotation program (*ChaWan*) to deal with conversational corpus, (3) HMM phoneme models to match the real world, such as office environments and telephone speech, (4) Windows version to deal with SAPI5.0, (5) adaptation programs to a task, a speaker, and environments, and (6) better user interface with hands-free and barge-in capabilities.

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 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).

5. REFERENCES

S.J.Young, (1996). A review of large-vocabulary continuous-speech recognition, IEEE Signal Processing magazine, 13(5), pp.45-57

H.J.M.Steeneken, V.Leeuwen, (1995). Multi-lingual assessment of speaker independent large vocabulary speech recognition system: the SQALE-project, Proc. EUROSPEECH

K.Itou, K.Takeda, T.Takezawa, T.Matsuoka, K.Shikano, T.Kobayashi, S.Itahashi, M.Yamamoto, (1998). Design and development of Japanese speech corpus for large vocabulary continuous speech recognition assessment, Proc.Oriental COCOSDA, pp.98-103, 1998

K.Itou, K.Takeda, T.Takezawa, T.Matsuoka, K.Shikano, T.Kobayashi, S.Itahashi, M.Yamamoto, (1998). The Design of the Newspaper-Based Japanese Large Vocabulary Continuous Speech Recognition Corpus, Proc. ICSLP, pp.3261-3264

T.Kawahara, A.Lee, T.Kobayashi, K.Takeda, N.Minematsu, K.Itou, A.Ito, M.Yamamoto, A.Yamada, T.Utsuro, K.Shikano, (1998). Common Platform of Japanese Large Vocabulary Continuous Speech Recognizer Assessment –Proposal and Initial Results --, Proc. Oriental COCOSDA, pp.117-122

T.Kawahara, T.Kobayashi, K.Takeda, N.Minematsu, K.Itou, M.Yamamoto, A.Yamada, T.Utsuro, K.Shikano, (1998). Sharable Software Respository for Japanese Large Vocabulary Continuous Speech Recognition, Proc. ICSLP, pp.3257-3260

N.Yodo, K.Shikano, S.Nakamura, (1998). Compression Algorithm of Trigram Language Models based on Maximum Likelihood Estimation, Proc. ICSLP, pp.1683-1686

S.Young, J.Jansen, J.Odell, D.Ollason, P.Woodland, (1995). The HTK BOOK

CMU-Cambridge,(1997). The CMU-Cambridge Statistical Language Modeling Toolkit V2

A.Lee, T.Kawahara, K.Takeda, K.Shikano, (2000), A New Phonetic Tied-Mixture Model for Efficient Decoding, Proceedings of ICASSP, 2000