Japanese Large-Vocabulary Continuous Speech Recognition System Based on Microsoft Whisper

Hsiao-Wuen Hon, Yun-Cheng Ju and Keiko Otani

Microsoft Research
One Microsoft Way
Redmond, Washington 98052, USA

ABSTRACT

Input of Asian ideographic characters has traditionally been one of the biggest impediments for information processing in Asia. Speech is arguably the most effective and efficient input method for Asian non-spelling characters. This paper presents a Japanese large-vocabulary continuous speech recognition system based on Microsoft Whisper technology. We focus on the aspects of the system that are language specific and demonstrate the adaptability of the Whisper system to new languages. In this paper, we demonstrate that our pronunciation/part-of-speech distinguished morpheme based language models and Whisper based Japanese senonic acoustic models are able to yield state-of-the-art Japanese LVCSR recognition performance. The speaker-independent character and Kana error rates on the JNAS database are 10% and 5% respectively.

1. INTRODUCTION

Large vocabulary dictation is one of the ultimate applications of speech recognition. Recently, many commercial products are beginning to emerge for English [2][3][4]. Input of Asian ideographic characters has traditionally been one of the biggest impediments for information processing in Asia. Speech is arguably the most effective and efficient input method for Asian non-spelling characters. This paper will explore the large-vocabulary continuous Japanese dictation system based on Microsoft Whisper technology [1][10]. We focus here on the aspects of the system that are language specific and demonstrate the adaptability of Whisper system to new languages.

Japanese is a mora-based language, where Kana's are the basic sound units and Kanji's usually carry semantic meaning. The Japanese writing system is a combination of Kana's and Kanji's without any space in between. In contrast to Western languages, the task of morphological analysis, which segments each sentence into small lexical units (morphemes) with part-of-speech tags and pronunciations (kana sequence), is an essential preprocessing step for both acoustic model and language model training. In this paper, we leverage the IME (input method editor) of Microsoft Japanese Window system to perform morphological analysis for any Japanese sentence.

Our experiments were carried out on JNAS (Japanese Newspaper Articles Sentences) database provided by ASJ (Acoustic Society of Japan). JNAS contains 306 speakers with 150 continuous sentences per speaker. We used a 298-speaker subset to build gender dependent acoustic models and the remaining 8-speaker subset for the test set. For language model, we use LDC Nikkei-Kyodo text corpora to train both trigram and bigram language models. We use the pronunciation/part-of-speech distinguished morphemes as the lexicon entries for training the language models. In a 50K-vocabulary system, the test set perplexity for the bigram and trigram language model is about 253 and 170 respectively.

Whisper uses a combination of sub-phonetic mixture Gaussian HMMs and decision-tree based senones for Japanese acoustic models. We use 29 phones to model Japanese acoustic. To train the state-cluster decision trees, we use 26 English phoneme-derived questions about the contexts. For our experiment, a 3000-senone decision tree is used and each senone is then modeled by a mixtures of 8 Gaussian distributions.

Because of the segmentation issues in Japanese writing system, word accuracy in a dictionary-based dictation is not so meaningful. Therefore, we use character accuracy (each Kana and Kanji is treated as a character unit) and Kana accuracy instead. Our speaker-independent recognition result shows 10% character and 5% Kana error rates for the 50k-dictation system. After speaker adaptation, the error rates are reduced by relative 26%/10% for male and female speakers respectively.

This paper is organized as follows. In Section 2 we describe the JNAS database used for this paper. In Section 3 we discuss the characteristics of Japanese and the morphological analysis for both acoustic and language modeling. In section 4 and 5 we discuss the acoustic and language modeling respectively. In section 6, we discuss the recognition results. Finally in section 7 we summarize our major findings and outline our future work.

2. SPEECH & TEXT DATABASE

Japanese Newspaper Article Sentences (JNAS) database [5] is a Japanese speech database designed for Japanese large-vocabulary continuous speech recognition (LVCSR) research, much like the Wall Street Journal (WSJ) database for English LVCSR research. It is designed and created by the Speech Database Committee of Acoustical Society of Japan (ASJ). It has been recorded in collaboration with 39 research institutions in Japan. A similar database was also constructed by other research institutions [6]

JNAS contains speech recordings and their orthographic transcriptions of 306 speakers (153 males and females) reading excerpts from the Mainichi Newspaper (100 sentences) and the ATR 503 PB- Sentences (50 phonetically balanced sentences). The corpus contains about 45,000 sentences and the total duration of the database is about 60 hours. Since our goal is to build a Japanese dictation system, we chose to use the close-talking version (recorded with Sennheiser HMD410/HMD25-1 microphone) of the database for our experiments. The speech waves were sampled at 16 kHz and quantized into 16 bits. We use a 298-speaker subset to build gender-dependent acoustic models while the remaining 8-speaker subset is used for testing. Only the 100 Mainichi Newspaper sentences of the testing speakers' database is used for testing because the 50 phonetically balanced sentences are not real sentences.

For text corpora, we use the LDC's Japanese Business News Text Corpora [8]. This corpus contains two parts. The first part is Nihon Kezai Shimbun 1993-1994. Nihon Keizai Shimbun is the highest ranked business daily in Japan and is viewed as the Japanese Wall Street Journal equivalent. It contains roughly 94 million characters. The second part is Dow Jones Telerate/Kyodo News Service 1994-1995. Kyodo News Service provide news service to staff members in large brokerage houses, banks, manufacturers who have Japanese management or other professionals who follow business news from Japan. It contains about 36 million characters.

3. CHARACTERISTICS OF JAPANESE

Japanese is a mora-based language, where each mora is represented by a kana to form the basic Japanese sound unit. There are about 80 different Kana's corespondent to 80 morae in Japanese. Kana's usually do not carry any semantic meaning except some of them function as particles that provide clues for Bunsetsu[1] mark. The Japanese writing system is a combination of Kana's and Kanji's without any space in between. The pronunciation of a Kanji usually depends on its meaning and often contains a few Kana's. Since each character (either Kana or Kanji) is usually neither a syntactic nor semantic unit, using character as the recognition unit for acoustic models or language models will be highly ineffective.

Similar to other Asian languages, like Chinese [7], one must segment Japanese sentences into word sequences [6] before constructing pronunciation sequences lexicons and language models. In Japanese, those linguistic lexical units are generally referred to as morphemes because they are the smallest meaningful parts in Japanese. However, each written morpheme in isolations usually has more than one possible meaning and pronunciation. It is common that some morphemes in isolation could have more than 5 different pronunciations or meanings. Particularly important for speech recognition purpose, it is critical to disambiguate multiple pronunciations during the process of segmentation. Such a process of segmenting into morpheme sequences and disambiguating multiple pronunciations is referred to an morphological analysis [9].

3.1 Morphological Analysis

Morphological analysis is typically a very sophisticated process which involves knowledge about bunsetsu, part-of-speech and morpheme lexicon. It is usually done with a combination of statistical modeling (N-gram approach) and natural language processing techniques.

One of the centerpieces of Microsoft's Japanese Window system is the IME (input method editor) module which can convert any Kana string into the corresponding written form (a combination of Kana's and Kanji's). The IME is a critical piece to input Japanese with keyboard via phonetic spelling. Speech dictation can then be viewed as a speech IME that will translate spoken Japanese into corresponding written form. Recently IME also provides morphological analysis (reverse conversion) which can convert any written form back to the Kana sequence (thereby pronunciation) with segmentation marks. The accuracy of the Microsoft IME morphological analysis on general Japanese sentences is about 95%. We use IME to segment the text corpus for language model construction and create corresponding pronunciations (Kana sequences) for acoustic model training. For JNAS database, since the correct Kana sequences are provided, we are able to check our Kana sequences against them and use the correct ones for training acoustic models.

4. ACOUSTIC MODELING

Microsoft's Whisper [1][10] engine offers general-purpose speaker-independent continuous speech recognition that can recognize unrestricted text and is effective for command and control, dictation and conversational systems. In order to achieve the goal, Whisper incorporates many state-of-the-art technologies, including normalized feature representations for improved robustness, senone models derived from inter- and intra-word context-dependent phonemes, semi-continuous or continuous density hidden Markov models implemented with the generic shared density function architecture, and efficient decoder algorithms.

We decide to model Japanese acoustics by using sub-phonetic models because senones have been shown to be very powerful for modeling multi-lingual acoustics. The following table illustrates the 29 base phones we used in our system.

Vowel	A I U E O
Doublevowel	AA II UU EE OO
Consonant	K S T N H M Y R W G Z D B P SH CH J
Nasal	NN
Glottal_stop	STOP

Table 1. Japanese base phone table

Readers will notice we did not model double consonants directly. It is because our earlier experiments of modeling double consonants as separate phones did not yield any improvement. Therefore, a double consonant is simply modeled by a glottal stop followed by a single consonant.

[1] A bunsetsu usually consists of more than one morpheme and represent a independent semantic unit. A bunsetsu is very similar to a phrase in Western languages.

Similar to Whisper's English system, we use CART [11] and a set of 26 phoneme categorical questions about the phonetic contexts to build senone decision trees. Most of these 26 phoneme categorical questions are directly derived from our corresponding English set used in Whisper. The most noteworthy additions are questions about double vowels and single/double vowel combination, and word-ending devoiced-vowel ("I" and "U").

The definition of morphemes in Japanese is usually very vague particularly because of compound noun combinations. The fact that there is no space between morphemes in Japanese writing also make it doubtful that the inter-morphemes co-articulation will behave the similar way as inter-word co-articulation in English. We thus decide not to follow the inter-word context modeling for English where the inter-word contexts are modeled separately from intra-word contexts. Based on the amount of our training data, a 3000-senone decision tree is used for our experiments and each senone is then modeled by either 5-codebook semi-continuous models or a mixtures of 8 continuous Gaussians distributions. Diagonal covariance is also assumed for all the Gaussians.

Since we have 50 ATR phonetically balanced sentences for each test speaker, we also conducted some speaker adaptation experiments using MLLR adaptation [12] and the results will be presented in the section 6.

5. LANGUAGE MODELING

As stated in section 2, the written morpheme form cannot uniquely identify its meaning or pronunciation. Building the lexicon or language model (LM) based on the written morphemes alone will need to accommodate many multiple syntactic/semantic and pronunciation representations. It will in general degrade the quality of both acoustic and language models.

In order to solve the issues above, we use pronunciation/part-of-speech distinguished morphemes as our basic lexical entries. For example, 一行 could be interpreted as "one line" (pronounced – ichigyou), "a group of people" (pronounced - ikkou) and "person's first name" (pronounced - kazuyuki). The lexicon will contain three occurrences for 一行 and each has its unique pronunciation and part-of-speech label.

Based on the pronunciation/part-of-speech labeled morphemic lexicon, we compile the lexical word frequency list for our text corpora. We then keep the top 50K words as our recognition lexicon. The OOV (out-of-vocabulary) rate for our training corpora is less than 1.4%. With this 50K lexicon, there are about 120 OOV words in our 8-speaker test set. Since we are merely interested in recognition performance in this paper, we add those 120 words into our testing lexicon to assure the test set is fully covered by the 50K lexicon. Table 2 shows the test set perplexity for the trigram and bigram respectively.

	Perplexity
Bigram	253
Trigram	170

Table 2 test-set perplexity

5.1 Conjugation

During preliminary error analysis, we observed that many errors are caused by the い-conjugation co-articulation. い-conjugation co-articulation occurs when the adjective/verb stem part ending with "I" is followed by the conjugation part beginning with "い". The two adjacent "I" vowels will be uttered as a double-vowel "II" rather than two independent "I" sounds because they are within one bunsetsu structure. For example, the acoustic characteristics of a double-vowel "II" (which sound like a long I vowel) in 良/い and 聞/いて is clearly distinct from two independent "I" vowels (like in "世界/一"). In order to remedy this deficiency in our acoustic modeling, we need to include those conjugation forms (e.g. 良い and 聞いて) in the lexicon. A post-processing after morphological analysis is thus performed to combine the adjective/verb stem part ending with "I" and the conjugation part beginning with "い" into a new morpheme lexical unit. This post processing can correctly model this cross-morpheme double vowel formation and therefore greatly reduces the recognition errors caused by い-conjugation co-articulation.

6. RECOGNITION RESULTS

Because of issues of the segmentation and no consensus of definite morpheme set for Japanese writing system, word accuracy in a lexical based dictation is not so meaningful. Instead, we use character accuracy (each Kana and Kanji is treated as a character unit) and Kana accuracy (Kanji is converted into its underlying Kana representation) to measure the recognition performance. In the nutshell, the kana accuracy is aiming at measuring the pure acoustic accuracy since it represents the mora (the basic acoustic unit) accuracy for Japanese. On the other hand, the character accuracy represents accuracy of writing units without taking into account the segmentation, which is not present in the writing form anyway.

Figure 1 shows the Kana error rate chart for various experiments: semi-continuous models w/ trigram, continuous models w/ bigram, continuous models w/ trigram and continuous w/ trigram and speaker adaptation.

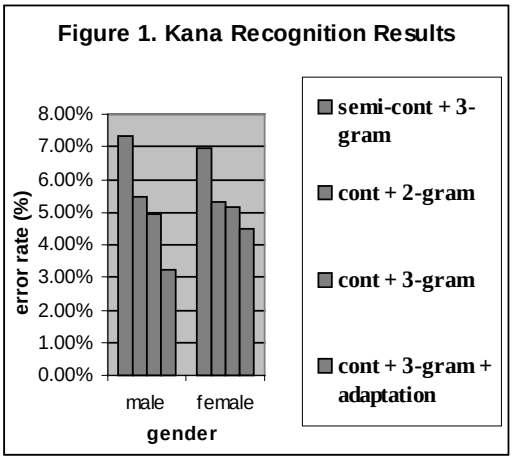

Figure 2 on the other hands shows the character error rate chart for corresponding experiments.

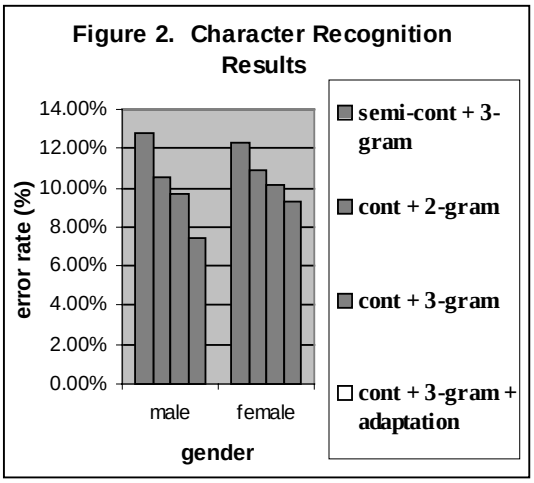

Table 3 shows the individual speaker performance for continuous models w/ trigram. As one can observe there is not much performance variation between speakers. However, comparing with adaptation results in Table 4, female speaker fb44 did not benefit from adaptation while most other speakers benefit significantly from adaptation. It will be interesting for further investigation.

	Kana error rate	**Kanji** error rate
mb39	4.55%	9.55%
mb40	4.50%	8.10%
mb40	5.90%	10.75%
mb40	4.75%	10.40%
fb43	5.90%	10.45%
fb44	5.55%	11.20%
fb45	5.75%	11.00%
fb46	3.45%	8.10%

Table 3 Recognition error rates for individual speaker

	Kana error rate	**Kanji** error rate
mb39	2.90%	7.35%
mb40	3.00%	5.85%
mb40	3.85%	8.10%
mb40	3.15%	8.35%
fb43	4.85%	9.40%
fb44	5.60%	11.25%
fb45	4.45%	8.80%
fb46	3.10%	7.65%

Table 4 Speaker-adaptation results

7. SUMMARY

We have described in this paper the importance of morphological analysis and the choice of appropriate acoustic/lexical units for Japanese large-vocabulary continuous speech recognition. We also demonstrated that our pronunciation/part-of-speech distinguished morpheme based language models and Whisper based Japanese senonic acoustic models are capable of yielding state-of-the-art Japanese LVCSR recognition performance.

Future work includes improving Japanese morphological analysis, pronunciation inspired morphemic lexicon units and investigation of inter-bunsetsu context dependency and intra-bunsetsu context dependency modeling

8. REFERENCES

[1] Huang X., Acero A., Alleva F., Hwang M.Y., Jiang L. and Mahajan M. "Microsoft Windows Highly Intelligent Speech Recognizer: Whisper". *IEEE International Conference on Acoustics, Speech, and Signal Processing. Detroit, May 1995.*

[2] L. Bahl, S Balakrishnan-Alyer., J. Bellgarda, M. Franz, P. Gopalakrishnan, D. Nahamoo, M. Novak, M. Padmanabhan, M. Picheny, and S. Roukos, "Performance of the IBM Large-Vocabulary Continuous Speech Recognition System on the ARPA Wall Street Journal Task", *vol 1, pp. 41-44, ICASSP-95*

[3] R. Roth, et al., "Dragon Systems' 1994 Large-Vocabulary Continuous Recognizer", *Proc. Spoken Language Systems Technology Workshop, Austin, January 1995*

[4] Microsoft SAPI 4.0 Speech SDK Suite - *http://research.microsoft.com/srg/sapisdk.htm*

[5] JNAS (Japanese Newspaper Article Sentence) database - *http://www.milab.is.tsukuba.ac.jp/jnas*

[6] T. Matsuoka, K. Ohtsuki, T. Mori, K. Yoshida, S. Furui, and K. Shirai, "Japanese Large-Vocabulary Continuous Speech Recognition Using a Business Newspaper Corpus", *ICASSP97, Munich, April 1997.*

[7] H. Hon, B. Yuan, Y. Chow, S. Narayan, and K. Lee, "Towards Large Vocabulary Mandarin Chinese Speech Recognition", *IEEE International Conference on Acoustics, Speech, and Signal Processing. Adelaide, April 1994*

[8] LDC Japanese Business News Text Corpus - *http://www.ldc.upenn.edu/ldc/catalog/html/text_html/jlfn.html*

[9] K. Lunde, "Understanding Japanese Information Processing", *O'Reilly & Associates, Inc. publisher1993*

[10] Whisper:*http://www.research.microsoft.com/research/srg/*

[11] Hwang, M.Y. and Huang, X. and Alleva, F. "Predicting Unseen Triphone with Senones". *ICASSP-93, Minneapolis, MN, pages 311-314. April, 1993*

[12] C. Leggetter, and P. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", *Computer Speech and Language, Vol 9, pp 171-185, 1995*