JAPANESE LVCSR ON THE SPONTANEOUS SCHEDULING TASK WITH JANUS-3

T. Schultz, D. Koll, and A. Waibel

Interactive Systems Laboratories
University of Karlsruhe (Germany), Carnegie Mellon University (USA)
{tanja,koll,waibel}@ira.uka.de

ABSTRACT

This paper presents our findings during the development of the recognition engine for the Japanese part of the VERBMOBIL speech-to-speech translation project. We describe an efficient method to bootstrap a large vocabulary speech recognizer for spontaneously spoken Japanese speech from a German recognizer and show that the amount of effort in developing the system could be reduced by using this rapid cross language bootstrapping technique. The Japanese recognizer is integrated into the VERBMOBIL system and shows very promising results achieving 9.3% word error rate.

1. INTRODUCTION

The overall goal of the first phase of the VERBMOBIL project is to build a speech-to-speech translation system from both German and Japanese spontaneously spoken input speech to English, German and Japanese output in an appointment scenario [1]. The Japanese recognizer described in this paper is beeing designed to be part of this translation system. Unlike Japanese dictation systems [2] there is no need for our recognizer to map the output onto Japanese *kanji/kana* characters. Due to the many homophones in the Japanese language this mapping requires syntactic and semantic knowledge about the uttered sentence. In the VERBMOBIL system this knowledge can be postponed to the syntactic and semantic analysis modules [3].

One peculiarity of Japanese is, that sentences are written in strings of *kanji* and *kana* characters without white space between adjacent words. Thus the Japanese language lacks the natural segmentation of speech into words, which can conveniently be used as basic units for recognition purposes, as known from Indoeuropean languages like English. In order to determine such basic speech units for the recognition system the transcribed speech data has been segmented by a semi-automatic morphological analysis program described in [4].

The recognition engine was evaluated based on the database described in the following section using the JANUS-3 Speech Recognition Toolkit in two passes: First a fast bootstrap technique from a German recognizer was applied on a subset of 190 dialogs of the training database. In the second step the training pass on the full database with 730 dialogs lead to the final system.

	Dialogs	Utts	Units	Vocab
Japanese Verbmobil Database				
Total	800	11308	322K	2668
Training data				
Bootstrap	190	3122	92K	1879
Final system	730	10278	305K	2598
Test data				OOV
Acoustic models	10	189	4K	0.35
Language model	70	1030	17K	0.6

Table 1: The Japanese spontaneous scheduling task

2. THE JAPANESE SPONTANEOUS SCHEDULING TASK

The domain of the VERBMOBIL scenario is limited to appointment scheduling. Native speakers were given a time schedule and were asked to schedule a meeting in a human-to-human dialog session. The speaking style of the dialog partners is not restricted, therefore spontaneous phenomena like noise, stuttering, false starts and nongrammatical sentences occur. A push-to-talk button and a close speaking microphone were used to avoid cross-talk effects.

The Japanese part of the VERBMOBIL database consists of 800 dialogs spoken by 324 different native speakers. On average the dialogs cover 14 utterances each of a length of about 30 words. All dialogs are collected by ATR Interpreting Telecommunication Laboratories and the University of Electro-Communications in Tokyo (Japan). Further information about the corpus and collection procedures are given in [4].

Human-to-human dialogs tend to be at a higher level of spontaneity than in human-to-computer scenarios, used for example in the ATIS-Task. A comparison of cross-talk vs push-to-talk scenario for the Spanish scheduling task showed that cross-talk is harder to recognize because it is more noisy [5]. On the other hand push-to-talk leads to longer and more complex utterances (38 vs 10 words per utterance on average), making this task more difficult for speech translation.

3. SYSTEM BOOTSTRAPPING

It has been shown earlier in [6] that for small vocabulary continuous read speech the cross language bootstrapping technique from English to Japanese leads to reasonable results. We expand this approach to large vocabulary spontaneously spoken speech.

3.1. Phoneme Set

Running an alignment of a German phoneme recognizer on Japanese input speech gave us the impression, that the Japanese phonetic is very similar to German. Therefore we decided to bootstrap the Japanese phoneme set with German models developed for the German VERBMOBIL recognizer system. Only 4 of the 31 phonemes required for acoustic modeling have no German counterparts. These were bootstrapped as follows:/4/ from /u/, /4:/ from /u:/, /&0/ from /s/, and /dZ/ from /tS/. To cope with the effects arising in the spontaneously spoken data, like i.e. stuttering, false starts, or mumbling, special noise models [7] were included. All together 44 different phonemes are used to model Japanese speech: 31 speech models, 11 noise models, 1 silence and 1 glottal stop (ref. table 2).

After training a Japanese system we determined the similarity between the Japanese and German phoneme set by performing the following experiment: we trained a SCHMM for the combined phoneme set with both German and Japanese input and ran a clustering procedure. This leads to the result that the most similar phonemes in this order are the consonant /z/ and /b/, the affricate /ts/ and the semi vowel /j/. Japanese short vowels are similar to the long version of their German counterparts.

3.2. Pronunciation dictionary

The Japanese syllable based writing system (*kana*) provides an almost phonetic transcription of the Japanese spoken language. Given a kana-transcription of the spoken dialogs the pronunciation dictionary required for training the recognition system could be built automatically using a simple mapping algorithm. The output was post-edited by native experts, adding pronunciation variants mainly for coping with differences in the speaking styles of male and female speakers (e.g. "atashi" as a variant for "watashi" in female speech for the English word "I")

Phone Type	with German counterparts	without counterparts
Vowels	a e i o a: e: i: o:	4 &0 4:
Semi vowels	j	
Consonants	b d f g h k l m n n2 p s S t v z	
Affricates	ts tS	dZ
Misc	1 Silence model 1 Glottal stop 11 noise models	

Table 2: Japanese phoneme models [Worldbet]

and with dialect specific variants. For full coverage of the 190 segmented training data a dictionary with 1879 word units was built. The segmentation approach we applied results in relatively small vocabulary growth rates compared to the not segmented data. The size of the dictionary is even smaller than for the German scheduling task as can be seen from figure 1.

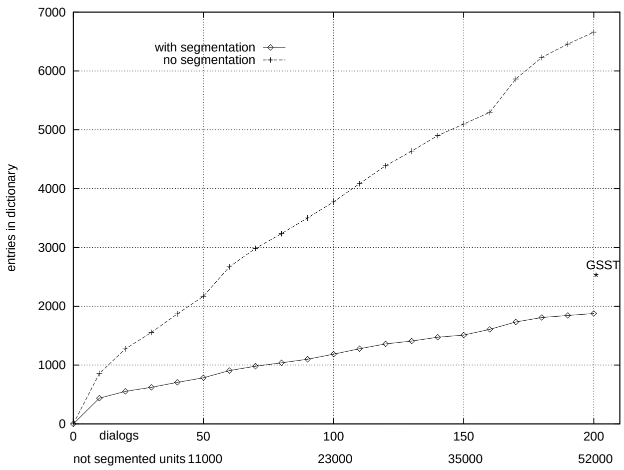

Figure 1: Vocabulary growth of segmented vs not segmented word units; label "GSST" gives the dictionary size for the German VERBMOBIL database

3.3. Acoustic Models Bootstrapping

To bootstrap the Japanese system we took a German context-independent 3-state HMM recognizer. Each state of the HMM is modeled by one codebook. Each codebook contains 16 mixture Gaussian distribution of a 32 dimensional feature space. 16 Mel-scale coefficients, power and their first and second derivatives are calculated from the 16kHz sampled input speech. Mean subtraction is applied. The amount of features is reduced to 32 coefficients by computing a Linear Discriminant Analysis (LDA).

We created a copy of the German recognizer containing the codebooks and distributions of the German counterparts for the Japanese models as described in table 2. Additional copies for codebooks and distributions from similar German phonemes are added for the 4 phonemes not found in the German phoneme set. As can be seen in table 3 this recognizer has a reasonable performance, though it has never seen any Japanese data. In the next step this bootstrapped version was trained 3 iterations. A new LDA had been calculated, and new codebooks were clustered. Another 3 training iteration made out the context-independent Japanese system, which leads to 20.9% word error rate.

Systems	Word Error
Japanese recognizer	
First bootstrapped version	65.0%
Context-independent system	20.9%
Context-dependent system	13.0%
Final system	9.3%
German recognizer	
Used for bootstrapping	38.1%
Currently best system	12.5%

Table 3: System performance

3.4. Further Improvements

After training and testing the context-independent system a context-dependent system was developed based on the JANUS-3 toolkit. Analyzing the scheduling database shows that only 54000 different sept-phones (a context of 3 phonemes to the right and to the left) could be found. The Japanese phonetic seems to be more restricted compared to German speech in the equivalent appointment scheduling task (200.000 quint-phones) and for English in the switchboard task (500.000 triphones).

We modeled 54000 sept-phones in the context-dependent Japanese system and clustered these sept-phones to 600 decision-tree-clustered polyphone models as described in [8]. The final context-dependent system has about 2000 distributions over the 600 polyphone models. The phonetic questions needed for the clustering procedure could be derived from the German phonetic questions.The resulting context-dependent Japanese speech recognition system achieves 13.0% word error rate.

4. FINAL SYSTEM

For the initial bootstrap only 190 out of the 800 dialogs were used for training the system. This lead to reasonable word accuracy on a system with comparatively few parameters in short time. The final system however made use of the full data-set for both language modeling and refining the acoustic models. Using all training data the dictionary size increased to 2598 words.

4.1. Acoustic modeling

Using the full data-set for the estimation of the HMM-parameters we changed the system structure to a fully continuous approach with an increased number of codebooks. The polyphonic tree of all occurring sept-phones (containing cross-word models with up to one phoneme lookahead to adjacent words) has been clustered to 2000 codebooks, each of which has been modeled as a mixture of 32 Gaussians with diagonal covariance. In order to increase recognition speed the dimensionality of the feature set was reduced to the first 24 LDA parameters of the feature set described in section 3.3. Label boosting (using supervised MLLR-adaption) was used to improve the quality of the database labeling.

4.2. Language Modeling

For the final system, trigram language models have been built on the 730 dialogs of the training set using a Kneser/Ney backoff scheme for unseen bi- respectively trigrams.

Our previous studies suggest that modeling noises like regular words improves the recognition performance moderately [7]. Breathing and key click noises are for example much more common at the beginning and the end than in the middle of an utterance. Therefore the noise events and hesitations (e.g. /eeto/ and /ano/) are modeled like regular words in computing their language model probabilities.

The morphological tagger used to segment the Japanese text transcriptions of the training dialogs lead to a reduction of the vocabulary size from roughly 10.000 to 2500. In doing so we get shorter, more frequent sequences. One drawback is however the reduction of the predictive power of n-gram models. Another is the necessity for checking the output of the segmentation tool which has to be done by native experts. In [9] a fully automatic statistical approach is described which find sequences of text that are both statistically important and semantically meaningful.

Whereas the data being sufficient for acoustic modeling, the corpus of approximately 300K words over a 2598 word vocabulary might still be a little small for accurate trigramm estimation. However the task limitation lead to a low perplexity of 12.8 and a slow growth of vocabulary as observed in figure 1, resulting in a OOV rate of only 0.6% on the 17K words test set.

4.3. Decoding

During recognition a two pass search is performed: in the first pass the search dictionary is organized as a phoneme tree, the second operates on a flat word list containing the most likely word-hypothesis of the first pass (relative error reduction: 6%). The resulting word-lattices are rescored, making full use of the trigram language model (relative error reduction: 10%). The recognition accuracy of the resulting final system using the full dictionary of 2598 words could thus be improved to 9.3% word error rate.

Table 3 compares the word error rate between the described Japanese systems and the German recognition engine of Karlsruhe University [10], which was the winner of the VERBMOBIL evaluation in 1996.

5. CONCLUSION

From the above experiments we conclude that the cross language bootstrapping of Japanese acoustic models from German models is a very efficient technique even for large vocabulary spontaneously spoken speech. Phonetic mismatches between the German and Japanese phoneme set are tolerable. A Japanese LVCSR system with good performance could be developed in short time. The resulting final system is now integrated into the VERBMOBIL speech-to-speech translation project. With a word error rate of 9.3% it achieves very promising results.

6. ACKNOWLEDGMENTS

The JANUS project is partly funded by grant 413-4001-01IV101S3 from the German Ministry of Science and Technology (BMBF) as a part of the VERBMOBIL project. We gratefully acknowledge support and cooperation with ATR Interpreting Telecommunication Laboratories and the University of Electro-Communications in Tokyo, Japan. The authors wish to thank all members of the Interactive Systems Laboratories, especially Michael Finke for useful discussion and active support.

7. REFERENCES

[1] T. Bub, W. Wahlster, and A. Waibel: *VERBMOBIL: The Combination of Deep and Shallow Processing for Spontaneous Speech Translation* in: Proc. of ICASSP, Munich 1997.

[2] T. Matsuoka, K. Ohtsuki, T. Mori, S. Furui, and K. Shirai: *Japanese Large-Vocabulary Continous-Speech Recognition Using a Business-Newspaper Corpus* in: Proc. of ICASSP, Munich 1997.

[3] M. Siegel: *Definiteness and Number in Japanese to German Machine Translation* in: Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference, Bielefeld (Germany) 1996.

[4] A. Kurematsu, M. Kitamura, T. Nakasuji, and A. Waibel: *Data Collection of Japanese Spontaneous Speech on Scheduling Task and Development of Speech Dictionary* in: Proc. of Eurospeech, Rhodes 1997.

[5] P. Zhan, K. Ries, M. Gavalda, D. Gates, A. Lavie, and A. Waibel: *JANUS-II: Towards Spontaneous Spanish Speech Recognition* in: Proc. of ICSLP, Philadelphia 1996.

[6] B. Wheatley, K. Kondo, W. Anderson, and Y. Muthusamy: *On Evaluation of Cross Language Adaptation for Rapid HMM Development in a New Language* in: Proc. of ICASSP, Adelaide 1994.

[7] T. Schultz and I. Rogina: *Acoustic and Language Modeling of Human and Nonhuman Noises for Human-to-Human Spontaneous Speech Recognition* in: Proc. of ICASSP, Detroit 1995.

[8] M. Finke and I. Rogina: *Wide Context Acoustic Modeling in Read vs Spontaneous Speech* in: Proc. of ICASSP, Munich 1997.

[9] L.J. Tomokyo-Mayfield and K. Ries: *What makes a word: Learning base units in Japanese for speech recognition* in: Proc. of the ACL Natural Language Learning workshop, Madrid July 1997.

[10] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and Martin Westphal: *The Karlsruhe-Verbmobil Speech Recognition Engine* in: Proc. of ICASSP, Munich 1997.