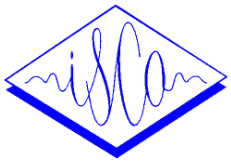

7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

MODELING FREQUENT ALLOPHONES IN JAPANESE SPEECH RECOGNITION

Long Nguyen, Xuefeng Guo, John Makhoul

BBN Technologies
10 Moulton St., Cambridge, MA 02138, USA
ln@bbn.com

ABSTRACT

In this paper, we describe a technique to model frequent allophones in Japanese speech recognition. The Consonant-Vowel syllabic structure (CV) is dominant in Japanese. Based on frequency, the distribution of CV pairs is rather skewed. Isolating out the most frequent allophones through the use of additional *phonemes* in acoustic modeling can achieve better recognition accuracy. By introducing ten new *phonemes* for the five most common CV pairs, we achieved a 30% relative reduction in word error rate for spontaneous speech and 6% relative reduction overall for all speech categories in a Japanese broadcast news transcription task.

1. INTRODUCTION

BBN Technologies has been an active participant in the DARPA Hub-4 Broadcast News task [1],[2]. We have broadcast news transcription systems in several languages, including English, Spanish, Arabic, and Chinese. Recently, we have been collaborating with NHK (or Japan Broadcasting Corporation) to use automatic speech recognition technology to provide live closed caption for NHK's TV news programs [3]. Fundamentally, BBN's broadcast news transcription systems use language-independent techniques and algorithms. However, for the Japanese language, we observe that there are certain language-specific characteristics that we can take advantage of to improve recognition accuracy further.

In the Japanese language, it is well-known that the CV syllabic structure is dominant. In addition, we observed that the distribution of all CV pairs, based on their occurrence frequency, is rather skewed. This phenomenon could pose a challenge for allophone clustering algorithms in most of the state-of-the-art speech recognition systems. Isolating out the allophones associated with the most frequent CV pairs through the use of additional *phonemes* in acoustic modeling can achieve better recognition accuracy. [In this context, *phoneme* is defined only as a phonetic unit used in acoustic modeling for speech recognition.]

The paper is organized as follows. In Section 2, we provide the background of the data corpus and the speech recognition system. In Section 3, we describe the analysis of our findings about the CV pairs in Japanese, and the formulation of the method that can take advantage of this observation. We then provide our experimental results in Section 4. We discuss and conclude in Sections 5 and 6.

2. BACKGROUND

This work is part of a long-term project to build a fast and accurate automatic speech recognition system to help in providing live closed caption for NHK's TV news programs. Research and development effort is carried out on the NHK Broadcast News Corpus, using a Japanese broadcast news transcription system developed at BBN Technologies.

2.1. NHK Broadcast News Corpus

The NHK Broadcast News Corpus [3] consists of 800 hours of speech selected from NHK news programs broadcast in the 1996-2000 period. Approximately, 600 hours were from male speakers, and 200 hours from female speakers. The Corpus also provides several date-dependent language models (DDLMs), in the form of trigram counts, calculated from 74M words of NHK's broadcast news manuscripts produced in the 1991-2000 period. Additionally, daily manuscripts are available for language model adaptation.

The development test set in the corpus is a collection of 3 hours of speech selected from NHK's news shows (*Good Morning Japan, News at Noon, and News at 7pm*) broadcast on June 1-7, 2000 (1-7jun00d). The entire test set is manually segmented into 1830 sentences and categorized into 7 narrow conditions: Studio News (C1), Studio Reports (C2), Spontaneous Speech (C3), Weather Forecast (C4), Sports News (C5), Noisy News (C6), and Field Reports (C7). Overall out-of-vocabulary rate is 0.5% when using a 62K-word lexicon. With DDLMs, trigram perplexities are remarkably low: less than 10 for read news (C1 and C6), 20-30 for reports (C2 and C7), 60-90 for weather forecast and sports reports (C4 and C5), and a little more than 100 for spontaneous speech (C3).

2.2. BBN Japanese Transcription System

The BBN Japanese transcription system [4] uses a two-pass N-Best recognizer: a fast-match pass to produce word endings and scores, and a second pass to generate N-best hypotheses. Recognition result is the top-1 hypothesis of the re-scored and re-ranked N-best hypotheses.

Acoustic models consist of phonetically tied-mixture (PTM) and state-clustered tied-mixture (SCTM) models [5]. The PTM model uses 41 mixture densities with 256 components each. The SCTM model uses about 4K mixture densities, each with 64 Gaussians.

Language models consist of a tree bigram language model (used in the fast-match pass), and a word trigram language model (for second pass decoding and N-Best rescoring). There are typically 62K words in recognition lexicons.

3. RATIONALE AND METHOD

Based on the fact that the CV syllabic structure is dominant in the Japanese language, we examined the frequency of CV pairs occurring in the 800-hour acoustic training data set of the NHK Broadcast News Corpus. As shown in Table 1, the top 5 most frequent CV pairs occur more than 150,000 times each, accounting for 22% of the total occurrences of all 219 possible CV pairs observed in this data set. Meanwhile, the median 5 pairs occur only within 1700-2000 times or 0.04% each; and the bottom 5 pairs occur 1 or 2 times each.

Rank	CV	Occurrence	Percentage
1	n-o	216426	4.97
2	sh-i	205191	4.71
3	k-a	204191	4.68
4	t-a	169938	3.90
5	n-i	156058	3.58
...			
110	b-aa	1954	0.04
111	k-uu	1880	0.04
112	j-a	1832	0.04
113	n-u	1803	0.04
114	z-ee	1714	0.04
...			
215	by-o	2	0.00
216	ts-oo	2	0.00
217	ry-aa	1	0.00
218	ky-aa	1	0.00
219	f-ee	1	0.00

Table 1: Occurrences of CV pairs in the NHK 800-hour Broadcast News Corpus ranked by frequency

For each of the top 5 CV pairs, the distribution of the pairing of a particular C and all possible V's is shown in Table 2, ordered by frequency. As can be seen from the table, more than 80% of the total occurrences for each consonant occur with only three or four vowels. In the case of *sh*-V pairing, the most frequent pair, *sh-i*, accounts for about 72% of the distribution, while the *sh-ee* pair occurs only 4 times!

In order to take advantage of the skewed nature of the CV distributions noted above, we decided to add the most common phonetic allophones to our set of phonemes in the system. (An allophone is a phoneme in a specific phonetic context.) For each CV pair, we actually add two phonemes, one to represent the right context and the other to represent the left context. For example, for the *n-o* pair, we add two phonemes: *nRo* and *nLo*. *nRo* is a specific allophone of *n* whose right context is constrained to be *o*, while *nLo* is an allophone of *o* whose left context is constrained to be *n*.

Another motivation for introducing additional phonemes was to take advantage of the large amount of training data available in the corpus. [We observed in the past that there was no significant recognition accuracy improvement when using 800 hours of speech data in comparison to 200 hours. We even doubled the number of Gaussians in our acoustic model at that time but obtained no gain.] We now hypothesized that by isolating the commonly occurring allophones, we would obtain better clustering for the less frequent allophones.

4. EXPERIMENTAL RESULTS

We carried out several experiments with a few different phoneme sets in which we added new phonemes derived from the most frequent CV pairs to the existing 41-phoneme list used in our baseline system. All phonetic dictionaries used in acoustic training and recognition were updated with these larger phoneme sets. Acoustic models were retrained from scratch because of new phonemes. Working with 800 hours of speech, especially training acoustic models from scratch, is typically not efficient for trying completely new ideas with many variations. We decided to experiment with only one gender that had lesser amount of training data: female, with 200 hours of speech. Language models used in these experiments are the trigram DDLMs briefly described in Section 2.1.

CV	Occur.	%	CV	Occur.	%
k-a	204191	30.26	n-o	216426	41.89
k-u	139978	20.74	n-i	156058	30.21
k-o	105798	15.68	n-a	105195	20.36
k-i	101590	15.05	n-e	23332	4.52
k-e	59673	8.84	n-oo	7922	1.53
k-oo	36663	5.43	n-aa	3080	0.60
k-ee	21798	3.23	n-u	1803	0.35
k-aa	2455	0.36	n-ii	1650	0.32
k-uu	1881	0.28	n-ee	1171	0.23
k-ii	786	0.12	n-uu	23	0.00
t-a	169938	33.31	sh-i	205191	71.65
t-o	151165	29.63	sh-oo	22076	7.71
t-e	135281	26.51	sh-a	20848	7.28
t-oo	36828	7.22	sh-o	13485	4.71
t-ee	11283	2.21	sh-uu	11234	3.92
t-aa	4043	0.79	sh-u	8856	3.09
t-ii	1340	0.26	sh-ii	4492	1.57
t-i	315	0.06	sh-aa	113	0.04
t-uu	12	0.00	sh-e	78	0.03
t-u	7	0.00	sh-ee	4	0.00

Table 2: Distributions of all observable right-context vowels for the 4 consonants *k, n, t,* and *sh* in the NHK Corpus

Test material is the portion of speech produced by female speakers in the 1-7jun00d test set, also described in Section 2.1. Overall, it consists of 49 minutes of speech for 573 sentences with 8460 words. The separation of numbers of sentences and words into focus conditions (C1-C7) are tabulated in Table 3.

4.1. Top-5 CV Pairs

In the first experiment, we took the 5 most frequent CV pairs in Table 1 and created the corresponding 10 new phonemes. As a result, the new phoneme list in our system increased from 41 to 51 phonemes.

As can be seen in Table 4, the overall word error rate (WER) for all conditions is 4.5%, a 6% relative reduction in WER over the 4.8% of the baseline system that used NHK's standard 41-phoneme set.

C1	C2	C3	C4	C5	C6	C7	All
71	27	15	259	175	12	14	573
2777	378	158	2685	1419	482	561	8460

Table 3: 49-minute test material separated into focus conditions. Top row: number of sentences; Bottom row: number of words.

Phs	C1	C2	C3	C4	C5	C6	C7	All
41	0.9	5.8	12.7	3.4	15.4	2.3	3.2	**4.8**
51	0.9	6.9	8.9	3.2	14.5	2.1	2.7	**4.5**
%rel.	0	+19	-30	-6	-6	-9	-16	-6

Table 4: WERs obtained by systems using either 51 or 41 phonemes; 10 new phonemes were derived from the 5 most frequent CV pairs. C1-C7 are focus conditions described in Sec. 2.

The most interesting result is that for spontaneous speech (C3) where we obtained a 30% relative improvement. We do not yet understand the reason for the increase in error rate for Studio Reports (C2). In terms of computation, using 51 phonemes increased the total (off-line) recognition time to 5.1 times real-time (xRT) versus 4.8xRT observed when using 41 phonemes. Timing was measured on machines equipped with Intel Pentium III 1GHz CPU, 2GB RAM, and RedHat Linux 7.2 OS.

We noticed that, by isolating out the allophones associated with the five most common CV pairs, our data-driven state clustering procedure produced slightly more mixtures, as shown in Row 2, Table 5, even though the minimum-occupancy threshold stayed the same. As a result, there were more Gaussians in the new acoustic model. However, the average number of Gaussians for each mixture was slightly less. Note that the number of components (i.e. Gaussians) of a mixture density in our system is determined first by the K-Means algorithm based on Euclidean distances of the samples, and then rearranged (i.e. reshaping, merging, and splitting) based on data likelihood in EM training [5]. Probably, where and how these new Gaussians were positioned in the acoustic space contributed to this gain.

Examining the allocation of mixture densities for the 7 phonemes associated with the 5 most frequent CV pairs used in this experiment revealed some interesting outcome, as shown in Table 6. Take phoneme *k* for an example. In the baseline acoustic model with 41 phonemes, *k* was allocated with 220 mixture densities to cover its acoustic subspace. When isolating out allophone *kRa*, by means of a new phoneme in the new system, the remaining allophones still required 187 densities to cover their *carved-out* subspace, while *kRa* was allocated with 54 densities. The more interesting candidate is phoneme *sh*. It used to have 130 densities in the baseline acoustic model. While in the new system, it was allocated with 171 densities in total, with 43 for allophone *shRi* and 128 for all others. Recall that out of 100 instances of *sh*, almost 72 of them are *sh-i*, as shown in Table 2. However, it requires

PhSet	#Mixtures	#Gaussians	#Gauss./Mixture
41phs	4128	259325	62.82
51phs	4345	270743	62.31
73phs	4620	287263	62.18

Table 5: Comparison of parameters in acoustic models trained with 41-phoneme set, 51-phoneme set, and 73-phoneme set

Ph	41-ph	51-ph		
		old	new	total
a	463	352	106	458
i	420	365	82	447
o	292	271	15	286
k	220	187	54	241
t	149	108	47	155
n	90	59	49	108
sh	130	128	43	171

Table 6: Numbers of mixture densities allocated for the 7 phonemes associated with the top-5 CV pairs in the 41-phoneme baseline system and the new 51-phoneme system

only 25% of the total densities (43/171) to cover *shRi*'s subspace. Meanwhile, other *sh* allophones, despite the *a priori* fact that they occur only 25% of the time, still require almost the same number of densities (128 vs. 130) to cover their scattering (but being carved-out) acoustic subspace.

We then tried to expand our experiments further by using more allophones as additional phonemes. However, so far, none of those experiments has been able to surpass the accuracy obtained with the top-5 CV pairs. We report only on two representative experiments here.

4.2. Top-16 CV Pairs

In the same manner as above, by lowering the occurrence threshold to 100,000 times, we selected 16 CV pairs. The phoneme list now consisted of 73 phonemes (41 + 32). We updated our training and recognition dictionaries, retrained the acoustic models, and ran recognition on the same test data. As shown in Table 7, we obtained the same 6% relative improvement overall while there were differences in each focus condition. Nevertheless, the condition where the biggest gain occurs is still spontaneous speech (C3). As expected, computation when having 73 phonemes in the phonetic inventory increased significantly to 6.3xRT; a 31% increase compared to the baseline system with 41 phonemes.

We also observed the same behavior about the parameters of the acoustic model. Since there are more phonemes, there are more mixture densities. However, there are slightly less number of components in an average mixture, as shown in Row 3, Table 5.

Phs	C1	C2	C3	C4	C5	C6	C7	All
41	0.9	5.8	12.7	3.4	15.4	2.3	3.2	**4.8**
73	1.0	6.1	9.5	3.3	13.8	2.1	3.4	**4.5**
%rel.	+11	+5	-25	-3	-10	-9	+9	-6

Table 7: WERs obtained by systems using either 73 or 41 phonemes; 32 new phonemes were derived from the 16 most frequent CV pairs.

Word	Pronunciation	Occurrence	%
の	n-o	493804	7.1
を	o	269065	3.9
に	n-I	261438	3.7
は	w-a	219535	3.1
が	g-a	218133	3.1
で	d-e	166400	2.3
と	t-o	161433	2.3

Table 8: The 7 most frequent words in the NHK Corpus. Each word consists of a single character. All, except 'o', follow the CV pattern.

4.3. Special Phonemes for Single-Character Words

In this experiment, we explored a different dimension. We observed that the 7 most frequent words in the NHK Corpus are single-character words, with 6 of them following the CV pattern. Statistics for these 7 words are tabulated in Table 8. Since short words are generally more difficult to recognize with high accuracy, we decided to assign special phonemes for the 7 single-character words in our phonetic dictionary. As a result, the new system had 54 phonemes (41 + 13).

As before, we updated the training and recognition dictionaries, retrained the acoustic models, and tested on the same material. Basically, the result, tabulated in Table 9, shows that there is no gain overall for modeling these specialized allophones, especially since we do not use cross-word allophones in these systems. However, it is worthy to point out that the WER for the spontaneous speech condition (C3) is the lowest among those obtained through the three experiments.

5. DISCUSSION

Even though the overall relative improvement is at a modest 6%, it is interesting to see that the Spontaneous Speech condition always gets the biggest gain. However, since the amount of spontaneous speech in the test material is rather small (15 sentences with 158 words), it would not be wise to conclude strongly. We plan to use the technique on the male speech with 600 hours of training data next, and probably on other languages as well (of course, with a modified formulation, if possible).

In a certain way, this technique of introducing additional phonetic units in the acoustic model can be thought of as an allophone clustering procedure with a strong *a priori* knowledge of their occurrence frequency. Would it be possible to extend the context broader? Or would it be possible to introduce additional phonemes based on frequent biphones, not necessarily restricted to the CV pattern? If so, the technique would then be language-independent.

Phs	C1	C2	C3	C4	C5	C7	C8	Overall
41	0.9	5.8	12.7	3.4	15.4	2.3	3.2	4.8
54	0.9	6.6	8.2	3.5	15.7	2.1	2.8	4.8

Table 9: WERs obtained by systems using either 54 or 41 phonemes; 13 new phonemes were derived from the 7 most frequent words.

6. CONCLUSION

We have presented a technique to model the frequent allophones in Japanese speech recognition. Based on the analysis of the distribution of the CV syllabic structures in Japanese through a sample of 800 hours of Japanese broadcast news speech, we formulated a method to model frequent allophones by the use of additional phonetic units in the acoustic models. By introducing ten new phonetic units derived from the five most commonly occurring CV pairs, we achieved a 30% relative reduction in recognition word error rate for spontaneous speech and 6% relative reduction overall for all speech categories in a Japanese broadcast news transcription task.

Acknowledgements

The authors wish to thank Drs. Akio Ando, and Toru Imai, and the NHK Speech Group for their close collaboration on this project and for providing all the necessary speech and language modeling data.

References

1. L. Nguyen, S. Matsoukas, J. Davenport, J. Billa, R. Schwartz, and J. Makhoul, "The 1999 BBN Byblos 10xRT broadcast news transcription system," in *Proc. NIST 2000 Speech Transcription Workshop*, http://www.nist.gov/speech/publications/tw00/pdf/bn20.pdf
2. S. Matsoukas, L. Nguyen, J. Davenport, J. Billa, F. Richardson, M. Siu, D. Liu, R. Schwartz, and J. Makhoul, "The 1998 BBN Byblos primary system applied to English and Spanish broadcast news transcription," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Herndon, VA, Mar. 1999, pp. 255-260.
3. A. Ando, T. Imai, A. Kobayashi, H. Isono, and K. Nakabayashi, "Real-Time transcription system for simultaneous subtitling of Japanese broadcast news programs," *IEEE Trans. on Broadcasting*, Vol. 46, No. 3, Sep. 2000, pp. 189-196.
4. L. Nguyen, X. Guo, R. Schwartz, J. Makhoul, "Japanese broadcast news transcription," in *Proc. ICSLP 2002*, Denver, Colorado, Sep. 2002, pp...
5. L. Nguyen, T. Anastasakos, F. Kubala, C. LaPre, J. Makhoul, R. Schwartz, N. Yuan, G. Zavaliagkos, and Y. Zhao, "The 1994 BBN/Byblos speech recognition system," in *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, Jan. 1995, pp. 77-81.