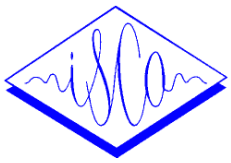

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

ON INTER-PHRASE CONTEXT DEPENDENCIES IN CONTINUOUSLY READ JAPANESE SPEECH

Kazuhiro Kondo[*], *Yu-Hung Kao*[**], *and Barbara Wheatley*[**]

Texas Instruments
[*]Tsukuba Research and Development Center, 17 Miyukigaoka, Tsukuba, Ibaraki 305, Japan
[**]Systems and Information Science Laboratory, P.O. Box 655474, MS 238, Dallas, Texas 75265

ABSTRACT

This paper investigates methods to model inter-phrase or word context for continuous Japanese speech recognition. It was found that by compiling a network of context-dependent phonetic models which models the inter-word or inter-phrase context, recognition error reduction by 32% can be achieved compared to models which do not account for inter-word context. However, this will significantly increase the number of phonetic models required to model the vocabulary. To overcome this increase, we clustered the inter-word/phrase context into only a few classes. Using one class for consonant inter-word context and two classes for vowel context, the recognition accuracy on digit string recognition was found to be virtually equal to the accuracy with unclustered models, while the number of phonetic models required was reduced by more than 50%.

1. INTRODUCTION

It has been shown previously that modeling between-word coarticulation in continuous speech dramatically improves the recognition accuracy[1]-[4]. However, by introducing additional phoneme variation models to account for these coarticulations, the required number of models increases significantly. This will mean more memory is needed to store these models, and more computation will be needed to match additional context with input speech. Efficient parsers which will reduce the added computation have been proposed previously[5][6]. In this paper, however, we propose to cluster the inter-word or phrase context into an extremely small number of clusters. We present results with the Japanese digit recognition task which prove that this clustering will have virtually no effect on the recognition accuracy. The required number of models was reduced by 50% compared to the unclustered case with the proposed scheme.

2. INTER-WORD/PHRASE CONTEXT-DEPENDENT MODELS

The word or phrase models used in this work were constructed by concatenating triphone models into a network which will model both the phonetic context within words, and the inter-word/phrase context. Unlike the methods described in [1], we did not distinguish triphones by its position. In other words, we did not distinguish triphones that are at the beginning, middle, or end of a word. Later on, however, when we start clustering the inter-word/phrase phones, we will distinguish these phones from triphones in the middle of a word.

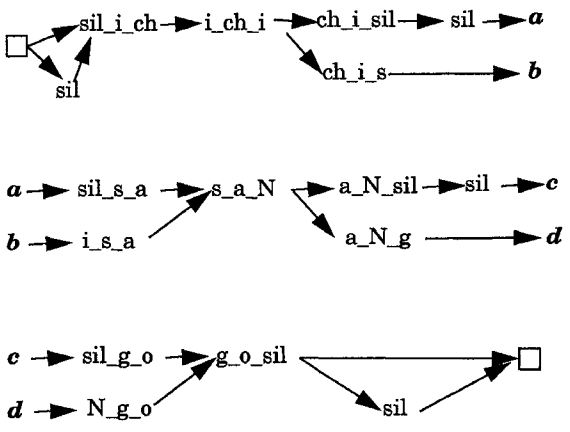

Figure 1 The sentence model for an example digit string "ichi(1) san(3) go(5)" for training.

Figure 1 shows the training sentence grammar for the Japanese digit string "ichi(1) san(3) go(5)." Each word has two paths coming into each word and going out. One path goes through

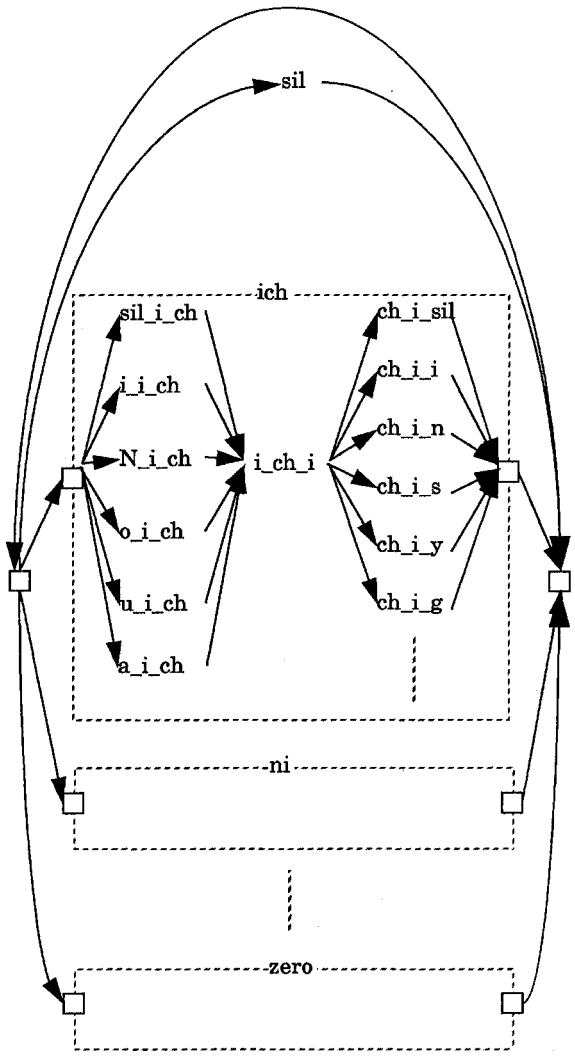

Figure 2 Example sentence model for Japanese digit string recognition.

an optional silence, while the other path connects directly to the neighboring word.

Figure 2 illustrates the recognition sentence grammar for the Japanese digit recognition task. This grammar allows any digit to follow each other, and is not length-constrained. Each word model is constructed by concatenating all valid triphone models with all possible right and left inter-word/phrase context in parallel, in addition to the conventional triphones which model phonetic context within words. The selection of the appropriate inter-word context is completely unsupervised. The context can be restricted according to the neighboring word in the search path, thereby reducing the search space, at a cost of added complexity. However, we found empirically that leaving this context modeling path unsupervised does not have significant impact on the performance, and thus we decided to leave the search unrestricted.

By introducing inter-word context dependencies, the number of required models increases significantly, as has been pointed out in earlier works [1]-[4]. For the digit recognition task, the number of triphone models increased from 96 for the inter-word context independent case, to 461 for the context dependent case.

To limit the increase in the required number of models to a minimum, we clustered the inter-word context into a very few classes, and found that this does not affect the recognition accuracy significantly. Detailed results will be given in the next session.

3. RECOGNITION TESTS

3.1 The Corpus

The corpus used in these experiments was collected for the voice dialing task [8]. The collected speech consisted of command phrases, such as "uchi e denwa (call home)", as well as continuous digit strings. Only the latter was used for the tests described here. A table-mounted linear microphone was used, and low to moderate level of noise was included in the background. There were 221 speakers, 112 men and 109 women. Each speaker spoke 100 sentences, of which half were digit strings. Approximately 80% of the data for each sex was used for training, while the remaining was used for testing. No overlap in speakers between the test set and training set exists. Overall, a total of 5447 male utterances and 5380 female utterances were used for training, and 2068 male and female utterances were used for testing.

3.2 Description of the Recognition System

The recognition system used here is an LPC-based HMM recognizer [7]. Speech is sampled at 8 kHz, LPC analysis is applied, and the LPC parameters are transformed into a feature vector. The feature vector is composed of spectral energy vectors output from a filter bank consisting of 14 mel-spaced filters, the short-term differences of these spectral energies, the speech level, and some voicing indicators. The total number of elements is 34. A linear transformation designed to normalize the

covariance statistics of the feature vector is applied, and the least significant 18 features are dropped, resulting in a vector dimension of 16. A unimodal Gaussian continuous distribution model is used along with a Viterbi-style maximum likelihood path scoring in the HMM model.

Table 1: Tested Clustering Schemes

Case	Inter-word Context	Context Position	Cluster	Member Phones
I	independent	right	all context	/i/, /n/, /s/, /y/, /g/, /r/, /h/, /k/, /z/, /m/, silence
		left	all context	/i/, /N/, /o/, /u/, /a/, silence
II	dependent	right	-	all 11 context modelled separately
		left	-	all 6 context modelled separately
III		right	silence	silence
			vowel	/i/
			bilabial	/h/, /m/
			dental-alveolar	/s/, /n/, /z/, /r/
			palatal-velar	/g/, /y/, /k/
		left	-	all 6 context modelled separately
IV		right	silence	silence
			vowel	/i/
			consonant	/h/, /m/, /s/, /n/, /z/, /r/, /g/, /y/, /k/
		left	-	all 6 context modelled separately
V		right	silence	silence
			vowel	/i/
			consonant	/h/, /m/, /s/, /n/, /z/, /r/, /g/, /y/, /k/
		left	silence	silence
			high vowel	/i/, /u/
			mid-low vowel	/a/, /o/
			nasal	/N/
VI		right	silence	silence
			all phones	/i/, /n/, /s/, /y/, /g/, /r/, /h/, /k/, /z/, /m/
		left	silence	silence
			all phones	/i/, /N/, /o/, /u/, /a/

The models were finite duration (no self-loops) models since we have observed consistently better performance with these models compared to infinite models with self-loops. The number of states in each model depends on the average duration of phone. The durations were computed from Viterbi aligned statistics with an initial set of monophones.

3.3 Test Results

Table 1 shows the various context clustering schemes tested, from inter-word context dependent models with no clustering (case II), to context dependent models with all phonetic contexts excluding silence context clustered into one (case VI). Test result for inter-word context-independent models (case I) were included for comparison. Table 2 shows the results for each case.

Results for cases I and II show that the introduction of inter-word context dependency decreases the word error rate from 2.5% to 1.7%, a 32% relative decrease in error rate. This comes with the cost of a five-fold increase in the number of models required to model the vocabulary. It is also interesting to point out that in these cases, as well as in other cases, most of the differences in the error rate can be seen in the substitution errors, not in the insertion or deletion errors.

Cases III through V compare the different clustering schemes. Results for case III shows that clustering of consonant contexts into a few classes will have no effect on word errors, while case IV shows that clustering of all consonant context will have only a minor increase in errors. The additional clustering of vowel context into two classes for case V did not show increase in word errors, and a slight increase in sentence error rate. The reduction in required number of models for case V compared to the unclustered case II was more than two fold, while the word error rate increase was kept within 0.1%. Finally, case VI shows that by just separating the silence context from other phone context, word errors can still be reduced considerably compared to the inter-word context independent models in case I.

Table 2: Test Results

Case	Number of Models	Error Rate				
		Substitution	Deletion	Insertion	Word	Sentence
I	96	1.6	0.4	0.4	2.5	18.1
II	461	1.0	0.4	0.3	1.7	13.3
III	320	1.0	0.4	0.3	1.7	12.9
IV	268	1.1	0.4	0.3	1.8	13.2
V	222	1.1	0.3	0.4	1.8	13.6
VI	146	1.2	0.4	0.5	2.0	15.1

4. CONCLUSION

We presented some results we obtained with inter-word context dependent models. The models were trained with a sentence grammar which supervises both the inter-word and within word phonetic context. The recognition grammar allows paths to all inter-word context dependent triphones in parallel, and poses no restriction on the search path. Even with this simple grammar, it was possible to reduce the error rate by 32% compared to models which do not model the inter-word context. We also proposed clustering schemes for the inter-word context. By clustering all consonants into one class and vowels into two classes, the total number of models required can be halved, while keeping the error rate increase within 0.1%.

Although the tests conducted here were for Japanese, we believe that similar methods will apply to other languages as well. The clustering scheme as well as its efficiency will differ, however.

The results shown here used phonetic models. However, the same inter-word context and its clustering scheme should also apply to other modeling units, such as word models. Unlike phone models, clustering of these context will not be as straightforward. We plan to investigate efficient methods for these units in the near future.

ACKNOWLEDMENTS

The authors would like to thank Nozomi Arai for her validation of the corpus, and Dr. Joseph Picone and Dr. Donald Shaver for their encouragements and discussions.

REFERENCES

[1]. M.-Y. Hwang, H.-W. Hon, and K.-F. Lee, "Modeling Between-Word Coarticulation in Continuous Speech Recognition," *Proc. Eurospeech '89*, Paris, France, Sept. 1989.

[2] D. B. Paul, "The Lincoln Continuous Speech Recognition System: Recent Development and Results," *Proc. DARPA Speech and Natural Language Process. Workshop*, Philadelphia, PA, Feb. 1989.

[3] R. Cardin, Y. Normandin, and E. Millien, "Inter-Word Coarticulation Modeling and MMIE Training for Improved Connected Digit Recognition," *Proc. IEEE International Conf. Acoust. Speech, Signal Process.*, Minneapolis, MN, Apr., 1993.

[4]. T. Watanabe, R. Isotani, and S. Tsukada, "Speaker-Independent Speech Recognition Based on Hidden Markov Model Using Demi-Syllable Units," *IEICE Trans. Part D-II*, vol. J75-D-II, no. 8, pp. 1281-1289, Aug. 1992.

[5]. W. Chou, T. Matsuoka, B.-H. Juang, and C.-H. Lee, "An Algorithm of High Resolution and Efficient Multiple String Hypothesization for Continuous Speech Recognition Using Inter-Word Models," *Proc. IEEE International Conf. Acoust. Speech, Signal Process.*, Adelaide, Australia, April, 1994.

[6] K. Itou, S. Hayamizu, and H. Tanaka, "Continuous Speech Recognition by Context-Dependent Phonetic HMM and an Efficient Algorithm for Finding N-Best Sentence Hypothesis," *Proc. IEEE International Conf. Acoust. Speech, Signal Process.*, San Francisco, CA, Mar., 1992.

[7] G. R. Doddington, "Phonetically Sensitive Discriminants for Improved Speech Recognition," Proc. ICASSP 89, May, 1989.

[8] K. Kondo, J. Picone, and B. Wheatley, "A Comparative Analysis of Japanese and English Digit Recognition," *Proc. IEEE International Conf. Acoust. Speech, Signal Process., Adelaide*, Australia, April, 1994.