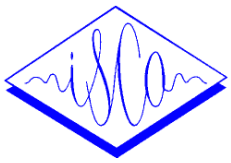

2nd European Conference on
Speech Communication and Technology
EUROSPEECH '91
Genova, Italy, September 24-26, 1991

WORD JUNCTURE MODELING USING INTER-WORD CONTEXT-DEPENDENT PHONE-LIKE UNITS

E. Giachin,† C.-H. Lee, L. R. Rabiner, A. E. Rosenberg and R. Pieraccini

Speech Research Department
AT&T Bell Laboratories
Murray Hill, NJ 07974, USA

ABSTRACT

We report on a speech recognition system that uses *inter-word context dependent units* (CDP) to model pronunciation variations at word boundaries. Word juncture coarticulation phenomena are a major source of acoustic variability for the initial and final parts of a word when spoken in fluent speech. In some cases, the alteration of a phone due to neighboring phones is comparatively small and the actual realization is perceived as a variation of the original phone. To cope with this type of variations, we design a set of inter-word CDP and modify the pronunciation of a word by replacing the word beginning and word ending phones with all possible inter-word units. By properly connecting all possible word beginnings and word endings, word boundaries are thereby better represented. Taking into account word juncture pronunciation changes, better models can be obtained in training. Such models have achieve better results in recognition.

1. INTRODUCTION

Coarticulation phenomena arising at the boundary between words are a major source of acoustic variability for the initial and final parts of a word when spoken in continuous speech. If this type of contextual variability is not adequately represented in a recognition system, errors are likely to occur. This is the case, for instance, when words in the dictionary are phonetically transcribed according to their pronunciations in isolation. One solution to this problem is to provide a more precise phonetic representation of the region across word boundaries. Different approaches can be taken to achieve this goal, depending on which type of coarticulation phenomenon is handled. In an earlier paper [1] two different types of pronunciation changes at word junctures, namely *soft* and *hard* changes, were characterized. In *soft* changes, the alteration of a phone due to neighboring phones is relatively small and the actual realization is perceived as a variation of the original phone rather than a transformation to a different phone. This alteration is of the same nature as the one observed in intra-word phones. By contrast, in *hard* changes, a boundary phone may undergo a complete deletion or a substitution by a totally different phone. It was experimentally shown [1] that simple phonological rules were effective in coping with errors attributed to hard changes. However for soft changes a different approach, based on the use of *context-dependent phones*, is required. Such units specify phones according to their context, i.e. the preceding and the following phones. Context dependent phones are not very effective for coping with hard changes because these changes are comparatively rare and the training material is not sufficient to model the units that would represent them. Soft changes occur more frequently in the training set and hence such changes can be modeled in an manner similar to the way that intra-word units are modeled.

In this paper we describe the application of inter-word context dependent phones to modeling soft changes at word boundaries. Taking into account word juncture coarticulation, improved models can be estimated in training, and such models have proven useful in recognition. We test the inter-word CDP approach on the DARPA Naval Resource Management task. Relative error reductions of up to 25% over results without inter-word units [2] are observed.

2. CONTEXT-DEPENDENT PHONE-LIKE UNITS

Context-dependent (CD) phones were proposed as a means of accurately representing highly variable sounds within words (e.g. [3]). The basic idea is to start from the definition of a set of *context-independent* (CI) phone units and to create, for each of them, a set of units representing all the possible contexts where the unit can occur in terms of more detailed CI units. Throughout this study, we use a set of 47 CI units [1-2].

Recently the use of context-dependent phones to model inter-word transitions in addition to intra-word transitions has been proposed by a number of research groups (e.g. [4]). In this case, the initial (final) phone of a word has to be specified for each possible final (initial) phone of the previous (following) word. Hence there generally are significantly more inter-word context-dependent phones than intra-word phones for a given vocabulary, and therefore the training issues outlined above become even more acute. In recognition, moreover, since it is not known *a priori* which words will precede or follow any given word, the words have to be represented with all their possible initial and final phones. The approach we adopted in this study follow closely with existing inter-word modeling approaches (e.g. [4]) in handling between-word silence and one-phone words. However, special attention is paid to address

† Now with CSELT, Torino, Italy.

issues related to using the continuously density HMM framework [2].

The key training issue is to have enough phone units to achieve a fine grained phonetic representation, but at the same time have enough training material to obtain statistically reliable models. In this study, a simple rule for generating CDP's has been followed [2] that insures that for each model the total number of occurrences in the training set never falls below a chosen threshold. The rule, which we will refer to as the *unit reduction rule*, provides a way to back off to *single-context units* if there are not enough occurrences of full-context units, and to *no-context* units if there are not enough occurrences of single-context phones. To describe the unit reduction rule, we assume that all the training sentences have been segmented in terms of CDP's, and let $C(x-y-z)$ be the number of occurrences of phone $x-y-z$ in the training set and let T be a count threshold. The rule is as follows:

For every $x-y-z$,

— If $C(x-y-z) > T$, then create $x-y-z$

— Otherwise, if $C(x-y-z) + C(x-y-\$) > T$, then create $x-y-\$$

— Otherwise, if $C(x-y-z) + C(\$-y-z) > T$, then create $\$-y-z$

— Otherwise, create $\$-y-\$$.

By varying the value of the threshold T, different phone inventories can be obtained and tested. Most existing implementations (e.g. [4]) use only double-context and no-context units to represent inter-word units. In order to be consistent with the unit reduction rule defined above, we also use single-context phone units in both training and recognition.

Training is based on a modified version of the segmental k-means training procedure [1-2, 5]. Segmentation is carried out on the finite state network (FSN) that represents each utterance in terms of phone units. Since context-dependent phone units are used to describe both word boundaries and word bodies, there is no discontinuity between words. However, optional silences are allowed between words.

3. RECOGNITION USING INTER-WORD CDP's

Recognition is based on a modified version of the well-known, frame-synchronous, beam search algorithm. In recognition it is not known *a priori* which words will precede or follow a given word. Hence the words have to be represented with all the possible initial and final segments. These segments include single- and no-context units as well as double-context units. As an example, the representation of the word "above" (transcribed as *ax b ah v* as an isolated utterance in terms of context-independent units) is shown in Figure 1, where the x_i represent the different possible left contexts for *ax* and the y_j the possible right contexts for *v*. Usually there are many possible word beginning and word ending units. Using the unit reduction rule described in Section 2, the average number of word beginnings or word endings is about 20. In cases with single-phone words, there is no word body and every possible word beginning also serves as word ending which make word connections more complicated.

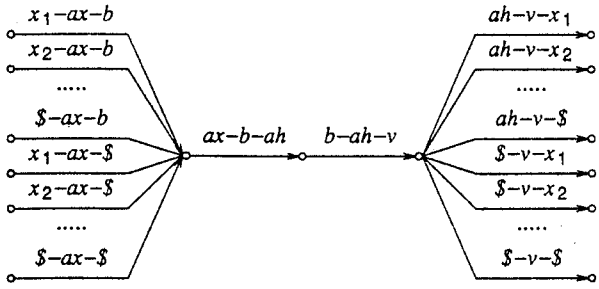

Figure 1. Phonetic transcription of word "above"

The above word transcriptions for all words in the vocabulary are inserted into the arcs of the FSN representing the lexicon. This situation is similar to the one described for the usage of phonological rules [1], however the inter-word connections are more complex. We now discuss issues with respect to the handling of single-phone words and between word silence.

3.1 Single-phone words

Single phone words have to treated as exceptions for several reasons:

1. They have no "body", or central core which is relatively insensitive to word juncture phenomena. Thus there is no inner node in which all segments merge. This complicates the control of the decoding algorithm.
2. Since the initial and final segments have to be combined in one single set, there are many more such segments than in the case of multi-phone words (as many as 10 times more). On the other hand, single-phone words are a small percentage of all words. This means that the structures used for multi-phone would not exploit memory efficiently if they were used for single-phone words also.

In the present implementation, each segment of a single-phone word is represented as a separate arc in the FSN. Also, special care is taken in the recognition algorithm to insure compatibility with words having more than two phones.

3.2 Silences between words

Optional silences between words could be represented as words on their own. However this would make it difficult to implement the word-pair grammar, because the check for legal pairs must ignore interword silence and hence must span three consecutive levels of the FSN instead of only two. In order to avoid this difficulty, silences are represented as segments that are added to final word segments. (There still exists a silence *word* that can be located at the beginning of the sentence.) There are two types of units that can be followed by silence:

— units having a silence as the right context. These units *must* be followed by silence, since they cannot be connected to any other unit.

— units having a "don't care" option ($) as the right context. These units can be followed by units different from silence; thus the segments corresponding to these units are duplicated when a silence is added to them. With a slight

redundancy, this greatly simplifies the treatment of silence.

4. EXPERIMENTAL RESULTS

The recognizer performance has been evaluated on the DARPA Naval Resource Management Task. All experiments were based on using the perplexity-60 word-pair (WP) grammar or the perplexity-991 uniform grammar. In this study, we report on results using three sets of test sentences, namely:

1. 150 sentences from 15 talkers, referred to as "150". This is the test set used for the initial evaluation of the SPHINX system.
2. 300 sentences from 10 talkers distributed by DARPA in February 1989, referred to as "FEB89".
3. 160 sentences randomly selected from the set of training sentences, referred to as "TRAIN".

Two training sets have been used. One is the 109-talker DARPA set, containing 3990 sentences. The other is a reduced 80-talker set containing 3200 sentences, with 72 out of 80 talkers included in the 109-talker set. This set has been used in the first series of experiments described below to facilitate comparison of results with earlier experiments. The "150" recognition set overlaps with the 109-talker training set but not with the 80-talker set.

The acoustic analysis is described in detail in [2] and is summarized as follows. Speech is filtered from 100 Hz to 3.8 kHz, sampled at 8 kHz, and blocked into 30 ms Hamming-windowed frames. The frame shift is 10 ms. A 10th order LPC analysis is then performed on the windowed data and 12 LPC liftered cepstrum and 12 delta cepstrum coefficients are computed and the 24-dimension vector is used as the feature vector for recognition.

4.1 Performance contribution by inter-word units

We present a comparison of the recognizer performance using inter-word units and of the baseline system that does not use inter-word units. Results for the word-pair grammar case, which are given in terms of word accuracy, are shown in Table 1. These results have been obtained with only one cycle of training (the least necessary to obtain the models), thus the benefits come mostly from the application of inter-word units in recognition. The 80-talker training set has been used, thus recognition results are presented for all test sets. Two types of models, 2502-I and 2502-II, have been used. Both of them have been obtained with a count threshold (T) of 10, giving rise to 2502 units. For the 2502-I model, the nominal number of mixture components per state is 8. For each model the mixture parameters (i.e. the weights, the means and the variances) were obtained independently, this model is called a CD1 model [2]. For the 2502-II model, the maximum number of mixture components for each state is 256; however, the mixture components are not obtained independently for each unit: this set has been developed starting from a set of context-independent 256-component units, and all the CDP's that are a specification of the same CI unit have the same means and variances; only the mixture weights are changed. These models are called CD2 models [2], and have been shown to be effective in modeling context using a small amount of training data [2]. They have some commonalities with the so-called semi-continuous or tied-mixture HMM (e.g. [6]); one important difference is that the models used here rely on a number of codebooks equal to the number of context-independent units (47), instead of using a single codebook covering the whole acoustic space. The 2502-II models have also been smoothed by averaging them with the CI units according to a technique similar to deleted interpolation.

In Table 1, we listed word accuracies with respect to the three testing sets in different testing conditions. System B in column 2 indicates the baseline system without using inter-word units for recognition and system IW uses the set of inter-word units for both word representation and word connection.

Model set	System	Test set & Word accuracy (%)		
		150	FEB89	Train
2502-I (CD1-8mix)	B	87.6	83.4	93.3
	IW	90.8	88.2	99.0
2502-II (CD2-256mix)	B	92.4	90.3	96.9
	IW	94.3	90.7	98.2

TABLE 1. Performance comparison using the WP grammar

The following can be observed from the results in Table 1:

— As expected, a much better recognition accuracy is obtained with the training sentences than with any of the test sets. Almost perfect result was achieved for the 2502-I model.

— The 2502-II set gives better performance, on the average, than the 2502-I set. This indicates that smoothing is helpful for obtaining robust models because the unsmoothed models correspond too closely to the training data. This is confirmed by the fact that, for the training sentences, the situation is reversed, and the unsmoothed models perform better than the smoothed ones. Note, however, that the relative error reduction for the unsmoothed models is higher than that of the smoothed ones.

— There are less errors on function words and short words in general ("a", "the", etc.). This happens because such words suffer from coarticulation more than the others and hence get a greater benefit from inter-word units.

— The sentence recognition likelihood (score) averaged over the three test sets, is higher with inter-word CDP's. Also, the difference between the score of incorrectly recognized sentences and the score of correct sentences (obtained by aligning the correct words on the acoustic data) is smaller. Both facts indicate a more precise phonetic representation of sentences.

The results using the uniform grammar case follow similar patterns as the results using the word-pair grammar. Both for the WP grammar and for the uniform grammar, there are considerable performance differences between the various test sets. One interesting result is that for the FEB89 testing set, the improvement from 65.5% to 70.4% word accuracy is more significant than for the case using the WP grammar.

4.2 The effect of single-context units on performance

All the results reported so far were based on the 80-talker training set and the threshold used in the unit reduction rule was set to a fixed value of 10. We now apply the same approach to the 109-talker training set and use a threshold of 30 in unit selection which results in a set of 1282 units when inter-word CDP's are used and a set of 1090 units when only intra-word units are used. We tested the two sets of models on the FEB89 test data using the WP grammar. The word accuracies are 93.0% and 91.3% respectively for testing with and without the inter-word units. This represents a 20% error rate reduction with the incorporation of inter-word units.

By comparing the two set the number of single-context units used in the 1282-unit set is rather limited. One simple method to expand the set of single-context units, called *model merging* method, was investigated in order to increase the number of single-context units in the inter-word unit sets. The "merging method" aims at augmenting an inter-word unit set with some single-context units drawn from the corresponding intra-word set. More precisely, the method is as follows. We compared the lexicon for the 1282 inter-word units with the lexicon for a set of 1090 intra-word units. Whenever a word body of the 1282-unit lexicon contained a no-context unit while the corresponding unit in the 1090-unit lexicon was a single-context unit, we added that single-context unit to the 1282-unit set and modified the lexicon accordingly. Eventually, 185 new units were added to the 1282-unit set, bringing the total number of units to 1467. In the entire lexicon 555 no-context units were replaced by single-context units. The new 1467-unit set was tested on the FEB89 test set again and a small improvement was obtained (93.2% word accuracy).

Most existing inter-word modeling approaches use only no-context and double-context units (e.g. [4]). The partial success of the model merging method suggests that the incorporation of single-context units contributed to an improved performance in two ways: one to better represent the lexicon and the other to better represent the inter-word connections. The success of the model merging method also suggests a second possible approach, in which one starts from an intra-word set and then adds inter-word units drawn from the corresponding inter-word set. This method should provide some insight into the way performance is influenced by unit clustering.

5. SUMMARY AND DISCUSSION

Inter-word context-dependent phones have been shown to be essential to improve the accuracy of phonetic transcriptions and consequently the performance of a continuous speech recognition system. Different sets of CDP's were tested, and a word accuracy of 93% was achieved for the FEB89 test set. Different approaches to using inter-word units in recognition and in training were studied, with results suggesting that only the most detailed boundary descriptions should be used in training, whereas the whole set of boundary descriptions gives the best performance in recognition.

More recently we have found that the spectral vectors, corresponding to the same speech unit, behave differently statistically, depending on whether they are at word boundaries or within a word. The results suggest that intra-word and inter-word units should be modeled independently, even when they appear in the same context. This so-called *position-dependent units* have been shown to improve separation between units and, as a result, the performance of the recognition system is also improved [7]. It was also shown that improved feature analysis leads to a significant improvement in performance for continuous speech recognition [7]. By incorporating inter-word units, improved feature analysis and position-dependent units into our system, we achieved a word accuracy of 95.4% for the FEB89 test set, using the word-pair grammar and the 109-talker training set. We also found the use of a combined female/male model [8] and the incorporation of state duration models improves performance. The best performance we have now for the FEB89 testing set using the WP grammar is over 96% word accuracy and over 80% sentence accuracy which demonstrates the effectiveness of acoustic modeling toward robust continuous speech recognition.

Since the use of inter-word context-dependent units mainly deals with soft changes at word boundaries, hard changes are not handled effectively. The success of the application of phonological rules [1] suggests that they should also be used together with context-dependent units. It is expected that both strategies will maintain their own corrective power, because they address different kinds of coarticulation phenomena. An integrated approach combining both phonological rules and inter-word units is now under investigation.

REFERENCES

[1] E. P. Giachin, A. E. Rosenberg and C.-H. Lee, "Word Juncture Modeling Using Phonological Rules for HMM-Based Continuous Speech Recognition," *Computer Speech and Language*, Vol. 5, pp. 155-168, 1991.

[2] C.-H. Lee, L. R. Rabiner, R. Pieraccini and J. G. Wilpon, "Acoustic Modeling for Large Vocabulary Speech Recognition," *Computer Speech and Language*, Vol. 4, pp. 127-165, 1990.

[3] R. M. Schwartz *et al*, "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech," *Proc. ICASSP 85*, Tampa, pp. 1205-1208, 1985.

[4] M. Y. Hwang, H. W. Hon, and K. F. Lee, "Modelling between-Word Coarticulation in Continuous Speech Recognition," *Proc. EuroSpeech 89*, Paris, September 1989.

[5] L. R. Rabiner, J. G. Wilpon, and B. H. Juang, "A Segmental K-means Training Procedure for Connected Word Recognition Based on Whole Word Reference Patterns," *AT&T Technical Journal*, vol. 65 no. 3, pp. 21-31, May/June 1986.

[6] X. D. Huang and M. A. Jack, "Semi-continuous Hidden Markov Models for Speech Signals," *Computer Speech and Language*, Vol. 3, No. 3, pp. 239-251, July 1989.

[7] C.-H. Lee, E. P. Giachin, L. R. Rabiner, R. Pieraccini and A. E. Rosenberg, "Improved Acoustic Modeling for Speaker-Independent Large Vocabulary Continuous Speech Recognition," *Proc. ICASSP 91*, Toronto, Canada, May 1991.

[8] J.-L. Gauvain and C.-H. Lee, "Bayesian Learning of Gaussian Mixture Densities for Hidden Markov Models," *Proc. DARPA Speech and Natural Language Workshop*, Monterey, CA., Feb. 1991.