Contextually-Based Data-Derived Pronunciation Networks for Automatic Speech Recognition*

Francine R. Chen
XEROX PALO ALTO RESEARCH CENTER
3333 Coyote Hill Road
Palo Alto, CA 94304

Abstract

The context in which a phoneme occurs leads to consistent differences in how it is pronounced. Phonologists employ a variety of contextual descriptors, based on factors such as stress and syllable boundaries, to explain phonological variation. However, in developing pronunciation networks for speech recognition systems, little explicit use is made of context other than the use of whole word models and use of triphone models.

This paper describes the creation of pronunciation networks using a wide variety of contextual factors which allow better prediction of pronunciation variation. We use a phoneme level representation which permits easy addition of new words to the vocabulary, with a flexible context representation which allows modeling of long-range effects, extending over syllables and across word-boundaries. In order to incorporate a wide variety of factors in the creation of pronunciation networks, we used data-derived context trees, which possess properties useful for pronunciation network creation.

Introduction

The context in which a phoneme occurs leads to consistent differences in how it is pronounced. Phonologists employ a variety of contextual descriptors to explain phonological variation. These descriptors are theoretically motivated by studies of different languages and are comprised of many factors, such as stress and syllable part. However, in current speech recognition systems, only a few contextual descriptors are employed when developing pronunciation networks. In these systems, generally the effects of only the preceding and following sounds, as in triphone models, or implicit within-word contextual effects, as in whole word models, are modeled.

Context-dependent triphone models have been found to be better models than context-free phone models, or monophones, and are used now by many recognition systems (e.g. Chow, 1986). More recently, Lee (1988, 1989) introduced the idea of clustered triphones, in which triphones exhibiting similar coarticulatory behavior are grouped together. Clustered triphones require less training data because there are fewer models, resulting in better training of the models which have been defined.

Paul (1988) has conducted studies comparing the recognition rates of whole word models and triphone models. In his studies, he found that whole word models provided somewhat better recognition rates over triphone models, and that triphone models have much better recognition rates than monophone models. However, in these studies, context across word boundaries was not modeled. In a subsequent study, Paul (1989) found that use of triphone models which modeled context across word boundaries provided significantly better recognition rates over word-internal triphone models. Extrapolating, one would expect even better performance with whole word models which also model contextual effects across word boundaries. However, the amount of data necessary to train such a model for a moderate size vocabulary could be prohibitive. Furthermore, addition of a new word to the vocabulary would require many new tokens of the word because at least one token of the word in each context in which it could appear would be required.

Instead of words, the use of a smaller representational unit, such as phones, with an enriched set of contextual descriptors can provide models which capture many of the contextual effects that whole word

*This work was sponsored in part by the Defense Advanced Research Projects Agency (DOD), under the Information Science and Technology Office, contract #N00140-86-C-8996.

contextual factor	values
preceding phoneme	(all phonemes) + sentence boundary
following phoneme	(all phonemes) + sentence boundary
preceding phone	(all phones) + deletion + sentence boundary
following phone	(all phones) + deletion + sentence boundary
syllable part	onset, nucleus, coda
stress	primary, secondary, unstressed
syllable boundary type	initial, final, internal, initial-and-final
foot boundary type	initial, final, internal, initial-and-final
word boundary type	initial, final, internal, initial-and-final
cluster type	onset, coda, nil
open syllable?	true, false
true vowel?	true, false
function word?	true, false

Table 1: Contexts used in pronunciation experiments

models capture. In addition, phone models have a smaller training data requirement and provide a more general framework for adding new words to the vocabulary.

Because more contextual effects can be captured by using a wide variety of factors, use of these factors when creating pronunciation networks from dictionary baseforms allows better predictions of pronunciation variants. Better modeling of observed phonological variation can result in better performance of speech recognition systems. A case in point is the work by Weintraub *et al.* (1989) who found that phonological modeling improved their recognition results. In their work, the phonological rules were derived by hand.

This paper describes the creation of pronunciation networks using a wide variety of contextual factors. We propose a systematic methodology for creating pronunciation networks which can capture the predominant contextual effects using a representation at the phoneme level. With the use of a phoneme representation, new words can be added without the need for additional training data, as would be necessary in whole word models. Our representation also allows us to capture long-range effects. In this study, some factors extend over syllables and across word boundaries.

Contextual Factors Represented in Context Trees

Linguists describe the context of a phoneme using many types of theoretically motivated *factors*, such as stress and word boundary. Each contextual factor describing the context of a phoneme has a *value*. For example, the factor stress may take on any one value of primary, secondary, or unstressed.

In this work, we describe the context of a phoneme using a set of linguistically motivated contextual factors. These factors and their corresponding values are listed in Table 1. Some of the factors are structures normally associated with several phonemes. For example, the factor syllable part may take on the value onset, which may be composed of up to three phonemes, as in the sequence /str/. In such cases, we assign the factor value to each phoneme within the structure. In our example, the exemplars /s/, /t/, and /r/ in an /str/ sequence would each be assigned a syllable part value of onset. This representation allows modeling of long-range contextual effects simultaneously with a local phoneme representation, which simplifies the addition of new words to a recognition vocabulary.

If the context of a phoneme is described by simultaneously using all the contextual factors listed in Table 1, a prohibitive amount of data would be required to form an adequate description of each phoneme in each context. Each contextual factor represents a separate dimension, and with such a large number of dimensions, the distribution of phonemes in a context will be sparse. One way to handle this difficulty is to build a "context tree," in which a subset of contextual factors is selected using a greedy algorithm which

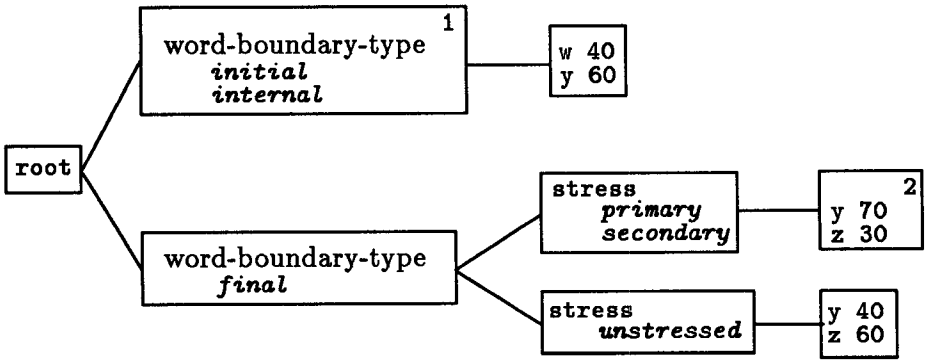

Figure 1: An illustrative context tree for some phoneme X

minimizes the loss of information at each selection. A combination of tree induction for selecting factors and clustering for grouping factor values is used to create a context tree. The number of leaves in the tree and branching of the tree is determined by the data. In the next paragraph we give a brief description of context trees. A more detailed description of context trees and how they are created is given in Chen and Shrager (1989). An alternate method for grouping contextual factors, based on binary splitting of subspaces until a preset number of clusters is formed, is given by Sagayama (1989).

A context tree is an n-ary decision tree which models the relationship between contextual factors and the allophones which occur in different contexts. We create context trees from a set of training exemplars using data derived from the hand-transcribed "sx" sentences of the TIMIT database (Lamel, *et al.*, 1986; Fisher *et al.*, 1987). An illustrative tree describing the distribution of allophones in context for some phoneme X is shown in Figure 1. The nodes of a context tree represent the values of a particular contextual factor. In the example, node 1 corresponds to the contextual factor word-boundary-type with the value initial or internal. Each leaf of a context tree encodes the distribution of allophones in each context. In general, more than one allophone occurs in a context because phoneme realizations are not deterministic. For example, leaf 2 of the example corresponds to the realizations of X, which is realized as the allophone y 70% and as z 30% of the time when in a word-final and either primary or secondary stress context.

The representation used in context trees permits flexible modeling of contextual effects. Since the context trees are derived from data in which phonemes are described by a set of contextual factors, long-range contextual effects are modeled. Also, since each factor represents a separate dimension, a factor ignores structures which are irrelevant to it. Thus, a contextual factor such as preceding phoneme extends across word boundaries.

Pronunciation Network Creation

This section describes a systematic method for creating pronunciation networks in which a wide variety of contextual factors are used. In addition to using more contextual factors, our method of network creation has a number of inherent advantages, such as the ability to estimate allophone distributions from partial contextual descriptions. This method is data-intensive, using the data to determine possible pronunciations as well as to estimate the probabilities associated with each pronunciation.

Mapping Dictionary Pronunciations

Dictionary pronunciations of words are relatively coarse-grained in that they do not indicate allophonic variation. Phone-based speech recognition systems generally represent words using allophones rather than

dictionary baseforms because the allophones of a dictionary "phoneme" may be very different, as measured by the acoustic similarity metrics commonly employed in recognition systems. The allophonically-based pronunciations are usually represented compactly in a pronunciation network. In this section, we describe the creation of pronunciation networks from dictionary baseforms. The networks are produced by *mapping* the dictionary pronunciation of a word into an allophonic representation specified by the context trees. The dictionary that we use was developed at Xerox PARC and is called the *X-Dictionary*[†].

The mapping from a dictionary baseform to a set of possible pronunciations is characterized by the substitution, deletion, and insertion of sounds. Each dictionary sound may be realized as an allophone or it may be deleted. Therefore, substitution and deletion of a dictionary phoneme may be treated identically in mapping a dictionary baseform into a network of allophones. A context tree, which we call a "phoneme" tree, is used to describe the observed substitutions and deletions of a dictionary phoneme in transcriptions of speech. One phoneme tree is created for each of the 45 dictionary phonemes and the data in each tree defines the set of allophones observed in each context for a dictionary phoneme.

In addition to modeling substitutions and deletions, as the phoneme trees do, pronunciation network creation also requires modeling of insertions. Insertions do not fit the substitution/deletion model in which a "phoneme" is realized as an allophone. Instead, insertions may occur between any pair of "phonemes." In addition, one must also model when insertions do not occur so that the probability of an insertion in any context can be predicted. These requirements are met by representing all insertions and non-insertions in one tree in which the contextual factors are redefined to be a set applicable to insertions. The contextual factors describing insertions do not describe the transformation of an underlying phoneme. Instead, the factors describe the context of the phonemes adjacent to where an insertion can occur. Contextual factors describing insertions are derived from factors describing the context of a phoneme by replacing each factor with ones describing the context of adjacent phonemes. For example, the factor stress is replaced with stress-of-preceding-phoneme and stress-of-following-phoneme. Since the new factors describe the context of adjacent phonemes, the value sentence-boundary is added to the allowable values of each factor to indicate the beginning or end of a sentence. In organizing the data to build an "insertion" tree, all pairs of phonemes in the training data are checked for whether or not an insertion occurred between them. If so, the context and insertion is noted; if not, the context and the fact that no insertion occurred is also noted. The "insertion" tree predicts when insertions can occur as well as what type of insertion can occur in a particular context.

Network Creation

Networks are created word by word and can be joined to produce a pronunciation network for a recognition system. Networks created using our method explicitly model cross-word-boundary contextual effects. If the context at the word boundaries is unknown, the possible allophones and corresponding probabilities are enumerated for each context value. Alternatively, if the context of the word is specified, only the allophones for the word boundary context are used. Since the context of the word is known, word-boundary phonemes can be treated the same as word-internal phonemes.

To create a word network, a two-pass process is used. First, each dictionary "phoneme" in a word is mapped to the allophone distribution represented by the leaf in a phoneme tree corresponding to the context in which the phoneme occurs, producing a sequence of allophones representing the sequence of phonemes (see Figure 2a). The context of a leaf in a phoneme tree is described by the contextual factor values encountered in traversing a phoneme tree from the root node to the leaf. Contextual constraints associated with the allophones from a leaf are matched to contextual constraints of adjacent allophones. If the phoneme is word initial or word final and the context at the word boundaries is not specified, then the allophones for each context must be incorporated into the network. Insertions are then added between the leaf values when the context for an insertion is compatible. Insertions are added after substitutions and deletions because

[†]The *X-Dictionary* has been checked for consistency and has been augmented from entries in standard dictionaries to include foot boundary indicators.

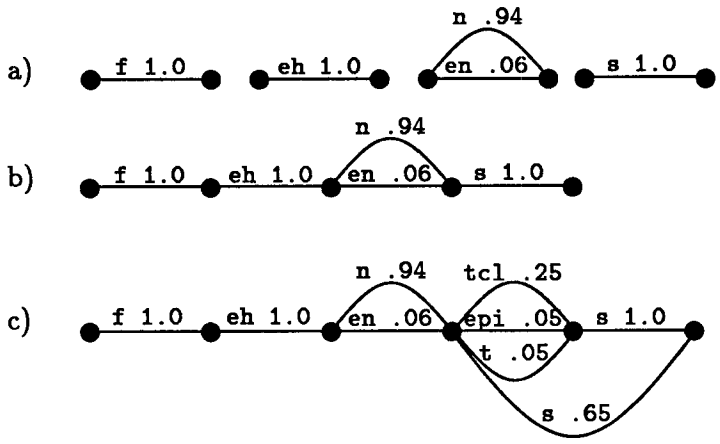

Figure 2: Pronunciation network for "fence": a) initial arcs b) arcs connected c) insertions added.

the context in which an insertion occurs is dependent upon adjacent phones, which is determined by the phoneme realizations.

Using our method based on context trees, the pronunciation network produced for the word "fence" is shown in Figure 2. In creating this network, we made the simplifying assumption that we would not use the contextual factors describing adjacent phones for modeling substitutions and deletions. This produces the simple network in Figure 2b in which each set of arcs from a leaf node are connected at the beginning and end. Addition of insertions, in which we do include the contextual factors describing adjacent phones, produces the network shown in Figure 2c.

In creating this network, we also assumed that the word was spoken in isolation and therefore preceded and followed by silence. Had we not specified a context, the boundaries of the word would be much more bushy with additional arcs representing the different possible allophones and probabilities in various contexts. For example, an /s/ in word-final position is more likely to be palatalized and pronounced as a [š] when followed by a /y/ or a /š/, as in "fence your" or "fence should," than when followed by a vowel, as in "fence any." When the context of a word is not specified, possible palatalization is modeled through the addition of two arcs which require that the following phoneme can cause palatalization, such as /y/ or /š/. One arc represents the [s] allophone and the other arc represents the [š] allophone; the probability of the two arcs sum to 1.0. The original [s] arc remains untouched with a probability of 1.0 and now has a constraint on its following context prohibiting any following phoneme which can cause palatalization of /s/. When all word boundary contexts are listed, unobserved cross-word-boundary contexts may be handled by including a "default" in which context is ignored. That is, the "default" context is composed of all observed allophones across all contexts. A more detailed estimate, based on a partial context specification, requires actual lookup in a context tree.

Because of limited training data, some of the words to be represented may contain a context value which has not been observed in the training data. However, each node of the tree contains the distribution of allophones for the partial context represented by the node. Thus, the allophones for unobserved contexts can be estimated from a partial context specification by tracing down the tree as far as consistent with the observed contextual factor values describing a phonemic baseform.

Properties of Context Trees

The context trees possess properties which are useful for producing pronunciation networks. As noted in the previous section, allophone distributions in unobserved contexts may be easily estimated from a partial

context specification by tracing down the tree until no context matches. Other properties of the context trees permit ease of context combination, specification of natural groups for tying (Jelinek and Mercer, 1980) in hidden Markov models (HMM), representation of allophone probabilities, and systematic reduction of network size.

In the tree-induction method, the leaves of the tree, by definition, represent mutually exclusive contexts. This simplifies the comparison and combination of contexts during network creation. For example, finding a set of arcs with a compatible context is simplified if one can assume that once a matching constraint is found, one need not look any further.

In HMMs, tying is used to constrain one probability in a network to be equal to another, the underlying idea being that equivalent items should be assigned the same probability. Each leaf of a phoneme tree represents all the allophones of that phoneme which have been observed in a particular context. Thus each leaf is a natural group for tying. In a network representation in which the labels are on the arcs, the probability assigned to an arc should be tied to all other arcs with the same label from the same leaf.

Many rule sets have been developed to describe phonological variation. However, by using a data-intensive approach, allophone probabilities in a context may be directly estimated. Furthermore, counts of allophones in context can be used to reduce the size of pronunciation networks by removing unlikely allophones.

The probabilities in a context tree can be further refined if the network is used in an HMM which is trained. The probabilities provide a good initial estimate, and more importantly, the absence of unlikely allophones in the network allow more robust training to be performed.

In the creation of pronunciation networks, it is hard to define an "optimum" number of pronunciations to represent. With only a few pronunciations, recognition performance may not be optimal because the modeling of pronunciation variation in words is left to the front-end. With many pronunciations, recognizer performance may be poor because the amount of training data is sparse and unlikely pronunciations may confuse the recognizer. This is the problem described by SRI as the problem of "many additional parameters to be estimated with the same amount of training data" (Weintraub *et al.*, 1989). SRI uses measures of coverage and overcoverage and accordingly modifies by hand the phonological rules they use.

With context trees, this problem can be handled at a phoneme level. Given a large data set, context trees tend to overgenerate pronunciations because each new allophonic realization of a phoneme in a context translates into another possible arc in a network. But because context trees contain count information on allophones in context, pruning can be used in a systematic way to remove the more unlikely pronunciations, thus reducing the number of pronunciations in a network. To remove unlikely allophones in a given context, pruning is performed on the allophone distributions within a leaf. Pruning can be based upon counts of an allophone in a leaf or upon the percentage of exemplars of an allophone in a leaf. In the first case, allophones are removed if they are based on only a few exemplars. In the second case, unlikely allophones are removed. In each case, the arcs representing the removed allophones are not created. In addition to reducing the number of pronunciations, pruning may also result in more robust predictions. For example, in a given context, one may observe just a couple exemplars of an allophone in several hundred realizations. These allophones may be due to transcription error, and so it is judicious to remove them.

Summary

This paper describes a systematic, data-intensive method for creating pronunciation networks. A phoneme representation with an enriched set of contextual descriptors is advocated for providing a general framework in which new words may be easily added to the vocabulary. A wide variety of factors is used to model contextual effects, including long-range and cross word boundary phenomena. The large number of dimensions entailed by a greater number of contextual descriptors is handled through the use of context trees for predicting allophonic variation. Context trees were shown to possess attributes, such as the ability to estimate distributions from partial contexts and the capacity to systematically reduce the size of a network based on the tree data, that make the trees a good representation from which to create pronunciation networks.

References

F. Chen and J. Shrager, "Automatic discovery of contextual rules describing phonological variation," *Proc. DARPA Speech and Natural Language Workshop*, pp. 284-289, Feb. 1989.

Y. Chow, R. Schwartz, S. Roucos, O. Kimball, P. Price, R. Kubala, M. Dunham, M. Krasner, and J. Makhoul, "The role of word-dependent coarticulatory effects in a phoneme-based speech recognition system," *Proc. ICASSP*, pp. 1593-1596, 1986.

W. Fisher, V. Zue, J. Bernstein, D. Pallett, "An acoustic-phonetic data base," *J. Acoust. Soc. Am.*, Suppl. 1, vol. 81, 1987.

F. Jelinek and R. Mercer, "Interpolated estimation of markov source parameters from sparse data," *Proc. Pattern Recognition in Practice Workshop*, E. Gelsema and L. Kanal, eds., North-Holland, 1980.

L. Lamel, R. Kassel, S. Seneff, "Speech database development: design and analysis of the acoustic-phonetic corpus," *Proc. DARPA Speech Recognition Workshop*, L. Baumann, ed., pp. 100-109, 1986.

K.-F. Lee, *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System*, Doctoral Dissertation, Carnegie Mellon University, Pittsburgh, PA, April 1988.

K-F. Lee, H-W. Hon, M-Y. Hwang, S. Mahajan, R. Reddy, "The SPHINX speech recognition system," *Proc. ICASSP*, pp. 445-448, 1989.

D. Paul and E. Martin, "Speaker stress-resistant continuous speech recognizer," *Proc. ICASSP*, pp. 283-286, 1988.

D. Paul, "The Lincoln robust continuous speech recognizer," *Proc. ICASSP*, pp. 449-452, 1989.

S. Sagayama, "Phoneme environment clustering for speech recognition," *Proc. ICASSP*, pp.397-400, 1989.

M. Weintraub, H. Murveit, M. Cohen, P. Price, J. Bernstein, G. Baldwin, and D. Bell, "Linguistic constraints in hidden markov model based speech recognition," *Proc. ICASSP*, pp. 699-702, 1989.