Syllable Network for Phonemic Decoding of Speech

V. Gupta, M. Lennig, J. Marcus, and P. Mermelstein**
Bell-Northern Research
3 Place du Commerce
Nüns' Island, Montreal, Quebec
Canada H3E 1H6

The decoding of speech into phonemes for large vocabulary speech recognition is made more reliable by restricting phoneme sequences to those which compose valid syllables. To apply this restriction when decoding a sequence of phonemes, we use a syllable network representing the valid syllables in Webster's 7th Collegiate dictionary.

Since major allophonic variants of a phoneme are determined by the phoneme's position within the syllable (e.g., prevocalic vs. postvocalic /r/), the syllable network can be used to represent allophonic variation by employing distinct allophone models of a phoneme in different positions within the network. A preliminary experiment using the syllable network in large vocabulary recognition to select appropriate Markov models for allophones shows promising results.

1.0 Introduction

In this paper, we describe the use of a syllable network when decoding speech as a sequence of phonemes in large vocabulary speech recognition. Phonemic decoding of speech without any restriction on valid phoneme sequences leads to a large number of hypotheses which do not obey the phonotactic constraints of the language. We have used a syllable network to restrict the possible phoneme sequences to correspond to sequences of valid syllables. The syllable network also serves to control the choice of positional allophones. Allophonic variation is represented by using different Markov sources (Bahl et al., 1983) for a given phoneme depending upon its position within the syllable network.

2.0 Syllable Network

A syllable network for English which generates all and only the 8157 English syllables is necessarily complex. Such a network can be obtained by first constructing a tree of all possible syllables and then merging the tree from both ends. Simpler networks overgenerate the English syllabary. We have constructed a syllable network of intermediate complexity to achieve a compromise between network complexity and overgeneration.

The syllabic onset, nucleus, and coda are the subunits of the syllable within which the tightest phonotactic constraints obtain (Selkirk, 1982). Thus, our syllable network includes separate subnetworks for each of these three subunits. The syllable network generates phoneme sequences of the form

$$(O_1(O_2(O_3)))N(C_1(C_2(C_3(C_4))))$$

where O_i stands for a consonant in the syllabic onset, N for the vowel in the syllabic nucleus, and C_j for a consonant in the syllabic coda. The parentheses imply that the segment is optional. Only the nucleus is compulsory in the syllable. The subnetwork for the onset allows a maximum of three consonants, while that for the coda allows a maximum of four.

The syllable network was created based on the 60,000 phonemic transcriptions contained in Webster's 7th Collegiate dictionary (henceforth, *the dictionary*). Starting with a rudimentary network, branches were added iteratively to account for syllables in the dictionary not generated by the network. The resulting network has 76 nodes and over 300 branches.

The phonotactic constraints can be tightened further by using a separate syllable network for each syllable position within the word. The maximum number of syllables for any word in the dictionary is 10 (except for one word which was excluded). The number of valid syllables decreases with increasing syllable position number within the word (Table 1). Note that the set of syllables which occur in the first position includes all syllables which can occur in any position.

Syllable position in word	Number of distinct syllables
1st	8157
2nd	6181
3rd	3931
4th	1718
5th	724
6th	306
7th	110
8th	36
9th	12
10th	2

Table 1. Number of distinct syllables possible at each position within the English word.

3.0 Use of the Syllable Network to Select Allophones

Allophonic variants of a phoneme are often determined by the phoneme's position within the syllable (e.g., prevocalic, postvocalic, intracluster). For example, the phonemes /l r w/ differ significantly in their prevocalic and postvocalic realizations. First and second formant trajectories move upward in most contexts when these phonemes appear in prevocalic position, while the formant trajectories move downward when these phonemes appear in postvocalic position. By using separate Markov sources for allophones which differ in position, we can account for such variation.

In some cases, allophones are conditioned by a more detailed positional specification. For example, the allophones of the nasal consonants which occur in the syllable-initial clusters /sm/ and /sn/ are realized as partially devoiced with a very short nasal murmur. Also, devoiced allophones of the phonemes /w j r l/ occur when preceded by a voiceless fricative as in *switch, few, three,* and *slide.* Allophones which are difficult to account for with the syllable network

* Also with INRS-Télécommunications, University of Quebec.

are those which depend on larger contexts than the syllable. For example, [ɾ], the flapped allophone of /t/, occurs ambisyllabically after a stressed and before an unstressed vowel, as in *butter*, pronounced [bʌɾɚ].

4.0 Preliminary Recognition Results

In a series of speaker-dependent, isolated word recognition experiments using the syllable network, the unknown word is decoded as a sequence of syllables, where each syllable corresponds to a path through the syllable network. Each of the syllable network's transitions is mapped to a Markov source allophone model. In the experiments we report, we vary this mapping. First, all occurrences of a phoneme are represented by a single Markov source. Then, separate Markov sources are used to represent a given phoneme occurring in the syllabic onset and in the syllabic coda. We use statistical decoding to compute between 200 and 600 most likely syllable sequences corresponding to words in the 60,000-word dictionary. Since our system does not employ a language model, all 60,000 words are assigned equal a priori probability. Thus, the perplexity of this task is 60,000.

The training set consists of 800 word tokens from arbitrary texts, 60 distinct words chosen to contain consonant clusters, and 100 distinct CVC words, where C stands for a stop or a liquid, i.e., one of the consonants /p t k b d g r l/.

Two test sets were used (see Appendices). The first, denoted *Chrysler*, is a 99-word automobile advertisement. The second is a 100-word list of CVC words where C is a stop or a liquid, having no words in common with the CVC training list. 59% of the words in the Chrysler test set and 6% of the words in the CVC test set are represented in the vocabulary of the training set. Training and test sets are disjunct.

Two experimental conditions are compared:

(1) One Markov source (one allophone) for each of the 39 phonemes in the syllable network.

(2) Stops and liquids are represented by two allophones each. One Markov source is used in the syllabic onset, the other in the syllabic coda. Other phonemes are represented by one allophone each.

The recognition results in Table 2 show the percent correct recognition in the top *n* phonetic transcriptions, where *n* is either 1, 5, 20, or 100. Use of distinct allophones for the stops and liquids as they occur in the syllabic onset and coda improves the performance only for the CVC test set.

test set	condition	$n=1$	$n=5$	$n=20$	$n=100$
Chrysler	(1)	60%	81%	91%	94%
	(2)	62%	81%	89%	94%
CVC	(1)	15%	36%	54%	67%
	(2)	21%	56%	77%	88%

Table 2. Percent correct recognition in top *n* choices.

5.0 Conclusions

The syllable network provides a convenient framework for the selection of different allophonic models depending upon a phoneme's position within the syllable. Separate allophones of stops and liquids for the syllabic onset and coda lead to a significant improvement in recognition of CVC words. The fact that no significant improvement is observed in recognition of arbitrary text suggests that a more general representation of allophonic variation in the multisyllabic environment and more complete training appropriate to that environment are necessary.

6.0 References

Bahl, L.R., Jelinek, F., and Mercer, R.L. (1983). "A Maximum Likelihood Approach to Continuous Speech Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI–5(2), 179–190.

Selkirk, E.O. (1982). "The Syllable," *The Structure of Phonological Representations*, (Foris Publications, Dordrecht, Holland), 336–383.

Appendix: Chrysler Test Set

begin paragraph here is the confidence of front hyphen wheel drive comma the security of advanced electronics and the quiet comma smooth ride you expect in a fine luxury car period begin paragraph and here are the luxuries you demand period automatic transmission comma power windows comma power steering comma power brakes comma power remote mirrors and individual reclining seats standard period begin paragraph and finally comma here is the new technology of turbo-power period more power to move you period to accelerate period to pass period to cruise in serene comfort ellipsis yet with remarkable fuel efficiency period.

Appendix: CVC Test Set

but could back write put god book cut dead pull bed role top bad deal date doubt care look rock lip tool lack pair tear cup pale load pour dare dear kick tip leap cop lobe rob rub cab tub gale gag tag pig log bog rogue gab goat guile ball lower bit roll bird beat cool tall root coat rout luck core cat rare tale Paul coal pike beer pot peer tail cape robe lab goad dug gape tug dip rot rat cot cod tight tide tuck tack lull roar lure rope ripe reap rip pile tile curd pearl.