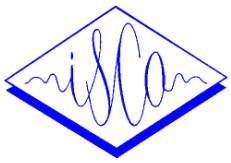

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

An Information Theoretic Approach to the Study of Phoneme Collocational Constraints[1]

Rob Kassel and Victor W. Zue

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 U.S.A.

ABSTRACT

In this paper we describe a lexical study of phoneme collocational constraints using a metric motivated by information theory. We used a pairwise, hierarchical clustering technique to combine phonemes into classes using a normalized measure of mutual information. The result of this clustering process can be displayed as a dendrogram, from which one may select an arbitrary number of equivalence classes. We have conducted a number of experiments investigating phoneme collocational constraints within pairs and triplets. In many cases we found that phonemes are organized into classes that share certain phonological features. In fact, phonemes that have similar *acoustic* properties often exhibit similar collocational constraints. We also compared the constraining power of our phoneme classes with those chosen by a phonological criterion, and found ours to be more than competitive. Based on our results, we conclude that our information theoretic metric is particularly useful as a description of lexical constraining power.

INTRODUCTION

Spoken language is limited not only by the inventory of sounds that a speaker can utilize but also by the permissible combinations of these sounds to form words. These constraints are usually localized, and phonologists describe the allowable combinations of phonemes using fairly explicit phonotactic rules. For example, the homorganic rule for American English states that the phonemes in a syllable-final nasal-stop sequence must agree on their place of articulation. Thus sequences like /send/ and /θɱk/ are permissible, whereas /linp/ and /fæmg/ are not. Such phonotactic, or collocational, constraints often are discernible when one contrasts a phoneme against another based on the statistical distributions of nearby phonemes. For example, Figure 1 displays the distributions of the phonemes that *follow* /n/ and /ŋ/. For the nasal-stop combinations, there is a predominance of /t/ and /d/ following /n/ and of /k/ and /g/ following /ŋ/. The above example also illustrates two important facts regarding phonotactic constraints. First, these rules are statistical in nature; rarely can we apply these rules universally. Second, collocational constraints often are expressed in terms of phoneme equivalence classes such as nasals, stops, and the like.

Phonotactic constraints are important for automatic speech recognition, since proper utilization of these constraints can help correct phonetic analysis errors while improving lexical access efficiency. Over the past decade there have been many lexical studies examining the constraining power of phoneme equivalence classes, demonstrating their utility for lexical access. Shipman and Zue [1] performed studies using a 20,000 word lexicon, showing that even broadly characterizing phonemes into six manner classes can provide substantial constraints for lexical access of isolated words. They used the average cohort size, defined as the average number of words that share a broad-class pattern, to quantify the constraints on allowable phoneme sequences. Huttenlocher [2] refined the study by incorporating a better metric, the expected value of the cohort size. Carter [3] subsequently suggested using word entropy over a lexicon's broad-class cohorts as a measure of the constraining power. Vernooij et al. [4] further refined this work by determining the best intermediate classes between five broad classes and the phonemes for distinguishing a specific word from its cohort's members. To do so, they use a constrained clustering technique with a metric based on the number of words uniquely identified by a proposed class set.

While there are good reasons to express these constraints using classes well-motivated by theory, as is the case with all the above studies, the phoneme space can be partitioned in many other ways. By allowing phonemes to form various sets of equivalence classes and quantifying the constraining power for each set, we can discover phoneme classes that provide the strongest constraints for lexical access.

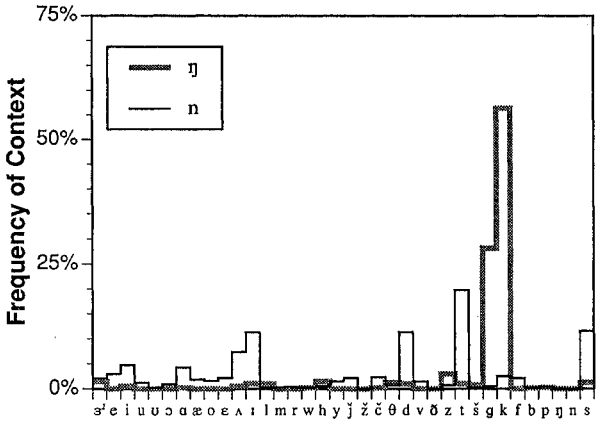

Figure 1: The distributions of phonemes following /n/ and /ŋ/, illustrating the homorganic rule. The data comes from the Merriam-Webster Pocket Dictionary containing nearly 20,000 words.

[1]This research was supported by DARPA under Contract N00014-82-K-0727, monitored through the Office of Naval Research.

This paper describes a lexical study of phoneme collocational constraints using a metric motivated by information theory. The inspiration for our investigation comes from recent work on discovering word equivalence classes by Jelinek [5] and word collocational constraints by Church [6]. Specifically, we developed a procedure that will enable us to partition the phoneme space using a normalized measure of mutual information. By using a hierarchical clustering algorithm, we may select an arbitrary number of equivalence classes. We also applied this metric to produce a rank ordering of the distinctive features [7], based on the constraining power that they offer. Finally, we used two different metrics to compare several different ways of forming equivalence classes.

METHODOLOGY

Let us assume that we have a lexicon containing words represented as a sequence of phonemes. We ask ourselves the following question: how can we group the phonemes into mutually exclusive equivalence classes such that the resulting lexical representations will provide the maximum constraints for lexical access? The solution to this problem must include a specification of the number and content of these equivalence classes, as well as the metric used for comparison.

Our approach is to use a pairwise, agglomerative, hierarchical clustering technique [8] to form phoneme classes. Initially, there are as many classes as there are phonemes, N. We combine classes, one pair at a time, to form $\binom{N}{2}$ sets of $N-1$ equivalence classes. Using a metric based on average mutual information we select the best of these formations. We repeat this process, each time reducing the number of equivalence classes by one. The process halts when a single class represents all of the phonemes. By retaining the resulting hierarchy, one can choose among many partitionings of the phoneme space.

Metric

For the sake of simplicity, we consider here only the constraints of one phoneme on its neighbor. The metric easily can be extended to more complicated cases, in which the context is more extensive.

Let us assume that we are interested in measuring the amount of dependency between phoneme x_i and x_{i+1}.[2] This can be quantified by the average mutual information:

$$I(X_i; X_{i+1}) = \sum_{X_i} \sum_{X_{i+1}} P(x_i, x_{i+1}) \log_2 \frac{P(x_i, x_{i+1})}{P(x_i)\,P(x_{i+1})}$$

We define a function Φ which maps a phoneme into its class. As we construct our hierarchy, i.e., define new classes, we change this function accordingly. We must apply this function to one or more of the phoneme positions used in our mutual information measure. For example, we denote measuring the mutual information between a phoneme's class and the following class by:

$$I(\Phi(X_i); \Phi(X_{i+1})).$$

In order to make it easier to compare results across data sets, we normalize the average mutual information into a dimensionless measure, known as the percentage of information extracted (PIE). Continuing the above example, we define PIE as:

$$\text{PIE} = \frac{I(\Phi(X_i); \Phi(X_{i+1}))}{I(X_i; X_{i+1})}$$

PIE is a ratio of the average mutual information before and after mapping a lexicon's phonemes into classes. At best this measure will be 100% since applying the map cannot extract additional information. Mapping all characters into a single class results in a PIE of 0%.

Lexicon

We based our studies on the Merriam-Webster Pocket Dictionary (MPD) containing approximately 20,000 words. Pronunciations in the lexicon have been refined by several MIT researchers over the past decade.

One of the goals of our study is to compare equivalence classes generated by self-organizing techniques to those motivated by the distinctive feature theory. To facilitate such a comparison, we make minor modifications to MPD such that each phoneme can be represented as a single feature bundle. Specifically, we applied the following set of rewrite rules to MPD:

- Schwas: /ə/→/ʌ/, /ɚ/→/ɝ/, /ɚ/→/ɝ/
- Syllabics: /l̩/→/ʌl/, /m̩/→/ʌm/, /n̩/→/ʌn/, /ŋ̩/→/ʌŋ/
- Diphthongs: /ɔʸ/→/ɔi/, /ɑʸ/→/ɑi/, /ɑʷ/→/ɑu/

The diphthongized vowels /iʸ/, /eʸ/, and /oʷ/ are represented by their monophthong counterparts, /i/, /e/, and /o/. We also remove all syllable and stress markers from the pronunciations.

RESULTS

In this section, we will present some results of our experiments. Due to space limitations, we can only highlight our findings with a typical example. We refer interested readers to Kassel [9] for more extensive descriptions. We will illustrate our self-organizing procedure for discovering phoneme equivalence classes by using the clustering metric based on the classes of a phoneme and its left and right neighbors. Our choice is motivated partly by the popular use of triphones, defined as a phoneme and its immediate context, as a unit for acoustic modeling in the speech recognition community.

Phoneme Clusters

The hierarchical clustering procedure iteratively constructs coarser phoneme equivalence classes, computing the resulting PIE at each stage. We can display such a hierarchy in the form of a dendrogram, as shown in Figure 2. The horizontal axis of the dendrogram lists all the phonemes, while the vertical axis of the dendrogram denotes the PIE level at which two classes merge. The higher the walls surrounding a cluster, the more robust the cluster is.

Perhaps the most striking aspect of this hierarchy is that the vowels and consonants are segregated completely from one another. This single distinction provides a large amount, roughly

[2] The subscripts denote the relative positions of phonemes within a word.

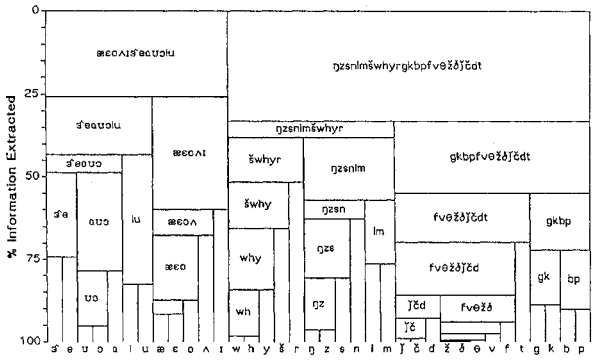

Figure 2: A dendrogram showing the phoneme class hierarchy derived from the Merriam-Webster Pocket Dictionary using the collocational constraints between a phoneme's class and its neighbors' classes.

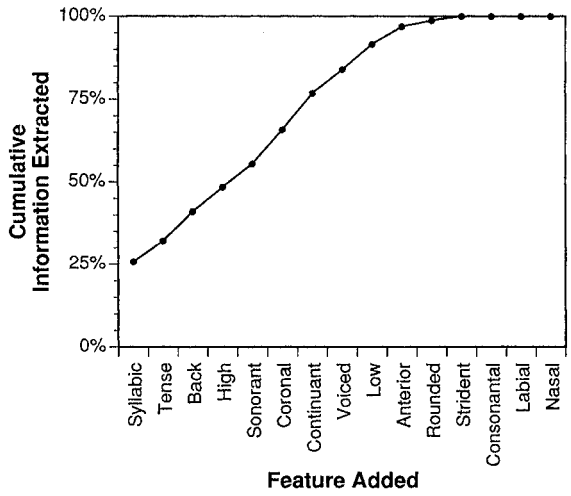

Figure 3: A feature ordering derived by a step-wise maximization of the information extracted from the Merriam-Webster Pocket Dictionary. The information measure is the same as that used in Figure 2.

25%, of the available information. At finer levels we see that the phonemes often are organized into linguistically relevant classes. On the vowel side, for example, /i/ and /u/ are high and tense, while /ʊ/ and /ɔ/ are back and rounded. We observe some similar organization within the consonants. With few exceptions, the consonants are divided roughly into four classes: stops, fricatives, nasals and semivowels. Many of the classes are similar acoustically even though no acoustic information is available to the clustering process. We suggest that this may be viewed as potential evidence of language's acoustic and contextual constraints evolving simultaneously.

Some of the clusters shown in Figure 2 do not conform to our linguistic intuition. Closer examination of the data reveals that these anomalies often can be attributed to two causes. First, minor differences in PIE for the cluster candidates can often change the structure of the dendrogram. This is particularly true for the initial clusters, some of which are not very robust. Use of a more sophisticated clustering technique might overcome this problem. Second, some of the peculiar clusters may be artifacts of the lexicon, which may contain a large number of regular prefixes and suffixes. A larger and more balanced lexicon may help to alleviate this problem.

Cluster Evaluation

One novel aspect of our procedure is that the number of equivalence classes is variable. We can choose any number of classes between 1 and N, the number of phonemes. For a given phoneme-class hierarchy, there are many ways we can select a particular number of equivalence classes. One possibility would be to select a particular set of equivalence classes by "slicing" the dendrogram at the appropriate PIE level. Thus, for example, we can obtain six equivalence classes if we choose to slice through the dendrogram at PIE=45% in Figure 2.

We compare our classes with four other sets. The first set is the six manner classes proposed by Shipman and Zue and studied by others [1,2,3,4]. The second set is a phoneme partitioning based on distinctive features [10]. The third set is a phoneme hierarchy obtained by measuring *acoustic* similarities of the phonemes [11]. The final set is formed by randomly clustering the phonemes into a hierarchy. We establish a baseline performance by computing 1000 such hierarchies and averaging their performance.

Before we can form a varying number of phoneme classes based on distinctive features, we must first determine an ordering of the features themselves. There are many ways one can establish such a feature hierarchy. We have chosen to do so by selecting the single feature which carries the most information according to our aforementioned metric. We then select the feature which, when used in addition to the first feature, provides the most information. By applying this procedure iteratively, we produce a ranking as shown in Figure 3. From the standpoint of lexical access, the features [Consonantal], [Labial], and [Nasal] are superfluous within the ranking we adopted.

We first evaluated these five phoneme partitioning methods using the expected cohort size as the metric. The results are shown in the upper graph of Figure 4. As the number of classes increases, the expected cohort size decreases. This is predictable, since a larger number of cohorts implies an improved ability to perform phoneme distinctions and so less lexical ambiguity. Using the six manner classes offers little advantage over other schemes. Surprisingly, the baseline classes obtained through random clustering generally perform *better* than the other sets of phoneme classes. We take this result as an indication that expected cohort size is a poor metric for evaluating the lexical constraining power of various phoneme-partitioning schemes.

As an alternative, we applied the mutual information metric described earlier as a measure of the lexical constraint provided by each of the five class sets. The results are shown in the lower graph of Figure 4. As expected, the classes formed through our clustering procedure extract information the fastest, thereby providing the most constraint for a given number of classes. Furthermore, the baseline classes extract information the slowest, ranking them below all others. The remaining classification schemes

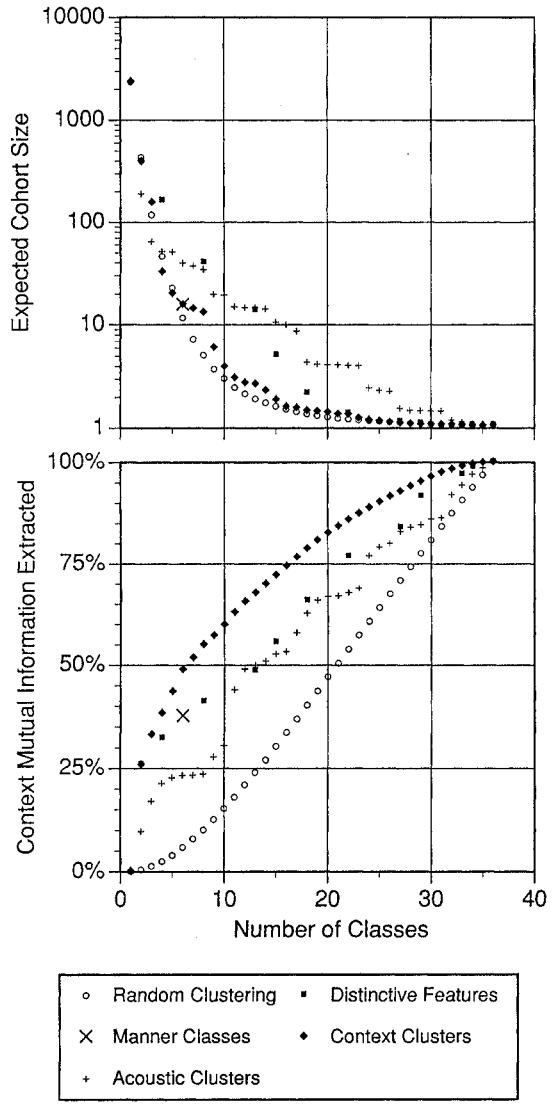

Figure 4: The lexical access constraining power of several phoneme class sets as measured by expected cohort size (top) and a mutual information measure (bottom).

fall between these two. The six manner classes perform as well as the distinctive features. The acoustic classes provide fewer lexical constraints, but its performance improves to match the features as we expand the number of classes used.

Note that there is a substantial difference in the amount of information extracted using two classes selected from the acoustic and context clusters. Closer examination of the two hierarchies indicates that the only difference between them is the way semivowels are grouped with other phonemes. The semivowels are more similar to vowels acoustically, but are more similar to consonants based on collocational constraints.

SUMMARY

In this paper we described a lexical study of phoneme collocational constraints using a metric based on normalized mutual information. Our procedure is different from that of previous lexical studies in that we make no use of preconceived notions regarding the number or the make-up of the equivalence classes. A comparison among various phoneme partitioning schemes led to the conclusion that expected cohort size may not be an appropriate metric.

While classes derived from distinctive features did not perform as well, one must keep in mind that we have established the feature hierarchy in the simplest way possible. With a feature hierarchy more motivated by linguistic theory, it is conceivable that the features can provide more constraint than we have been able to determine.

Finally, we have thus far motivated these classes strictly from the standpoint of lexical constraints. To fully assess their utility for speech recognition, we must also investigate how to detect a set of classes from the speech signal reliably.

REFERENCES

[1] Shipman, D.S. and V.W. Zue, "Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1982, pp. 546–549.

[2] Huttenlocher, D.P., "Acoustic-Phonetic and Lexical Constraints in Word Recognition: Lexical Access Using Partial Information," S.M. Thesis, Dept. Electrical Engineering and Computer Science, Massachusetts Institute of Technology, May, 1984.

[3] Carter, D.H., "An Information-Theoretic Analysis of Phonetic Dictionary Access," *Computer Speech and Language,*, 1987, pp. 1–11.

[4] Vernooij, G.J., G. Bloothooft, and Y. van Holsteijn, "A Simulation Study on the Usefulness of Broad Phonetic Classification in Automatic Speech Recognition," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1989, pp. 85–87.

[5] Jelinek, F., "Self-Organized Language Modeling for Speech Recognition," in (Eds. Waibel and Lee) *Readings in Speech Recognition*, Morgan Kaufmann, 1990, 450–506

[6] Church, K., W. Gale, P. Hanks, and D. Hindle, "Parsing, Word Associations and Typical Predicate-Argument Relations," in *Proc. DARPA Speech and Natural Language Workshop*, October, 1989, 75–81.

[7] Chomsky, N. and M. Halle, *Sound Patterns of English*, Harper & Row, 1968.

[8] Duda, R. O. and P.E. Hart, *Pattern Recognition and Scene Analysis*, John Wiley and Sons, 1973.

[9] Kassel, R., "An Information-Theoretic Approach to Studying Phoneme Collocational Constraints," S.M. Thesis, Dept. Electrical Engineering and Computer Science, Massachusetts Institute of Technology, May, 1990.

[10] Stevens, K., "An Approach to Lexical Access Based on Distinctive Features," *Proc. Second Symposium on Advanced Man Machine Interface Through Spoken Language*, November 1988, pp. 10:1–10:23.

[11] Glass, J.R., "Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition," Ph.D. Thesis, Dept. Electrical Engineering and Computer Science, Massachusetts Institute of Technology, May, 1988.