ON THE ROBUSTNESS OF PHONETIC INFORMATION IN SHORT-TIME SPEECH SPECTRA

Meg Withgott and Marcia A. Bush[1]

Stanford University, Center for the Study of Language and Information, Stanford, California 94305, USA

Schlumberger Palo Alto Research, 3340 Hillview Avenue, Palo Alto, California 94304, USA

Abstract: Speech recognition techniques which take fixed-time slices as input to a matcher face the task of mapping from arbitrary pieces of the physical signal to abstract linguistic units. This paper examines the reliability with which individual vector-quantized LPC spectra can be mapped to various sets of acoustic-phonetic classes. The database for the experiments consisted of approximately 130,000 spectra from a pre-labeled corpus of 616 5-digit strings, and classification was performed on the basis of a maximum likelihood decision rule. Classification accuracy, when the same database was used for training and testing, ranged from 94.0% for a simple voiced-voiceless distinction to 42.7% for a set of 45 acoustic-phonetic classes used in earlier connected digit recognition experiments [1,2].

Introduction

It is commonly accepted that the variability inherent in speech makes it difficult to recognize linguistic units such as allophones directly from sequences of short-time spectra. This observation has, in part, motivated work on broad phonetic classification schemes, in which an initial labeling of the recognition vocabulary is made on the basis of presumably robust acoustic-phonetic categories which then is used to identify subsets of the vocabulary for more detailed acoustic processing. Studies have shown that, for instance, a coarse-grained classification based on manner of articulation reduces a 20,000-item wordlist into approximately 100 phonetic cohorts (i.e., wordlist sublists) [3]. Relatively little quantitative data are available, however, to determine whether classification strategies designed and tested on the basis of abstract phonetic or phonemic considerations are actually useful in labeling large corpora of speech signals. Similarly, little is known about trade-offs between classification accuracy and the granularity of the labeling scheme.

This paper examines the reliability with which individual vector-quantized LPC spectra can be mapped to three types of acoustic-phonetic classes: one based on manner of articulation; a second based on multidimensional distinctive features (see e.g. [4]); and a third "system-specific" type influenced both by knowledge of the classifier's front end and of acoustic characteristics of individual classes in the recognition vocabulary.

Procedure

The database for the experiments consisted of 129,812 spectra from a pre-labeled corpus of 616 5-digit strings. The connected-speech utterances were spoken by 56 adult talkers (27M, 29F) from 22 geographically defined dialect groups, and form a subset of the training portion of Texas Instruments' connected digits database [5]. The initial label set comprised 45 acoustic-phonetic classes used in earlier connected digit recognition experiments [1,2]. Labeling was done primarily by hand, with simple durational rules for automatically dividing diphthongs and certain sonorant and word-boundary regions.

Signal preprocessing consisted of digital downsampling of the TI data from 20 KHz to 8 KHz (i.e., a 4 KHz bandwidth) and preemphasis by first-differencing. Short-time spectra were computed using an 11-pole LPC analysis, with a 25.6 msec Hamming widow and a 10 msec frame rate, and were vector quantized to a size 1024 codebook.

Classification of spectra was performed using a maximum-likelihood decision rule and, in these preliminary experiments, the same database was used for training and testing.

Classification Schemes

As noted above, three classification schemes were examined. Each involved grouping the initial 45-label set into smaller numbers of acoustic-phonetic categories. The grouping was complicated slightly by the fact that the initial labeling of the data was partially automated and thus not completely phonemic (e.g., glides typically included a short portion of the adjacent vowel). Such phenomena were uniform, however, across the three classification schemes.

With respect to the first classification, based on manner of articulation, label sets of size 4 (silence, fricative, nasal, vowel) and 6 (silence, weak fricative, strong fricative, nasal, glide and vowel) were used.

The second, multidimensional classification employed diverse distinctive features so that a given label represents a vector of cross-classified values. In contrast, manner forms a unidimensional classification. Figure 1 shows a distinctive feature tree corresponding to the complete [-sonorant] subset of the distinctive feature categories. Such trees yield relatively coarse-grained classes at the top nodes and finer-grained classes as the tree is descended. A binary partitioning of the initial label set led to the [+/-sonorant] distinction.

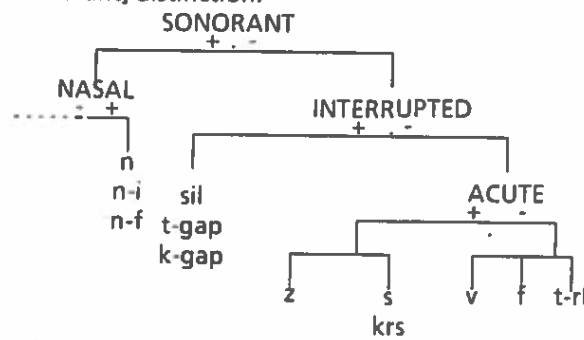

Figure 1: Distinctive Feature Tree for "consonants"

A partial tree for the third scheme, which is system-specific and multidimensional, is shown in Figure

2. As noted above, this classification strategy takes into account both characteristics of front-end processing and acoustic characteristics of individual acoustic-phonetic classes in the recognition vocabulary. For example,

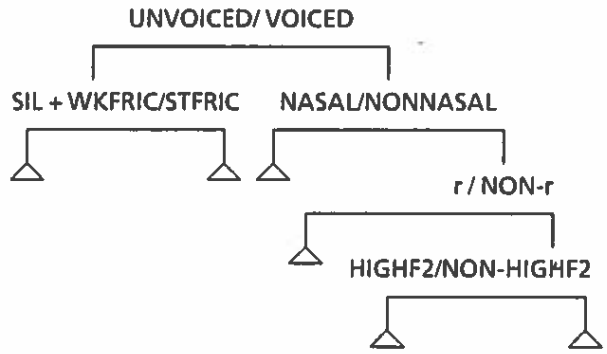

Figure 2: Partial Tree for the system-specific classification

weak fricatives and silent intervals are collapsed into a single class because they are difficult to discriminate on the basis of LPC spectra alone. On the other hand, the release portions of the [t]'s in the digits 2 and 8 are classified as strong and weak fricatives, respectively, on the basis of context-dependent acoustic manifestations.

Results and Discussion

Figure 3 shows overall classification accuracy (i.e., the percentage of short-time spectra correctly classified) as a function of number of acoustic-phonetic categories for the three classification schemes. Percentages are similar across the classification schemes when small numbers of categories are used. (For the purpose of comparison, a fourth classification with arbitrary six-way partitions was created and found to exhibit classification accuracy of 48.4%).

Number of Categories	manner	multiple features	task-specific
2		93.5	94.0
4	84.6	84.6	87.0
6	79.0	73.7	79.2
10		67.4	73.5
21			64.3
45			42.7

Figure 3:. Overall classification acuracy (percent correct) versus number of acoustic-phonetic categories for the three classification schemes.

An advantage of multidimensional classifications, such as the feature-based and system-specific classifications, as opposed to a unidimensional classification such as manner, is that they support a selective traversal down one or more branches of a classification tree. The choice of whether to collapse or differentiate categories can therefore be determined on the basis of the lexicon, or the discriminability of individual classes.

Figure 4 shows overall classification accuracy as a function of the branch traversed for the system-specific scheme, and shows, for example, that a 9-way classification determined by a broad unvoiced class being more finely-differentiated was equal to the performance of a 6-way classification when the voiced branch was descended. The same advantage does not

Number of Categories	branch traversal	
	unvoiced	voiced
3	89.0	92.0
4	84.6	88.8
6		82.3
9	83.0	
10		74.2

Figure 4. Overall classification accuracy (percent correct) for system-specific scheme as a function of the branch traversed.

show up in a 3-way or 4-way comparison, and thus classification accuracy depends both on how categories are sub-divided and on how many sub-divisions are formed. We are also able to note that combining categories representing relatively broad classes with categories containing a single segment type which proves to be highly discriminable in the vocabulary of interest (e.g., the early vocalic region in 4 (AOR1) in this database) can be advantageous.

Summary

Multidimensionality appears to be a desirable trait of classification systems for applications in automatic speech recognition. This is because the identity and grain-size of the classes can be determined freely both by what features are the most useful for discriminating lexical items, and by what classes prove to be the least confusable for a particular classifier.

1. After Aug 86: Division of Engineering, Box D, Brown University, Providence, RI 02912, USA

References

[1] Bush,M. 'Durational constraints for network-based connected digit recognition,' (This volume).

[2] Bush,M., and G. Kopec. 'Network-based connected digit recognition,' submitted to *IEEE Transactions on Acoustics, Speech and Signal Processing*, March 1986.

[3] Shipman,D.W. and V.W. Zue. 'Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems,' *1982 IEEE International Conference on Acoustics, Speech and Signal Processing*, Paris, France.

[4] Fant, G. *Speech Sounds and Features*. MIT Press, 1973.

[5] Leonard, G. 'A database for speaker- independent digit recognition,' *1984 IEEE International Conference on Acoustics, Speech and Signal Processing*, San Diego, CA.