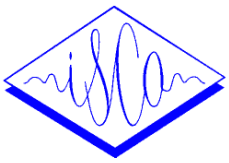

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

# TOWARDS THE PERFORMANCE LIMITS OF CONNECTIONIST FEATURE DETECTORS

Gary Bradshaw and Alan Bell

Beckman Institute, University of Illinois, Urbana, Ill. 61801
Department of Linguistics, University of Colorado, Boulder, Colo. 80309

## ABSTRACT

Attempts to improve the performance of a connectionist network trained to detect selected phonetic features in multispeaker connected speech indicate some of the limitations on the information available at a peripheral level of speech analysis. The three-layer feedforward network has 12 detector outputs, and is trained over large subsets of sentences from the MIT Ice Cream database. Its input consists of smoothed spectral vectors sampled at 15 msec intervals. Little contextual information is available to the detectors, since each vector has an effective window of about 30 msec. Overall, the detector network generalizes very well to new speakers and new sentences: average a' drops only to .93 on test data from .94 on training data. Frequently occurring features like sonorance are better discriminated than infrequent ones like rhotic, mainly because the learning algorithm gives greater weight to the many negative training instances than to the few positive ones; discrimination is improved by weighting the learning rate for positive and negative vectors in inverse proportion to their frequency of occurrence. Performance was also improved modestly by adding preceding and following vectors to the input. Several other modifications yielded little or no performance improvement.

## INTRODUCTION

One critical problem in automatic speech recognition is to recode the complex acoustic signal onto linguistically relevant dimensions. We have been developing a general feature detector system to accomplish this recoding. The system uses a neural network to map a short segment of speech onto a set of corresponding linguistic features.

## NETWORK STRUCTURE AND DATABASE

Our current connectionist network consists of an input layer of 18 nodes, a hidden layer of 30 nodes, and an output layer of 12 nodes, one for each feature. (We have discussed networks with slightly different sizes of layers elsewhere [1,2]; such differences are not important for the results reported here.) Activation propagates through the network in a feedforward fashion. Each node of a layer projects to all nodes in the next layer. As the network is trained, errors in the output nodes are used to adjust the weights of the connections by a backpropagation algorithm.

### Training and Test Databases

We used large subsets of the MIT Ice Cream Database to train the network and to test its performance. The network was trained over 240 utterances of 120 distinct sentences, each of 40 men and 40 women pronouncing three sentences each. The network was tested on a disjoint subset of sentences and speakers, consisting of 200 utterances of 100 distinct sentences, each of 10 men and 10 women pronouncing 10 sentences each. Acoustic segments of the sentences are labelled according to a system similar to the one used in the TIMIT database; a certain amount of sub-phonemic detail is included, such as stop closures, stop releases, taps, and some voiceless vowels.

### Input to the Network

The input to the network was a short-term spectral vector representing the acoustic properties of a brief period of speech. The spectral vector was derived by various transformations from 128-point FFT of a 16 msec sample of speech convolved with a Hamming window. Separate transforms were computed at 3 msec intervals. The end result was 17 spectral values spaced at one-third Bark intervals up to 6.4 kHz. The values were smoothed over time and frequency by a Gaussian filter with a standard deviation of 9 msec in time and a standard deviation of two Barks in frequency [3]. These values were differenced over frequency to normalize for overall intensity. In addition to the 16 spectral values left after differencing, the input vector included coefficients of amplitude change and of spectral change derived from adjacent spectral vectors.

The network was trained and tested on vectors taken at 15 msec intervals from the database sentences. The feature detection system thus operates essentially without temporal context other the minimal amount provided by the smoothing process, i.e an effective window of approximately 30 msec for each vector. The training set consisted of 34,279 vectors over 240 sentences. Output errors were determined for each input, then an average error was determined for all training inputs. Weight adjustments were made at the end of each epoch. The test set was made up of 28,359 vectors over 200 sentences.

### Output features

For this research we did not pretend that we would be able to anticipate what the best choice of features would be in a particular application. Our choice was rather designed more to inform us about the performance of connectionist detectors over a range of possible features. In keeping with the restricted temporal context in the input, and our conception that feature detectors would be followed by other higher-level modules in recognition systems, we stuck to features with a fairly clear local acoustic content. Given these guidelines, we then chose some very broad features, like sonorance, common to many segment classes, and some narrow features, like high tonality, found in much smaller classes of segments. We also chose features that we thought might be relatively easy to detect with temporally independent inputs, like sonorance and sibilance, as well as others that we thought would be more difficult, like frication. The labelling conventions of the database also limited our choice to some extent. While some of the labels represent low-level phonetic events, e.g. the distinction made between a stop closure and a stop release, others represented more abstract phonological events, e.g. voiced and voiceless fricatives, either of which in English may or may not contain a voicing pulse. Thus voicing was not included in our output features.

The 12 output features, with a brief characterization of the categories that they represent, are listed in Table 1. In some cases details of the feature definition may not be obvious, e.g. silence not only includes onset, offset, and

10.21437/ICSLP.1992-116

pausal segments without speech energy, but also stop closures with and without a voicing pulse.

*Table 1. Output features*

| | |
|---|---|
| silence | sentence onset/offset, pause, stop closures |
| frication | fricatives, stop bursts, aspiration, devoiced vowels and resonants, h |
| sibilance | s, sh, z , zh, fricative parts of ch and j |
| nonsib. fric. | other frication |
| sonorance | vowels, nasals, laterals, semivowels |
| resonant | nasals and laterals |
| vocalic | vowels and semivowels |
| high | \| |
| low | \| |
| front | tonality features of vocalic inputs |
| back | \| |
| rhotic | \| |

## MEASURES OF PERFORMANCE

Finding appropriate measures of performance is a crucial part of system module design. Overall system accuracy is unlikely to be the best measure of the effects of changes in a module's design; interactions with other parts of the system may mask changes in the module's performance. Instead of a system measure like word recognition accuracy, we could have used an analogous module measure of feature categorization accuracy. This has been the approach of some other feature detection studies [4]. Such measures are inappropriate as primary performance measures, because feature detectors are fundamentally discriminators, not categorizers.

### Discrimination versus Categorization

The output of each feature detector is a value between 0 and 1. In Figure 1 we can see for the feature vocalic, how the network learns to assign higher output values to the

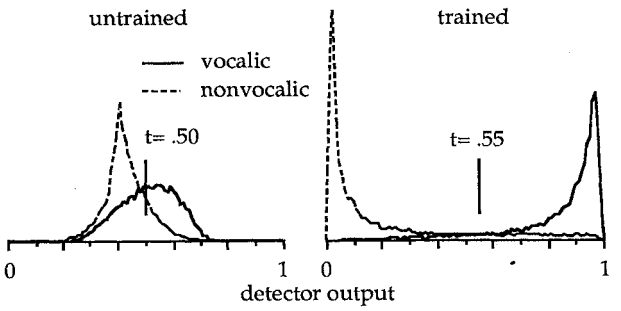


*Figure 1. Outputs of the untrained and trained vocalic detectors. The dashed curves show the probability distribution of outputs for nonvocalic inputs; the solid curves, for vocalic inputs. The vertical bars show the thresholds for maximum accuracy.*

vocalic inputs than to the nonvocalic inputs. The values themselves do not yield a decision whether the feature was present or absent in an input. Such a decision can of course be obtained by adding a decision procedure, e.g. if the output is greater than a threshold of 0.5, assign the feature to the input. One could then measure the performance improvement by calculating and comparing the accuracy of such decisions. Still further procedures are required to determine what category an input belongs to. For example, a fricative is not only characterized by the presence of the feature frication, but by the absence of the features sonorance and silence. The entire vector of feature outputs is in principle relevant to categorization, and there are a wide variety of possible categorization procedures. The point is that **any** accuracy measure, whether of feature presence or of the more complex phonetic categorization, depends on particular assumptions about procedures which are not part of the feature detection process itself.

It would nevertheless be worth using accuracy measures derived from simple categorization procedures if they appropriately represented detector performance. In a number of circumstances, unfortunately, they do not. Compare the performance of the sibilance detector in Figure 2 with that of the vocalic detector in Figure 1.

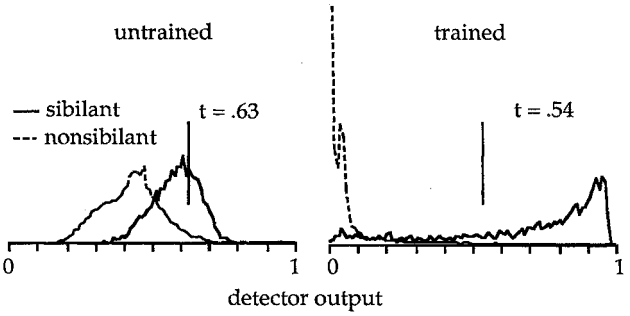


*Figure 2. Outputs of the untrained and trained sibilant detectors. The dashed curves show the probability distribution of outputs for nonsibilant inputs; the solid curves, for sibilant inputs. The vertical bars show the thresholds for maximum accuracy.*

The untrained response distributions for the two features are similar; there is some separation of the target and nontarget inputs, but still considerable overlap. We get very different accuracies, however, if we use the optimum thresholds to decide whether the feature is present. The categorization accuracy of the untrained vocalic detector is 0.71, but that of the untrained sibilant detector is 0.90. The sibilant accuracy is so high because only 11 percent of the test set are sibilant vectors. Since an accuracy of 0.89 can be attained by simply saying all inputs are nonsibilants, high thresholds that exclude most nonsibilants will always give accuracies this high. For vocalic inputs, on the other hand, which make up 47 percent of the test set, this strategy is not available. We obviously need performance measures that do not depend on the proportions of target and nontarget inputs.

### Signal Detection Measures

Its concern with discrimination between signal and noise distributions makes signal detection theory a useful framework for evaluating feature detection performance [5]. Essentially, we would like to have measures of the separation of a detector's target and nontarget distributions.

*Table 2. Performance measures of the vocalic and sibilant detectors in Figure 1 and Figure 2. p is the accuracy obtained by categorization at the optimum threshold.*

| | vocalic | | sibilant | |
|---|---|---|---|---|
| | untrained | trained | untrained | trained |
| a' | .74 | .96 | .88 | .98 |
| g' | .54 | .82 | .58 | .82 |
| p | .71 | .90 | .90 | .96 |

One such discrimination measure is denoted by a'. Let it suffice for brevity's sake to say that a' is a nonparametric index ranging from 0 to 1, with chance performance at 0.5. Another measure we have found valuable is g', the area between the cumulative distribution of nontarget outputs

and the cumulative distribution of the target outputs, normalized to the same range as a'. These measures and the optimum categorization accuracy p are compared in Table 2 for the detectors in Figures 1 and 2. While it is important to use an array of both discrimination and categorization measures to assess performance, we will focus on comparisons of a' and g' in this report.

## BASELINE PERFORMANCE

The performance of the 12 detectors is presented in Figure 3, based on the discrimination of the detectors on test data after 5000 epochs of training. Average a' of the detectors is .93; average g' is .75; average p is .92. An important characteristic of the network is that it generalizes very well to new speakers and sentences: average a' over training data is .94, only 0.01 greater than the test average. Notice that the broadest features on the left in Figure 3 all perform very well. In the next group of features, sibilant and vocalic perform well, but resonant and nonsibilant frication more poorly. The narrowest features, the vocalic tonality features on the right, are as a group the weakest.

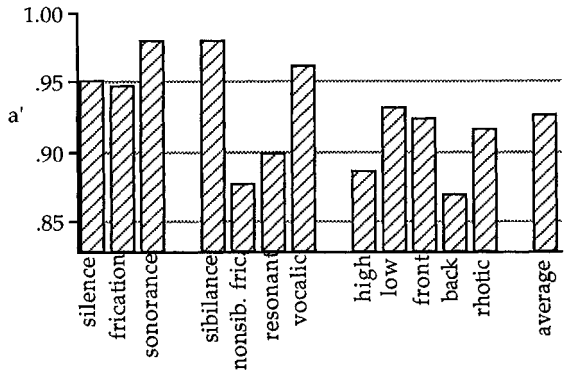


*Figure 3. Baseline discrimination performance of the feature detectors. The average a' is .93.*

## RESULTS OF DESIGN PARAMETER CHANGES

Both to understand why some detectors perform better than others and to discover what limits their performance, we have begun to explore the feature detector design space systematically. After a fair amount of experience, we have been surprised that it yields improvements so grudgingly.

**Training weights of positive and negative instances**

An early observation was that it appeared to be harder to detect the features that had fewer instances in the training (and test) data. See Figure 4. We explored several ways in which lower prior expectations of a feature might have affected the training algorithm. One possibility was that it simply took longer for such features to reach their asymptotic performance level. There is no more than a minor effect of this sort; the relative performance of the detectors changed little after a few hundred epochs. Another possibility is that the 12 detectors compete for the resources of the common hidden layer, and that the learning algorithm favors detectors of more frequent features as it adjusts the network weights. We tested this hypothesis by training networks with single output nodes for the features vocalic, sibilant, resonant, and high. The asymptotic a' discrimination of these single feature networks was almost identical to that of the 12-feature network.

The main reason that infrequent features are more poorly discriminated is that the learning algorithm gives

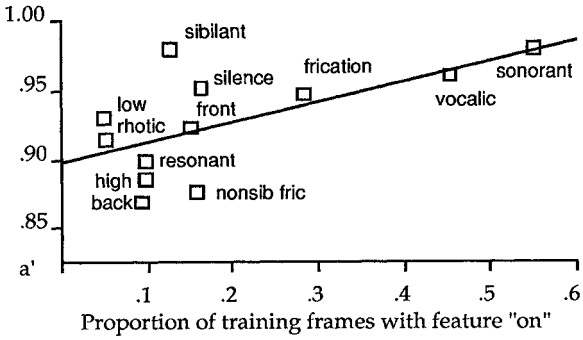


*Figure 4. Performance of the 12 detectors as a function of the proportion of positive instances in the training dataset. The line is a linear fit to the points.*

greater weight to their many negative training instances than to the few positive ones. Discrimination is improved by weighting the learning rate for the positive and negative vectors during training in inverse proportion to their frequency of occurrence. The improvement in average a' is in fact less than .01, but it shows up more clearly in individual features and in g'. Average g' improved from .75 to .79, with gains for all but the three most frequent detectors (sonorant, vocalic, and frication).

**Duration of feature strings**

The edge vectors at the beginning or end of a string of features have less information about the feature than the core vectors in the middle of a string. This is both because edge vectors straddle adjacent acoustic event (vowel and fricative, say) and because of inherent transitional and coarticulatory effects. Typically longer classes of segments, e.g. sibilants, will have more core vectors than edge vectors in the datasets, which will be easier to discriminate. Figure 5 shows that the proportion of core vectors is indeed strongly related to a feature's performance. Note that the five tonality features form a separate set whose lower performance is presumably due more to other factors. The relationships in Figure 5 help us understand some of the differences in performance of the detectors. For instance, its relatively low proportion of cores is presumably one of

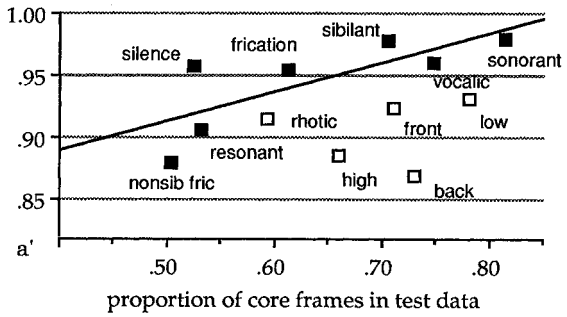


*Figure 5. Performance of the 12 detectors as a function of the proportion of their core vectors in the test data. Unfilled squares represent the tonality detectors. The line is a linear fit to the remaining detectors, represented by filled squares.*

the reasons for the poorer performance of the resonant detector, and the good performance of the silence detector in spite of a low core proportion suggests that this is an inherently more discriminable feature than others. How

the detectors perform on core vectors compared to edge vectors is also revealing. For example, the vocalic, resonant, and sibilant detectors did substantially better, 0.03 to 0.05 for a', 0.06 to 0.07 for g', for core vectors than for edge vectors. But the nonsibilant frication detector did not; a' was only .01 higher and g' .02 higher. It is evident that its high proportion of edge vectors is not mainly responsible for this detector's poorer performance.

**Input representation**

We are continuing to assess the effect of changes in the input representation. Technical improvements in the smoothing filters used to compute the input vector were responsible for raising the previously reported a' performance of 0.87 to the present 0.93. [1,2] Narrower frequency smoothing filters of one Bark standard deviation, and narrower temporal smoothing filters at higher frequencies had little effect, both yielding an average a' about 0.01 lower than the two-Bark x 9-msec filters.

**Temporal context**

To test the limits imposed by the system's restricted local temporal context, we compared two additional networks, one in which the input was augmented by the preceding and following vector, and one in which the two preceding and two following vectors were added. There was remarkably little improvement. Average a' for the 3-vector network improved to 0.94, and g' improved rather more, from 0.79 to 0.84, with virtually all the improvement found in the seven poorer detectors. Neither a' nor g' improved further for the 5-vector network.

**Database errors**

A careful examination of the database revealed that some five percent of the labels were incorrect in location or identity. To assess the effect of the errors we corrected about 10 percent of the original training and test sets and compared the performance of the corrected and original versions. The performance of the resonant and rhotic detectors improved, as we had expected, since a high proportion of the errors were found in laterals, nasals, and schwar. Overall performance did not improve, surprisingly, since the performance of several other detectors unaccountably dropped. This may have been partly due to the smaller datasets. It is also likely that that the effect of labelling errors on performance is diminished because they mainly occur in edge vectors, which are already less well discriminated. And it is of course plausible that connectionist detectors tolerate a certain level of error in training sets fairly well. These results suggest that the resonant and rhotic detector performance would improve if the errors were removed, but that they are probably not a significant factor for the other features.

## CONCLUSIONS

Using "off-the-shelf" connectionist technology, we have been able to construct a high-quality feature detector system that makes rapid judgments about the features present in small segments of speech. The broadest features–sonorance, frication, and silence–are detected with great accuracy, as are the narrower features vocalic and sibilant. Resonant, nonsibilant frication, and the tonality features are detected with lower, yet still good, accuracy.

Our experience suggests that the network is functioning near the limit of information available in the input. That this is the case for the five best performing features is strongly supported by the their lack of improvement with additional temporal context. While we think that there may still be ways to improve the performance of some of the other detectors, it must be acknowledged that a simple acoustic input may not contain sufficient information to unambiguously cue some of the distinctions we are asking the network to make. This is especially true for brief events that are subject to the strongest coarticulatory effects. Furthermore, learning about rare events is harder because of the need to experience speech across the full range of variation. For these reasons, brief, rare events, typical of the heterogeneous collection of phonetic segments assigned to nonsibilant frication, but found to some extent in all feature inputs, pose the most difficulty for the network. We can partially compensate for these problems, but cannot expect to correct them.

If we assume the information about many brief events will be of moderate quality at best, the solution appears to lie in changing the prior probability as a function of context: Providing additional acoustic context, as we did with the 3-vector and 5-vector networks, is one such approach, whose potential, however, is severely limited. Since the addition of acoustic vectors to the network inputs increases the number of possible input states exponentially, the network is faced with an increasingly difficult learning problem. Other forms of context such as lexical or syllabic do not depend on acoustic variation (e.g. speaker gender), and so are simpler than acoustic contexts.

A network that included a lexical or syllabic input into the feature detection process violates our modular design approach to the problem. Must a speech recognition system be constructed such that the very highest levels like word recognition influence the most peripheral processes such as feature detection? Diehl [6] reviewed findings of human speech perception that argued against a separate stage of feature detection precisely because there are so many high-level influences on feature detection. In response, we note that the connectionist network produces graded outputs, not binary decisions. To the extent that these outputs preserve the ambiguity present in speech, as signal detection analysis shows that they largely do, a higher-level decision system can interpret the ambiguity appropriately. Although modest improvements may be possible, other components of a speech recognition system ought to expect some ambiguity in the input, and be able to use alternate sources of knowledge to correctly interpret the signal.

## REFERENCES


[1] Gary Bradshaw and Alan Bell. *J. Acoustical Society of America*, vol. 89, p. 1891, 1991

[2] Gary Bradshaw and Alan Bell. *Robust feature detectors for speech.* Cognitive Science Technical Report UIUC-BI-CS-91-17. Urbana: Beckman Institute. 1991.

[3] Gary Bradshaw and Elizabeth Richards. *Speech as eyes seize it: Optical filtering of speech waveforms.* Final Report to the Colorado Institute for Artificial Intelligence. 1989.

[4] Kjell Ellenius and György Takács. Acoustic-phonetic recognition of continuous speech by artificial neural networks. *Speech Transmission Laboratory QPSR*, vol. 2-3/1990, pp. 1-44, 1990.

[5] J. P. Egan. *Signal Detection Theory and ROC Analysis.* New York: Academic Press. 1975.

[6] R. Diehl. Feature detectors for speech: A critical reappraisal. *Psychological Bulletin*, 89, 1-18, 1981.