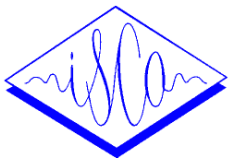

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Speaker-Independent English Alphabet Recognition: Experiments with the E-Set [1]

Mark Fanty and Ron Cole

Oregon Graduate Institute
Beaverton, Oregon, 97006

ABSTRACT

As part of an effort to do high-accuracy speaker-independent recognition of the English alphabet, we focus our attention on the E-set, a very difficult subset of nine letters: B,C,D,E,G,P,T,V,Z. By adding knowledge-based features and finding the most suitable spectral representation, we are able to reduce the E-set error rate by 20% compared to a baseline system that uses features designed for the whole alphabet. The resulting E-set classifier was successfully added to the full-alphabet system, to be invoked whenever the first answer is any member of the E-set.

INTRODUCTION

For the past two years, the speech group at OGI has been investigating neural network approaches to computer speech recognition, with specific emphasis on speaker independent recognition of spoken English letters. We have extended this technology to recognition of names spelled with brief pauses between letters. Name retrieval accuracy, from over 1000 spellings produced by 34 speakers, was 95% when using a database of the 50,000 most common last names in the USA [1].

The success of our approach depends upon the ability to accurately locate and classify individual letters. Our English Alphabet Recognizer (EAR) recognizes letters spoken in isolation, and has achieved 96% speaker-independent recognition of the full alphabet [2,3], when trained on 120 speakers and tested on a different set of 30 speakers. This level of accuracy was obtained by training neural network classifiers with knowledge-based features. Our research has shown that a system trained with selected features, known to be important for discriminating among the speech sounds in the English alphabet, performs better than a system trained with raw data such as spectral coefficients.

A large number of the errors in EAR are from the E-set—B,C,D,E,G,P,T,V,Z. These letters present a difficult challenge to computer speech recognition. A number of fine phonetic distinctions are required, such as B vs. D, B vs. P, D vs. T, T vs. G, C vs. Z and V vs. Z. Successful recognition of the E-set requires the ability to discriminate among the minimal sound units of the language, a fundamental problem in computer speech recognition.

[1]This research was supported by a grant from Adaptive Solutions Inc., Beaverton, OR and DARPA grant MDDA972-88-J-1004 awarded to the Department of Computer Science, Oregon Graduate Institute. The authors wish to thank Vincent Weatherill for recruiting and recording most of the speakers.

The research reported here attempts to build a classifier just for the members of the E-set by adding features and customizing the representations for these discriminations. Letters in the E-set which are misclassified are usually labeled as another member of the E-set. The ultimate goal was to use the E-set classifier to reclassify any utterance which EAR places in the E-set. The following sections give a brief overview of our E-set classifier, describe some of the experiments leading up to our best system, and relate some further experiments on the utility of various recognition features and strategies.

OVERVIEW OF THE RECOGNITION SYSTEM

Speech is recorded using a Sennheiser HMD 224 noise-canceling microphone, lowpass filtered at 7.6 kHz and sampled at 16 kHz with 16 bit resolution. Data capture is performed using the AT&T DSP32 board installed in a Sun 4/100. Each utterance is a single letter, and is recorded in a two second buffer.

A number of representations are computed every three msecs on the utterance, including a 256 point FFT (128 real numbers) computed on a 10 msec Hanning window, the zero crossing rate averaged over a 10 msec window, the peak-to-peak amplitude in a 10 msec window (largest minus smallest value in the window), the peak-to-peak amplitude in a 10 msec window in the waveform lowpass filtered at 700 Hz, and the mean square spectral difference in adjacent 12 msec intervals. A neural network pitch tracker [4] is used to locate peaks beginning pitch periods. The location of pitch peaks is used to calculate the median pitch of the speaker as well as to indicate voicing.

Fundamental to our approach is the segmentation of the utterance into the broad phonetic categories CLOS (closure or background noise), SON (sonorant interval), FRIC (fricative) and STOP. For the experiments reported here, we used a hand-crafted, rule-based segmenter, described in [2].

After segmenting the utterance, we extract features from the most important regions. For the E-set, the most important information is contained in the STOP or FRIC before the SON, and in the formant transitions at the onset of the SON. In addition, the presence of voicing before the initial consonant is useful for discriminating voiced from unvoiced consonants. The features used are described in the next section.

The features are input to a fully connected, feedforward network trained with back-propagation using conjugate gradient descent [5]. The output unit with the highest response is the classification response of the network.

SYSTEM DEVELOPMENT

Our approach was to analyze recognition errors, hypothesize new features or representations and test them by measuring differences in recognition accuracy. We used the ISOLET database [6], which consists of two utterances of each letter produced by 150 speakers. For the experiments during system development, the ISOLET training set was cut in half: 60 speakers were used to train the networks and 60 were used to test. We did this to avoid customizing the features for the official test set. The final system was trained on all 120 training speakers and tested on the 30 test speakers for the first time.

Baseline System

We began with the feature set of the EAR system, used for classifying all 26 English letters, and removed features we believed to be unimportant to E-set discrimination (e.g. features following the sonorant used to distinguish F, S, H and X). Experiments verified there was no loss of performance after removing these features. The baseline system, against which experimental feature sets were compared, contained the following subset of the EAR features:

- Consonant spectrum from 0 to 8 kHz, represented by first averaging the coefficients in thirds of the consonant, then reducing the number of coefficients per averaged segment by taking the maximum of every four: the first feature is the maximum of the lowest four coefficients, the second feature is the maximum of the next four coefficients, etc. (96 inputs).

- Sonorant spectrum from 0 to 4 kHz, represented by the averaged coefficients from the first, second and third seventh of the sonorant (i.e. almost the first half in total). The coefficients in each averaged segment are reduced by taking the maximum of every two (96 inputs).

- Sonorant onset spectrum from 0 to 4 kHz, represented by the spectrum at 6 and 15 msec into the sonorant. The coefficients are reduced by taking the maximum of every two (64 inputs).

- Several parameters for the interval 198 msec before the sonorant onset are represented by their average in 11 segments of 18 msecs each. The same parameters are given for the first half of the sonorant in 6 equal segments. The parameters are peak-to-peak amplitude of the waveform (filtered 0-700 Hz and unfiltered), zero-crossing rate, and spectral difference, which is represented as its maximum value in the intervals rather than the average, since we don't want to smooth out the peaks (68 inputs).

- F0, represented by the median distance between pitch peaks found by the pitch tracker (1 input).

- The duration of the sonorant and consonant (2 inputs).

- The broad-category label of the segment preceding the sonorant (3 inputs).

- The number of pitch peaks found in the consonant (1 input).

- The number of pitch peaks found before the consonant (1 input).

- The largest spectral difference value in the interval from 100 msec before the sonorant onset to 21 msec after (1 input). (This value is normalized to emphasize the difference between B and V.)

- The size of the biggest jump in the peak-to-peak amplitude in the 30 msec interval around the consonant onset (1 input).

- The peak-to-peak amplitude around the sonorant onset at high resolution: 5 segments from 12 msec before to 30 msec after the sonorant onset (5 inputs).

The performance of the baseline system was 91.9/92.1%. The two numbers are the best test performance from two different nets, each begun with different initial weights. The network was fully connected. It had 339 inputs, 27 hidden and 9 output units. Conjugate gradient descent back propagation was used for training. The same number of hidden units was used for all experments reported below unless otherwise specified.

New Features

We proceeded by looking at the spectrogram, waveform, segmentation, and feature values for utterances that were misclassified by the baseline classifier. We were especially interested in unreasonable errors (i.e. errors which we would never make as spectrogram readers). When the system made an unreasonable error, we analyzed the utterance to find important information which may be missing or poorly represented. On the basis of this analysis, we performed a series of experiments in which new features were added to the network. In most cases, the experimental feature or representation was added to the baseline system, not to the previous experimental system. The final system incorporates all successful changes.

One of the most obvious groups of errors was classifying an E as a B or D or vice versa. An E spoken in isolation will typically not be realized as silence followed by a sonorant. There is often a glottal stop or high frequency breath noise before the sonorant onset. Our segmenter usually labels this interval as a stop. The most salient cue for distinguishing

glottal stops and breath noise from /b/ and /d/ is the change in the spectrum after the sonorant onset. For E, the spectral energy before the sonorant occurs at roughly the same frequencies as in the sonorant. Since we did not sample the sonorant spectrum above 4 kHz, we reasoned that the classifier did not have sufficient information to determine if high frequency energy abruptly stops at the sonorant boundary (as in a /b/ or /d/) or whether it continues smoothly (as in a noisy E).

As a direct measure of spectral continuity, we added a spectral derivative measure for each 1000 Hz range from 0 to 8000 Hz to highlight the difference in average energy in the two frames (3 and 6 msec) before and the two frames after the sonorant onset. The results were 92.5/91.7%, a small gain. B and E showed a 20% and 40% reduction in error on the first run, but a much smaller reduction on the second.

A second feature that was useful for detecting glottalization was the "quick change" spectral difference measured for all adjacent frames in the 18 msec before the sonorant onset. This feature differs from the other spectral difference features because it does not average adjacent frames before taking the difference. The result was a slight improvement to 92.2/92.5%.

We also extended the two spectral slices at 6 and 15 msecs after the sonorant onset from 0 – 4 kHz to 0 – 8 kHz, so a comparison with the consonant spectrum could be made. The results improved slightly: 92.1/92.3%.

When we removed the averaged spectra representing the first, second and third seventh of the sonorant, performance improved slightly to 92.1/92.3%. With no sonorant spectrum after 15 msec, the net cannot detect a gradually rising second formant, so we later put back one spectral slice from the middle of the sonorant. This resulted in a 12% error reduction (the change was not made to the baseline system).

We tried many other features which did not help. In particular, we wanted to reduce the number of P/T confusions by directly representing the rising (over time) lower limit of spectral energy typically found in /p/. We were not successful.

Spectral Representation

For our classifier, inputs should generally fall in the range (-1.0, 1.0). We have often found that the way in which feature measurements are mapped to this range (normalized) is important. For all our spectral measurements, the normalization of coefficients was done on a per-slice basis. The max and min coefficients are found in each slice. The values in (min,max) are mapped linearly to (0, 1.0). This emphasizes the spectral shape within the frame but hides the *relative* energy of this frame compared to other frames in the utterance (the peak-to-peak amplitude is a rough measure of energy). When we replaced min and max in the slice with the 10th and 90th percentile values in the whole utterance, the testing performance was 92.6/92.7%.

Several experiments were run to find the best frequency resolution for spectral input. The best results were obtained with a 4-to-1 reduction (from 62.5 Hz to 250 Hz) in the consonant above 4 kHz, and a 2-to-1 reduction everywhere else. Rather than reducing the number of coefficients by taking the maximum of adjacent coefficients, we tried a narrow-window weighted average. This improved performance as well. After all the spectral representation changes, the testing performance was 93.0/93.9%, a much bigger jump than achieved with the new features. (Using a weighted average instead of max to reduce the resolution of spectral coefficients helped when put in the 26-letter EAR system, but the global normalization hurt.)

Shifting Boundary

Virtually all of our feature extraction is determined by the broad-category segmentation. Many features are especially sensitive to the placement of the of the sonorant-onset boundary. A /b/ may be only 9 msecs long and the stop-sonorant boundary is often ambiguous. We examined the effect of moving the sonorant onset boundary, and found that small fixes or seemingly arbitrary one-frame shifts could change the classification.

In order to make our system more robust, we tried some shifting boundary experiments. None resulted in improved performance. We tried shifting the sonorant boundary on the training data +1 or +2 and -1 or -2, yielding 3 times as many training vectors. This had no effect. We tried shifting on test, taking as the classification response whatever boundary placement yielded the most confident response. This hurt performance significantly. In particular, a shifted B often made a "good" E.

Fixed Spectral Sampling

Our whole approach relies on the careful selection of features to present the classifier. Since feature measurements rely on boundary locations provided by the segmenter, errors in the segmentation are propagated to the classification. We tried to reduce our dependence on segmentation by ignoring the consonant segmentation and taking a large number of spectral samples before and just after the sonorant onset. The best such system sampled the spectrum at 150, 83, 71, 69, 57, 45, 33, 21, 15, 9, 6 and 3 msecs before the sonorant onset, and at 0, 3, 6, 12, and 30 msecs after. It did only 91.3% even after we added additional hidden units.

Segmenter Adjustments

All errors due to poor segmentation in our original list (15% of the total) were debugged, resulting in some fine tuning and bug fixes in the rule-based segmenter. This resulted in a small improvement.

Final E-set Classifier

Most of the improvements described above were added independently to the baseline system. The final system incorporates all of them, adding to the baseline system the new features: spectral derivative in 1000 Hz increments around sonorant onset and adjacent frame spectral difference in the 18 msec before sonorant onset. We replaced the three averaged sonorant spectral segments with a single slice from

the middle. The improved spectral representation was used. The new segmenter was used. The results were 93.3/93.7%, a 19% reduction in error from the baseline system.

When we increased the number of hidden units to 52 from 27, the performance went up to 93.6/94.4%. A further increase to 80 hidden units had no effect.

The net was retrained on two utterances each from all 120 speakers in the training set. Training proceeded until the best generalization performance was achieved on a separate cross-validation set. This was repeated with another random seed. Then the best net was tested on two utterances each from 30 new test speakers. The performance was 95% (see Table 1). When the baseline system was trained and tested in the same way, the performance was 93.7%. The new features account for a 20% reduction in the error. This is a worthwhile reduction, although we had hoped for more. We were especially frustrated in out effort to improve the recognition accuracy for P and T. When the E-set net was added to EAR as a post-processing stage, the performance on the E-set went from 93.9% to 94.6% (not 95% because there were errors outside the E-set).

We used the E-set features in the above system to train a network which just classifies B, D, E and V. When trained on the 120 training speakers and tested on the 30 test speakers, the performance was 94.2%.

FURTHER RESULTS

The usefulness of a feature when added to a baseline system does not necessarily reflect its usefulness when added to some other system. Two features may be redundant, or may only work well in combination. Although we cannot try all possible subsets, we did try some interesting combinations of features from our final E-set classifier. On the data sets used for experiments, with 27 hidden units, the full feature set performs at about 93.5%. With only the spectral features, the performance is about 89.5%. With only the non-spectral features (but including the spectral difference measures), the performance is about 87%.

Table 1: Confusion matrix for best network.

label	B	C	D	E	G	P	T	V	Z	Accuracy
B	54	.	1	1	.	1	.	3	.	90.0%
C	.	60	.	.	.	.	.	.	.	100.0%
D	3	.	56	.	.	.	.	1	.	93.3%
E	.	.	.	60	.	.	.	.	.	100.0%
G	.	.	.	.	60	.	.	.	.	100.0%
P	1	.	1	.	1	55	.	2	.	91.7%
T	.	.	1	.	2	3	54	.	.	90.0%
V	1	.	1	2	.	.	.	56	.	93.3%
Z	.	.	.	1	1	.	.	.	58	96.7%

We tested all the EAR features we kept in the baseline E-set classifier one at a time by removing them and retraining the network. For several, there was no loss of performance. When we removed all apparently unneeded features at the same time, however, performance went down.

When we use a network with no hidden layers (i.e. a perceptron), the performance is 93%.

CONCLUSION

Although the use of different data sets makes comparison difficult, these results are the best reported on the E-set. Brown [7] reported 92% multi-speaker recognition (same training and test speakers) on the E-set using hidden Markov models. Lang et al. [8] reported 93% multi-speaker on the letters B,D,E,V using a neural network classifier with only spectral input.

The improvement in E-set classification which resulted from these experiments was small but significant. We feel our approach of carefully selecting and testing knowledge-based features for input to neural network classifiers has been further supported. Our full feature set is significantly better than spectra alone.

The final network was installed as a second stage classifier in EAR, which recognizes all 26 letters. If the main net has as first choice any member of the E-set, the specialized E network is invoked and its response is used. This post-processing improved our score slightly and helped us achieve 96% overall performance [2].

BIBLIOGRAPHY

[1] Cole, R. A., M. Fanty, M. Gopalakrishnan and R. Janssen, "Speaker-Independent Name Retrieval from Spellings using a Database of 50,000 Names," Submitted to ICASSP 1991.

[2] Cole, R. A., M. Fanty, Y. Muthusamy and M. Gopalakrishnan, "Speaker-independent recognition of spoken English letters," *International Joint Conference on Neural Networks*, 1990.

[3] Cole, R. and M. Fanty, "Spoken Letter Recognition," DARPA Proceedings Speech and Natural Language Workshop, June 1990.

[4] Cole, R., E. Barnard, M. Vea and F. Alleva, "Classification of pitch periods using expert knowledge and neural net classifiers," *Journal of the Acoustical Society of America*, **84**, 1988.

[5] Barnard, E and R. Cole, "A neural-net training program based on conjugate-gradient optimization," Technical report CSE 89-014, Computer Science Department, Oregon Graduate Institute, 1989.

[6] Cole, R. A., Y. Muthusamy and M. Fanty, "The ISOLET spoken letter database," Technical report CSE 90-004, Computer Science Department, Oregon Graduate Institute, 1990.

[7] Brown, P. F., "The acoustic-modeling problem in automatic speech recognition," Doctoral Dissertation, Carnegie Mellon University, Dept. of Computer Science.

[8] Lang, K. J., A. H. Waibel, and G. E. Hinton, "A time-delay neural network architecture for isolated word recognition," *Neural Networks*, **3**, pp. 23–43, 1990.