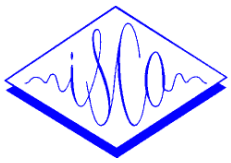

2nd European Conference on
Speech Communication and Technology
EUROSPEECH '91
Genova, Italy, September 24-26, 1991

ENGLISH ALPHABET RECOGNITION WITH TELEPHONE SPEECH

Ronald Cole, Krist Roginski and Mark Fanty

Oregon Graduate Institute of Science and Technology
19600 N.W. Von Neumann Dr., Beaverton, OR 97006

ABSTRACT

We describe database and system development for speaker-independent recognition of telephone speech. The telephone speech database contains about 4,000 callers from the U.S.A. and Canada each of whom provided several utterances, including city names, first and last names, spelled names, and answers to yes/no questions. About 1,000 of the callers recited the English alphabet with pauses between letters. A portion of the database has been verified and phonetically labeled, and this portion was used to develop a baseline system that recognizes names spelled with pauses between letters. The system uses a neural network to segment speech into a sequence of 24 phonetic categories. The phonetic categories are used to hypothesize a sequence of letters which are then reclassified using a second neural network. First choice letter recognition accuracy was 87.6% in the best condition. First choice name retrieval was 85.5% for 200 spelled names retrieved from a database of 50,000 common last names.

1 Background

The English alphabet is a challenging vocabulary for computer speech recognition because of the acoustic similarity of many letter pairs, such as B/V, T/G and M/N. During the past two years, our group has developed speaker-independent systems that combine speech knowledge and neural network classification to achieve accurate spoken letter recognition using high quality speech [2, 4]. Our first system, called EAR, recognizes letters spoken in isolation. The system (a) segments the letter into broad phonetic categories; (b) uses the category boundaries to compute a set of empirically-derived feature measurements; and (c) classifies the letter using a three-layer feed-forward neural network. The system was trained on two tokens of each letter produced by 120 speakers. First choice accuracy was 96% when tested on two tokens of each letter from a different 30 speakers.

We extended isolated letter recognition to recognition of words spelled with brief pauses between the letters [4, 1]. This task is more difficult than recognition of isolated letters because there are "pauses" within letters, such as the closures in "X," "H" and "W," which must be distinguished from the pauses that separate letters, and speakers do not always pause between letters when asked to do so (about 10% of the time in our database). The system (a) uses a neural network to segment speech into a sequence of five broad phonetic categories (closure, sonorant, stop, fricative and glottalization); (b) individual letters are located by applying rules to the broad category labels; (c) classification is performed using the same feature measurements and neural network architecture used for isolated letters, and (d) the letter scores are used to retrieve the best scoring name from a database of 50,000 last names. First choice name retrieval was 95.3%, with 99% of the spelled names in the top three choices.

This paper extends our research to telephone speech. We describe a baseline system that recognizes names spelled with pauses between letters by unknown talkers. Additional problems encountered in this task are the limited telephone bandwidth, variability due to different telephone handsets, increased speaker variability due to the large number of talkers of various ages and dialects who contributed to our database, and variability due to background and channel noise. This paper reviews (a) the data collection procedures, (b) the system modules that locate letters, classify letters, and retrieve names from the letter scores, and (c) preliminary recognition results.

2 Data Collection

The data collection effort was promoted under a "donate your voice to science" theme. Callers were solicited through local newspaper and television coverage, and notices on computer bulletin boards and news groups. Callers had the choice of using a local phone number or toll-free 800-number.

A Gradient Technology Desklab attached to a UNIX workstation was programmed to answer the phone and record the answers to pre-recorded questions. Three thousand callers were given the following instructions, designed to generate spoken and spelled names, city names, and yes/no responses.

1. What city are you calling from?
2. What is your last name.
3. Please spell your last name.
4. Please spell your last name with short pauses between letters.
5. Does your last name contain the letter "A" as in apple?
6. What is your first name.
7. Please spell your first name with short pauses between letters.
8. What city and state did you grow up in?
9. Would you like to receive more information about the results of this project?

In order to achieve sufficient coverage of rare letters, the final 1000 speakers were asked to recite the entire English alphabet with brief pauses between letters.

Approximately 900 utterances have been verified by three listeners. These consist of 300 complete alphabet recitations produced by 150 female and 150 male talkers, and 600 last names spelled with brief pauses between letters produced by 300 female and 300 male talkers. (Talkers identified as "children" or "unknown gender" were excluded from this study). When verifying spelled last names, the listener was also presented with the spoken name to aid identification.

Time-aligned phonetic labels have been assigned to 324 first and last names and 38 alphabets, using the following labels: cl bcl dcl kcl pcl tcl q aa ax ay b ch d ah eh ey f iy jh k l m n ow p r s t uw v w y z #h h#. This label set represents a subset of the TIMIT labels sufficient to describe the English alphabet.

3 System Overview

We have produced a baseline recognition system that accepts telephone speech and performs speaker-independent English alphabet recognition and directory name retrieval. The system modules are shown in Figure 1.

We view this as a baseline system since the feature set used to locate and classify letters is limited to PLP coefficients and one duration feature. Subsequent systems will include empirically derived features which have been shown to improve performance for high quality speech.

Data Capture. Telephone speech is sampled at 8 kHz at 14-bit resolution.

Signal Representation Signal processing routines perform a seventh order PLP (Perceptual Linear Predictive) analysis [5] every 3 msec using a 10 msec window. This analysis yields eight coefficients per frame, including energy.

Frame-based Phonetic Classification. Frame-based phonetic classification provides a sequence of phonetic labels that can be used to locate and classify letters. Classification is performed by a fully-connected three-layer feed-forward network that assigns 22 phonetic category scores to each 3 msec time frame. The 22 labels provide an intermediate level of description, in which some phonetic categories, such as [b]-[d], [p]-[t]-[k] and [m]-[n] are combined; these fine phonetic distinctions are performed during letter classification, described below.

The input to the neural network classifier consists of 56 features representing PLP coefficients in a 168 msec window centered on the frame to be classified. The manner in which PLP coefficients were averaged across 3 msec time frames is shown in Figure 2.

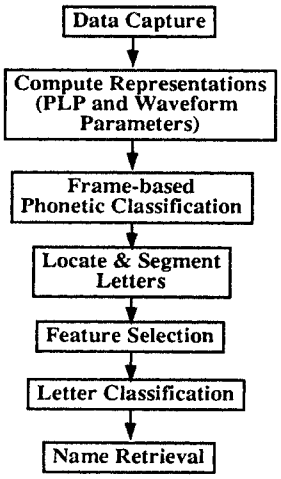

Figure 1: The modules in the telephone name retrieval system.

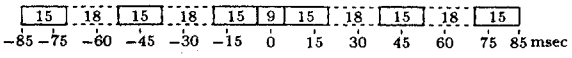

Figure 2: PLP coefficients represent near and far context. The solid boxes indicate the intervals over which PLP coefficients are averaged. Dashed boxes indicate intervals that are skipped.

Letter Segmentation. The frame-by-frame outputs of the phonetic classifier are converted to a sequence of phonetic segments corresponding to a sequence of hypothesized letters. This is done with a Viterbi search that uses duration and phoneme sequence constraints provided by letter models. For example, the letter model for N consists of optional glottalization (MN-q), followed by the vowel [eh] (MN-eh), followed by the nasal murmur (MN-mn). Because background noise is often classified as f-s or m-n, a noise

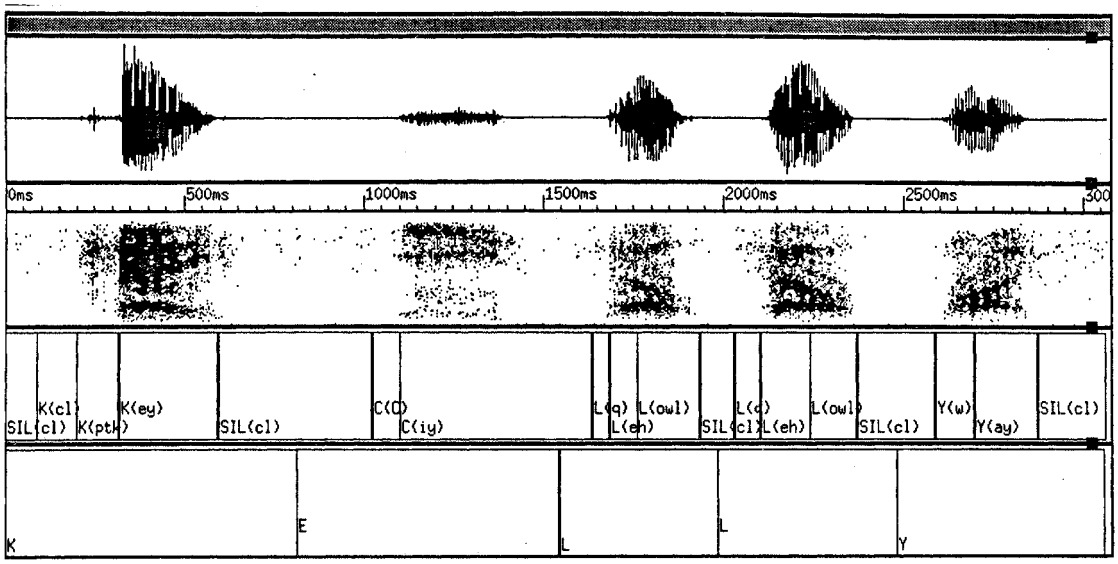

Figure 3: Waveform, DFT, phonetic alignment and letter classification for the letter string K-E-L-L-Y. Letter reclassification corrected the E–C error from segmentation. Noise segments were removed for readability.

"letter" model was added which consists of either of these phonemes. Figure 3 shows the alignment for the letters in K-E-L-L-Y.

Letter Classification. Once letter segmentation is performed, a set of 105 features is computed for each letter and used by a fully-connected feed-forward network with one hidden layer to reclassify the letter. Feature measurements are based on the phonetic boundaries provided by the segmentation. At present, the features consist of duration of the initial consonant plus PLP coefficients averaged over thirds of the initial consonant (fricative or stop), sevenths of the first sonorant, and the first three 70msec intervals after the first sonorant. The outputs of the classifier are the 26 letters plus the category "not a letter."

Name Retrieval. The output of the classifier is a score between 0.0 and 1.0 for each letter. These scores are treated as probabilities and the most likely name is retrieved from the database of 50,000 last names. The data base is stored in an efficient tree structure. Letter deletions and insertions are allowed with a penalty.

4 Performance

4.1 Frame-Based Phonetic Classification

The phonetic classifier was trained on selected speech frames from 260 speakers. About 200 speech frames were selected from 50 different occurrences of each phonetic category. Phonetic segmentation performance on 164 test utterances was evaluated by comparing the first-choice of the classifier at each time frame to the label provided by a human expert. The agreement was 76% (before the Viterbi search). After the Viterbi search, inspection of the phonetic segmentation on test utterances revealed accurate location of most boundaries.

4.2 Letter Classification

Data Generation. In order to avoid hand-segmenting training data for letter classification, an automatic procedure was used. Each utterance was listened to and transcribed manually. Segmentation was performed as described above, except the Viterbi search was constrained to match the transcribed letter sequence.

To generate training data for the "not a letter" category, a second Viterbi search was run without the forced alignment. Any "letters" found by the unconstrained search which correspond to noise or silence from the constrained search are used as training data for the "not a letter" category. To summarize, there are two ways noise can be eliminated: It can match the noise model of the segmenter during the Viterbi search, or it can match a letter during segmentation, but be reclassified as "not a letter" by the letter classifier.

Effect of Training Set Size Spelled alphabets from 300 speakers (equal number of male and female speakers) were used to determine the effect of training set size on network performance. The test set was fixed at 100 alphabets. The network was trained on 1 (i.e. just 26 inputs) to 200 alphabets. The results are summarized in table 1.

Table 1: Effect of training set size on performance.

No. of Alphabets	Test set %
1	45.0
2	56.8
6	72.0
10	74.0
26	80.4
50	83.2
100	86.3
200	87.9

These scores include performance on "not a letter" which was typically around 94%. Removing this category from the average drops the score in the 200 training alphabet case from 87.9 to 87.6.

Since chance classification is around 4%, the performance achieved after training on a single male speaker is quite high. Since performance has not leveled off, more training data should continue to improve performance.

4.3 Name Retrieval

The best network from above (200 training alphabets) was used in a name retrieval system tested on 200 last names spelled by 200 callers. All the last names are in a list of 50,000 common last names. The results were: 85.5% first choice; 96% in the top three. The raw letter classification (no grammar) was 81.5% correct—a significant drop from the alphabet test set. There were 1.6% letter insertions and 0.4% deletions; the letter accuracy (insertion penalty applied) was 79.9%.

When the training set of 200 alphabets was augmented by the letters from 400 additional callers spelling their last names, the letter accuracy on the same 200 name test set rose to 84.4% (83.1% with insertion penalty), but the first choice name retrieval dropped to 82.5%. With a training set of 100 alphabets plus the letters from 400 last names, the letter accuracy was 83.8% (82.5%); name retrieval was 85.5% first choice, 92.5% top three.

5 Discussion

The baseline recognition system described in this paper classifies letters of the English alphabet produced by any speaker over telephone lines at 88% accuracy for spelled alphabets. Fully automatic name retrieval performance of the baseline system on test speakers using a database of 50,000 names is 85.5% first choice accuracy, and 96% for the top three choices in the best condition.

There are at least two reasons why letter recognition performance was higher for the alphabet test set: the segmentation was performed with knowledge of what letter it was (not true for the 200 name test set), and some high scoring letters, like Q and W, are less common in names.

It is not clear why adding letters from spelled names to the alphabet test set resulted in better letter classification but worse name retrieval. It could be an artifact of the particular test set; a larger test set is desirable. Also, the distribution of letters in names is unbalanced resulting in a bias towards the more common letters. This would help letter classification in names but may not help name retrieval where uncommon letters are not likely to cause confusion with other names.

Euler et al. [3] report 93% classification for a 36 word vocabulary, including the letters, spoken over local phone lines. It is hard to compare the results directly, since they include digits which have a high recognition rate and use the same speakers in the training and test sets.

In the next several months, we hope to verify an additional 3,000 spelled names and 600 alphabets to greatly increase our training and test sets. We will develop more powerful feature measurements and work on noise robustness.

New performance statistics will be generated for phonetic segmentation, letter classification and name retrieval. Based on our past experience using high quality speech, we can expect that increasing the training data and improving feature measurements will result in steady improvements in system performance.

6 Acknowledgements

Research supported by U.S. WEST, APPLE Computer, NSF and ONR.

References

[1] R. A. Cole, M. Fanty, M. Gopalakrishnan, and R. D. T. Janssen. Speaker-independent name retrieval from spellings using a database of 50,000 names. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1991.

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

[3] S. A. Euler, B. H. Juang, C. H. Lee, and F. K. Soong. Statistical segmentation and word modeling techniques in isolated word recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1990.

[4] M. Fanty and R. A. Cole. Spoken letter recognition. In R. P. Lippman, J. Moody, and D. S. Touretzky, editors, *Advances in Neural Information Processing Systems 3*. San Mateo, CA: Morgan Kaufmann, 1991.

[5] H. Hermansky. Perceptual Linear Predictive (plp) analysis of speech. *J. Acoust. Soc. Am.*, 87(4):1738–1752, 1990.