ROBUST FEATURE EXTRACTION FOR ALPHABET RECOGNITION

Montri Karnjanadecha and Stephen A. Zahorian

Department of Electrical and Computer Engineering
Old Dominion University
Norfolk, VA 23529, USA

ABSTRACT

Spectral/temporal segment features are adapted for isolated word recognition and tested with the entire English alphabet set using Hidden Markov Models. The ISOLET database from OGI and the HTK toolkit from Cambridge university were used to test our feature extraction technique. With our feature set we were able to achieve 97.3% recognition accuracy on test data with one pass using a whole word based recognizer. Gaussian noise was also added to evaluate robustness of the feature set. We were able to obtain recognition accuracies of 49.6% and 84.3% at SNR of -10dB and 0dB, respectively. Linear discriminant analysis was also applied to the initial feature set for a number of feature configurations and noise levels but, generally, the performance was not improved. We conclude that the initial feature computations used are both very efficient (best results obtained with 50 total features) and robust in the presence of noise.

1. INTRODUCTION

Although automatic speech recognition performance has improved substantially over the past several years, automatic methods are still far inferior to most human listeners for most tasks. Automatic performance is quite good for limited tasks with clean speech, but often degrades under adverse conditions, or for the case of difficult phonetic distinctions. For example, stop consonants extracted from continuous speech can be identified with over 95% accuracy [5], whereas the best reported machine performance for the same task is about 82% [7]. Similarly, human listeners are generally able to better discriminate members of the "E" set (b, c, d, e, g, p, t, v and z) than are the best machine algorithms. In this paper, we adapt and modify feature extraction techniques which we previously found to work well for phonetic classification [10], to isolated word recognition and test these with the entire English alphabet set.

In this work, each feature vector was comprised of 50 terms which encode the trajectory of short-time cepstral coefficients over overlapping intervals 200 ms long. Additionally, Linear Discriminant Analysis (LDA), a linear transformation technique that has been shown to improve recognition performance of many speech recognizers [4], was applied to these features and transformed terms were used for recognition. For the case of LDA, the number of terms used was varied from 5 to 50. We compared recognition results obtained with these two set for features for varying signal-to-noise ratios.

The ISOLET spoken letter database from OGI was used in all experiments presented in this paper [1]. This database suits the objectives of evaluating our signal modeling/feature computation methods in many aspects. It is spoken in isolation which was more convenient for testing than would have been a continuous speech recognition system. It is comprised of 7800 spoken letters uttered by 75 males and 75 females and is suitable to train and test a speaker independent recognizer. It is a small but difficult task. It was also previously used by several researchers which makes it easier to directly compare our results with published work.

This paper is organized as follows. Section 2 presents the speech analysis procedures used in the study. Experiments and results are given in Section 3. In Section 4, results are discussed and, in Section 5, final conclusions are given.

2. ANALYSIS PROCEDURES

For all experiments, every speech file from the database was analyzed by an endpoint detection program in order to locate more accurate endpoints. An endpoint detection scheme proposed in [2] was used to locate initial endpoints, which were then extended 30 ms in each direction (i.e., backward in time for the onset and forward in time for the offset) to allow for some inaccuracies in the original detection, and also to include a small amount of silence at the beginning and end of each utterance. The endpoint detection algorithm from [2] can be summarized briefly as follows. First the speech signal is pre-emphasized to eliminate the DC component and to emphasize the higher frequency components prior to background noise estimation. Then, at each endpoint, a location of low energy area is detected using energy thresholds derived from the estimated background noise. The energy is computed using 80ms non-overlapped frames. Then a shorter frame, 30ms, is used to locate the endpoint more precisely by finding a location where the maximum change of energy content of two adjacent frames occurs. In this step the frame is shifted one sample at a time.

We adapted the endpoint algorithm of [2] slightly as follows, to be more suited to the ISOLET data. One fundamental difficulty was that only short silence intervals were available at the beginning and end of each utterance. Therefore , we used a frame size of 20ms and 10ms for the long frame and short frame, respectively. Also, for about 10% of the utterances, the algorithm did not satisfy certain threshold criteria to indicate a reliable result (often due to insufficient silence intervals for the background noise estimates). For those cases, we used the

endpoints in the original database. Although every speech file in the database was already endpoint detected, our pilot tests indicated that the recognition performance on test data could be improved by a small amount (0.3%) if our endpoint detection program was applied.

For signal modeling/feature extraction, we use a variation of methods previously presented for phonetic classification [10]. Summarizing briefly, after second order pre-emphasis, Kaiser-windowed 20-ms speech frames are analyzed with a 512-point FFT every 5 ms. For the results given in this paper, a Kaiser window beta of 8 was used, corresponding to a somewhat "smoother" window than the more typically used Hamming window. Using 10 basis vectors over frequency, which incorporate a bilinear frequency warping, 10 modified cosine terms over frequency were computed for each spectral frame. These 10 terms, very similar to cepstral coefficients, were computed with a bilinear warping factor of .45 over the frequency range of 60 Hz to 7600 Hz. These 10 terms in turn were each represented by a 5 term modified cosine expansion over time, using a "block" window with variable length. Thus each block was represented by 50 spectral/temporal features.

For the experiments reported in this paper, special attention was given to adapting the block length to the position within the utterance. In particular, at the beginning of an analyzed token, a block size of 6 frames (i.e., a 45 ms total duration, including end effects of the analysis frames) was used. As the analysis window moved forward, the block size increased until a maximum of 40 frames (215 ms total duration) was reached. The block size was then fixed at 40 frames until the end of the token. Time "warping" was also applied to each block, again using a Kaiser window, but for this case the Kaiser window beta was 5.0 for the 40 frame blocks. The Kaiser window beta also varied from 0 for the 45 ms blocks up to 5.0 for the maximum block length. Thus, the features gave better time resolution for the onset portion of each word, and less time resolution in later portions of each word. The block features were recomputed every 10 ms. No manual segmentation or phonetic labeling was required or used.

Although the features described above perform well for both phonetic classification and speech recognition, we investigated the effectiveness and robustness of these features by transforming them with LDA. Both feature sets were also tested with clean speech and noisy speech.

We began with clean speech utterances (as distributed by the LDC) and added various levels of Gaussian noise to them before the feature extraction step was performed. In this work speech signals with signal-to-noise ratios (SNR) of -10dB, 0dB, 10dB, 20dB, and 30dB were evaluated.

In our implementation of the LDA technique, we compute two covariance matrices, B and W. The between class covariance B is estimated as the grand covariance matrix of all the training data (the same as for a principal components analysis). The within class covariance W is estimated by computing the average covariance of time aligned frames of data belong to the same class. Time alignment is accomplished using dynamic time warping to first determine a “target” for each word by successively aligning and averaging all tokens of that word in pairs until only one token remains. Covariance contributions are then computed as variations about the target, after another time alignment to that target. These two matrices are then used to create a linear discriminant analysis transformation which maximizes the ratio of between-to-within class covariance. Our implementation of this technique is similar to what is presented in [8].

3. EXPERIMENTS

The entire ISOLET database was used to test the recognizer. The speech waveform was sampled at 16000 Hz with 16-bit quantization. This database contains 5 subsets: ISOLET1, ISOLET2, ISOLET3, ISOLET4, and ISOLET5. Each subset is comprised of speech utterances pronounced by 15 males and 15 females. Each speaker utters the same word twice. In Experiments I and II, all speech utterances from ISOLET1-4 (6240 total tokens) were used for training and all speech utterances from ISOLET5 (1560 total tokens) were used for testing. In Experiment III, training and test data were organized into 5 groups by rotating the test set. Each group has one database subset as test data and the remaining subsets as training data. Recognition results were averaged over all groups. Note that all database arrangements were done in a speaker independent fashion.

For some cases, as shown below, additive Gaussian noise was added to the speech files in order to test the robustness of the features to noise at various levels.

In all cases the HTK toolkit version 2.1 from Cambridge university [9] (distributed by Entropic Cambridge Research Laboratory Ltd.) was utilized to provide a whole word HMM based speech recognizer. Continuous density, full covariance HMM’s were used, with 3 Gaussian mixture components, 5 states, and only self transitions and transitions to the next state allowed.

With the HTK toolkit, initial HMM models are estimated by uniformly segmenting each training token to have an equal number of states (frames). Model parameters are computed based on segments of all training tokens. Viterbi decoding is then carried out to determine the most likely segment boundaries of each token and new boundaries are assigned to it. Model reestimation is performed after every token had been resegmented. These steps are repeated until the estimate does not change or a specified number of iterations is exceeded. By default the number of iterations is 20. The HKT toolkit also provides the Baum-Welch reestimation program but we found that in most cases our test results were superior with initial models (about 0.1% to 0.4% higher without reestimations). Since the toolkit was mainly designed for continuous word recognition, recognition scores given in this paper are based on tokens that yield the highest probability of the most likely state sequences obtained by Viterbi decoding. Thus Baum-Welch restimation was not used at all in our reported work

3.1. Experiment I

The objective of this experiment was to evaluate the performance of our overall system with and without LDA analysis in the presence of noise on both training and test speech. For these tests white Gaussian noise was added to

the speech signal such that the overall signal-to-noise ratio varied from -10 dB to 30 dB. Also performance with clean speech was tested. Test results for 50 initial features and 30 LDA transformed features are given.

Signal to Noise ratio (dB)	Results with 50 original features (%)	Results with 30 LDA features (%)
-10	46.1	49.6
0	82.5	84.3
10	92.9	92.0
20	96.6	96.3
30	96.6	96.2
clean speech	97.3	96.5

Table 1: Test recognition results for various signal-to-noise ratios, using original features or LDA transformed features.

Discussion

The result of 97.3% (42 error tokens) with 50 original features for clean speech is the baseline result for this work. Note that this result was obtained after numerous pilot tests to used to adjust parameter values. It can be further broken down in terms of alphabet subsets as follows. The accuracy on the E set portion of the alphabet (B, C, D, E, G, P, T, V, and Z) is 96.1%. The performance on the M and N alphabet is 90.0% and performance on the rest of the alphabet is 99.1%. Note that performance is poorest for the recognition of M and N. Although some attempts were made to implement a two pass recognizer, with the goal of the second pass to improve performanceon the E set and the M/N pair similar to [6], none of our attempts improved this baseline result.

The total number of errors for the test data was so few (42 tokens in error out of 1560 total), that is was very easy to visually and auditorily inspect all error tokens in detail. Based on this inspection, we concluded that all remaining errors are due to three sources: 1, In some cases the endpoints are still not correct with either extra noise, or truncation problems; 2, Some of the tokens are simply extremely difficult to discriminate, even with careful listening; and finally 3, some of the error tokens, although clearly recognizable, were pronounced in kind of slow "drawl," and thus much longer than average.

3.2. Experiment II

For this case, recognition performance of the alphabet set experiments was examined on test data as the number of LDA transformed features varied from 5 to 50 on clean speech and degraded speech (SNR=15dB). In both cases, test performance is quite stable from 15 features to 50 features, with slight degradation in performance with fewer than 15 features.

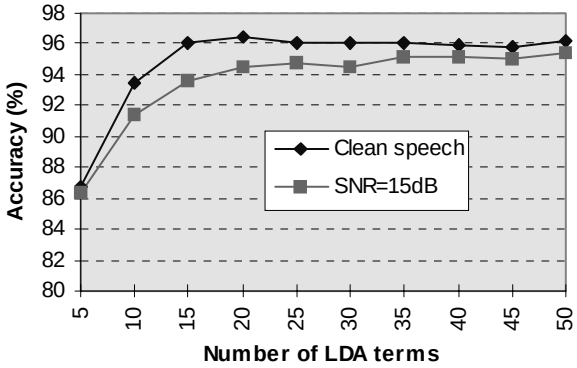

Figure 1: Test recognition results with various number of LDA terms on clean speech and degraded speech.

3.3. Experiment III

We conducted this experiment in order to evaluate whether or not our very high results were due to unfair tuning of our parameters to the test set ISOLET5. Therefore, for this experiment each database subset was selected as a test set and the remaining data were used for training our recognizer. For example if ISOLET1 was used for testing then ISOLET2-5 were used for training. The following table depicts recognition accuracy on each set of test data. Average recognition results are also shown in the last row. Note that these experiments were all performed with our initial features (50 terms).

Test set	Recognition Accuracy (%)
ISOLET1	98.0
ISOLET2	97.0
ISOLET3	97.6
ISOLET4	97.7
ISOLET5	97.3
	Average = 97.5%

Table 2: Test recognition results on various test sets.

The results indicate that the features described in this paper perform very well on all subsets of the ISOLET database.

4. DISCUSSION OF RESULTS

The best speaker independent performance on OGI's ISOLET database with the same test data as used for our work was obtained using a 2-stage, phoneme-based, context-dependent HMM recognizer [6]. The result reported was 97.37%. The next best reported result of 96.0% was obtained using 617 features and a neural network approach [3]. Our recognizer was able to achieve 97.3% of accuracy with a simple one pass, word-based recognizer. We also use a relatively low number of features (50) which are computed in a straight forward manner. We believe our methods would be much easier to duplicate and to apply to other tasks than an isolated word

alphabet recognizer than would be the methods reported in the other two studies mentioned in this paragraph.

One puzzling result, at least to us, was that LDA did not improve recognition performance for most cases. It only improved performance slightly at very high noise levels, -10dB and 0 dB. In the developmental stages of this work, we did observe, however, that for feature parameter values that did not result in optimal performance, LDA did usually result in improved accuracy. We hypothesize that if original features are very good, LDA is not particularly beneficial. We further hypothesize that the feature modeling techniques described in this paper are already beyond the point that LDA can result in further improvements.

However, as shown by Experiment II, LDA can be used to reduce the number of features by a factor of about 3 with only modest decreases in performance. Results of Experiment II show that 15 LDA terms could be used for recognition without losing much accuracy. Thus the feature reduction capability of LDA still holds true.

5. CONCLUSIONS

Methods for compactly representing the spectral temporal structure of speech have been applied to recognizing the letters of the English alphabet set. The use of LDA does not result in any improvement with either clean or noisy speech. We conclude that the signal modeling techniques described in this paper, and also as reported in more detail previously, apply very well to difficult isolated word recognition tasks. The best result obtained from this test is as good as the best reported result for this database.

Although the signal modeling methods used in this study are very similar to the commonly used cepstra and delta cepstra features, there are also many important differences which, we believe, lead to improved performance. The most obvious difference is the greater consideration paid to temporal information in the feature set. In particular, as compared to most reported work, our features use more trajectory terms and a much longer time interval.

6. ACKNOWLEDGMENTS

Portions of this work were supported by NSF grants IRI-9217436 and BES-9411607.

7. REFERENCES

[1] Cole, R., Muthusamy, Y., and Fanty, M., "The ISOLET spoken letter database," Tect. Rep. 90-004, Oregon Graduate Inst., 1990.

[2] Dermatas, E. S., Fakotakis, N. D., and Kokkinakis, G. K., "Fast Endpoint Detection Algorithm for Isolated word Recognition In Office Environment," *Proc. ICASSP'91, Toronto, Canada, MAY 1991, pp. 733-736.*

[3] Fanty, M., and Cole, R., "Spoken letter recognition," *in Proc. Neural Inform. Processing Syst. Conf., Nov. 1990, pp. 220-226*.

[4] Hunt, M. J., and Lefebvre, C., "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," *in Proc. ICASSP'89, Glasgow, Scotland, May 1989, pp.262- 265*

[5] Lamel, L. F., "Identification of stop consonants from continuous speech in limited context," *J. Acoust. Soc. Am. Suppl. 1 82, S80., 1987.*

[6] Loizou, P. C., and Spanias, A. S., "High-Performance Alphabet Recognition," *IEEE Trans. Speech and Audio Processing. Vol. 4, no. 6, pp. 430-445, 1996.*

[7] Nossair, Z. B., and Zahorian, S. A., "Dynamic Spectral Shape Features as Acoustic Correlates for Initial Stop Consonant," *J. Acoust. Soc. Am. 89, pp. 2978-2991, 1991.*

[8] Parsons, T., *Voice and Speech Processing*, Mc-Graw Hill, New York, 1987.

[9] Young, S. J., Odell, J., Ollason, D., Valtchev, V., and Woodland, P., *Hidden Markov Model Toolkit V2.1 reference manual*, Technical report, Speech group, Cambridge University Engineering Department, March 1997.

[10] Zahorian, S. A., Silsbee, P. L., and Wang, X., "Phone Classification with Segmental Features and a Binary-Pair Partitioned Neural Network Classifier," *Proc. ICASSP97, Munich, Germany, Apil. 1997, pp. 1011-1014.*