Minimal Consonant Pair Discrimination for Speech Therapy
Using an Expanded Feature Set and Pattern Element Selection in Time and Frequency

Martin S. Glassman* and Mary Beth Starkey**

*Siemens Corporate Research and Technology Laboratories, Princeton, New Jersey 08540
**Reading Rehabilitation Hospital, Reading, Pennsylvania 19607

ABSTRACT

 Systems for speech training have concentrated almost exclusively on presenting analysis of stationary speech parameters (formants, vocal tract shape, pitch, etc.) in visual form, or giving feedback based on vowel or utterance recognition [1]. Consonants are, however, usually given more attention during therapy. This is primarily due to their greater complexity (increased articulator agility is required to produce consonants) and their importance for intelligibility.

We have developed a system based on dynamic selection from a variety of feature generation options, and a detailed selection of regions over time and frequency using an error rate criterion which allows the system to focus on aspects of speech patterns that are most effective for phoneme pair discrimination. Tests with a set of 14 minimal consonant pairs resulted in discrimination error rates of 0.4% for this system vs. 2.7% for a knn classifier and 5.7% for a neural network trained with backpropagation. Prototype application with a version of this algorithm utilizing only the expanded feature set indicates clinically useful performance with head injury and stroke patients.

INTRODUCTION

Speech therapy which focuses on contrasts, rather than absolute targets, is an effective tool to aid patients in achieving usable communication, relying on the ability of the listener to accommodate perceptually to the speech patterns of the patient. The use of minimal pairs allows patients to concentrate on one aspect of production and on the specific feature which is in error. It also constrains pattern recognition decision boundaries to agree more closely with perceptual boundaries. Unfortunately, automatic analysis and recognition of consonants is generally more difficult since many are transient events, or low amplitude signals tending to be less well articulated. Perception of consonants is complex, involving multiple cues, so that automated recognition based on a relatively small set of parameters is not likely to replicate human performance. This system is designed to utilize a large number of pattern measurements in a way that allows training and recognition in real time.

BOUNDARY ESTIMATION

In our work, each consonant is used in a cv syllable with the same vowel. An estimation technique is used to identify a region at syllable onset designed to include the consonant in a robust way. Since the consonants both as a group and individually have more variability in articulation, the importance of reliable pattern generation is critical for template based techniques. Robustness is gained through use of an ensemble of boundary points combined in such a way as to reduce the impact of large deviations of individual boundaries in response to incremental changes in the signal. A simple heuristic weighting function further deemphasizes the contribution of boundary points at "non speechlike" portions of the signal, adding immunity to certain types of interference. Time domain analysis is performed at each of a set of amplitude levels spanning the minimum to maximum signal level. An "estimate" of the output boundary point for each level is initialized to the location at which the signal first drops below that level from the major signal peak in the region. This "estimate" is updated using a weighted contribution from each successively encountered boundary point defined whenever the signal again crosses below the analysis level. This weighting is based on heuristics defining the likelihood that the next peak belongs to the peak being analyzed. A final begin or endpoint is obtained from the mean location of the resultant boundary points for each level, offset by the mean absolute deviation, plus a fixed offset to insure capture of any additional relevant data.

FEATURE EXTRACTION

A set of simple feature extraction processes operating on log scaled bandpass filterbank data in both time and frequency domains generates a corresponding set of template types for the initial consonant region of each utterance. [This data is 12 bit, sampled every 2.5 ms, with center frequencies from 260 to 4400 Hz.] Improved performance is achieved through the availability of alternate features for discrimination, since both time and frequency domains carry important information for discrimination. Discrimination information is also found on different scales - i.e., local or global measures. This particular set of 18 features is defined by first and second difference operations on the input data template at different smoothing levels (scales) in both time and frequency. The output is 16 templates ranging in size from 16 x 1 to...
in the given case. The effect of using alternative features and groups of features is included in the following plot, which shows a geometric mean reduction in error rate by 61% when using the best feature type as opposed to the default type. Use of the predicted feature group reduces the error rate to 52%; it is felt that in practice most of the gain by using multiple features is in robustness. 

This system has been undergoing clinical evaluation during the past year, and has achieved generally acceptable discrimination performance over the set of minimal pairs seen in therapy to date, although extensive testing is needed.

**USE OF FEATURE ELEMENT SELECTION OVER TIME AND FREQUENCY**

In order to accommodate a wider range of articulatory variability over the range of disorders seen in therapy, it was desired to use a wider time window to capture the CV segments. This generally results in a greater portion of the data in the window to be nondiscriminatory (i.e., silence and vowel segments). It was found that there was significant discriminatory power in individual time frame
feature vectors, and that selection of a group of the best performing time frames within each feature type could also be made based on a rank ordering of the training data error rates for each frame. Incorporating time frame selection resulted in a reduction to 24% of the original error rate when using the best feature type, and to 15% when using feature groups [see previous plot]. Discriminatory power exists over frequency as well, and similar results were obtained using selection of "frequency frames". Further investigation along these lines produced the somewhat surprising result that the most potent discriminatory information for this data is generally found at the level of the individual matrix elements, as illustrated in the following figure. Training data error rates for each matrix element for the p/b discrimination task are shown above for four feature types using a grey scale. It is typical of these matrices that the error rates range from close to (or) zero to maximum (since a small distance ratio threshold must be exceeded), and moreover that a generally poor feature can have small regions where performance is excellent, as seen for feature 1. Feature 15 exhibits multiple discriminatory regions; the effectivenes of temporal over spectral features is also evident - particularly comparing types 11 and 4.

The danger is that at this level of detail, performance on training data will not be a good predictor of performance on test data. The element selection procedure uses an error rate threshold which is the sum of the minimum error rate per feature type plus an "error rate window" specification. As in the frame selection technique, the feature types are rank ordered using the error rates obtained for matrices of the selected elements only. Utilizing matrix element selection produces a significant reduction in error rate, and demonstrates that this is the most effective component of the algorithm; element selection using the default feature results in a reduction to 0.73% of the error rate for the full default matrix, and use of a group of the best features further reduces the error to 0.45% of the reference rate [previous plot]. This is indirect evidence that there is good correlation between training and test data discriminability for individual elements. The system error rate is also not overly sensitive to the error rate window parameter. In addition to improved performance, this technique also reduces computational load, using only a small percentage of the original elements to achieve this level of performance. The following plot shows the relative error rate reduction by consonant pair for each technique, and demonstrates the superiority of matrix element selection for each pair.

Since it is impractical to measure the cross correlations between 2336 individual feature elements during a therapy session, we have taken the expedient of using groups of elements jointly by feature type (as a next step we would like to cluster groups of elements within each feature type and measure correlation between clusters). As a compromise between independent and joint pattern matching over the set of feature types, we have defined and tested a hybrid decision rule. In joint use of feature types a single weighted k-nn distance is obtained to each consonant for the defined group of types. Distances to each token are the sum of distances for each type to that token; distances are normalized using the average distance for that type on training data. The feature types are again used in rank order, but the distance ratios associated with each type are now the D1/D2 and D2/D1 resulting when that type is used jointly with the preceding types in rank order. This allows the better features to have a greater impact on the distance ratio sum (since closer matches and higher ratios will be generally obtained when using fewer, better features), while poorer features contribute in a more controlled fashion; thus the best features are used more independently, and poorer features must work in conjunction with better features. Test results are shown comparing independent ("ratio sum after k-nn"), joint ("normalized distance sum before k-nn"), and hybrid use of feature types. The superiority of joint over independent use appears as the poorest features are included in the decision, but close-

![Performance with multiple feature types](image)

The upper curve is produced using full matrices with independent use of feature types, and shows the effectivenes of individual feature element selection in utilizing discriminatory information from otherwise poor feature matrices, since the dip in the (independent decision) curve is much broader when using element selection.

If, as previously mentioned, features (cues) are uncorrelated, using feature templates independently allows pattern matching to occur without interference from other features (which may match best to different subsets of training tokens). Conversely, we want to use correlated features jointly in pattern matching for the ability to separate clusters in multidimensional pattern space.
TEST DATA AND RESULTS

Test results have been collected using a set of 14 consonant pairs most often used during therapy [t/d, k/g, m/n, p/b, s/z, f/v, f/th, s/sh, sh/ch/zh (as in treasure), w/f/i]. Twenty repetitions of each consonant pair were recorded in a single session to simulate use in therapy. Data from a total of 2240 utterances from 1 male and 3 female (unimpaired) speakers was recorded for off-line experiments. The training data is defined as the first set of 10 repetitions of each syllable with the test data as the second set; the roles are then reversed to double the number of available trials. The geometric mean of the ratio of change in error rate from the reference system are computed over measurable change in results for the individual consonant pairs. This measure is used in developmental testing in order to prevent the few high error rate pairs from dominating test results. In all recognition tests, error rates for feature, frame, and element ranking are obtained using all training data as previously described, but the "single training token test" is most often used to generate sufficient errors for statistical reliability. In this case tokens from the test set are constrained to match to single tokens from the training set - removing the benefit of the knn rule and multiplying the available number of trials to a total of 181,440 for this data set; data from this test is shown in all plots except the following. Matching to the full training token set allows 2240 trials, with results comparing the reference system (default feature, full matrices) to a system using matrix element selection and feature groups using the "ratio sum after knn" decision rule showed below:

FULL TRAINING DATA SET TEST

![Graph showing error rates for different systems.]

The unacceptable error rates for f/th and sh/ch have been reduced to usable levels, and the total number of errors is reduced by a factor of 6.75. It should be pointed out that in early clinical trials pairs such as f/th and sh/ch were not even attempted, and difficulties were seen with pairs such as p/b and m/n, which do not show high error rates for the test data from unimpaired speakers. This indicates the danger of drawing conclusions as to performance improvement in the clinical setting, and the importance of acquiring a data base from the patient population for testing and further development.

As an additional point of reference, a fully interconnected feedforward neural network using backpropagation training was also tested on this data. This network has been used for experiments with speech and other acoustic data in house. Best results to date are not as good as the reference system (5.7% vs. 2.7% errors). Networks reported in [2] generally achieve comparable performance with knn classifiers; this network may need further tuning, or specific architectural modifications for this type of data.

CLINICAL EXPERIENCE

During training, the patient is prompted to produce each target combination a number of times. The system only registers those productions/approximations judged to be good by the therapist. This constitutes the training data. When the previously described error rate analysis is complete the system enters the practice mode. In the unprompted mode patients interact with the system by saying a target sound/syllable/word. The system responds by showing which sound it "heard", rating the production with a number of stars based on the distance ratio sum. Other rewards/indicators in development include both explicit and implicit prompting - a horse race game, maze game, and a version of "breakout".

In the practice mode patients tend to produce the target syllable very rapidly, slowing down only when the system cannot "recognize" the sounds. At that time they appear to concentrate on production/placement/features of the target sound, speeding up once more when the system again recognizes and rewards a properly produced sound. Currently, application of the speech recognizer has been done with the speech pathologist present and rating the patients productions as well as providing intervention and guidance as necessary.

Preliminary application shows good agreement with clinical analysis of production as well as greatly increased motivation for articulatory drill. It further shows promise for clinical application as a means of therapeutic drill with minimal supervision.

References


EUROSESPCH '89, Paris, France, September 1989 2276