IMPROVED SPEAKER INDEPENDENT IWRS FOR SMALL Vocabularies

E Dermatas, N Fakotakis, G Kokkinakis

Laboratory of Wire Communications, University of Patras

A speaker-independent isolated word recognition system (IWRS) using pattern matching is presented. The novelty of this new system lies in its structure and in the introduction of new and improved algorithms in several stages. A short description of the overall system is as follows:

An improved endpoint detection algorithm is applied to the input signal to estimate with accuracy the ends of the word to be recognised. The signal is then divided into a fixed number of frames in order to avoid the deforming of the linear time warping during the classification procedure. Three parameter vectors are extracted from these frames: An energy vector, a zero crossings vector and a PARCOR coefficients vector. The latter is formed through quantization of the PARCOR coefficients vector of each frame. In the recognition stage, three distances are calculated corresponding to the three parameter vectors and the final decision is taken by using a decision function based on the three distances and on the intra- and inter-word distance distributions of the above vectors. In the training procedure multiple templates are created for each word of the vocabulary, taking into account not only the training patterns of this word but also that of all other words.

The improved algorithms used in the presented IWRS are the following:

i. Endpoint detection algorithm: The signal is scanned forward and backward with a certain frame and the corresponding time functions of the energy ratio of two successive non-overlapping frames are estimated. The beginning of the word is set at the point between the two successive frames whose forward energy ratio gives the first peak above a prescribed threshold. Scanning the signal backward, when the energy ratio of two successive frames is found to be above a given threshold, the zero crossings in the right frame are measured. If they are below a certain number, the end of the word is set at the point between the two frames assuming that the right frame contains noise. Otherwise, the endpoint is set at the right end of the right frame assuming that this frame contains speech.

ii. The codebook design algorithm: The codebook is estimated from a large set of speech frame parameter vectors. These are first non-linearly transformed to give a new set of vectors which are furthermore separated in clusters. Then, a modified k-means with splitting algorithm is applied and an initial codebook is estimated. After that a simple k-means iterative algorithm leads to the final codebook. This procedure gives better results in accuracy and in the number of empty cells than other algorithms, eg the LBG and MKM.

iii. The reference templates creation algorithm: For each word multiple templates are created taking into account not only the training patterns of this word but also that of all other words. This algorithm is based on conventional k-means iterations and averaging techniques. The optimal number of templates per word is calculated using intra- and inter-cluster distance statistical parameters and a given maximum classification error. The advantages of this algorithm are a smaller number of reference templates, less computation...
time during the recognition phase and higher accuracy.

iv. The time warping path estimation algorithm: This algorithm is based on dynamic programming (DP) and ensures that the continuity conditions of DP are satisfied. This leads to a better matching of the word patterns and a better distance measure. The latter is calculated as the sum of the frame distances along the warping path.

v. Classification algorithms: The recognition phase has three discrete levels: rejection, comparison and decision. In the rejection level, the energy and the zero crossings vectors are used to exclude words of the reference set which show great dissimilarities to the unknown pattern. In the comparison level, the distances between the unknown pattern and the non-rejected reference patterns are calculated for all three parameter vectors. In the final level, the three distances are used to calculate a decision criterion for classifying the unknown pattern as follows: Assuming that the probability density functions of both intra- and inter-cluster distance distributions of all words are exponential, the ratio of the intra- to the inter-cluster distance probability is calculated from the values corresponding to the distance between the unknown pattern and the reference pattern for each parameter vector. The product of the three ratios gives the decision criterion. The unknown pattern is classified in the word with the highest product.

The presented IWRS has been implemented in C-language on a Micro-Vax II computer. Its performance was tested in a number of experiments with 34 Greek words, (the 10 digits and the 24 letters of the Greek alphabet) and 10 male speakers. In all tests, the performance of the stages and that of the whole system were measured and compared to the performance achieved by applying other commonly used algorithms. The results have shown a better accuracy and a smaller response time of the proposed IWRS while its training and is more time consuming.