

A SELFORGANIZING CLUSTERING TECHNIQUE FOR VECTOR QUANTIZATION IN SPEECH RECOGNITION

A. Aktas, L. Gläßer, B. Kämmerer, W. Küpper.*

ABSTRACT

For the sake of data reduction in automatic speech recognition often vector quantization based on a previously generated code book is performed. In the approach described here the necessary code book is set up by means of a selforganizing clustering technique. It takes the shape of a two-dimensional array of feature vectors. Phonetically similar vectors are also arranged in geometrical vicinity. The definition of a new distance measure suitable for this so-called phonotopic map is introduced. The procedure has been implemented for an isolated-word recognition system for large vocabularies (1,000 words). From a small number of phonetically balanced training utterances (17 words) a map of size 10x10 is built. A recognition rate of more than 98 per cent is achieved with single training of the lexicon when the phonotopic map is used as code book in combination with the proposed distance measure.

INTRODUCTION

For the purpose of speech recognition a rather large degree of data reduction is mandatory in order to accommodate memory size and computational capacity of the available hardware. After a first data reduction during signal analysis and an optional one by temporal segmentation of the utterance, often vector quantization is chosen. In this approach the total feature space spanned by speech is mapped to a limited number of feature vectors, a *code book*. The clustering techniques which are usually employed for its generation are based on minimizing the average distortion between feature vectors of a large training set and the corresponding code book vectors. Investigations with *adaptive arrays* led Kohonen (ref 1, 2) to develop a method for creating a map of feature vectors which is selforganized according to phonetic properties. Based on this work, we demonstrate in this paper the utilization of such a map as code book for vector quantization and propose a distance measure derived from its properties.

FUNDAMENTAL CONCEPT

Consider a two-dimensional adaptive array $\mathbf{A}(I,J)$. To each grid point (i,j) of this array a weight vector $W_{i,j}(t)$ is attached which is initialized with random numbers.

The learning stage consists of two phases. In both phases each input vector $V(t)$ is compared with all lattice points. By means of a simple distance measure the point of the array which most closely corresponds to the input vector is determined. The weights of the chosen point and its surroundings are adapted to the input vector. The adaptation factor $a(t)$ decreases with the number t of adaptation steps already performed.

*Siemens AG, Corporate Research and Technology Division,
Corporate Laboratories for Information Technology, Signal Processing,
Otto-Hahn-Ring 6, D-8000 Munich 83, West Germany.

The prescription for adapting the weights in each step t is as follows:

$$W_{i,j}(t+1) = \begin{cases} W_{i,j}(t) + a(t) [V(t) - W_{i,j}(t)] & \text{for } (i,j) \text{ within the adaptation radius,} \\ W_{i,j}(t) & \text{for } (i,j) \text{ beyond the adaptation radius.} \end{cases} \quad (1)$$

The adaptation factor $a(t)$ decreases linearly with the number t of adaptation steps as:

$$a(t) = \begin{cases} c_1 (1 - t/T_1) & \text{during the first phase } (0 < t \leq T_1), \\ c_2 (1 - t/T_2) & \text{during the second phase } (T_1 < t \leq T_2). \end{cases} \quad (2)$$

The adaptation radius $r(t)$ is reduced linearly during the first learning phase and stays constant during the second phase:

$$r(t) = \begin{cases} R + (1 - R) t/T_1 & \text{during the first phase,} \\ 1 & \text{during the second phase.} \end{cases} \quad (3)$$

Thus the array \mathbf{A} is adapted rather quickly and strongly during the first learning phase. The selforganizing aspect of the method is emphasized here, and at the end of the first phase the gross structure and the orientation of the map are fixed, while during the second phase the fine structure of the weight vectors is generated without alteration of the spatial ordering of the map.

APPLICATION TO SPEECH PROCESSING

If feature vectors of analyzed speech are chosen for input vectors $V(t)$, the resulting array \mathbf{A} can serve as a code book for vector quantization. Feature vectors with similar properties (e.g. vectors of a single phoneme or of similar phonemes) are associated with neighbouring regions of the code book, and the extension of these regions corresponds to the statistics of the speech material employed for training. In this way we obtain a *phonotopic map* (ref 2) for the feature space spanned by speech which reflects the acoustic-phonetic properties and statistical information of the training material.

For the training of this map the utilization of a few words chosen from a phonetic point of view is sufficient. In each adaptation step a word is randomly chosen out of this small training set. By offering a whole word at a time the statistical dependence of the temporal sequence of vectors is taken into account, while by randomly choosing succeeding words an unbalanced adjustment is avoided.

DERIVED DISTANCE MEASURE

The ability of selforganization suggests a new measure of similarity for the feature vectors. The distance between two vectors can be defined as the geometrical distance of the corresponding points in the phonotopic map. Contrary to the Euclidian or city-block distance, with this distance measure a homogeneous distance distribution is obtained. For an appropriately determined map the new distance measure also reflects the statistical relationship between the ordered feature vectors. Apart from this, the new distance can be calculated much faster, since the computational effort is linearly proportional to the dimensionality of the vectors to be compared, and this is of course smaller for the two-dimensional map than for the original multi-dimensional feature space.

RESULTS

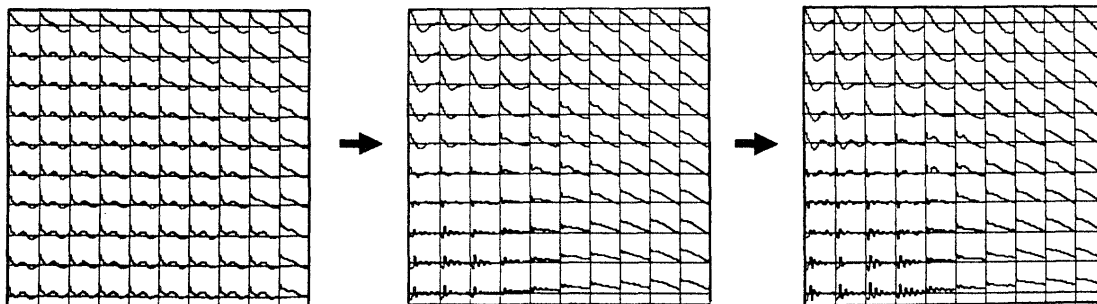
For the following investigations the speech signal was sampled with 12.5 kHz and every 8 ms a short-time autocorrelation analysis with a window size of 10 ms was carried out. The first 16 coefficients, which are normalized to the signal energy, form the feature vector for the corresponding time frame. For the non-linear time alignment between test and reference word the technique of Dynamic Time Warping (DTW) on the base of a Dynamic Programming algorithm was used. These basic characteristics correspond to the recognition system described in (ref 3).

The size of the phonotopic map was chosen to be 10×10 ; the maximum adaptation radius $R = 10$ corresponded to the grid size of the map. The number of adaptation steps in the first learning phase was $T_1 = 10,000$ with an initial factor $c_1 = 0.1$; the corresponding values for the second learning phase were $T_2 = 90,000$ and $c_2 = 0.008$.

The training set consists of merely 17 words chosen by means of phonetic considerations; it contains all phonemes of the German language.

Some snapshots during the generation of a phonotopic map are shown in fig. 1. The leftmost array displays the weight vectors shortly after the beginning, the middle one at the end of the first learning phase. The rightmost array shows the final map after learning has been completed.

Fig. 1: Examples of weight vector values during training of the map



Number of adaptation steps performed:

$t = 1,000$

$t = 10,000$

$t = 100,000$

In a first test we investigated the ability of the proposed method to generate a code book from a given training set which is also well suited for other words. For this purpose, 50 words not contained in the training set were vector quantized and the average distance between original feature vectors and corresponding code book representatives was considered. The investigation was performed for several speakers and yielded small distortions in the case of both the phonotopic map and the code book based on k-means clustering. For a code book generated by a hierarchical clustering technique, however, considerably larger distortions were found.

Recognition tests were performed with a technical vocabulary of 1,000 German words. The lexicon words were trained once and stored away as quantized sequences of feature vectors. The test utterances were also vector quantized in turn, thus rendering possible the employment of the proposed distance measure for similarity evaluation. For several speakers an average recognition rate of more than 98 per cent was achieved. In order to obtain the same recognition accuracy in the case of hierarchical clustering and utilization of the city-block distance, one has to use a code book of a size larger than 1,000.

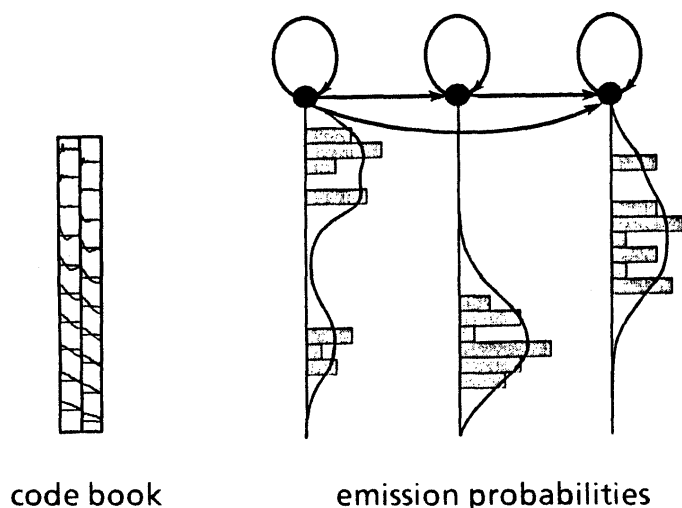
CONCLUSION AND FURTHER WORK

The introduction of a selforganizing clustering technique leads to a well-suited code book for vector quantization in speech recognition. A new distance measure developed according to the properties of the code book yields both high recognition accuracy and fast response time.

Future work will be concerned with the optimal design of the phonotopic map for speech. The required number of dimensions has to be investigated, since a one-dimensional array may well be sufficient. On the other hand, the question of the boundary of the map is not settled. In the two-dimensional case, for example, a closed surface (e.g. a sphere) can be imagined, which may lead to a better representation because of the resulting larger number of neighbourhoods.

Apart from the use of the clustering technique for speech recognition systems based on pattern matching by DTW, the utilization for discrete hidden Markov models (HMM) is imaginable. During the training stage, a modification of vanishing emission probabilities for these HMM is necessary. Usually all probabilities are forced to exceed a small value (epsilon modification). With a one-dimensional phonotopic map for the code book, the modification may be achieved quite easily and consistently using a Gaussian filtering for the emission probabilities. An example of this procedure is shown in fig. 2. Here the emission probabilities have been smoothed, producing meaningful emission probability values also for code book entries not available during the training of the HMM.

Fig. 2: Utilization of a one-dimensional map as code book for discrete HMM



REFERENCES

1. T. Kohonen: "Clustering, Taxonomy, and Topological Maps of Patterns", Proc. IEEE ICPR Munich (1982) 114.
2. T. Kohonen, K. Mäkisara, T. Saramäki: "Phonotopic Maps. - Insightful Representation of Phonological Features for Speech Recognition", Proc. IEEE ICPR Montreal (1984) 182.
3. A. Aktas, B. Kämmerer, W. Küpper, H. Lagger: "Large-Vocabulary Isolated Word Recognition with Fast Coarse Time Alignment", Proc. IEEE ICASSP Tokyo (1986) 709.