A Neural Network-Based Speaker-Independent System for Word Recognition in Romanian Language

Dragoş Burileanu, Mihai Sima, Corneliu Burileanu, and Victor Croitoru

«Politehnica» University of Bucharest, Faculty of Electronics and Telecommunications, Blvd. Iuliu Maniu 1-3, sect. 6, Bucharest 77202, Romania, phone: (+40.1) 410.64.45, fax: (+40.1) 411.11.87, bdragos@mESsnet.pub.ro, sima@ADComm.pub.ro, cburileanu@mESsnet.pub.ro, croitoru@ADComm.pub.ro

Abstract. This paper describes the design philosophy of a speaker-independent system for isolated-spoken word recognition in Romanian language. First, two main methods utilized for speech recognition are briefly discussed. We then present the structure of the proposed system, based on a modified Learning Vector Quantization algorithm combined with a Dynamic Programming technique. The paper ends with a comparative set of experiments, conclusions and intended further work.

1 Introduction

Between the numerous models proposed for the task of automatic spech recognition in the last ten years, neural networks (NN) take an important place. Widely used in several domains, including speech processing, neural algorithms offer alternatives to classical techniques and have an important potential for implementing discrimination, nonlinear feature extraction, or classification based on the distance to learned reference patterns. Based on the fact that speech recognition is essentially a pattern recognition problem, which obviously requires the processing of massive amount of data, the performed computations being repetitive and parallel, it is a natural idea to use the NN paradigm for solving such a problem. Many experiments were made ranging from isolated-spoken word recognition on small vocabularies to continuous speech recognition [2], [4].

However, training a NN to recognize by itself the speaker-independent speech is not a trivial task, due to the large amount of labelled speech data to be handled. On the other hand, a classical technique like the Dynamic Programming (DP) algorithm, even if it has been proven successful in modeling the temporal structure of the speech signal, it is not capable of representing the wide variety of spectrum pattern variations. So, the key idea of our work is a combination between the high time alignment capabilities of DP with the flexible learning function of the neural paradigm in an attempt to increase the temporal invariance of isolated templates and consequently achieve shift-invariant speaker recognition task. Several activities have already been reported regarding this subject, most of them performing the recognition at the phoneme level [3], [4], [8].

2 Speech Recognition Methods

Speech recognition systems generally consist of four stages [4]: *signal processing*, *feature extraction*, *time alignment and pattern matching*, and *speech recognition decision*.

The signal processing stage processes the sampled speech signal in order to perform basic spectral and temporal measurements. The feature extraction stage computes a set of parameters, usually at fixed-time intervals.

Time alignment and pattern matching algorithms – the most important part of a general speech recognition model – attempt to match a spoken word against a given representation of that word. Time alignment refers to the alignment of acoustic or phonetic events which have been modeled, to those that are *time warped* in the utterance due to changes in speaking rate. The simplest way to recognize an isolated-spoken word is to compare it against a number of stored word templates and determine which is the *best match*. But the rate of speech may not be constant throughout the word, the optimal alignment between a template and the speech sample being a nonlinear function.

Dynamic Time Warping (DTW) is an efficient algorithm for finding this optimal alignment [6]. The algorithm finds a global alignment path by computing locally optimized segments at each step. The resulting path may be stored by means of a table (Fig. 1). An optimal alignment path is computed for each reference word template, and the one with the lowest cumulative score is considered to be the best match for the unknown speech sample.

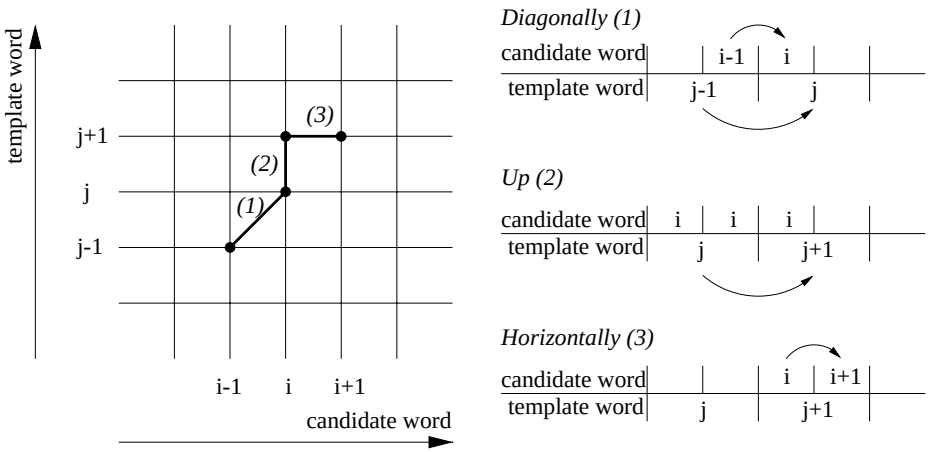

Fig. 1. The process of computing and storing the DTW path

Learning Vector Quantization (LVQ) is a pattern classification method and can also be used in this stage of a speech recognition system, as a connectionist approach. In a LVQ network, each output unit represents a particular class.

Based on a nearest neighbours principle, LVQ is similar to vector quantization; therefore, the weight vector for an output unit is often referred as a *reference* or *codebook vector* for the class that the unit represents. Recognition is performed by computing the nearest reference to the unknown pattern which will be classified according to this reference vector [1], [6].

3 The Basic Structure of an Isolated Digit Recognition System

At present, our speaker-independent system for isolated-spoken word recognition is utilized to recognize the ten digits in Romanian language. A separate network is created for each of the ten digits (Fig. 2). Each network has only one output unit which is scaled by the number of input units, to provide a normalized output score (Fig. 3). Therefore, one cannot talk about a classical LVQ algorithm, but rather a *modified LVQ* algorithm.

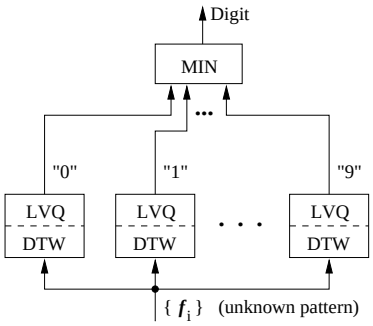

Fig. 2. The structure of the speaker-independent system for isolated-spoken word recognition

The DTW module transforms the variable-length feature set into the fixed number of LVQ network inputs defined during the training phase. Each LVQ network define a word template and calculates the Euclidian distance along the DTW path.

The final recognition decision is performed by the *minimum detector* (Fig. 2).

Two tasks have to be done during training: *initialization* of the LVQ networks (both defining the number of input units for each network and initializing the weight vectors) and *updating* the reference vectors.

Initialization of each LVQ Network. To determine the number of input units we computed the medium length of all the utterances corresponding to the digit being modeled in the data-base, spoken with medium speed. Then we chose the utterance with the length closest to that medium length. The *number*

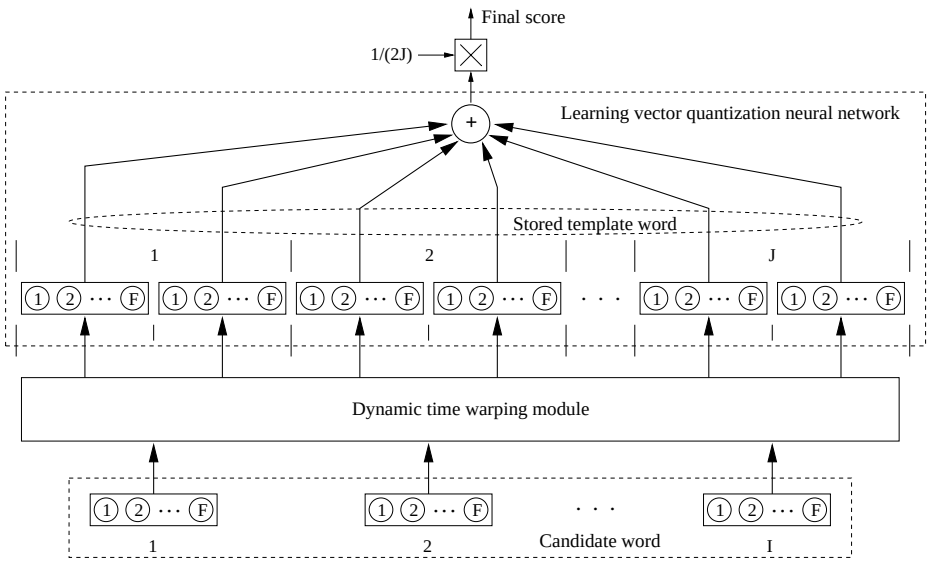

Fig. 3. The hibrid LVQ+DTW module

of the input units for the network is set up at a value of twice the number of frames of the chosen utterance.

To initialize the weight vectors $\boldsymbol{t}_i$, we take twice each parameter vector $\boldsymbol{f}_j$ of the chosen utterance and use them as weight vectors:

$$(\boldsymbol{t}_1\, \boldsymbol{t}_2\, \boldsymbol{t}_3\, \boldsymbol{t}_4\, \ldots\, \boldsymbol{t}_{2J}) = (\boldsymbol{f}_1\, \boldsymbol{f}_1\, \boldsymbol{f}_2\, \boldsymbol{f}_2\, \ldots\, \boldsymbol{f}_J\, \boldsymbol{f}_J) \quad . \tag{1}$$

Updating the Reference Vectors. Updating the reference vectors is performed as much as once every training epoch. During an epoch, each network is tested with an exemplar of all other words in the vocabulary. If the right network corresponding to the exemplar gives the lowest score, no reference pattern is modified. If an incorrect network gives the lowest score, then the reference patterns are changed in two networks. Each of the corresponding spectral parameters is shifted further from the exemplar in the network that wrongly received the best score:

$$\boldsymbol{t}\,(\text{new}) = \boldsymbol{t}\,(\text{old}) + \alpha\,[\boldsymbol{f} - \boldsymbol{t}\,(\text{old})\,] \tag{2}$$

and closer to the exemplar in the network that did not achieve the best score:

$$\boldsymbol{t}\,(\text{new}) = \boldsymbol{t}\,(\text{old}) + \alpha\,[\boldsymbol{f} + \boldsymbol{t}\,(\text{old})\,] \quad . \tag{3}$$

4 Experimental Results

As we have mentioned, a way to recognize an isolated word sample is to compare it against a number of stored word templates and determine what is the *best match*.

Based on a previous work describing the feature modification in Romanian spoken words (typically a change in speaking rate, which leads to a time axis modification, and three main variations in prosodic contour of the ten digits pronunciation), we conceived a database including speech data uttered by 14 speakers (7 males and 7 females). Every speaker uttered each of the 10 digits with 3 speaking rates (slow, medium, high), and 3 prosodic countours (up, horizontal, down).

The data-base was recorded with a close-speaking microphone in a room with normal ambiental noise. Recordings were digitized at a sampling rate of 8 kHz. We performed an overlapping analysis with a window duration of 32 msec and a frame duration of 16 msec.

We used several methods to determine speech parameters [7], only the cepstral analysis being presented here. Performing a cepstral analysis, we obtained 10 Fourier-Transform derived cepstral coefficients every 16 msec. To reduce the variability of the data and achieve a speaker normalization, we also performed a classical *prewhitening transform*. Due to the fact that cepstral parameters can be regarded as uncorrelated [5], the transform reduces to a diagonal matrix. Therefore, by applying this transform, each of the cepstral parameters, after subtracting their mean value, is scaled by its standard deviation.

We performed three sets of experiments under rather different experimental conditions. In the first set, we used a classical DTW algorithm along with 10 cepstral coefficients (CEPC-10). In the second set we used the DTW algorithm along with 10 cepstral coefficients which were previously scaled by their standard deviations (SDCEPC-10), and in the last one we used the discussed DTW+LVQ algorithm along with the 10 scaled cepstral coefficients.

The scores are shown in Tables 1, 2, and 3.

5 Conclusions and further work

The experimental results demonstrate that the combination of DTW for aligning the exemplars in time, and the NN learning process for distinguishing the exemplars in frequency and creating discriminatory template patterns for each of the words in vocabulary, provide the basis for an effective speech recognition system.

The hybrid proposed approach lead to a recognition rate of about 94% simultaneously with a significant reduction of the computing time, crucial for a speaker-independent recognition task.

An important effort was directed to identification of optimum choice at different levels of processing; we think that only in this way good performances can be obtained (i.e., real time processing and higher recognition rate). For example,

Table 1. CEPC-10 / DTW algorithm

Digit	0	1	2	3	4	5	6	7	8	9	General score
Score (%)	95	89	83	87	94	99	85	74	89	90	88.3

Table 2. SDCEPC-10 / DTW algorithm

Digit	0	1	2	3	4	5	6	7	8	9	General score
Score (%)	95	90	90	85	94	99	82	82	94	90	90.2

Table 3. SDCEPC-10 / LVQ + DTW algorithm

Digit	0	1	2	3	4	5	6	7	8	9	General score
Score (%)	97	94	94	91	97	99	89	90	92	96	93.9

we showed that a carefully conceived data-base may dramatically improve the overall performance of the system.

We recently started a set of experiments of word recognition for a Romanian voice dialing system. Noticing that for such a system the pronounced digits are actually semi-connected, we are about to establish an algorithm based on a time-domain analysis, to extract an isolated digit from a pronounced telephone number. We also aim to extend the proposed algorithms on a vocabulary of about 20 words.

References

1. Fausett, L.: Fundamentals of Neural Networks – Architectures, Algorithms, and Applications. Prentice Hall, Englewood Cliffs, New Jersey (1994)
2. Haton, J.P.: Les modèles neuronaux et hybrides en reconnaissance automatique de la parole: état des recherches. In: Méloni, H. (coordinateur): Fondements et perspectives en traitement automatique de la parole. Aupelf-Uref (1996) 263–281
3. Ikuta, H., Ishida, Y., Honda, T., Arai, Y.: Spoken Word Recognition Using LVQ Based on DP Matching. Proceedings of the International Conference on Signal Processing Applications & Technology. Boston (1996) 1755–1758
4. Morgan, D., Scofield, C.: Neural Networks and Speech Processing. Kluwer Academic Publishers, Boston Dordrecht London (1994)
5. Picone, J.: Signal Modeling Techniques in Speech Recognition. Proceedings of the IEEE, Vol. 81, No. 9. IEEE (1993) 1215-1247
6. Rabiner, L.R., Juang, B-H.: Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs, New Jersey (1993)
7. Sima, M., Croitoru, V., Burileanu, D.: Performance Analysis on Speech Recognition using Neural Networks. Proceedings of the International Conference on Development and Application Systems. Suceava, Romania (1998) 259–266
8. Yoshimura, M., Honda, T., Arai, Y.: Japanese Spoken Word Recognition Using LVQ Based on Phonetic Units. Proceedings of the International Conference on Signal Processing Applications & Technology. Boston (1996) 1750–1754