THE APPLICATION OF NEURAL NETWORK PARADIGMS IN SPEECH RECOGNITION FOR A ROMANIAN VOICE DIALING SYSTEM

Dragoş BURILEANU[*], Mihai SIMA, Corneliu BURILEANU, Victor CROITORU

I Introduction

The development of robust, speaker independent, speech recognition systems that perform well over dialed-up telephone lines has been a topic of interest for over a decade. This work has progressed from systems that can recognise a small number of vocabulary items spoken in isolation, to systems that can recognise medium size vocabulary sets spoken fluently [7].

A basic idea is that even if Dynamic Programming (DP) or hidden Markov Models (HMM) have been proven successfully in modeling the temporal structure of the speech signal, they are not capable of assimilating a wide variety of speaker-dependent spectrum pattern variations. For example, *the first-order assumption* — which says that all probabilities depend solely on the current state — and *the independence assumption* — which says that there is no correlation between adjacent input frames — of HMMs, are false for speech applications.

On the other hand, it has been reported that some neural network models (LVQ, SOM) are effective for clustering of phonemes; however, the correct

[*] All authors are with «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, e-mail: {bdragos,cburileanu}@mESsnet.pub.ro, {sima,croitoru}@ADComm.pub.ro.

segmentation of phonemes is necessary and it is obvious that is very difficult to segment speech data into phonemes.

The main idea of our approach is to generate fixed-dimensional acoustic feature vectors from variable-dimensional speech signal and to classify these static time-normalized acoustic feature vectors with a discriminative classifier. The time-normalized vectors are created by segmentation based on a Dynamic Time Warping (DTW) algorithm. The temporal information of speech is therefore preserved. The classifier is based on a Learning Vector Quantization (LVQ) distance classifier.

II A Voice Dialing System in Romanian

The key idea of our work is that for a standard voice dialing system, the pronounced digits are actually semi-connected, the task being easier than a fully connected word recognition application. Our experiments proved that for such a system the pronounced digits are usually separated by silent periods between 30-100 msec. Consequently, we adapted an algorithm [2] based on a time-domain analysis, in order to extract isolated digits from a pronounced telephone number. The algorithm uses two time-domain measurements — energy, and zero-crossing rate — for the start-point and end-point detection of each word.

II.1 The overall system architecture

The Voice Dialing System contains a separate DTW/LVQ module for each of the ten digits. The DTW module transforms the variable-length feature set into a

fixed number of LVQ network inputs. Each LVQ network defines a word template and calculates the Euclidian distance along the DTW path. The recognition decision is performed by a minimum detector.

In the last time, we experimented several methods to determine proper speech parameters for the recognition task [4,6]. In this paper we propose to use the first-order cepstral parameters in addition to zero-order cepstral parameters.

For this purpose, we calculated first-order time derivative of the cepstral coefficients. Making use of the *Taylor's formula* [1,6] for a function *f(x)*, we can write:

$$\left.\begin{aligned} f(x+h) &= f(x)+\frac{h}{1!}f'(x)+\frac{h^2}{2!}f''(x)+\frac{h^3}{3!}f'''(x)+\cdots \\ f(x-h) &= f(x)-\frac{h}{1!}f'(x)+\frac{h^2}{2!}f''(x)-\frac{h^3}{3!}f'''(x)+\cdots \end{aligned}\right\} \Rightarrow$$

$$\Rightarrow f(x+h)-f(x-h) \approx 2hf'(x)+2\frac{h^3}{3!}f'''(x)$$

and

$$\left.\begin{aligned} f(x+2h) &= f(x)+\frac{2h}{1!}f'(x)+\frac{(2h)^2}{2!}f''(x)+\frac{(2h)^3}{3!}f'''(x)+\cdots \\ f(x-2h) &= f(x)-\frac{2h}{1!}f'(x)+\frac{(2h)^2}{2!}f''(x)-\frac{(2h)^3}{3!}f'''(x)+\cdots \end{aligned}\right\} \Rightarrow$$

$$\Rightarrow f(x+2h)-f(x-2h) \approx 4hf'(x)+\frac{16h^3}{3!}f'''(x)$$

Then

$$f'(x) \approx \frac{2}{3h}\left[f(x+h)-f(x-h)\right]-\frac{1}{12h}\left[f(x+2h)-f(x-2h)\right]$$

Multiplying the above equation by 12*h* to reduce the computations, we approximate the first-order time-derivative cepstral coefficients by:

$$\dot{c}_k[n] \approx 8\left[c_k(n+1)-c_k(n-1)\right]-\left[c_k(n+2)-c_k(n-2)\right], \quad 0 \le k \le 9$$

Therefore, differenced cepstral coefficients consist of 16x2=32 and 16x4=64 msec differences.

The overall architecture is presented in Fig. 1.

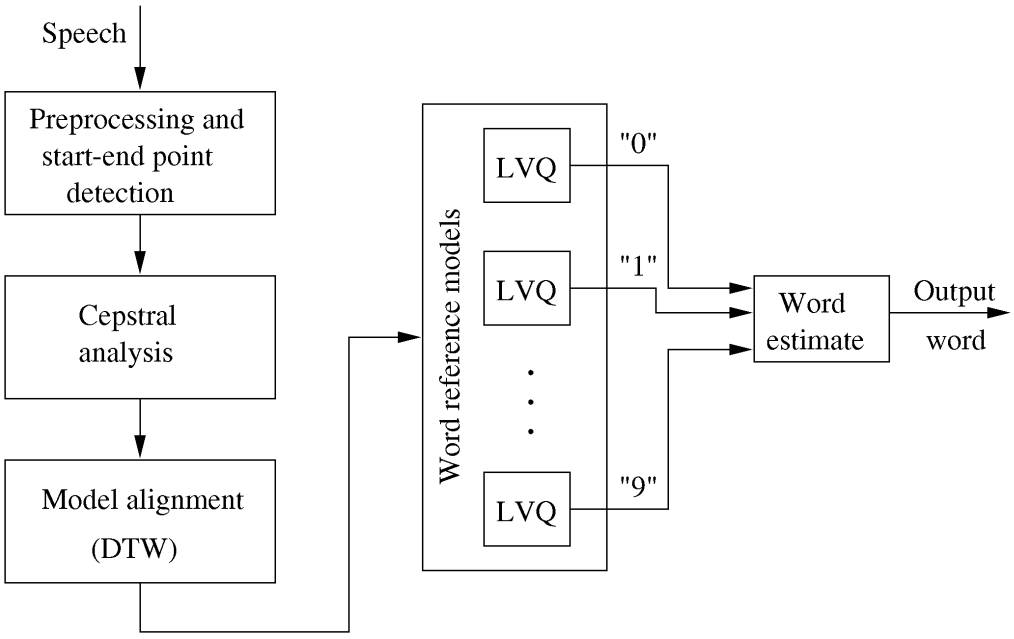

Figure 1 – Block diagram of overall recognition system

II.2 The Training Data-Base

In order to train the LVQ networks, we conceived a data-base 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 contour (up, horizontal, down) [11].

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 computed 10 zero-order and 10 first-order Fourier-derived cepstral coefficients every 16 msec.

III Experimental Results

We performed four sets of experiments under rather different experimental conditions. In the first set, we used a classical DTW algorithm along with 10 cepstral coefficients (DTW/CEPC-10) [6]. In the second set we used a DTW algorithm along with 10 cepstral coefficients which were previously scaled by their standard deviations (DTW/SDCEPC-10) [4], and in the third set we used a DTW+LVQ algorithm along with the 10 scaled cepstral coefficients (DTW+LVQ/SDCEPC-10) [4]. In the last one we used the above discussed DTW+LVQ hybrid system along with 10 scaled cepstral coefficients and 10 scaled first-order time derivative cepstral coefficients (DTW+LVQ/SDCEPC-10+SDΔCEPC-10).

The system has been tested with a separated testing data-base consisting of seven digit telephone numbers uttered by 20 different speakers. Table 1 shows the medium scores for the testing data-base at the word level in the first column, and the medium scores obtained at the string level in the second column.

Table 1. The experimental results for a Voice Dialing System

Experimental conditions	Scores at word level (%)	Scores at string level (%)
DTW/CEPC-10	88.3	-
DTW/SDCEPC-10	90.2	-
DTW+LVQ/SDCEPC-10	93.9	-
DTW+LVQ/SDCEPC-10+SDΔCEPC-10	97.1	...

4 Conclusions

A standard algorithm based on a time-domain analysis provided good performance to transform the semi-connected digit recognition task for a voice-dialing application in a standard isolated word recognition one.

The experimental results demonstrate that the hybrid approach DTW+LVQ and also the using of high-order cepstral parameters provide the basis for an effective speech recognition system.

Bibliography

[1] Bauche, É., Gajic, B., Minami, Y., Matsuoka, T., Furui, S.: Connected Digit Recognition in Spontaneous Speech. Proceedings of Eurospeech97. Rhodes, Greece (1997) 923-926

[2] Bucur, C.M.: Metode numerice. Editura Facla, Timişoara (1973)

[3] Burileanu, C.: PhD dissertation. «Politehnica» University of Bucharest (198…)

[4] Burileanu, D., Sima, M., Burileanu, C., Croitoru, V.: A Neural Network-Based Speaker-Independent System for Word Recognition in Romanian Language. Proceedings of the "Text, Speech, Dialog" Conference. Brno, Czech Republic (1998)

[5] Sakoe, H.: Dynamic Programming-Based Speech Recognition Algorithms. In: Furui, S., Sondhi, M.: Advances in Speech Signal Processing. Marcel Dekker, New York Basel Hong Kong (1992) 487-508

[6] 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

[7] Wilpon, J., Rabiner, L., Lee, C-H., Goldman, E.: Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models. In IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 38, No. 11 (1990) 1870-1878