komunikácie
COMMUNICATIONS

Matúš Jurečka *

FEATURE EXTRACTION USING PULSE-COUPLED NEURAL NETWORK IN ISOLATED SPEECH RECOGNITION

This article presents achieved results concerning an feature extraction in isolated speech recognition problem using the Pulse-Coupled Neural Network (PCNN) approach. PCNN based feature extraction is analyzed for a direct Pulse Coded Modulation (PCM) input and a Fast Fourier Transform (FFT) coefficients input.

1. Introduction

The speech recognition problem may be interpreted as a speech-to-text conversion problem. A speaker wants a voice to be transcribed into text by a computer. Automatic speech recognition has been an active research topic for more than four decades. With the advent of digital computing and signal processing, the problem of speech recognition was clearly posed and thoroughly studied. These developments were complemented with an increased awareness of the advantages of conversational systems. The range of the possible applications is wide and includes: voice-controlled appliances, fully featured speech-to-text software, automation of operator-assisted services, and voice recognition aids for the handicapped persons. Different approaches in speech the recognition have been adopted. They can be divided mainly into two trends – hidden Markov model (HMM) and artificial neural network (ANN).

2. Speech signal processing

Speech acquisition begins with a person speaking into a microphone. Generally, a speech signal is converted onto a digital form using the pulse coded modulation (PCM). This means of speech signal representation is not so suitable for a pattern recognition. However, it can be represented by a limited set of features. There are several methods available for features extraction and dimension reduction. The dimension reduction is a transformation of an input signal space into a feature space with a lower dimension. The goal of the dimension reduction is to obtain significant features for a unique pattern representation. Classical methods of dimension reduction include Karhunen – Lo_ve transform [11], singular value decomposition (SVD), etc.[5] Dimension reduction methods based on ANN are for example Kohonen Self-Organized maps [6] or principal component analysis (PCA neural networks) [10, 9].

Classical methods of features extraction in digital signal processing for speech recognition include coefficients of discrete Fourier transform, linear predictive coefficients (LPC), filter bank, mel scale frequency cepstral coefficients (MFCC) etc. [7,8] The method scoring growing interest for a dimension reduction & feature extraction in a field of image processing is the Pulse Coupled Neural Network (PCNN). My work focused on the feature extraction in isolated speech recognition process using the PCNN.

3. A PCNN structure

The structure of a standard PCNN comes out from the structure of an input pattern which will be processed. Let us consider that the input pattern is a matrix of values for input of an isolated word. The PCNN is a single layered, two-dimensional, laterally connected neural network of pulse coupled neurons connected with values of an input matrix. Each input matrix value is associated with a pulse coupled neuron of a specific structure. The PCNN neuron consists of an input part, linking part and a pulse generator. The neuron receives the input signals from feeding and linking inputs. The feeding input is a primary input from the neuron's receptive area. The neuron receptive area consists of neighboring values of the corresponding value in the input matrix. The linking input is a secondary input of lateral connections with neighboring neurons. The difference between these inputs is that the feeding connections have a slower characteristic response time constant than the linking connections. The standard PCNN model is described as iteration by the following equations:

$$F_{ij}(n) = S_{ij} + F_{ij}(n-1) \cdot e^{-\alpha F} + V_F \cdot (M * Y(n-1))_{ij} \quad (1)$$

$$L_{ij}(n) = L_{ij}(n-1) \cdot e^{-\alpha L} + V_L \cdot (W * Y(n-1))_{ij} \quad (2)$$

$$U_{ij}(n) = F_{ij}(n) \cdot (1 + \beta \cdot L_{ij}(n)) \quad (3)$$

$$\Theta_{ij}(n) = \Theta_{ij}(n-1) \cdot e^{-\alpha\Theta} + V_\Theta \cdot Y(n-1) \quad (4)$$

$$Y_{ij}(n) = \begin{cases} 1 \text{ if } U_{ij} > \Theta_{ij} \\ 0 \text{ otherwise} \end{cases} \quad (5)$$

* **Matúš Jurečka**
Department of Technical Cybernetics, Faculty of Management Science and Informatics, University of Žilina, Veľký diel, 010 26 Žilina, Slovakia, jurecka@frtk.fri.utc.sk

Where F_{ij} is the feeding input, L_{ij} is the linking input, n is an iteration step, S_{ij} is a value at i,j coordinates in the input matrix. W and M are the weight matrices, is the convolution operator, Y_{ij} is the output of the neuron at i,j coordinates, V_L and V_F are potentials, α_L and α_F are decayed constants.

Single signals of the linking input are biased and then multiplied together. Next, the input values F_{ij}, L_{ij} are modulated in the linking part of a neuron. We also obtain internal activity of the neuron U_{ij} in the specific iteration step. If internal activity is greater than dynamic threshold Θ_{ij}, then the neuron generates output pulse. Otherwise, the output equals to zero. The neuron output Y_{ij} does not necessarily need to be binary. It is possible to use a sigmoid pulse generator where the neuron takes the analogue value from 0 to 1. The input matrix is transformed through the PCNN into a sequence of temporary binary matrixes. Each of these binary matrixes has the same dimension as the input matrix. The sum of all activities in a specific iteration step gives one value representing one feature for the classification. If we have N iteration steps, we obtain N features. The one-dimensional time signal generated from the values of the output matrix $Y_{ij}(n)$ in every iteration step n can be defined as follows:

$$G(n) = \sum_{i,j} Y_{ij}(n) \tag{7}$$

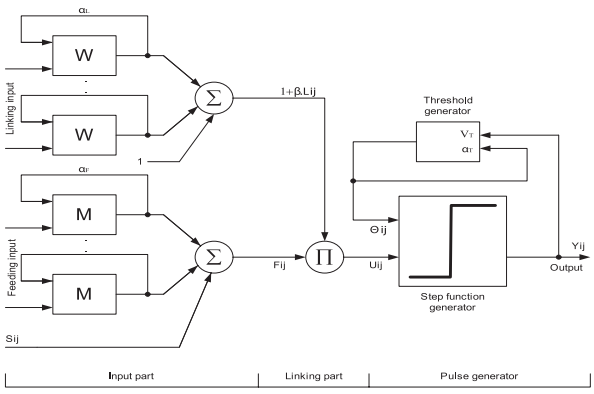

Fig. 1: PCNN neuron structure - taken from [4]

Significant advantage of the PCNN, which is useful mainly in image recognition, is the invariance of a generated time signal to rotation, dilatation or translation of images [4]. Therefore, the PCNN is advisable for the feature generation and pattern recognition in the classification tasks using conventional neural networks or other methods. Thanks to translation invariance of generated features, the PCNN method used in speech recognition does not relay on an outstanding endpoint word detection. It is evident that the PCNN is not the neural network in the term of classification. It is only a means of feature extraction for a pattern classification using conventional neural network models, like that of multi-layer perceptron. Several models of the PCNN have been developed. The most used PCNN models are, for example, a PCNN with modified feeding input [1], fast-linking PCNN [3] or feedback PCNN [2].

4. Experiments and Results

The following experiments with feature extraction using the PCNN were made in my testing database consisting of 36 isolated Slovak words uttered once by 23 different speakers:

The abovementioned PCNN approach was applied directly to a sequence of PCM of an isolated word. I used 16-bit PCM with f_s = 8kHz. The PCNN with 200 iteration steps produced 200 features for every isolated word. The 200x1 feature vectors $G_{ab}(n)$ where a is the word index ($1 \le a \le 36$) and b is the speaker index ($1 \le b \le 23$), were than divided by their maximum values for certain normalization reasons. Figure 2 shows the mean courses of $G_a(n)$ functions for all 36 input words, the mean course of $G_a(n)$ function for the input word a was computed as follows:

$$G_a(n) = \frac{1}{23} \sum_{b=1}^{23} G_{a,b}(n) \tag{8}$$

In my next experiment I used the Fourier transform for the PCNN input coefficients. The sequence of PCM values was partitioned into the sequence of consecutive frames. The frame length was chosen as 128 samples with 64 samples overlap. After applying the Hamming window function which prevents some spectral leakage, the fast Fourier transform (FFT) was computed in these frames. The input matrix was formed by the first 64 FFT coefficients (as the product of FFT is symmetrical) from every time frame. The PCNN with 200 iteration steps produced 200 features for every isolated word. The feature vectors were then divided by their maximum values for certain normalization reasons. Figure 3 shows the mean courses of $G_a(n)$ functions for all 36 input words computed similarly as in my first experiment.

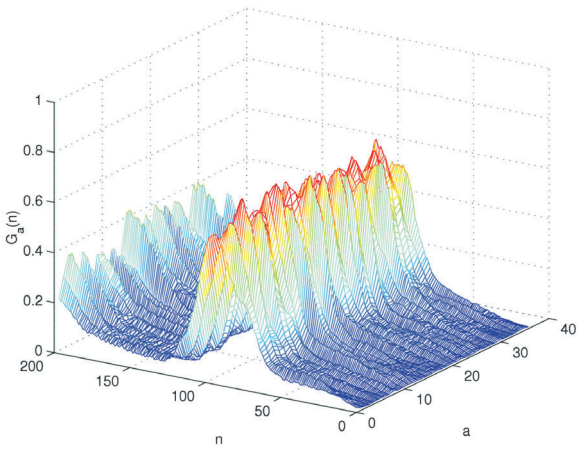

Fig.2 Mean courses of $G_a(n)$ functions for PCM input

Figure 4 shows the variance of the mean courses of $G_a(n)$ functions for input words No. 4, 5, 6, 7 with their 95% reliability level for FFT coefficients input. The experiments described were carried out in the MATLAB environment.

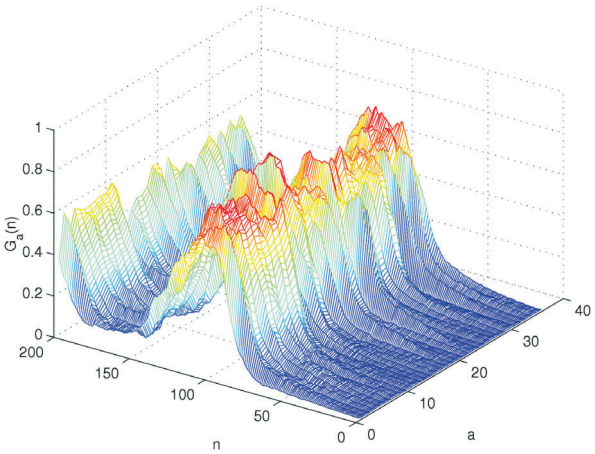

Fig. 2 Mean courses of $G_a(n)$ functions for PCM input

5. Conclusion

It is clear that for recognition it is very important for the feature vectors $G_{ab}(n)$ to be as similar as possible to the same word uttered by different speakers and at the same time these vectors should be as different as possible in different words. Difference measure can be understood, for example, as the Euclidean distance between the feature vectors. As it can be seen from the above mentioned results, the feature vectors for different words do not differ too much. On the other hand, the variance of these vectors for the given input word is much bigger than the difference measure of distinct words. It clearly shows that the speech recognition system which will rely on the introduced PCNN based feature extraction will fail. The PCNN approach has been very successful in the field of image recognition, but there is still hope that some other methods of speech signal preprocessing will be helpful in dealing with speech recognition.

Acknowledgement
This work was supported by the institutional grant FRI 2/2006.

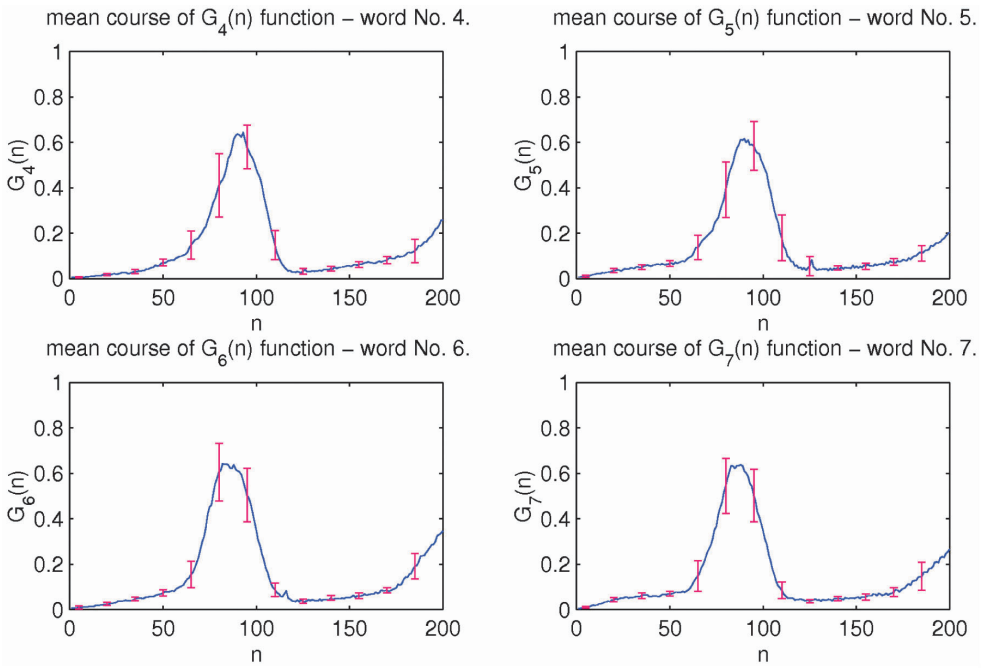

Fig.4 Mean courses of $G_a(n)$ functions for input words 4, 5, 6, 7 with their 95 % reliability level

References

[1] RANGANATH, H. S., KUNTIMAD, G.: *Iterative Segmentation using Pulse Coupled Neural Networks,* SPIE, Vol. 2760, pp. 543–554
[2] JOHNSON, J. L., PADGETT, M. L.: *PCNN Models and Applications,* IEEE Transaction on Neural Networks, Vol. 2, No.3, 1999, pp. 480–498
[3] KISER, J. M., JOHNSON, J. L.: *Implementation of Pulse-Coupled Neural Networks in the CNAPS Environment,* IEEE Transactions on Neural Networks, Vol. 10, No. 3, 1999, pp. 584–590

[4] FORGÁČ, R., MOKRIŠ, I.: *Artificial neural networks for feature space dimension reduction and classification (in Slovak),* Banská Bystrica: Univerzita Mateja Bela, 2002, ISBN 80-8055-743-8
[5] ORAVEC, M., POLEC, J., MARCHEVSKÝ, S.: *Neural networks for digital signal processing (in Slovak),* Bratislava: FABER, 1998
[6] WASSERMAN, D. P.: *Neural Computing Theory and Practice,* New York: VNR, 1989
[7] PSUTKA, J.: *Communication with Computer by Speech (in Czech),* Praha: Academia, 1995
[8] PROAKIS, J. G., MANOLAKIS, D. G.: *Digital Signal Processing,* New York: MPC, 1992
[9] KHANNA, T.: *Foundations of Neural Networks,* New York: Addison-Wesley Publishing Company, 1990. ISBN 0-201-50036-1
[10] HAYKIN, S.: *Neural Networks – A Comprehensive Foundation,* Macmillan College Publishing Company, New York, 1994
[11] ROSENFELD, A., KAK, A. C.: *Digital Image Processing,* John Wiley & Sons, New York 1982.