https://doi.org/10.26552/com.C.2013.2.13-18

Stefan Badura – Stanislav Foltan – Martin Klimo *

FUZZY LOGIC NETWORKS FOR SPEECH RECOGNITION

This paper proposes a massive fuzzy logic network which can be considered as a novel model of pattern classification network. Our approach introduces fuzzy logic circuits fulfilling the function of a binary classifier at first, which are connected into fuzzy logic networks with fuzzy flip-flop circuits as memories. Genetic programming is used as a circuit designing method. In order to establish design methodology, experiments aimed at testing the suitability of fuzzy logic operation sets, fitness functions and parameters of genetic algorithm were carried out. From trained circuits a hierarchical layered structure is built, where single layers consisting of given circuits are contextually dependent. Experiments with fuzzy logic circuits and fuzzy flip-flop network show some valuable results especially in the task of audio and visual speech recognition.

Keywords: *Fuzzy logic, speech recognition, genetic programming, binary classifier, memristor, lip-reading, structure, network.*

1. Introduction

Speech is the most natural form of human communication. No wonder that with the development of technology, a man has come with an idea to communicate with the machine or a computer. Since then, the history of speech recognition started [1]. Initially, it was recognition of isolated words, later the development of systems recognizing continuous speech. All these systems are based on the acoustic representation of speech [2]. Many approaches exist also for visual speech recognition. Most of them use artificial neural networks (ANN) or hidden Markov models (HMM). Recurrent neural networks are often used for time series recognition. If we consider just visual speech recognition, then in [3] authors recognize silence and vowels, where an Elman topology of ANN is utilized and that is constructed from 3 layers. In [4, 5, 6] a time delay NN (TDNN) is used. In [7] a modified TDNN is introduced for the same purpose. Many researchers resorted to the hidden Markov model (HMM) since it performs well also in audio speech recognition [7, 8, 9].

In recent years, some works describing the recognition from an entirely different perspective appeared. This is a view of the brain response to the received perception. In late 2008 work describing the recognition of simple black and white images based on the brain activity, scanned by functional magnetic resonance imaging (fMRI) was published [10]. In 2011, scientists from UC Berkeley announced reconstruction of the video [11]. In early 2012 work describing the reconstruction of audio perception [12] was published. Inspired by these publications we attempt to simulate the activity of the nervous system and we try to design which is suitable for audio or visual speech recognition.

We propose a novel network model where we combine fuzzy logic circuits with a fuzzy flip-flop memory described in [13]. Our goal is to interconnect two class classifiers into a robust network that can be considered as a universal multiclass classifier for dynamic data. Hierarchical organization and layered structure introduce a contextual modeled system which is not difficult to understand and which provides suitable abilities for a general task of speech recognition. Speech is a non-stationary process; our goal is to model this property using mentioned memories and fuzzy logic memories.

This paper is organized as follows. The second part describes initial identification of simple speech features in the input signal. This identification is realized in a form of binary classifiers that decide whether there is a specific feature detected at the input or not. The classifiers are represented by fuzzy logical functions where their output's values correspond to the level of the fired feature. The third part presents a hierarchical structure of the network which is responsible for identification of larger units of the input signal (time sequence). The fourth part describes briefly inputs used for the network evaluation. The fifth part describes some results and the final part concludes this paper.

2. Fuzzy logic circuits

The structure of the logical function representing a single classifier is designed by means of genetic programming [14], and it consists of different sets of fuzzy logical functions where some of them allow even hardware implementation with memristors [15, 16].

* **Stefan Badura, Stanislav Foltan, Martin Klimo,**
Department of Info-Com Networks, Faculty of Management Science and Informatics, University of Zilina, Slovakia, E-mail: baduras@itall.sk

The example of an individual is shown in Fig. 1. The structure of the chromosome for this individual can be encoded to string "0(1(2(.3),.248)),0(1(.72),.40)))))", where in this example "0" represents function NOT, "1" – MIN, "2" – MAX, ".XYZ" – XYZ-th value of input spectrum, ")" – creation of left child node, ")" – termination of a sub-tree and transition to the parent, and "," – creation of the right child node.

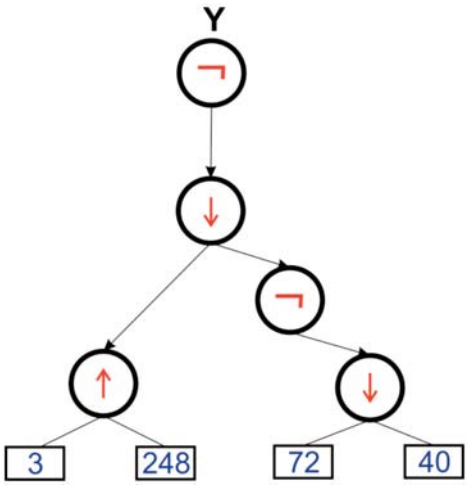

Fig. 1 This image shows an example of generated structure and its prescription can be written as: Y=NOT(MIN(MAX(spectrum[3]; spectrum[248]), NOT(MIN(spectrum[72]; spectrum[40]))))

2.1. Fuzzy logical operations

For realization of a decision process we use a fuzzy logical function. There are many types of fuzzy logic, like Zadeh logic, probabilistic logic, and Lukasiewicz logic. Since there is no clear answer to which of these is the most appropriate representation of human reasoning, we formed several different sets of logical operations consisting of each of these three logics. Valuable results were obtained by means of these two sets of functions:

Zadeh logic:

$$F_{AND}(x,y) = \min(x,y)$$

$$F_{OR}(x,y) = \max(x,y) \tag{1}$$

Lukasiewicz logic:

$$F_{\rightarrow}(x,y) = \min(1 - x + y,1) \tag{2}$$

The advantage of Zadeh logic over remaining logics lies in the possibility of hardware implementation with memristors. In our experiments we used Lukasiewics's logic for the network design evaluation. Zadeh's logic was left for further investigation.

2.2. Mechanism of individual's evaluation

The individual's evaluation is based on the output of the logical function for any input sample. Fig. 2 illustrates histograms of output values divided into 101 bins; approximated by A) Gaussian functions and by B) fuzzy-like functions, where T1 is the lowest output value for the sample of class i and T2 is the

$$T = \frac{T1 + T2}{2}$$

highest output value for the sample of class *j*.

The sample is considered as correctly recognized at A) if

$$f(x_i; \mu_i; \sigma_i^2) > f(x_i; \mu_j; \sigma_j^2),$$

or

$$f(x_j; \mu_i; \sigma_i^2) < f(x_j; \mu_j; \sigma_j^2), \tag{3}$$

where

$$f(x;\mu;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}. \tag{4}$$

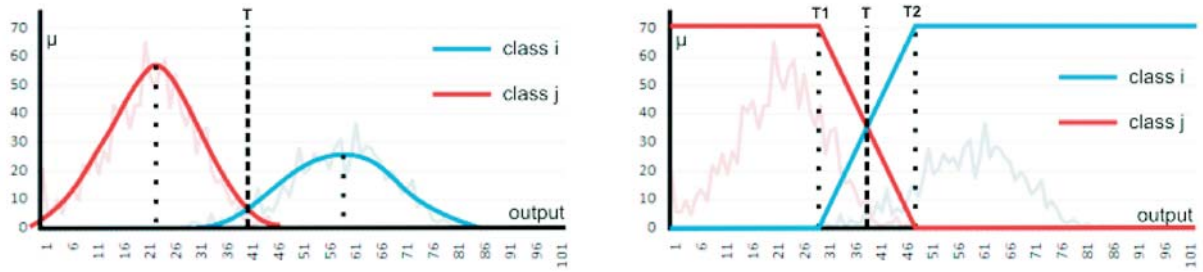

Fig. 2 Image on the left shows histograms of output values approximated by Gaussian functions and image on the right shows histograms of fuzzy-like functions

The sample is considered as correctly recognized at B) if $f(x_i) > T$, or $f(x_j) < T$.

Our aim is to find out a logical function which recognizes correctly as many input samples as possible. To meet this objective, we use the following fitness functions:

1) The first fitness function tends to make output values (see Fig. 2 histograms) the most distinguishable by increasing the distance between the outputs of each class:

$$F_1 = \mu_i - \mu_j \tag{5}$$

where μ_i is the mean of class i outputs and μ_j is the mean of class j outputs.

2) The second fitness function corresponds to the total probability of a correct recognition:

$$F_2 = \left(\frac{|good_i|}{|N_i|}\right)\left(\frac{|N_i|}{|N_i + N_j|}\right) + \left(\frac{|good_j|}{|N_j|}\right)\left(\frac{|N_j|}{|N_1 + N_j|}\right) \tag{6}$$

where $|good_x|$ is the number of correctly recognized samples from class x and $|N_x|$ is the size of class x.

3) The third fitness function maximizes the number of true positives and minimizes the number of false negatives:

$$F_3 = \frac{2 * P * R}{P + R} \text{ where } P = \frac{N_{correct}}{N_{detect}},\ R = \frac{N_{correct}}{|N|} \tag{7}$$

where $N_{correct}$ is the number of correctly recognized samples from class i, N_{detect} is the number of all samples detected as i, and $|N|$ is the size of class i.

3. Hierarchical network design

In the previous section we described the basic principle of training and evaluation of single circuits (structures). In this section we propose a hierarchical network's model. The proposed network is aimed to be used in the speech recognition task, deeper description can be found in [17].

The topology consists of 2 layers where each layer has its purpose, see Fig. 3.

- *The 1st layer* – this layer tries to indicate simple properties. In the task of lip reading one property could be, e.g., the mouth's position (open-close position).
- *The 2nd layer* – this layer tries to consider time dependence of detected properties.

Both layers consist of trained fuzzy logic circuits. Each structure is trained on one property against other properties as it was described in the previous section.

As it was already mentioned, the first layer tries to indicate some properties. In our case we define property as a number of a single cluster. A set of vectors from training samples, which will be introduced in the 4th section, is grouped with Ward method into several clusters. Then each cluster represents one property. The number of groups was chosen as 15 (see the dendogram in Fig. 4 for given input data) where a slice for 15 groups is shown.

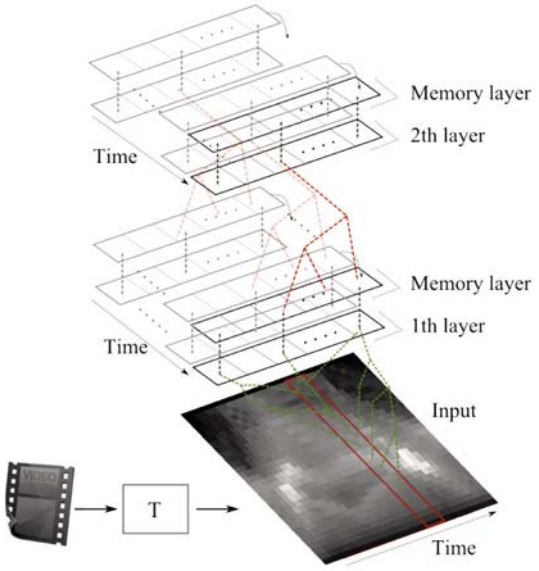

Fig. 3 General network model. Layer 1 and 2 consit of trained structures; the block T presents an image transformation into feature space

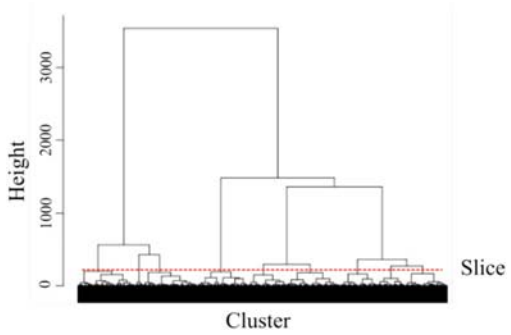

Fig. 4 Dendogram for input data obtained with Ward clustering method; the count of chosen clusters is decided based on a slice. In our case we use 15 clusters

The 1st layer consists of 15 structures where each structure is trained for one property (one data cluster). Structures for the second layer are trained for the output values of the first layer. We are using 23 vowels in our experiments, so the 2nd layer is generated from the structures trained on output sequences from the first layer for each vowel. The topology of the second layer is more complicated than the first layer, but the training principle of the second layer is very similar to the first one. The difference is in an objective

function where it is evaluated after a time period for one vowel when training structures.

3.1. Memory

Memory is an important part when the time dependence of input data is modeled. In [13] a fuzzy flip-flop that provides abilities for sequential remembering of the input signal was presented. If a signal is close to value 1, the flip-flop can remember its value – it is excited. If a higher signal is proposed for a longer time period, the output stays excited also for a longer time. In our experiments we used basic flip-flop as it is shown in Fig. 5.

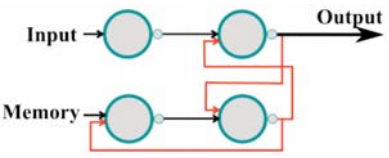

Fig. 5 Basic flip-flop used as memory designed from NAND fuzzy logic operation

Above each of the two introduced layers; a memory layer can be placed, which is designed from fuzzy flip-flops. As it was already described, the basic idea behind the memory is as follows: if the signal is high enough (close to 1) for a longer time period, the memory supports it and if it is not strong enough, the memory suppresses it. Using the memory serves also for modeling the time dependence. From other point of view at memory it can be said; that memory enlarges the gap between stronger and weaker signals. In our experiments we use basic flip-flop memory as Fig. 4 shows. The main goal of using the memory is to provide ability for continuous speech recognition. We are interested in the state of the network after a time period. Structures, which are strongly fired, are indicating inputs for which they were trained. In next experiments the memory is placed just above the second layer. We executed experiments without using memories also. In this case we do not consider the time dependence and the results are as expecting less satisfied.

4. Used inputs

In our experiments two datasets were tested. The first dataset is used for the logic circuits evaluation and the second one is used for the hierarchical structure verification. The database of spectra was created in our department from audio recordings of the Slovak book "Cukor a sol'", written by Keleova and Vasilkova, published by Ikar, Bratislava, 2004. At the recording, the book was read by one woman (sampling rate 22050Hz, 16 bit per sample, mono), the position of each phoneme was marked, and spectrum of the window (512 samples) centered on the mark was computed by means of R-software (functions *spec.pgram()* and *log10()*). The final spectrum is represented by 256 values. The database consists of more than 133000 samples of 60 phonemes. At current experiments, 1000 randomly selected spectra of each vowel (a, e, i, o, u) are used. This database is referred as db1 in proposed experiments.

A set of vowels extracted from video sequences is used as inputs for the lip reading task. The feature extraction process is shown at the flow chart in Fig. 6. Fig. 7 shows an example of an extracted feature for different video sequences. Median sieves are used for scale, space invariant feature extraction. Each column in each image in Fig. 7 represents the feature vector extracted from one video frame..

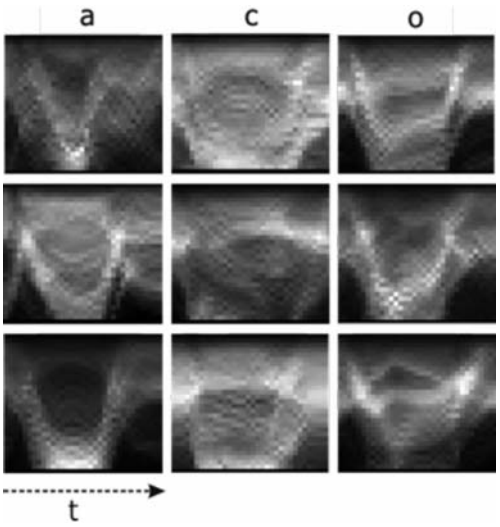

Fig. 7. An example of extracted features for the process of lip-reading. Columns in each image represent feature vector extracted from one frame of video sequence for vowels "a,c,o".

This database is referred as db2 in our experiments.

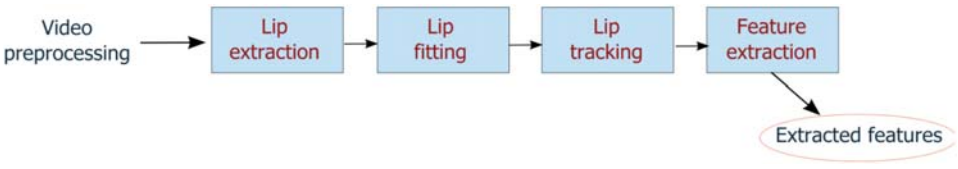

Fig. 6 Flow chart for the feature extraction process; the input is a video sequence; lips are extracted in the first frame; the lips'model is found (the lip fitting process) in next step; the 3th phase represents the lip's tracking system and the final stage is the process of feature extraction

5. Experiments

For the logic circuits' evaluation, a set of experiments was executed which is described in next text. All the experiments were tested on the audio database, referred as db1.

5.1. Expriments with audio

The experiment is focused on the accuracy of recognition when pairs of phonemes were used. In all the scenarios 1000 vs. 1000 samples of phonemes, divided with ratio 7:3 into a learning/test set were used. The total number of binary classifiers at each scenario was 10 (a/e, a/i, a/o, ..., o/u). Table 1 contains the results of this experiment. It is visible that the best accuracy was obtained with Lukasiewicz logic (L6), and it was around 93% no matter which fitness function was used. The results achieved with Zadeh logic (L2) were approximately at the same level with any fitness function.

Results obtained from the recognition between pairs of phonemes; F1, F2 represent fitness functions and L2, L6 are logics which were used — Table 1

	a	e	i	O	u	L: μ(S)	T: μ(S)
F_1L2	87.1%	77.4%	87.6%	84.9%	85.2%	84.5%	83.4%
F_1L6	95.5%	92.8%	96.8%	95.2%	92.2%	94.5%	93.1%
F_2L2	89.8%	84.0%	92.7%	84.1%	86.8%	87.5%	83.5%
F_2L6	95.6%	92.5%	98.5%	96.1%	94.8%	95.5%	93.0%
F_3L2	92.3%	83.3%	89.3%	86.1%	84.9%	87.2%	83.6%
F_3L6	97.0%	93.2%	96.5%	95.3%	94.9%	95.4%	92.7%

5.2. Experiments with video

For the evaluation of the network design we processed a series of experiments with the lip reading data (referred as db2). A database of Slovak vowels was collected for our purposes. Together, a set of 23 different vowels was recognized and each vowel was recorded 54 times. Before experiments the whole dataset was divided into training and testing subsets with the ratio 6:4. All the structures at the first and the second layer were trained for the first subset. The Ward's clustering algorithm was used for defining properties at the first layer. This method labels time rows of each feature vector with a group number, which represents an interesting property. Together 15 different labels (groups) were used in proposed experiments. For evaluating of the vowel recognition, two different objective functions were examined:

- U1 – the cumulative objective function (sum of values in time).
- U2 – this function takes the value in the last considered time.

When using U2 function, another parameter was examined, and it was the fall time (the fall time is time where the network is feeded by 0 values as inputs). Table 2 shows some experimental results of these experiments.

Experimental results - positive recognition rates in % for different settings and objectives functions of given network design; rows present results depending on the fall time, U1, U2 are objective functions used for results evaluation. — Table 2

Memory	0.05		0.15		0.25	
Fall time	U1	U2	U1	U2	U1	U2
0	16.41	14.16	15.84	12.90	18.51	12.34
5	16.41	14.16	16.54	14.58	18.51	13.74
10	16.12	15.28	16.83	14.02	17.67	12.34
15	16.12	14.58	17.11	14.44	17.25	10.93
50	16.12	15.00	16.97	6.45	17.67	4.347

6. Conclusion

In this paper a novel approach for speech recognition was introduced. It shows a good behavior for the task of lip-reading. The final result obtained from experiments were around 18–19% for a positive vowel recognition which is considered as satisfied because it shows good behavior for sequential remembering of time values. For an effective lip reading task it must be enhanced in the future.

Interesting results were obtained when the memory was used. In the experiments it was shown that using memories leads to the modeling of dynamic properties in the input signal. Future experiments should cope with the memory itself and its structure because the memory does not consider time occurrence of excitation in the described experiment. Other experiments which can be executed can concentrate on the objective function in the training phase, especially for the second layer. The future work can be directed to increasing the number of recognized classes (consonants), the recognition of speech for different speakers, and for the proposal of a strategy for the recognition of larger units of speech (i.e. words). At the level of logical circuits' training a different logic can be tested for example as it is shown in [18].

It is important to say that the aim of this paper was not to present a perfect speech recognition system. The traditional methods of speech recognition are much more sophisticated and successful. Unlike the traditional methods, the main advantage of proposed approach lies in its possibility of hardware implementation (which as far as we know has not been presented so far). As it is suggested in the paper, this is possible by using memristors.

Acknowledgment

This work was partially supported by the Slovak Research and Development Agency under the contract No. VMSP-II-09.

References

[1] JUANG, B. H., RABINER, L. R.: *Automatic Speech Recognition – A Brief History of the Technology Development,* Georgia Institute of Technology, Atlanta, 2004

[2] GUBKA, R., KUBA, M.: *Audio Patterns Searching and Retrieval*, 21st Intern. Conference Radioelektronika, 2011

[3] WILLIAMS, R.J., ZIPSER, D.: A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation.* 1989

[4] DUCHNOWSKI, P., HUNKE, M., BUSCHING, D., MEIER, U., WAIBEL, A.: *Toward Movement-Invariant Automatic Lip-reading and Speech Recognition,* 1995

[5] DUCHNOWSKI, P., MEIER, U., WAIBEL, A.: *See Me, Hear Me: Integrating Automatic Speech Recognition and Lipreading.* Proc. of the ICSLP, 1994

[6] BREGLER, C., OMOHUNDRO, S.: Nonlinear Manifold Learning for Visual Speech Recognition, Proc. IEEE ICCV, pp. 494–499, 1995

[7] GOLDSCHEN, A. J.: *Continuous Automatic Speech Recognition by Lipreading,* Ph.D. dissertation, George Washington Univ., Washington, DC, Sept. 1993

[8] POTAMIANOS, G., COSATTO, E., GRAF, H.P., ROE, D. B.: *Speaker Independent Audiovisual Database for Bimodal ASR.* Proc. European Tutorial Workshop Audiovisual Speech Processing, Rhodes, 1997

[9] POTAMIANOS, G., VERMA, A., NETI, C., IYENGAR, G., BASU, S.: *A Cascade Image Transform for Speaker Independent Automatic Speechreading.* Proc. IEEE Int. conf. Mulitmedia, New York, 2000

[10] YOICHI, M., HAJIME, U. et al.: *Visual Image Reconstruction from Human Brain Activity Using a Combination of Multiscale Local Image Decoders,* [online] www.sciencedirect.com/science/article/ pii/S0896627308009586, 2008

[11] NISHIMOTO, S., VU, A. T., NASELARIS, T. et al.: *Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies,* [online] http://www.sciencedirect.com/science/article/pii/ S09609822 11009377, 2011

[12] PASLEY, B. N., DAVID, S. V., MESGARANI, N., FLINKER, A., SHAMMA, S. A., et al., Reconstructing Speech from Human Auditory Cortex. PLoS Biol 10(1): e1001251.doi: 10.1371/journal.pbio.1001251, 2012

[13] KLIMO, M., BORON, J.: *Dynamicke vlastnosti pravdepodobnych fuzzy klopnych obvodov / Dynamic Properties of Probabilistic Fuzzy Flip-flops.* Proc. of ITAT (Informacne technologie – aplikacie a teoria), 2009

[14] KOZA, J. R.: Genetic Programming: A Paradigm for Genetically Breeding Populations of Computer Programs to Solve Problems, Stanford University Computer Science Department, 1990

[15] STRUKOV, D. B., SNIDER, G. S., STEWART, D. R., WILLIAMS, R. S.: The Missing Memristor Found, *Nature* 453, pp: 80–83, doi:10.1038/nature06932, 2008

[16] KLIMO, M., SUCH, O.: *Memristors Can Implement Fuzzy Logic,* in CoRR, [online] http://arxiv.org/ abs/1110.2074v1, 2011.

[17] BADURA, S., KLIMO,M., SKVAREK, O.: *Lip Reading Using Fuzzy Logic Network with Memory,* AICT, Georgia, Tbilisi, ISBN: 978-1-4673-1740-5, pp:35-38, 17-19 Oct., 2012

[18] SUCH, O.: *Phoneme Discrimination Using KS-algebra I.,* [online] http://arxiv.org/abs/1302.6031, 2013