Towards Speaker Identification System based on Dynamic Neural Network

E. Ivanovas[1], D. Navakauskas[1]
[1]*Department of Electronic Systems, Vilnius Gediminas Technical University, Naugarduko St. 41-413, LT-03227, Vilnius, Lithuania, phone: +370 612 03253*
edgaras.ivanovas@vgtu.lt

***Abstract*—The conventional, Finite Impulse Response and Lattice-Ladder multilayer perceptron (MLP) structures with 4, 8 and 16 hidden neurons were verified for speaker identification. The experiments were performed on 10 speakers, 3 Lithuanian words, 7 sessions' database. Identification performance was compared against two baseline methods: Vector Quantization (Linde-Buzo-Gray) and Gauss Mixture Models (Expectation Maximization). Increase of neuron number in hidden layer has led to smaller mean square errors on training dataset. A Finite Impulse Response MLP showed smaller mean square errors values. The results of experimental investigation show that neural networks can be used for speaker identification system as they outperform baseline methods. The best identification rate was archived by a multilayer perceptron with 4 hidden neurons and Finite Impulse Response MLP with 8 hidden neurons.**

***Index Terms*—Speech processing, neural networks, speaker recognition, multilayer perceptrons.**

I. INTRODUCTION

Human-machine interaction gets more popular every day and probably the most acceptable method for human being interaction is speech. There are numerous reasons stimulating research in this area starting with increased efficiency via saving human resources in customer support business areas to user-friendliness in controlling domestic appliances, cars, mobile phones and computers. As some of the mobile gadgets may not have enough processing power some custom designed hardware or FPGA usage could be a solution [1].

Human speech carries not only a plain text message, but also a lot of very distinct information such as: language being spoken, physical and emotional states, accent as well as identity of the speaker. A human being always tries to consider all this information before taking action, which is really not an easy task for a machine [2]. As it was already demonstrated, Lithuanian has specific phonetic, syntactic and lexical properties along with specific accentuation. Therefore, a machine we want to control also should work taking in to account other information contained in speech. The most important task after speech recognition is to determine the identity of the speaker as it would let to ignore the commands given by people, who should not have the authority to work with the device. This could prove useful in systems described in [3].

Therefore, speaker identification also receives a lot of research effort. However, neither optimal set of features nor the most suitable classifier has been agreed on. The experiments usually show different results depending on speaker database used [4], which can be influenced by different recording hardware, surroundings, language, and speaker origin and so on.

In this short paper we compare two implementations of classical baseline methods – VQ-LBG [5] and GMM-EM [6] – against three artificial neural networks – MLP [7], FIR MLP [8] and Lattice-Ladder MLP [8] – presenting experimental results of their use for speaker identification. Our aim is to incorporate national speech aspects in system thus Lithuanian words uttered by mother-tongue speakers were used for the experiments.

II. SETUP FOR SPEAKER IDENTIFICATION

A. Signals and features used

The records of 10 speakers in 7 sessions pronouncing 3 Lithuanian words – "turėti", "nebūti" and "mokykla" – are used in the experiments. These words have been chosen as they include all vowels of Lithuanian language. Accentuation of these words is also specific: first two have the second syllable stressed, while the third has its last syllable stressed. Selected words do not include diphones. There exists 1542 diphones [9] and it would be very difficult to cover them all.

The sampling frequency of the records is 11 kHz, the records are saved in 16 bit PCM format as "WAV" files. Recordings have been done in silent environment using personal computers. The stationary noise including the interference of electric mains and other noises produced by non-professional equipment were removed by Wiener filter as 10 s of pure silence containing stationary noises has been recorded before pronouncing the words.

Mel-frequency cepstrum coefficients' (MFCC) feature space was selected to be used for transformation of the signals. The amplitudes of pronounced words were normalized making the maximal value of the signals equal. The signals were framed taking 256 samples per frame (23.22 ms) with an overlap of 100 samples (11.03 ms). The frame energy and zero-cross-ratio were calculated. If

Manuscript received March 19, 2012; accepted May 12, 2012.

threshold values of the frame or its 10 neighbours from each side were not exceeded the frame was dropped. 20 MFCC were extracted from each of the frames. The first coefficient was discarded resulting in 19 MFCC used to form a feature vector.

MFCC vector contains only information within one frame and does not give any information on changes from previous frames. In order to take dynamic information into account special classifiers such as HMM based are required. The other common way is to form delta or even delta-delta coefficients by subtracting MFCC vector values of previous frame from the values of the current frame.

B. Types of artificial neural networks selected

Artificial neural networks are investigated in order to see how well various structures can evaluate dynamics of MFCC vectors and perform speaker identification task.

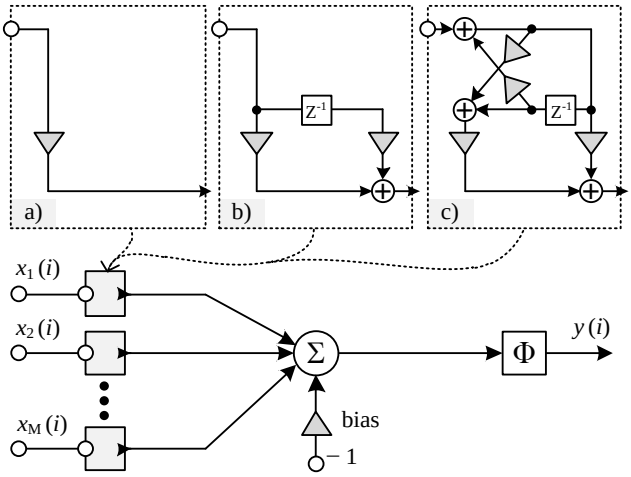

Fig. 1. A universal representation of artificial neuron used in hidden layer for structures of: a) MLP, b) FIRMLP, c) LLMLP.

Following artificial neural networks are investigated:

1) Multilayer Perceptron (MLP);

2) Finite Impulse Response Multilayer Perceptron (FIRMLP);

3) Lattice-Ladder Multilayer Perceptron (LLMLP).

FIRMLP and LLMLP (see Fig. 1) looks very similar to a Multilayer Perceptron, however, weights of neuron synapses in hidden layer are changed with Finite Impulse Response filters or Lattice-Ladder filters, correspondingly.

FIRMLP structure is considered to be more powerful than MLP due to capabilities to process time-dependent signals. LLMLP structure outstands because in it is easy to track the stability of the filter during the training procedure – the stability of lattice-ladder filter is guaranteed if absolute value of any lattice coefficient does not exceed 1.

C. Method used for dataset construction

The available data is separated into 3 groups. 3 sessions have been used for training, 1 for validation and 3 for testing. MFCC feature vectors from the 3 words of one session are combined into one set of feature vectors for each speaker. Further, the data of different speakers are combined into one for each session. Afterwards, the resultant vectors from sessions 1, 2 and 3 are combined into training dataset, from session 4 into validation dataset and from sessions 5, 6, 7 into testing dataset.

A separate network for each of the speakers is constructed, so 10 networks must be trained separately. 10 different desired output signals must be formed for each of the network. The desired output is set to "1", for the MFCC feature vectors of speaker we want to identify, whereas feature vectors of other speakers are marked by "– 1" in the desired output signal. A problem arises in such construction, because the ratio of feature vectors belonging to the speaker to be identified and other speakers is 1 : 9. This causes a bias of the pattern classification and unseen feature vectors of a true speaker are classified incorrectly more often. The problem is solved by using the same signals each shifted by ten samples for extraction of additional feature vectors. Data composition of one session is graphically depicted in Fig. 2. As a result a different training, validation and testing datasets for each of the 10 networks is composed.

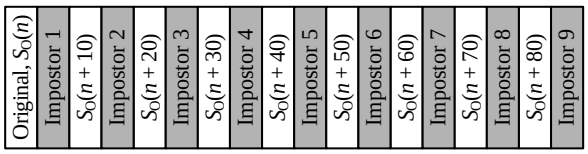

Fig. 2. A dataset formed from one session data for each of the 10 speakers. Each block consists of MFCC feature vectors extracted from 3 words. White blocks depicts a feature vectors extracted from the same shifted signal produced by speaker to be identified, gray blocks depicts feature vectors of other speakers.

D. Method used for neural network training

The training of the neural networks is performed changing μ training parameter using Levenberg-Marquardt training algorithm

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \left(\mathbf{J}^T\mathbf{J} + \mu\mathbf{I}\right)^{-1}\mathbf{J}^T\mathbf{e}, \qquad (1)$$

where $\mathbf{w}_k$ – weights' matrix at k-th instance; $\mathbf{J}$ and $\mathbf{I}$ – Jacobian and identity matrixes; $\mathbf{e}$ – error vector. The optimization criterion is mean square error (MSE): $E = \sum_{i=1}^{F}\left(d(i) - o(i)\right)^2 \Big/ F$, with $d(i)$ and $o(i)$ as desired and actual network output; F – the total number of feature vectors. The standard algorithm requires a change, because it tends to overshoot and to make filter unstable while μ is small (performing similar to Newton algorithm). The stability of the filters is tested and μ is multiplied by μ_{inc} in case of instability. Increasing the μ Levenberg-Marquardt algorithm performance become more steepest-descent like.

The sigmoid activation functions are used in hidden and output layer neurons, the initial ladder coefficients are initialized randomly whereas lattice coefficients and biases are set to 0, $\mu = 0.001$, $\mu_{inc} = 5$, $\mu_{dec} = 0.15$. Training is stopped if any of criteria is met: number of iterations is more than 25; MSE reaches 10^{-6}, gradient value drops below 10^{-4}, or μ gets greater than 10^{16}. Values of filter coefficients are saved after each iteration and the ones with the smallest MSE value on validation dataset are taken.

E. Structures of neural networks considered

The performance of all chosen types of neural networks

has been tested under the following structural constrains:

1) *19 inputs – 4 hidden – 1 output neurons;*
2) *19 inputs – 8 hidden – 1 output neurons;*
3) *19 inputs – 16 hidden – 1 output neurons.*

In order to be consistent in the comparison, 1-st order for FIRMLP and LLMLP filters was chosen.

Finding the global minimum of MSE is not guaranteed and the outcome depends on initial values of the trained network coefficients. Thus the experiments have been repeated 10 times, where only the best solution with the smallest validation error has been chosen, for each of 10 networks for each speaker. Analysis of 3 different types and 3 different structures led to 900 experiments in total.

F. Alternatives

For comparison purpose classical baseline algorithms have been used [5], [6]:

1) VQ method with Linde-Buzo-Gray training algorithm (VQ-LBG), calculating sum of minimum Euclidean distances from 16 centroids;

2) GMM trained by EM algorithm (EM-GMM) using 16 mixtures.

TABLE 1. MEAN SQUARE ERROR VALUES OF VARIOUS NEURAL NETWORK STRUCTURES ON TRAINING AND TESTING DATASET.

Type	**MLP**						**1-st order FIRMLP**						**1-st order LLMLP**					
Structure	**19–4–1**		**19–8–1**		**19–16–1**		**19–4–1**		**19–8–1**		**19–16–1**		**19–4–1**		**19–8–1**		**19–16–1**	
NN\Data	**TR**	**TE**	**TR**	**TE**	**TR**	**TE**	**TR**	**TE**	**TR**	**TE**	**TR**	**TE**	**TR**	**TE**	**TR**	**TE**	**TR**	**TE**
1	0.362	0.485	0.182	0.550	0.043	0.465	0.286	0.514	0.099	0.612	0.042	0.495	0.348	0.758	0.288	0.842	0.198	0.858
2	0.206	0.390	0.094	0.362	0.010	0.398	0.141	0.362	0.032	0.336	0.003	0.397	0.120	0.044	0.120	0.048	0.098	0.052
3	0.351	0.505	0.207	0.536	0.043	0.484	0.275	0.569	0.092	0.547	0.024	0.478	0.263	0.530	0.231	0.442	0.123	0.527
4	0.141	0.241	0.043	0.244	0.015	0.205	0.095	0.259	0.010	0.188	0.003	0.278	0.208	0.491	0.049	0.495	0.005	0.555
5	0.362	0.479	0.229	0.383	0.094	0.426	0.310	0.484	0.167	0.455	0.086	0.422	0.352	0.440	0.336	0.432	0.139	0.431
6	0.212	0.405	0.100	0.378	0.019	0.374	0.158	0.397	0.026	0.349	0.018	0.347	0.081	0.219	0.108	0.341	0.090	0.215
7	0.243	0.319	0.113	0.276	0.037	0.266	0.188	0.311	0.051	0.302	0.004	0.249	0.191	0.350	0.115	0.347	0.052	0.403
8	0.188	0.479	0.072	0.471	0.008	0.485	0.220	0.434	0.026	0.483	0.002	0.465	0.328	0.343	0.094	0.243	0.093	0.259
9	0.236	0.513	0.127	0.467	0.028	0.393	0.184	0.461	0.039	0.393	0.003	0.338	0.306	0.423	0.124	0.270	0.092	0.190
10	0.071	0.219	0.016	0.157	0.000	0.174	0.029	0.215	0.000	0.187	0.000	0.243	0.086	0.204	0.062	0.180	0.020	0.140
Average	0.237	0.403	0.118	0.382	0.030	0.367	0.188	0.401	0.054	0.385	0.018	0.371	0.228	0.380	0.153	0.364	0.091	0.363

Note: Structure: number of neurons in Input–Hidden–Output layers; **TR** – training dataset; **TE** – testing dataset

III. RESULTS OF SPEAKER IDENTIFICATION

A. Results for individual speakers

Resulting mean square errors on training and testing datasets are given in Table I for all structures. As mentioned earlier, MSE values for validation datasets were used for prevention of network over-fitting and for picking the best solution out of 10 tries.

Unfortunately, the smallest validation error does not guarantee the best network performance on the testing dataset as it can be seen in Fig. 3a, where minimum validation error is at iteration 5, minimum testing error at iteration 9. Mostly, this has been seen during training of the networks for the first speaker identification. Other networks showed smaller shape differences between validation MSE and testing MSE curves (Fig. 3b). This could be explained by bigger differences in feature vectors of Speaker 1 speech sessions. In this case, taking coefficients at iteration 9 instead of 5 would reduce MSE of testing set by more than 34 %. However, testing set is believed to be data unseen by the network during learning procedure and probably the possible solution would be to collect more data from Speaker 1 for validation.

A closer look at results of mean square errors for training datasets given in Table I reveals that increasing number of hidden neurons decreases the errors in all three types of networks (values are in grey background) with the exception of the LLMLP networks for Speaker 6 identification (value underlined). Comparing the performance of the network types with the same number of hidden neurons shows marginal advantages of FIRMLP networks with an exception for Speaker 8 identification underlined by wiggled line. The demonstrated performance of the LLMLP structure networks seems to suffer from the curse of dimensionality which probably could be solved by initializing the ladder and bias coefficients from the FIRMLP structure and letting the learning process to adapt the lattice coefficients initialized as zeros.

B. Generalized results

Unlike MSE values on training dataset, values produced by testing dataset seem to have no trends. Speaker identification results are given in Table II. The given values show how many mistakes the networks have done identifying the speakers in a closed set test. The signals from session 5, 6 and 7 were used. This results in 3 words for each of the ten speakers (30 per session). The speaker identification was performed feeding each word into the networks and the decision from the outputs of the networks has been done in two ways:

$$D_1 = \arg\max_p \sum_{i=1}^{F} \Phi\left(o_p(i)\right), \tag{2}$$

$$D_2 = \arg\max_p \sum_{i=1}^{F} o_p(i), \tag{3}$$

where op(i) – the i-th value of output of p-th artificial neural network; F – the total number of frames in utterance; $\Phi(x) = \{1,\ x > 0;\ -1,\ x \le 0\}$ – decision function.

A closer look at Table II shows that two architectures perform best on test signals: MLP with 4 hidden neurons and FIRMLP with 8 hidden neurons.

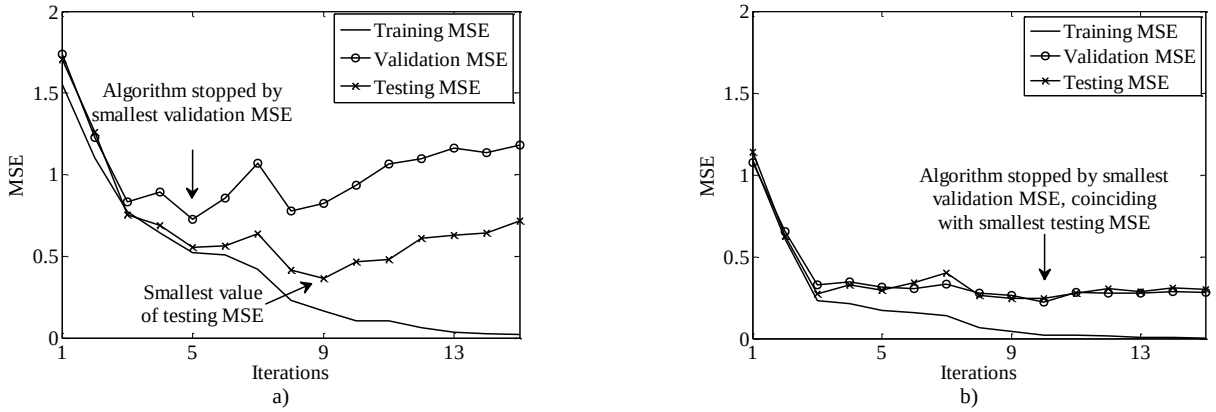

Fig. 3. Mean square errors of the first order FIRMLPs for different speaker identification: a) Speaker 1, b) Speaker 10.

TABLE II. SPEAKER IDENTIFICATION ERRORS.

Method/ Utterance	MLP*			1-st order FIRMLP*			1-st order LLMLP*			VQ-LBG	GMM-EM
	19–4–1	19–8–1	19–16–1	19–4–1	19–8–1	19–16–1	19–4–1	19–8–1	19–16–1		
"turėti"	0 | 0	0 | 0	0 | 0	1 | 1	0 | 0	0 | 0	1 | 0	0 | 0	0 | 0	0	0
"nebūti"	0 | 0	0 | 2	1 | 0	1 | 1	0 | 0	1 | 1	2 | 2	2 | 3	2 | 2	2	1
"mokykla"	0 | 0	1 | 2	4 | 1	1 | 3	0 | 0	2 | 2	3 | 2	0 | 0	4 | 4	2	2
Total	0 | 0	1 | 4	5 | 1	3 | 5	0 | 0	3 | 3	6 | 4	2 | 3	6 | 6	4	3

Note: * – results are presented as A | B, here A results are calculated by (2); B results – by (3)

Furthermore they outperform both used baseline methods. It is also worth noticing, that all methods have achieved better identification results for word "turėti". The worst identification rate (except for LLMLP with 8 hidden neurons) has been achieved using word "mokykla". The possible cause of this phenomena could be that the long vowel "y" in the latter word is not in the stressed syllable, whereas long vowels "ė" and "ū" in other two words are in stressed syllables. First decision function (2) showed marginally better results, however more experiments need to be performed to confirm that.

IV. CONCLUSIONS

The results of experimental investigation show that neural networks can be used for speaker identification system as they outperform classical baseline methods. The best identification rate was achieved by a MLP with 4 hidden neurons and FIRMLP with 8 hidden.

Mean square error values on testing dataset have provided no information about performance of neural networks on speaker identification. The use of a loss function, taking speaker identification rate into account, would probably improve the results.

REFERENCES

[1] G. Tamulevicius, V. Arminas, E.Ivanovas, D. Navakauskas, "Hardware accelerated FPGA implementation of Lithuanian Isolated Word Recognition System", Elektronika ir Elektrotechnika (Electronics and Electrical Engineering), no. 3, pp. 57–62, 2010.

[2] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, C. Wellekens, "Automatic speech recognition and speech variability: A review", *Speech Communication*, no. 10-11(49), pp. 763–786, 2007.

[3] R. Maskeliunas, R. Simutis, "Multimodal Wheelchair Control for the Paralyzed People", Elektronika ir Elektrotechnika (Electronics and Electrical Engineering), no. 5, pp. 81–84, 2011.

[4] C. Hsu, S. Yang, W. Wu, "Implementing Speech-Recognition Microprocessor into Intelligent Control-System of Home-Appliance", in *Proc. of the Asia-Pacific Services Computing Conference*, 2008, pp. 881–885. [Online]. Available: http://dx.doi.org/10.1109/APSCC.2008.75

[5] Y. Linde, A. Buzo, R. M. Gray, "An Algorithm for Vector Quantizer Design', *IEEE Transactions on Communications*, vol. 1, no. 28, pp. 84–95, 1980. [Online]. Available: http://dx.doi.org/10.1109/TCOM.1980.1094577

[6] D. A. Reynolds, R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 3, pp. 72–83, 1995. [Online]. Available: http://dx.doi.org/10.1109/89.365379

[7] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 1999, p. 700.

[8] D. Navakauskas, 'Artificial Neural Networks for the Restoration of Noise Distorted Songs Audio Records", Ph.D. dissertation, VGTU, Vilnius, 1999, p. 103.

[9] P. Kasparaitis, "Diphone Databases for Lithuanian Text-to-Speech", Informatica, vol. 2, no. 16, pp. 193–202, 2005.