
Production Systems and Information Engineering
Volume 7 (2015), pp. 5–14

AUTOMATIC GESTURE GENERATION

László Czap
University of Miskolc, Hungary
Department of Automation and Infocommunication
czap@mazsola.iit.uni-miskolc.hu

Roland Kilik
University of Miskolc, Hungary
Department of Automation and Infocommunication
kilik@mazsola.iit.uni-miskolc.hu

[Received September 2015 and accepted November 2015]

Abstract. The usage of virtual agents make necessary to produce natural motions and gestures. The paper introduces a novel solution which generates vertical head motion from the speech. The method uses neural networks for estimating the positions of the head. The outputs of the proposed algorithm have validated by human feedback. The results show that the naturalness of the generated motions is similar to the original ones.

Keywords: gesture generation, neural networks

1. Introduction

Our initial assumption, according to which both the intensity of the sound, and its pitch have a given time domain relationship with head gestures, was proved by correlation methods in our prior publications [3].

At the phase of the research that described in this paper our goal was automatic head gesture generation from the sound, based on the prior results. We tried to give a solution not only for the automatic generation of vertical and horizontal movements, but the blinks as well.

The most advanced part of the research is the field of generating the vertical movements, and the paper mostly describes its methods, procedures and results.

The teaching and testing samples for our two neural networks used for the vertical gestures, were alltogether 100, one sentence video files, what were

downloaded from the web, or made by the authors with starring different subjects.

The topic of the paper is automatic gesture generation. After mentioning the results of the related researches, the paper describes the methods, procedures and current results of generating vertical movements of gestures from the sound, that developed and reached by the authors. After showing the current stage of the validation, the research's other phases are mentioned in the paper.

2. Related work

A partial basis of our research was the result of other works, that stated for example that there is a relationship between different gestures (for example movements of the hands), or the movments of the lips and the speech features [1, 2]. However, these researches did not examine the time-domain relationships between those, and produced head movements with very different approach and with restrictions, furthermore did not produce blinks from the sound.

In [2] the authors use *Hidden Markov Model*. Our proposed solution is based on *Artificial Neural Network* and this way it is able to avoid some post processing steps. In the case of HMM implementation the algoritm results motion segments and necessary to join them in a further step. The neural network is a better choice in this sense because it can determine the positions while processing the given segment of the speech.

3. Teaching and testing samples

The teaching videos for the first network were fully spontaneous, while the ones for the second network were not, but still mainly performed by the announcer's own words. The testing videos were spontaneous speeches.

4. Preprocessing

Both the neural network teachings, and testings were performed with sentences, that were created from the above mentioned video files.

The movement vectors of the sample sentences (for the testing and the training) were produced partially by a Java program that created by the authors that followed the eyes of the subjects, and by manual determination of the eye positions as well.

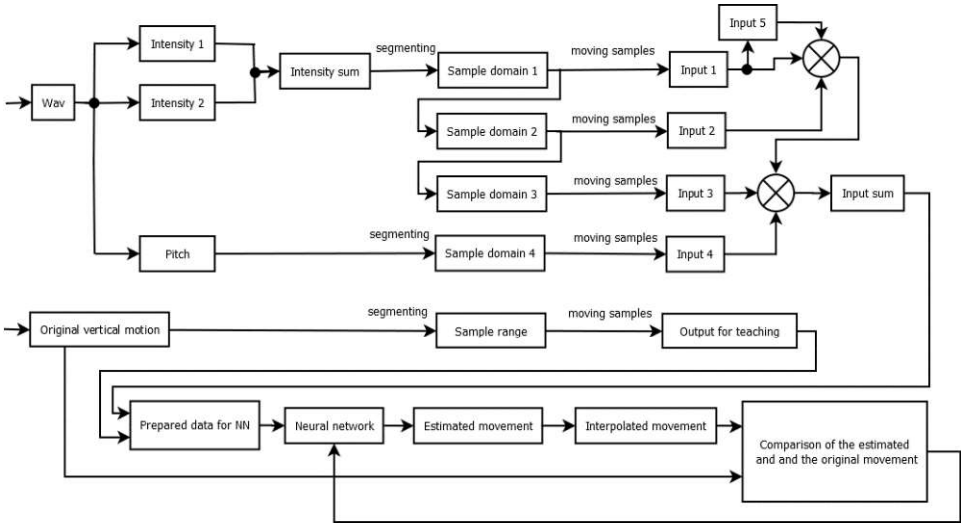

Figure 1. The teaching procedure and the input structure of the first neural network

The pitch vectors of the sound were produced by Praat software, and the intensity vectors (intensity1 on the figure 1) were calculated by the following formula with MATLAB:

```
intensity1(i) = sqrt(sum(window .^ 2));
```
(4.1)

The elements of the vector are the square root of the sum of the square of every element of the window, where the window was always a frame-sized part of the amplitude vectors.

As the first figure shows, the initial inputs of our first neural network were mainly derivatives of the intensity vectors of the given teaching samples, and the input of the other neural network was only a pitch vector.

As it can be seen on the Figure 1, a second form of the intensity was also used (`intensity2`), which was calculated on a vector that we got by a 25 window size smoothing mean on the amplitude vector. The calculating program code was:

```
intensity2(i) = mean(window .^ 2);
```
(4.2)

Every element of the vector is the mean of square of every element of the window, where the window here was a frame-sized part of the previously smoothed and meaned amplitude vector.

Finally, `intensity1`, and `intensity2` were summed by the following formula (after simplification):

```
intensity_sum = intensity1 * 0.735 + intensity2;
```
(4.3)

The segmented version of this summary vector (sample domain 1) was one of the five for neural network no. 1. The second (sample domain 2) vector was produced by subtracting the values of the previous element from each element's value in the sample domain 1 vector. The third (sample domain 3) vector was calculated by subtracting the values of the previous element from the value of each element in the sample domain 2 vector. The fourth sample domain vector was the segmented version of the pitch vector. The segmentation was carried out by having the syllable starting times in each sentences (that were manually marked), and calculating the segmentation time by having the time of the maximum amplitude between every two syllable starts in the voice.

As teaching and testing movement vectors, vertical movements were used. These movement vectors were the differences between the vertical position coordinates of the left eye on every frame, finally divided by the mean of them. These were also segmented with the same procedure as the sound-originated vectors, and used as sample range vectors.

Atfer preparing the sample domain vectors, corresponding input vectors were created from them, and from the segment range vector, by a moving sample procedure. This procedure prepared the data for prediction, by pairing the following element from the movement vector (sample range), to every consecutive two elements from the given sample domain vector. The procedure made 4 input vectors from the 4 sample domain vectors, and a 5 th input, which was created by elements made from shifting elements of the input1 by one to the left). An output vector was also created by the procedure from the sample range vector for teaching the neural network.

For the second neural network, the first sample domain vector was the pitch vector (Figure 2). At this stage there was only one input vector, which was created by the same moving sample function as used for the first neural network.

The neral networks were nonlinear autoregressive networks with feedback [4], trained in open loop mode, and then closed for the testings and used in closed loop state. The structure of the networks can be seen on Figure 3 and Figure 4.

As can be seen from the figures, the mainly intensity-based first neural network has an input delay 4, and a layer delay 4, while the second, pitch-based network has input delay 2, and a layer delay 2.

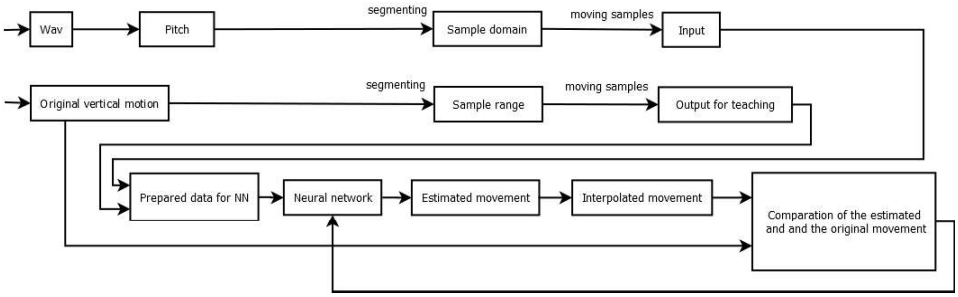

Figure 2. The teaching procedure and the input structure of the second neural network

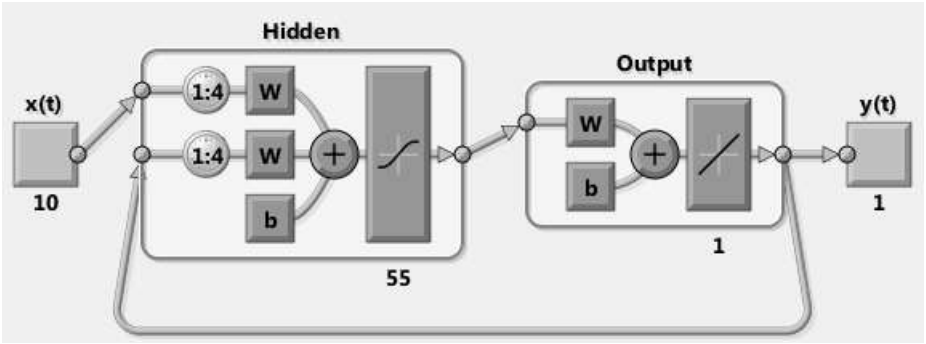

Figure 3. The structure of the first neural network

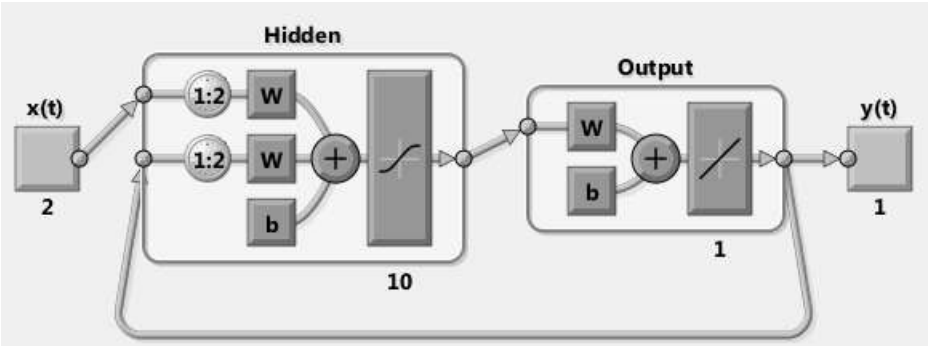

Figure 4. The structure of the second neural network

Not only does MATLAB'S data preparing function order the input and output due to the delays of the network, but it also cuts the same number of

elements from the input and the target as the number of the delay. Therefore, at first, we supplemented the beginning of our vectors by zeros for these to be cut, in order to avoid the shortening of our useful sample sizes. This also made possible to have predicted motions even at the beginning of the sentences.

5. Postprocessing

After the predictions of the two networks, the amplitudes of the outputs of the neural networks were corrected. At this stage it was accomplished only by constants. The values of the multipliers were determined by the volunteers, as they saw results produced by different multiplier values, and chosen those for the best. The outputs of the intensity-input neural network were multiplied by 2 (and resulted in a high amplitude, frequent movement), and the outputs of the pitch-based neural network were multiplied by 4 (and resulted a lower amplitude, less frequent movement). This method was suitable for most of the videos, but for others, the intensity-based neural network's results had to be divided by 4 in amplitude, and the results of the other network were not modified. There is an ongoing procedure that could predict two things from some of the sound features. One of them is about which of the 2 neural network's output would be more natural by the opinion of the viewers .The other procedure is concerned with how much amplitude correction is would be the best for the given example.

As both the inputs and outputs were segmented vectors, the outputs of the neural networks had to be finally interplolated to have the movement for every frame.

The corrections of the amplitude, the inputs of the networks, and the suitability of the two selected networks (from the plenty of trained and tested ones) were mostly determined by the authors and users. The method was to compare the naturalness of the generated movements with the original and random movement, they could saw on an announcer's picture being moved vertically by those movements. Before this, as a precondition, an acceptable similarity had to exist with the selected network between the original and generated movement vectors.

On Figure 5 some of the segmented, and concatenated teaching sample's movement vectors and the first neural network's predicted output vectors can be seen. It can be observed, that the similarity is not tight, but the goal of the research is to produce a result movement for sounds that seems natural to the viewer rather than producing a movement vector that is similar to the original one.

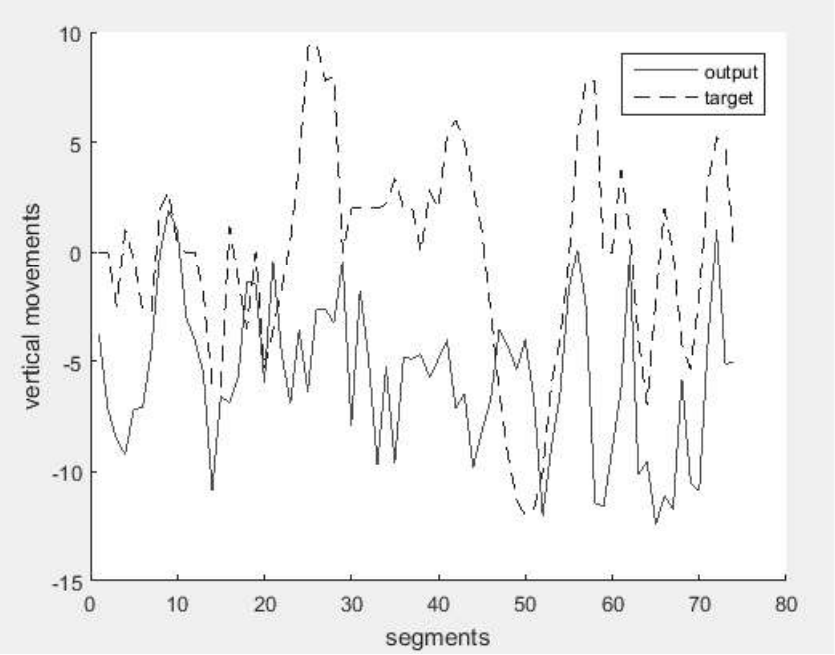

Figure 5. Output and target for a part of teaching samples for neural network 1.

On Figure 6, what is the original movement and target movement of some of the second neural network's teaching samples, the above statements are also true.

6. Testings and results

After the authors had chosen some of the previously created networks performing better in the objective similarity than others, 10 test videos were tested by 6 volunteers at this stage. At this phase of tests, only a still frame of an announcer was moved with the given frame rate while playing the sound of the sentence at the same time. For every test, a moving picture was created this way with the original movement, with random movement with the same co-domain as the original, and two with the neural network generated outputs.

These tests resulted in choosing the two best neural networks, refinements of the amplitude corrections, a predicted value correcting function (with an output-correcting neural network), and hand-made corrections in the results, each according to the authors' and the volunteers' opinions. Having summed

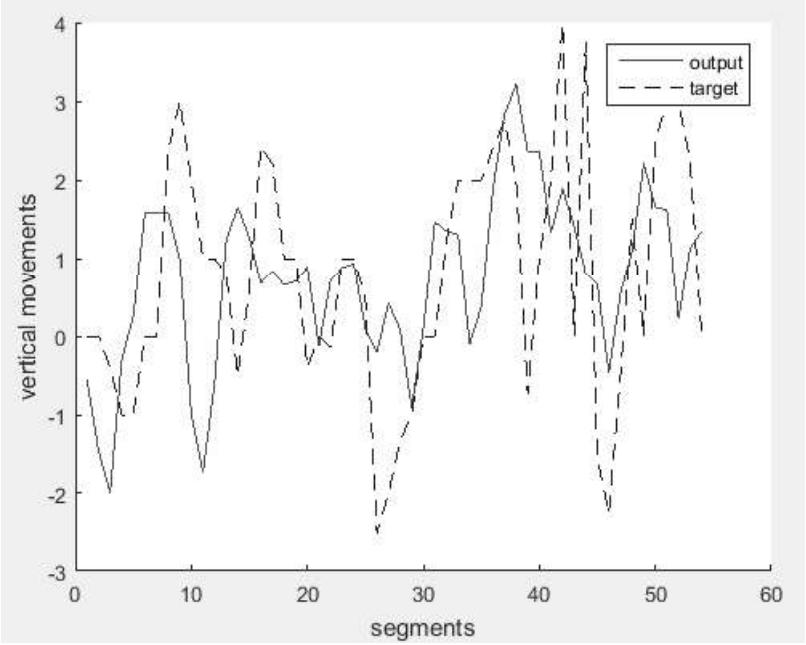

Figure 6. Output and target for a part of teaching samples for neural network 2.

those corrections, a neural network retraining was carried out with the modified teaching target diagram to make the output similar to this, without having the correctional network and hand-made corrections.

As a result of this, 9 from the 10 testing samples described both of the two final neural network's result as more natural than the random movement in almost every cases. About half of them, most of the volunteers stated that the video with neural network generated movement was even more natural, than the one with the original movement.

On Table 1 some of the results can be seen.

In the table each cell contains the rank of the naturalness of the given movement variant for the video, by the (A, B, C or D) volunteer. Examining the first table, it can be seen for example for the first test video, that first neural network's result described as the most natural one by A, B, and D volunteers, and the least natural (4-th place) by one volunteer D. Neur. 2/a means the results with the older amplitude corrections to the results of the second neural network (which were related to the original amplitude), and 2/b version means the results with an amplitude correction that is independent of the original movement's amplitude.

	random				original				neur 1.				neur 2./a				neur 2./b			
volunt.	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
video 1	4	4	1	4	2	1	2	2	1	1	4	1	3	3	3	3	-	-	-	-
video 2	4	4	3	4	3	1	2	2	2	3	4	1	1	2	1	3	-	-	-	-
video 3	4	1	4	4	2	2	1	2	3	4	3	3	1	3	2	1	-	-	-	-
video 4	4	4	4	4	3	3	3	3	2	2	1	2	-	-	-	-	1	1	2	1
video 5	3	5	5	5	2	3	3	2	1	4	1	4	5	1	3	1	4	2	2	3
video 6	5	4	5	5	2	1	3	4	1	2	1	1	4	5	2	2	3	3	4	3

Table 1. Some user opinions on some videos

Since the purpose of the tests was mainly to promote the improvement of the methods of the generations and corrections, the volunteers were not asked to give their opinions at the same time, and not every video was required to face opinions. The table contains only those videos that the most of the volunteers were asked to give opinion about.

As stated above, the similarity of the original and generated movements was not a goal. A verification of this approach is for example the result of a video that can be seen on figure 7, (video 6 on Table 1) where there are many differences between the original and the generated version of movement, however the generated one was even more natural according to the opinions.

7. Future works

It is important to note that the purpose of these opinion tests of users was mainly to help in choosing the proper networks and creating the necessary corrections. The real tests, with a larger number of volunteers, and using a virtual speeching head, will be the scope of the next phase of the research, as also the correction and the test cases of the horizontal movements and blink generation.

Acknowledgements

The research has beeen carried out within the framework of Mechatronics and Logistics Centre of Excellence operating as one of the strategic research

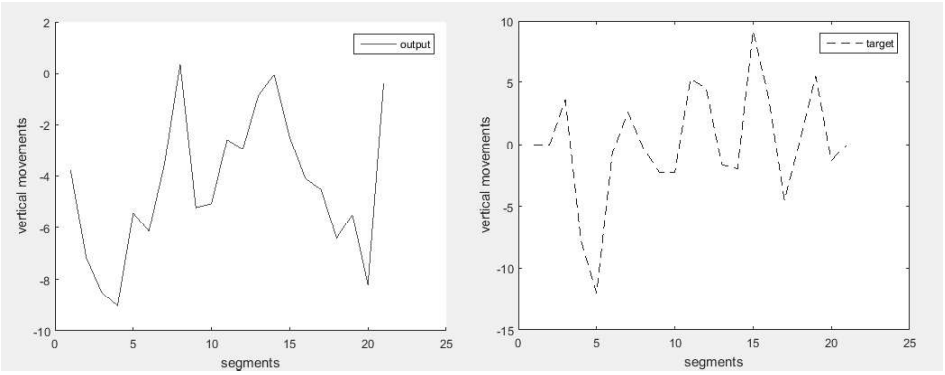

Figure 7. Neural network's estimated movement and the original movement for a given test video

area of the University of Miskolc and as the part of the TÁMOP-4.2.2.C-11/1/KONV-2012-0002 project funded by the European Unio, cofinanced by the European Social Fund.

The presented research work was partially supported by the grant TÁMOP-4.2.2.B-15/1/KONV-2015-0003.

REFERENCES

[1] Busso C.; Deng C.; Zhihang, Narayanan, S.; Neumann, U., : *Learning expressive human-like head motion from speech, Data-Driven 3D Facial Animation*, Springer, 2008.

[2] Rubin, P; Yeiha, H.; Vatikiotis-Bacon, E.: *Quantitatice association of vertical tract and facial behavior*, Speech Communication 26., 23-43., 1998.

[3] Czap, L. and Kilik R.: *Prelimenary Examinations For Automatic Gesture Generation*, Alkalmazott Nyelvészeti Közlemények, Miskolci Egyetemi Kiadó, 105-113., ISSN 1788-9979, 2015.

[4] Billings, S. A.: *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio- Temporal Domains*, Wiley, ISBN 978-1-1199-4359-4, 2013.