Annales UMCS Informatica AI 1 (2003) 125-131

Annales UMCS
Informatica
Lublin-Polonia
Sectio AI

http://www.annales.umcs.lublin.pl/

# Speech syllabic structure extraction with application of Kohonen network

Elżbieta Smołka[a*], Wiesława Kuniszyk-Jóźkowiak[a], Waldemar Suszyński[a], Mariusz Dzieńkowski[b]

[a]*Institute of Physics, Maria Curie-Skłodowska University, Pl. Marii Curie-Skłodowskiej 1, 20-031 Lublin, Poland*
[b]*Computer Systems Division, Management Department, Technical University of Lublin, Nadbystrzycka 38, 20-618 Lublin, Poland*

**Abstract**

Kohonen network was applied in the analysis of fluent and non-fluent utterances of any length. The input vector consisted of sound level values measured in dB(A) in 1/3-octave bands. Fragments of stuttering people's and fluent speakers' utterances were analysed. The optimum neural network was chosen of a 5×5 rectangle topology. Time arrangement of winner neurons in the network extracts single disfluencies, or, in the case of fluent courses, reflects the syllabic structure of an utterance.

## 1. Introduction

In automatic speech and speaker recognition systems and in the research of speech disturbances, what is used are not solely the acoustic characteristics resulting from the articulation process, but also the features of the perception process. Artificial neural networks are frequently used then [1-5]. It is justified to apply self-organising maps, and especially Kohonen network, which reflect information processing in the brain best [2]. The network may implement initial processing or play the role of the classifier. In the works published so far, time changes of input n-component spectral vectors, consisting of amplitudes in the chosen frequency bands, created trajectories of the winner neurons on the topological map during the learning process. Disturbed sounds resulting from e.g.: cleft palate [6-7], voice disorder [8-9] and incorrect articulation [10] created a greatly different image from that created for undisturbed speech on a two-dimensional map.

* Corresponding author: *e-mail address*: esmolka@tytan.umcs.lublin.pl

The authors of the present article applied the Kohonen network to reflect the time structure of fluent fragments of stuttering people`s and fluent speakers` utterances. Stuttering people`s utterance fragments containing a word with disfluency, reflected in the same way, were also observed [11]. In the case of speech impediment it is the word structure that is disturbed and not the structure of single sounds in a three-dimensional space (time, frequency, amplitude) [12-17].

## 2. Methodology

Stuttering people`s speech samples were obtained with the use of control recordings, which are always done before the beginning of therapy as well as later, in various periods during its course. During such a recording, stuttering people read the same text twice, first on their own (with the simultaneous auditory feedback), and then with the echo (delayed auditory feedback). In the situation of speaking in chorus with the echo, stuttering intensity usually diminished [18], which made it possible to choose fluent counterparts of the words which had been uttered in a non-fluent way. Fluent speakers were also recorded reading the same text and samples were chosen, which were needed for comparison. The examined people were placed in a soundproof booth, in which there was a microphone. Sound signals were transformed into digital signals with the use of a Sound Blaster card, with the sampling frequency 22050 Hz and sampling precision of 16 bits, and they were saved on a computer hard disk. With the use of a 21 digital 1/3-octave filters of center frequencies between 100 and 10000 Hz and an A-weighting filter, which describes the sense of hearing (such a solution allows for analysis which is comparable to that done by the human ear [19, 20]). The sound levels which occurred on the 1/3 octave filter outputs created N 21-component vectors, which were given as inputs in the Kohonen network.

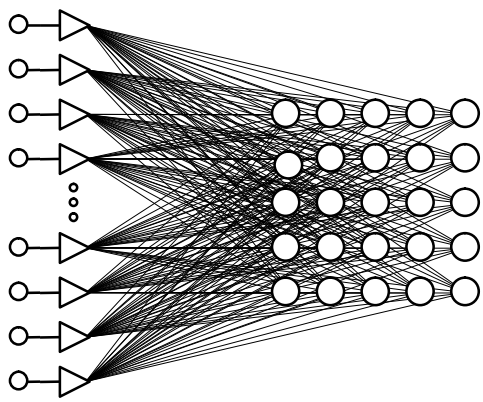

Fig. 1. Kohonen network of 21 inputs and 25 output neurons reflecting the fragments of fluent speakers` and stuttering people`s speech

Fig. 1 presents the architecture of the applied network, it has 21 inputs and contains 25 neurons in the output layer, arranged in a 5×5 square. The size of the map was experimentally determined, by testing the range of values from 3×3 to 10×10. Also, different parameters characterising the learning process were tested. Finally, the network was taught with the following parameters: training time – 100 epochs, learning rate – 0.1 and neighbourhood – 1. The result of the network operation was presented not as trajectories of the winner neurons on a 5×5 topological map, but in the form of graphs presenting the winner neurons (y axis) in subsequent moments of the utterance duration (x axis).

## 3. Results

Fig. 2 presents the time courses and corresponding arrangements of the winner neurons in the Kohonen network of the same fragment of a fluent speaker`s (a) and stuttering person's (b) utterances. The network distinguished particular syllables precisely for both speakers. While comparing the times corresponding to the beginning and end of particular syllables measured from the oscillograms and spectrograms those ones indicated by the Kohonen network, it was observed that the differences do not exceed 20 ms. Stop consonants are reflected in a characteristic way, the burst being distinctly marked, which follows the silence corresponding to the closure of the articulatory organs. It should be noticed that the network distinguishes even syllables which are not separated. It is so for the linked syllables *demu* (in the word *kaʒdemu* ), the border between the two being barely noticeable on the oscillogram. However, the network does separate them distinctly. While analysing the obtained relations, one might venture to say that the Kohonen network creates in time an image of acoustical information which is fed into it, the image resembling to some extent "the auditory sensation" which is created in the auditory cortex after the neural signals with the encoded acoustical information coming from ears have been processed [6, 19, 21-22].

In the second part of the research the subject of observation was the way the image obtained in Kohonen network changed when a disfluency occurred in an utterance. Of the numerous kinds of errors characteristic of stuttering, stop and affricate repetitions were chosen, which usually occur in a fixed form of this speech disturbance [14, 17].

Fig.3 a, b presents, respectively, the oscillogram and the winner neurons in the time function in a 4-second fragment of an utterance *im mɲej t͡ʃ͡wovjek poɕada*, where the word *t͡ʃ͡wovjek* is uttered in a non-fluent way. The *t͡ʃ͡* sound is repeated three times at the beginning of the word (at time moments: 1.51; 1.73; 1.94 s). Each short repetition is registered in the Kohonen network (Fig. 3a) as a great change in the position of the winner neuron from 1 to 21, 21 and 17.

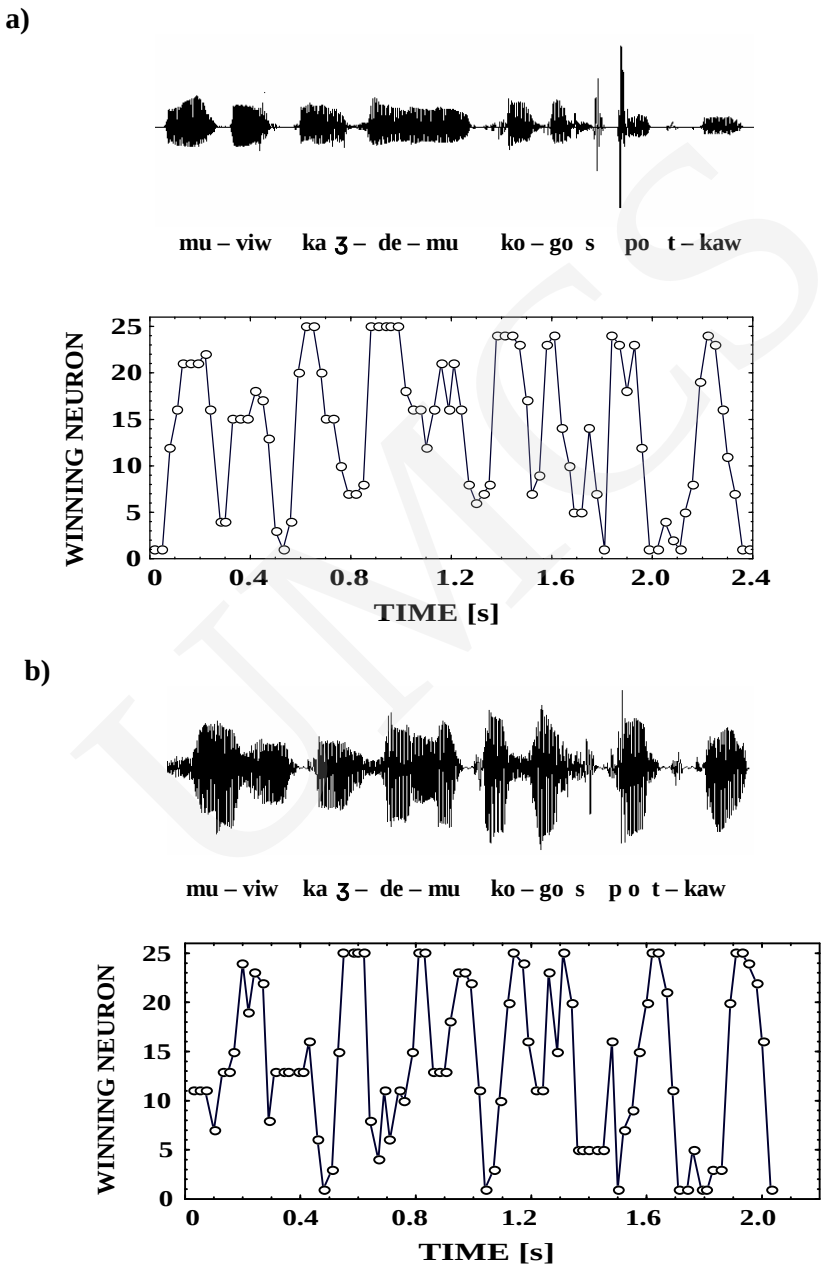


Fig. 2. Oscillograms and time arrangements of the winner neurons in fluent articulations of the phrase *muviw kaʒdemu kogo spotkaw* of a) a fluent speaker and b) a stuttering person

Fig. 3b presents, for comparison, the respective characteristics for the same fragment uttered by the same person fluently with the echo (delayed auditory

feedback). The positions of the winner neurons reflect here also the syllabic structure of an utterance with the prolongation of particular syllables resulting from the influence of the echo.

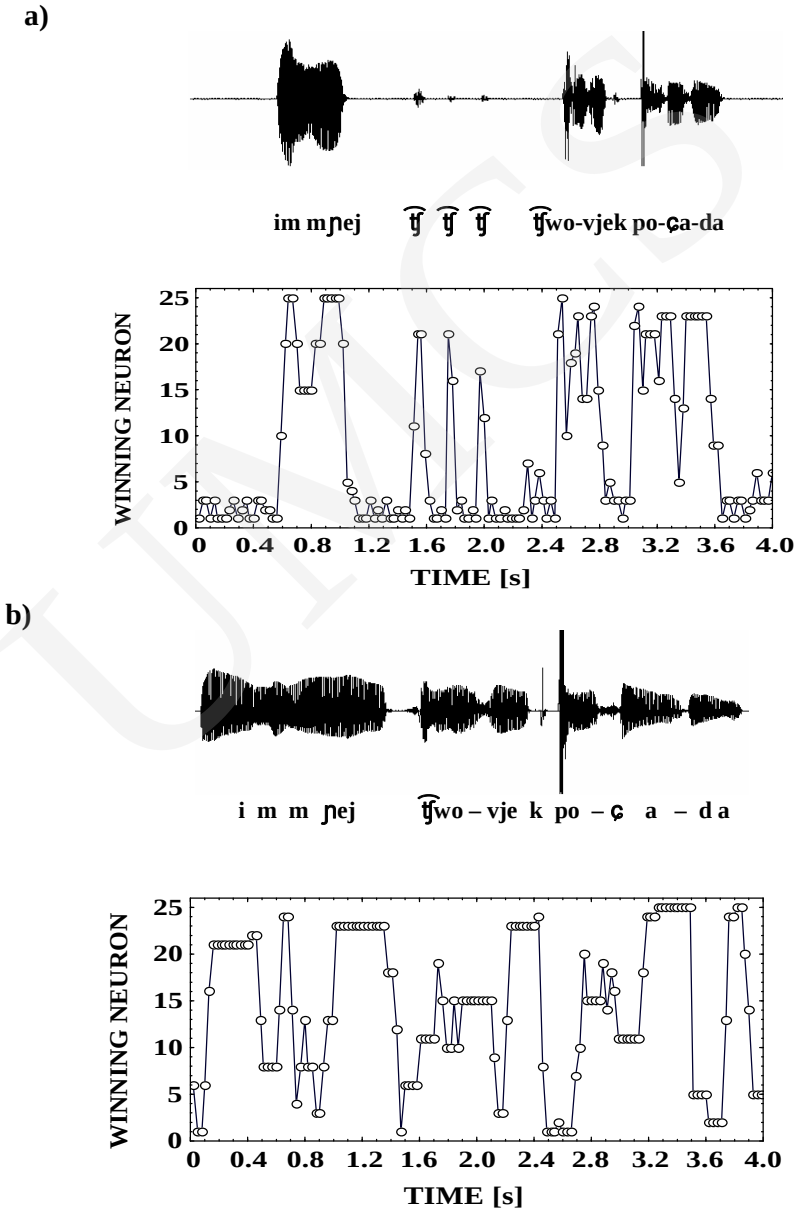


Fig. 3. Oscillograms and time arrangements of the winner neurons in the articulation of the phrase *im mɲej t͡ʃwovjek poɕada* by a stuttering person: a) non-fluent articulation, b) fluent articulation in the situation of speaking with the delayed auditory feedback

## 4. Conclusion

The Kohonen network, which was applied in this research, models the perception process of single disfluency episodes and syllabic structure of words. The input vector consisting of amplitudes in 1/3 octave frequency bands with the additional correction with the use of an A-weighting filter may be treated as an output signal from the inner ear. The network reflects the aspect of the reception and processing of signals from the auditory receptor, variable in time, which relates to distinguishing non-fluently articulated sounds and division into syllables without paying attention to specific information which carries the distinctive features of particular phonems.

## Acknowledgements

The research was supported by Grant No. KBN 4 T11E 035 22 from the State Committee for Scientific Research in Poland.

The authors wish to thank Natalia Fedan for translation of the paper into English.

## References

[1] Leinonen L., HiltunenT., Torkkola K., Kangas J., *Self-organized Acoustic feature Map in Detection of Fricative-vowel Coarticulation,* J. Acoust. Soc. Am., 93 (1993) 3468.

[2] Skubalska-Rafajłowicz E., *Self-Organisig Maps.* in: Biocybernetics and Biomedical Engineering – t. 6**,** Neural Networks, Academic Publishing House EXIT, (2000) 179, in Polish.

[3] Tadeusiewicz R., *Neural Networks,* Academic Publishing House, Warszawa, (1993), in Polish.

[4] Tadeusiewicz R., *Speech Recognition with Application Neural Networks,* Seminar Polish Phonetical Society, Warszawa, (1994) 137, in Polish.

[5] Żurada J., Barski M., Jędruch W., *Artificial Neural Networks,* Polish Scientific Publishers, Warszawa, (1996), in Polish.

[6] Izworski A., Wszołek W., *Artificial Intelligence Methods in Diagnostics and Processing of the Pathological Acoustics Signals,* in: Speech and Language Technology, t. 3, Poznań, (1999) 299, in Polish.

[7] Wszołek W., Tadeusiewicz R., *The Evaluation of Effectiveness of Various Neural Network Types in Pathological Speech Analysis,* XLVII Open Seminar on Acoustics OSA`2000, Rzeszów – Jawor 2000, II (2000) 721, in Polish.

[8] Leinonen L., Kangas J., Torkkola K., Juvas A., *Dysphonia Detected by Pattern Recognition of Spectral Composition,* J. Speech Hear. Res., 35 (1992) 287.

[9] Leinonen L., Hiltunen T., Laakso M. L., Popius H., *Categorization of Voice Disorders with Six Perceptual Dimensions,* Folia Phoniatr. Logop., 49 (1997) 9.

[10] Mujunen R., Leinonen L., Kangas J., Torkkola K., *Acoustic Pattern Recognition of /s/ Misarticulation by the Self-Organsing Map.* Folia Phoniatr., 45 (1993) 135.

[11] Smołka E., Kuniszyk-Jóźkowiak W., Suszyński W., *Reflection of Fluent and Nonfluent Words in Kohonen Network,* XLIX Open Seminar on Acoustics OSA`2002, Warszawa–Stare Jabłonki, (2002) 371, in Polish.

[12] Kuniszyk-Jóźkowiak W., *Acoustical Analysis and Stimulation of the Speech Process,* Publishing Company UMCS, Lublin, (1996), in Polish.

[13] Kuniszyk-Jóźkowiak W., Suszyński W., Smołka E., *Acoustical Methods in Diagnosing and Therapy Speech Disfluencies,* in: Biocybernetics and Biomedical Engineering – t. 2, Biomeasurements, Academic Publishing House EXIT, (2001) 251, in Polish.
[14] Kuniszyk-Jóźkowiak W., Smołka E., Suszyński W., *Acoustical Characteristics Alteration in Persons who Stutter Resulting from Therapy,* in: Structures-Waves-Biomedical Engineering, X(2) (2001) 57.
[15] Kuniszyk-Jóźkowiak W., Smołka E., Suszyński W., *Computer Visual-auditory diagnosing of speech disfluency,* XII National Scientific Conference "Biocybernetics and Biomedical Engineering", Warszawa 2001, II (2001) 758, in Polish.
[16] Suszyński W., Kuniszyk-Jóźkowiak W., Smołka E., *Acoustic Mehods in Diagnosing and Therapy of Speech Disorders,* Maintenance and Reliability, 7 (2000) 19.
[17] Suszyński W., *Acoustical Analysis of the Disfluencies in Stutterers` Speech,* XLVII Open Seminar on Acoustics OSA`2000, Rzeszów – Jawor 2000, II (2000) 715.
[18] Adamczyk B., Kuniszyk-Jóźkowiak W., Smolka E., *Correction effect in chorus speaking by stuttering people,* XVIth World Congress of the International Association of Logopedic and Phoniatrics, Interlaken: Karger Basel, (1976) 1.
[19] Jorasz U., *Lectures on Psychacoustics,* Scientific Publishing Company UAM, (1998), in Polish.
[20] Tadeusiewicz R., Wszołek W., Izworski A., *Neural Networks as Tool for Simulation of Processing of Acoustical Information from Auditory System,* X National Scientific Conference "Biocybernetics and Biomedical Engineering" Warszawa 1997, II (1997) 801, in Polish.
[21] Grossberg S., Myers ChW., *The resonant dynamics of speech perception: interword integration and duration-dependent backward effects,* Tech. Rep. CAS/CNS–TR–99-001, address: http://cns-ftp.bu.edu/pub/Diana/GroMye00/Gro/mye00.html
[22] Jorasz U., *Why sound is hearable?,* XLIX Open Seminar on Acoustics OSA` 2002, Warszawa – Stare Jabłonki, (2002) 91, in Polish.