Annales UMCS Informatica AI 5 (2006) 291-299

Annales UMCS
Informatica
Lublin-Polonia
Sectio AI

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

Vowel recognition in continuous speech with application of MLP neural network

Elżbieta Smołka[1*], Wiesława Kuniszyk-Jóźkowiak[1], Mariusz Dzieńkowski[3], Waldemar Suszyński[1], Marek Wiśniewski[2]

[1]Institute of Computer Science and [2]Institute of Physics, Maria Curie-Skłodowska University, Pl. Marii Curie-Skłodowskiej 1, 20-031 Lublin, Poland
[3] Department of Informatics Systems, Lublin University of Technology, Nadbystrzycka 38, 20-618 Lublin, Poland

Abstract

The aim of the present work was to find the answer to the question: To what extent can the multilayer perceptron be applicable in the automatic vowel recognition process in any given fragments of a particular speaker?

Initial research was carried out with the use of recordings of 3 adult people's speech. Vowel recognition was performed with the application of multilayer perceptron. On the input of the network, N-element vectors were fed, which consisted of sound levels values obtained every 0.02s as a result of spectral analysis. Each created network was taught to recognise 6 vowels – a, e, o, u, i, y as well as one pattern including all other fragments of an utterance – consonants and pauses.

The networks in which a result of over 90 % correct classifications for all the time moments was obtained were used to carry out a test on a completely different set of data. The best result in that part of research was 92% vowel recognition. At the same time, only 50% time moments, which made up these vowels, were correctly recognised. The other half was recognised as other vowels or a different fragment of the utterance. There also occurred 15% incorrect recognition of time moments making up consonants or pauses.

1. Introduction

Speech recognition with the application of artificial neural networks has had quite a long history. In the research the following types of networks were put to use: multilayer perceptron (MLP), time delay neural network (TDNN), Kohonen network, probabilistic and others. The subjects of recognition were: single sounds, strings of letters articulated one after another, words [1-7]. Various methods of signal parameterisation were applied, on the basis of various types of its analysis: e.q. the FFT (Fast Fourier Transform), LPC (Linear Predictive Coding), homomorphic analysis. Frequently the input vectors consisted – for

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

example – of melscale spectral coefficients [1,3-4,6]. For vowels, frequently the values of 2, 3 or 5 formants completed e.g. with formant range widths were applied, thus correlating their recognition with the shape of the vocal tract during articulation [2,7]. The authors of the present work, in their previous trials involving neural networks, have used raw data from the Fourier analysis with the application of 1/3-octave bands and A-weighting filter, which makes an approximation of the processes that occur in a human ear [8,9]. At the same time we have tried to recognise vowels (their 100ms fragments were chosen) uttered in isolation, in simple syllables CVC, CVCV, where C-consonant, V-vowel. One vowel was always represented by one vector consisting of data characterising 4 consecutive time moments [10,11].

In the present work, an attempt is made at vowel recognition in continuous speech (text reading), which means that the vowels were of various lengths, and, what results from that, represented by various numbers (from 2 to 8) input vectors. Here one input vector always corresponded to one time moment. Below, initial processing of the examined signal, structure of input vectors as well as an output vector, the architecture of the applied MLP networks and the obtained results are presented.

2. Methodology

In the research recordings of three adults' speech were used – 1 woman and 2 men. Their task was to read a chosen fragment of a story a few times. All the trials were recorded with a computer sound card, using 16-bit quantisation of the amplitude, with 22050Hz sampling frequency. The recordings were monophonic. They were used to prepare the data fed on the input and output of an artificial neural network – in this case it was the multilayer perceptron. When preparing input data, recordings of each person were divided into fragments containing the same text. Each of the fragments was the subject to spectral analysis (FFT) with the application of twenty-one 1/3 octave bands of center frequencies from 100 Hz to 10,000 Hz and A-weighting filter. In this way, every (approximately) 0.02 s consecutive 21-element vectors were obtained, which consisted of sound level values, fed on the network input (fig. 1 and fig. 5). On the other hand, in the network output there must be the data which will be the perceptron's „teacher". In order to achieve this for every text fragment subject to examination, using signal visualisation (oscillogram, spectrogram – fig. 2) and human hearing, vowel locations in the file (their beginning, end, duration time) were determined exactly. On the basis of these data, there was determined which time moments belong to the vowels and which are other parts of the utterances. 7 pattern were planned: 6 syllabic vowels: a, e, o, u, i, y and 1 patterns containing consonants, transition states, silence, breathing, and that pattern was marked „x" (fig. 5).

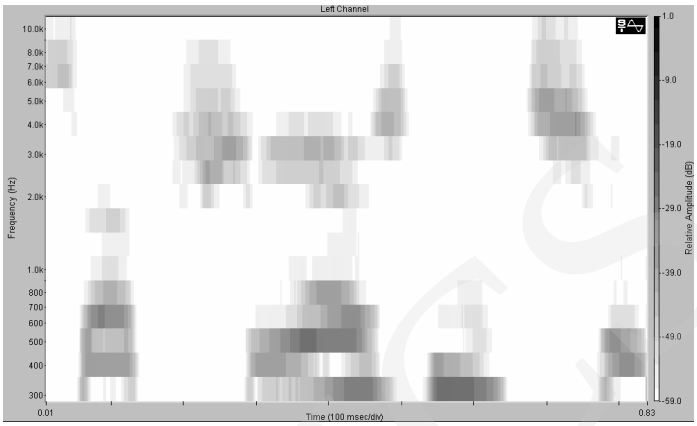

Fig. 1. Spectrogram of an utterance fragment „zaprzyjaźnić się" after FFT with the application of twenty one 1/3-octave bands and A-weighting filter

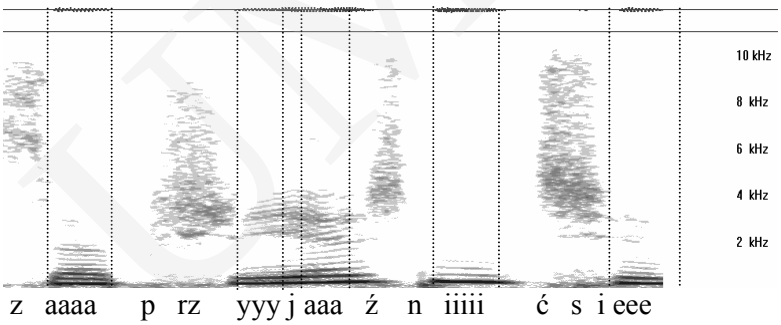

Fig. 2. Spectrogram supporting human hearing in vowel location recognition in an utterance. The figure shows "zaprzyjaźnić się" fragment – ę was here pronounced as e

The examples of vowels and other part vectors are shown in Figs. 3 and 4. In these, one time moment for „e", „i", „y" and one time moment for silence, as well as „s" and „n" consonants were characterised.

The longest utterance fragment was used to teach the network. For each speaker, approximately 20 networks were taught. Among them there were perceptrons with one hidden layer or two hidden layers, containing various numbers of neurones. In the input layer there were 21 neurones, and in the output – 7. The numbers remained unchanged during the experiment (Fig. 5). The networks were taught with two methods: in the initial phase – with the method of back propagation, and then – with the conjugate gradient descent method. It is more effective in the case when the network has to distinguish more groups.

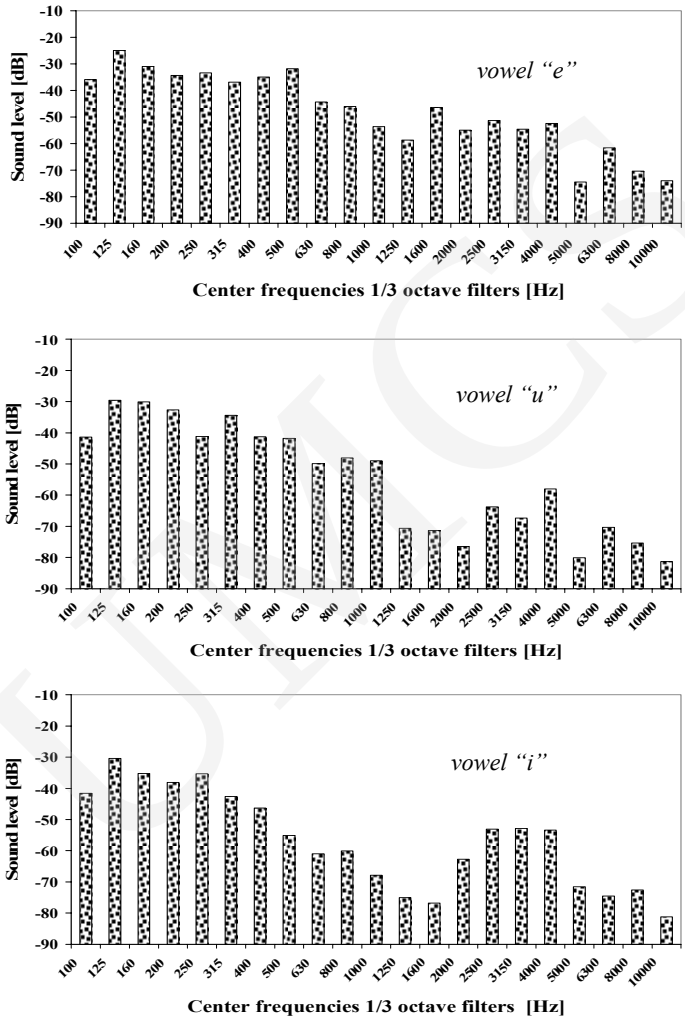

Fig. 3. Examples of input vectors representing „e", „u", „i" vowels

Subsequently, the networks where chosen – 10 per a speaker – for which the level of learning was high (over 90% of correctly recognised time moments of an utterance) and tests were carried out on a different fragment of an utterance which had not been used to teach or verify the network. The network was considered to have the „best recognition" when it selected the highest percentage of vowels occurring in the tested utterance. At the same time, a vowel was considered recognised when the networks correctly qualify at least one time moment.

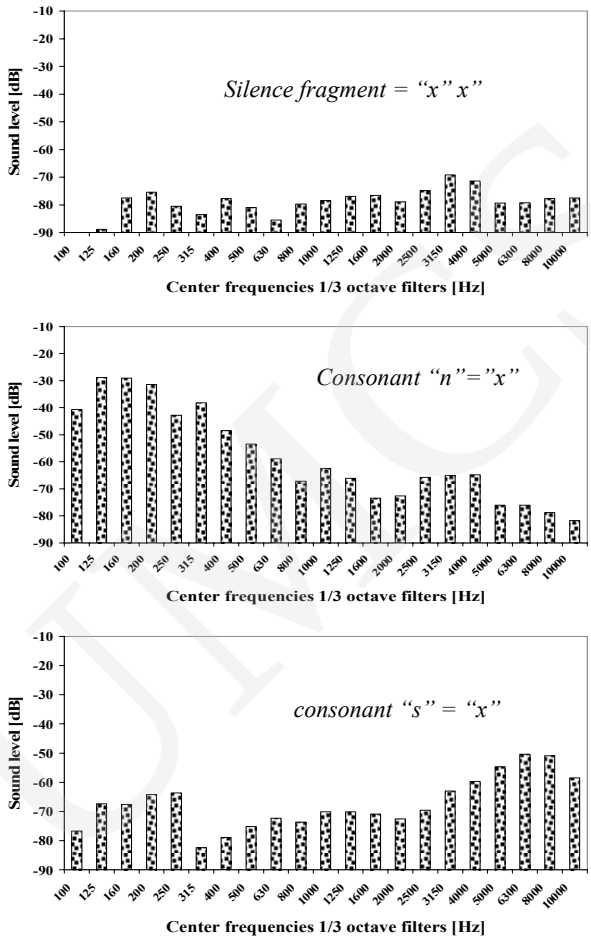

Fig. 4. Examples of „x" input vectors – corresponding to a moment of silence, „n" and „s" consonants

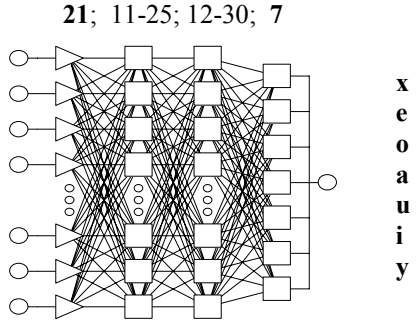

Fig. 5. Diagram of MLP network with two hidden layers

3. Results

In the performed trials the best results, both in the learning process and testing, were obtained for a perceptron of two hidden layers (Fig. 5). In Table 1 the networks are characterised for each speaker by means of whom the most vowels were recognised. The test results are also given. In both cases the chart distinguishes: 1) % of correctly recognised all time moments of a signal, 2) % of vowels in which 1 or more time moments were correctly recognised and 3) % of all correctly recognised time moments for vowels.

What is the most interesting in the experiment is the result of the test. It shows that the percentage of correctly recognised vowels is quite high and for each speaker it exceeds 85%, however, the percentage of correctly recognised time moments representing vowels is much lower and its values are 58.7, 48.1 and 54.1% (Tab. 1). It means that the remaining part of time moments was incorrectly recognised by the neural network: as other vowels or as „x". After averaging the percentage for all 3 speakers the values amounted to approximately 52 and 48% of all incorrectly recognised vowel time moments. It can be thus claimed that according to the neural network, the vowels in the tested fragment seem much shorter than it would follow from an expert's opinion supported by a spectrogram (Fig. 6a. b). For M1 speaker (Fig. 6.b.) the neural network recognised as many as 10 vowels on the basis of one time moment. For F1 speaker there were 7 such vowels, for M2 – 9, while in reality both in the speech of M1 and the other speakers such short vowels were practically not noticeable (Fig. 6.a).

Table1. Results of vowel recognition in a new fragment of text by the previously taught network

	Numbers of neurones in hidden layers and results of teaching a recognition network - % of all correctly recognised:				Test on unknown data – % of all correctly recognised		
Speaker	Hidden layers	time moments	vowels	time moments in vowels	time moments	vowels	time moments in vowels
F1	25; 30	95.9%	98.7%	89.4%	72.3%	85.1%	58.7%
M1	19; 19	95.6%	98.7%	89.3%	73.7%	92.3%	48.1%
M2	12; 12	93.6%	100.0%	86.1%	82.2%	85.7%	54.1%

As far as the remaining fragments of utterances are concerned, there also occurred cases of incorrect recognition. „x" time moments were recognised as vowels – 19.6% for F1 speaker, 15% for M1, and 5.4% for M2. The errors occurred nearly in all the cases when the utterance contained a voiced consonant.

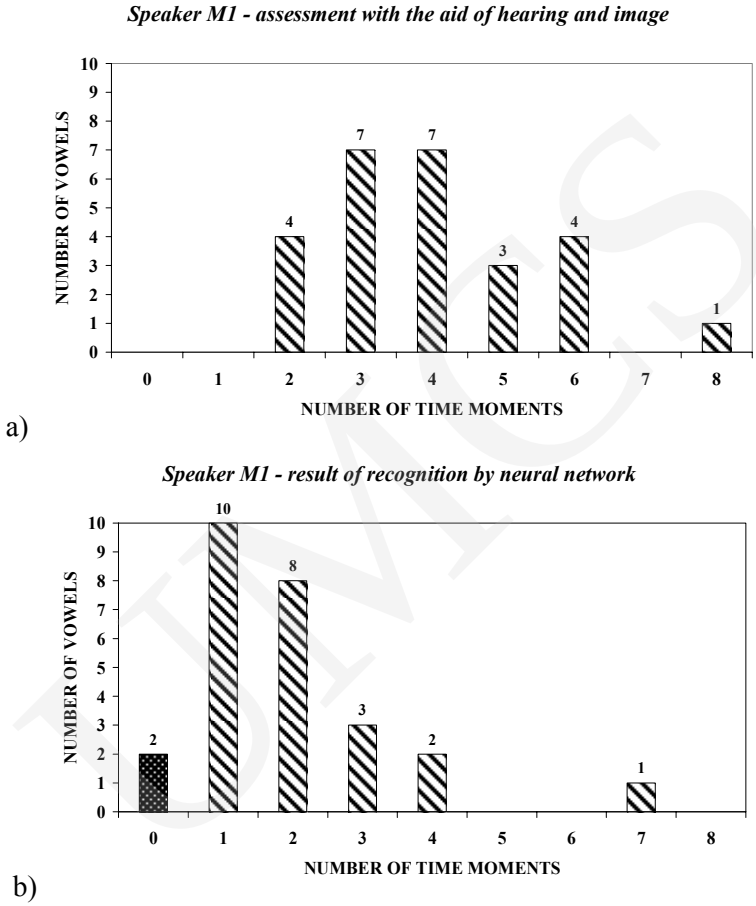

Fig. 6. Vowel time moments distribution in a fragment of a tested utterance by M1 speaker assessed by an expert (a) and neural network (b). The black column in fig. 6b represents the number of unrecognised vowels

Conclusions

If assessing the research results we take into account only the percentage of correctly recognised (according to the definition given above) vowels by the neural networks which tested new fragments of speech by a given speaker, then the obtained results are good – 85%, 92%, 86%. However, false recognition occurs, which significantly distorts the results. Their number should be decreased. In order to facilitate the analysis, they can be divided into 3 groups: 1) qualification of some or all time moments of one vowel as characteristic of another vowel; 2) recognition of vowel time moments as an „x" fragment; 3) recognition of „x" time moments as characteristic of various vowels. As far as the first group of errors is concerned, we may attempt at their elimination by

increasing the number of parameters characterising speech signals, e.g. by exchanging 1/3 octave bands with 1/6 octave bands, thus exposing more differences between particular vowels.

Thorough analysis of the second group of errors shows that some „x" moments are located at the beginning or at the end of a vowel. It may mean that there occurs a transition state between a consonant and a vowel and that the network „assessed" that fragment differently from the person who listened to that fragment. Here, thorough analysis of the spectrogram of this part of the signal and examination of network activation on its output may effectively diminish the number of errors.

The third group of errors may probably be decreased by the introduction of a distinct pattern for voiced consonants. It seems that the application of one „x" pattern for all parts of an utterance other than vowels was too big a generalisation, although the percentage of correctly recognised „x" moments was already surprisingly good.

If these solutions allowed for decrease in the number of errors in vowel recognition by a neural network in any given utterances of a given speaker, then it could be said that it is applicable in automatic vowel recognition. However, the authors are interested, as their ultimate goal, to prepare a set of data from many speakers – optimised as far as its size, such that a network taught on it could recognise vowels in any given text and independently from the speaker. That would, in turn, allow for practical application of the multilayer perceptron in image management in the visual speech echo-corrector designed for stuttering people [12,13].

Acknowledgements

The research was partially supported by Grant of Deputy Rector for Science of Maria Curie-Skłodowska University.

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

References

[1] Deller J. R. Jr., Hansen J. H.L., Proakis J. G., *Discrete-time Processing of Speech Signals,* Wiley-Interscience - IEEE, Nowy Jork., (1999) second edition.

[2] Grigore O., Gavat I., *Neuro-fuzzy Models for Speech Pattern Recognition in Romanian Language,* European Symposium on Intelligent Techniques, (1999), http://www.erudit.de/erudit/esit99/12581-p.pdf

[3] Hampshire J.B., Waibel A.H., *A novel objective function for improved phoneme recognition using time-delay neural networks,* Neural Networks, IEEE Trans. Neural Syst., 2 (1990) 216.

[4] Hild H., Waibel A.H., *Speaker-Independent Connected Letter Recognition with a Multi-State Time Delay Neural Network*, 3rd European Conference on Speech, Communication and Technology, Berlin, 2 (1993) 1481.

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

[6] Tebelskis J., *Speech Recognition using Neural Networks,* PhD Thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, (1995).
[7] Jassem W., Grygiel W., *Pattern-Based Classification of Polish Vowel Spectra Using Artificial Neural Networks,* Speech and Language Technology, Poznań, 6 (2002) 13, in Polish.
[8] Moore B.C.J., Peters, R.W., Glasberg, B.R., *Auditory filter shapes at low center frequencies*, The Journal of the Acoustical Society of America, 88(1) (1990) 132.
[9] Ozimek E., *Sound and Perception, Physical and Psychoacoustic Aspects,* Polish Scientific Publishers, Warszawa-Poznań, (2002), in Polish.
[10] Smołka E., Kuniszyk-Jóźkowiak W., Dzieńkowski M., Suszyński W., Świetlicki M., *Vowel Recognition in isolation and in continuous Speech with Application of Multi-Layer Perceptron,* Structures – Waves – Human Health ed. Panuszka R., Kraków, XIV(2) (2005) 143, in Polish.
[11] Smołka E., Kuniszyk-Jóźkowiak W., Dzieńkowski M., Suszyński W., Świetlicki M., *Vowel Recognition with Application of Muli-Layer Perceptron Neural Network,* Speech and Language Technology, Poznań, 8 (2005) 237, in Polish.
[12] Kuniszyk-Jóźkowiak W., Smołka E., Adamczyk B., *Effect of acoustical, visual and tactile echo on speech fluency of stutterers,* Folia Phoniatrica et Logopedica, 48/4 (1996).
[13] Smołka E., *Visual feedback in stuttering therapy,* Proc. SPIE 3054, Jankiewicz Z. (ed.), (1997) 235.