ISSN 1392-1215 ELECTRONICS AND ELECTRICAL ENGINEERING 2006. No. 8(72)
ELEKTRONIKA IR ELEKTROTECHNIKA

T170 ELECTRONICS
ELEKTRONIKA

Automatic Segmetation of Phonemes using Artificial Neural Networks

J. Kamarauskas
Institute of Mathematics and Informatics,
A. Goštauto str. 12, LT–01108 Vilnius, Lithuania, e-mail: j.kamarauskas@ltec.lt

Introduction

When pronouncing words, each word consists of certain parts – phonemes, therefore some sequence of the phonemes forms a word. Automatic segmentation is often used in speech, speaker recognition, and in the speech synthesis tasks. It is worth mentioning, that there are no mathematically based methods developed for this purpose as yet, and many techniques of automatic segmentation refer to heuristic methods. We could mention one of the proposed techniques of automatic segmentation that uses the maximum likelihood and least squares segmentation of autoregressive random sequences with abruptly changing parameters. The objective function was modified into the form suitable for applying the dynamic programming method in its optimization. Expressions of Bellman functions were obtained for this case [1]. But this method is complicated enough.

The other proposed methods of segmentation could be mentioned too: statistical method of segmentation of continuous speech [2], where the main idea is to model a signal by the statistical simulation and to use test statistics to sequentially detect changes in the parameters of the model. Another proposed statistical method of segmentation is used in the recognition of the connected words. The estimation algorithm based on quadratic polynomials is used that sets limits of the word [3].

In speech and speaker recognition systems various features are used that are calculated in short intervals - frames (duration is approximately 25ms) of speech signals. These frames overlap one another. The purpose of this research is to find out, how artificial neural networks that can assimilate linear and non-linear connection of the pattern, will be able to classify different frames which correspond to the different phonemes. In this paper, automatic finding of the start and end points of the words will be considered too. This problem can be solved in different ways [4], for example, by calculating signal energy in the frame and comparing it with the threshold: if energy exceeds the threshold, the frame can be classified as voiced, otherwise, it can be considered as noise. But this method makes gross mistakes, especially when the level of noise is high or non-voiced phonemes (for example s) must be separated out of the noise.

Features of the speech signal

In speech and speaker recognition systems various features are used [5], calculated from the short intervals (named as frames) of the speech signal: Linear prediction coding (LPC) coefficients, Fourier spectrum, cepstral coefficients, mel-cepstrum coefficients, cepstral coefficients, calculated according to the bark scale, and so on. It can be mentioned that the same features are often used in speech and speaker recognition systems.

Now we consider some features of the speech signal, which were used in this research.

In the model of linear prediction (LPC) coding, the speech signal is shown as an autoregression sequence. It is considered that in short time intervals the vocal tract is time-invariant, therefore the value of the signal can be approximately predicted having a certain count of the previous signal values in their linear combination

$$\hat{s}[n] = \sum_{i=1}^{p} a_i \hat{s}[n-i] + Gu[n], \qquad (1)$$

where $\hat{s}[n]$ is the nth predicted value of the signal, u[n] is an error signal, G is the amplification coefficient which makes the energy of the actual and the predicted signal equal, p is the order of the LPC model.

Coefficients of the predicted filter a_i are calculated by minimazing energy of the error signal. For this purpose the Durbin algorithm is often used [6].

Cepstral coefficients are obtained after implementing an inverse Fourier transform to the logarithm of the spectrum, that has been calculated using the Fourier transform. It is shown in Fig. 1–3.

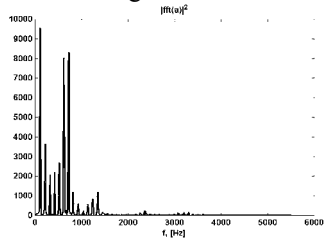

Fig. 1. Fourier transform of the frame of the signal

It is considered that some of the first cepstral coefficients can represent the vocal tract (after implementing their Fourier transform, we get the approximation of the logarithm of the Fourier transform). It is shown in Fig. 4.

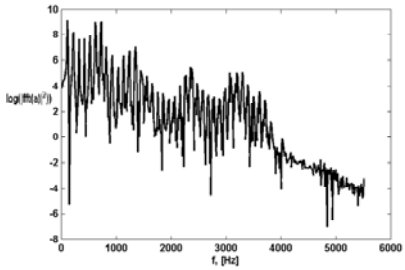

Fig. 2. Logarithm of the Fourier transform

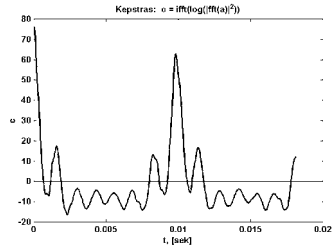

Fig. 3. Sequence of cepstral coefficients

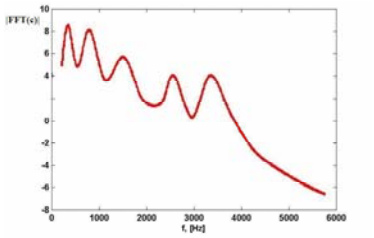

Fig. 4. Fourier transform of cepstral coefficients

Automatic segmentation

In this research the experiments of automatic setting of the start and end points of the words and automatic segmentation of the phonemes have been done, using artificial neural networks. Often there arise problems when we want to find the start and end points of the words automatically, and we have to distinguish between noise and non-voiced sounds (e.g. phoneme s), because for both the energy level is low and the spectrum of non-voiced sounds and that of noise are similar. The purpose of this research is to find how artificial neural networks will distinguish between noise and different phonemes i.e. automatically set the start and end points of the words, and how they will separate different phonemes, using various features of the speech signal.

Structure of artificial neural network (ANN)

The structure of an artificial neuron is shown in Figure 5. It is a very simplified model of the biological neuron. The output of the artificial neuron is formed using formula (2), every signal in the input is multiplied by the weighted coefficients and summed. Afterwards the value of the weighted threshold is subtracted from the sum and the signal obtained is affected by some activation function F.

$$b_j = F\left(\sum_{i=1}^{n} a_i w_{ij} - w_{0j}\Theta_j\right) \quad (2)$$

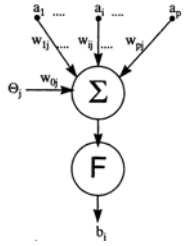

Fig. 5. Structure of an artificial neuron

Activation functions used in artificial neurons are shown in Table 1.

Table 1. Activation functions of the neurons

Function	Definition	Range
Identity	x	(-inf,+inf)
Logistic	$\frac{1}{1-e^{-1}}$	(0,+1)
Hyperbolic	$\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$	(-1,+1)
Exponential	e^{-x}	(0, +inf)
Softmax	$\frac{e^{x}}{\sum_i e^{x_i}}$	(0,+1)
Unit sum	$\frac{x}{\sum_i x_i}$	(0,+1)
Square root	$\sqrt{x}$	(0, +inf)
Sine	$sin(x)$	[0,+1]
Ramp	$\begin{cases} -1 & x \le -1 \\ x & -1 < x < +1 \\ +1 & x \ge +1 \end{cases}$	[-1,+1]
Step	$\begin{cases} 0 & x < 0 \\ +1 & x \ge 0 \end{cases}$	[0,+1]

The layers are formed by artificial neurons which are connected and artificial neural networks are obtained.

During this research the perceptron and back-propagation artificial neural networks were used. Structures of these networks are shown in Fig. 6 and Fig. 7.

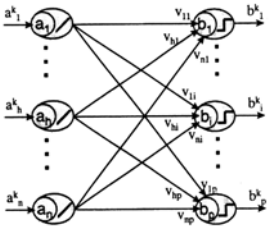

Fig 6. Structure of the perceptron neural network

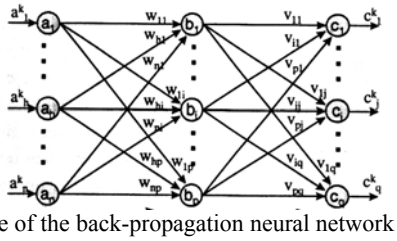

Fig 7. Structure of the back-propagation neural network

The count of neurons in the input layer depended on the used features and was equal to the count of the components of the features. Perceptron ANN's consist of two layers, and that of back-propagation of three layers. The count of neurons in the hidden layer was the same as in the first layer. Count of neurons in the output layer was n, where 2^n is the count of classes into which the input data had to be classified.

ANN training

The training algorithm depended on a chosen network: in the case of perceptron, the perceptron error correction procedure was used [7]:

- Input example A^k (feature vector calculated from the signal frame) is given to the input layer of the perceptron network, and output B^k of neurons is obtained in the output layer.
- Error D^k is calculated between the actual B^k and desired B'^k output of the neurons, $d^k=b'^k-b^k$
- If the error is rather great, the weighted coefficients are modified:

$$\Delta\omega_{i,j}^k = \lambda \cdot a_i^k \cdot d_j^k, \quad (3)$$

$$\Delta\Theta_j^k = \lambda \cdot d_j^k, \quad (4)$$

where λ is the constant of the training rate, a_i is the input signal of the neuron. In the case of the back-propagation neural network, the back-propagation training algorithm was employed [7]. For training and classifying the phonemes, Fourier transform coefficients (energy density spectrum), coefficients of the linear prediction coding (LPC) of the 12-th order and the cepstral coefficients of the 22-nd order have been used. Doing the first experiment (trying to set the start and end points of the words), the words „vienas", „du", „trys", „keturi", „penki", „šeši", „septyni", „aštuoni" were recorded. Features from the frames of the words „vienas", „penki" „šeši" and noise were used for artificial neural network training. The training results are shown in Tables 2 and 3. During the second experiment – automatic segmentation of the phonemes, 124 frames corresponding to every phoneme A, E, I and noise (496=124×4 frames all in all) were manually selected for training of neural networks. Artificial neural networks were trained until they could distinguish between the training data without errors.

Table 2. Training of perceptron ANN

Feature vectors	Number of epochs	Count of errors
Energy density spectrum	638	0%
LPC parameters	392	0%
Cepstral coefficients	1010	11%

Table 3. Training of back propagation ANN

Feature vectors	Number of epochs	Count of errors
Energy density spectrum	2872	7%
LPC parameters	2400	0%
Cepstral coefficients	2000	10%

The training results are shown in Tables 4 and 5.

Table 4. Training of the perceptron ANN

Features	Count of epochs	Count of errors
Energy density spectrum	48	0%
LPC parameters	58	0%
Cepstral coefficients	9	0%

Table 5. Training of the back-propagation ANN

Features	Count of epochs	Count of errors
Energy density spectrum	1600	0%
LPC parameters	600	0%
Cepstral coefficients	600	0%

Experimental results

During the first experiment, after artificial neural network training, various words were presented for setting the start and end points. In Figure 8, the signalogram and determined limits for words „du" and „trys" are shown, using the perceptron neural network and the energy density spectrum as features. The word „trys" ends with phoneme „s" and as we can notice, the perceptron network distinguishes the frames of phoneme „s" and noise. (Experiments were done using the automatic segmentation software **DNTsegm.exe**).

During the second experiment, after training the artificial neural networks, for each of them ten times recorded separated phonemes A E and I (feature vectors, corresponding to these phonemes) were presented. The results of correct classification, expressed in percents of the frames, corresponding to A, E, I phonemes and noise, using different feature vectors are shown in the tables below. The same phonemes were presented for each neural network.

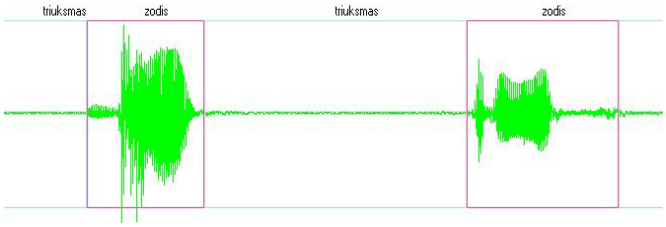

Fig. 8. Signalograms and set limits for words „du" and „trys"

Table 6. Recognition results of phoneme A frames

Features	Perceptron ANN	Back propagation ANN
Energy density spectrum	84,3%	83%
LPC parameters	86,2%	85,7%
Cepstral coefficients	91%	93%

Table 7. Recognition results of phoneme E frames

Features	Perceptron ANN	Back propagation ANN
Energy density spectrum	81,1%	83,2%
LPC parameters	78,5%	88,5%
Cepstral coefficients	85,4%	88,5%

Table 8. Recognition results of phoneme I frames

Features	Perceptron ANN	Back propagation ANN
Energy density spectrum	86,9%	88,5%
LPC parameters	87,4%	85,4%
Cepstral coefficients	91%	85,4%

Table 9. Recognition results of noise frames

Features	Perceptron ANN	Back propagation ANN
Energy density spectrum	95,2%	94,7%
LPC parameters	96,4%	96,9%
Cepstral coefficients	96,1%	97,8%

Conclusions

1. Artificial neural networks can classify different phonemes using various features of the speech signal: coefficients of Fourier transform, coefficients of linear prediction coding (LPC), and cepstral coefficients.
2. Artificial neural networks can be used to distinguish between the voiced frames and noise or to set the start and end points of the word. For this purpose the coefficients of the Fourier transform (energy density spectrum) are most suitable.
3. The error rate of recognition of the back-propagation artificial neural network is lower than that of perceptron, but its training is much longer.

References

1. **Lipeika A**. Segmentation of random sequences // INFORMATICA. – 2000. – Vol. 11, No. 3. – P. 243–256.
2. **Regine Andre-Obrecht**. A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals // IEEE Transactions on Acoustics. Speech and Signal Processing. – 1988, Vol 36, No.1. – P. 29–40.
3. **Zelinski R., Class F.** A Segmentation Algorithm for Connected Word Recognition Based on Estimation Principles // IEEE Transactions on Acoustics, Speech, and Signal Processing, 1983. – Vol. ASSP-31, No.4. – P. 818–827.
4. **Zilca R. D., Pelecanos J. W., Chaudhari U. V., Ramaswamy G. N.** Real Time Robust Speech Detection for Text Independent Speaker Recognition. Proceedings of Odyssey-04, The Speaker and Language Recognition Workshop. – Toledo, Spain, 2004. – P. 123–128.
5. **Picone J.** Signal Modeling Techniques In Speech Recognition // Proceedings of the IEEE. 1993.
6. **Rabiner L. and Juang B. H.** Fundamentals of Speech Recognition. – Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1993.
7. **Navakauskas D.** Skaitmeninio signalų apdorojimo priemonės. Dirbtinių neuronų tinklai. – Vilnius: Technika, 2000.

Submitted for publication 2006 04 01

J. Kamarauskas. **Automatic Segmentation of the Phonemes using Artificial Neural Networks // Electronics and Electrical Engineering. – Kaunas: Technologija, 2006. – No. 8(72). – P. 39–42.**

Automatic segmentation of phonemes is often used in speech technology. The purpose of this research is to find how the perceptron and back-propagation artificial neural networks (that can assimilate linear and non-linear connection of the pattern) distinguish between different phonemes, using various features of the speech signal used in speech or speaker recognition tasks: coefficients of linear prediction coding (LPC), cepstral coefficients, and coefficients of the Fourier transform (energy density spectrum). Artificial neural networks can be used for setting the start and end points of the word, too. They can separate not only voiced frames of the signal from noise, but also non-voiced, whose spectrum and that of noise are similar. Experiments were carried out and we can affirm that in order to segment the phonemes all the feature vectors used are suitable. However, if we want to separate different phonemes out of noise by automatically setting the start and end points of the word, the coefficients of the Fourier transform are most suitable, meanwhile cepstral coefficients do not fit. Ill. 8, bibl. 7 (in Lithuanian; summaries in English, Russian and Lithuanian).

Ю. Камараускас. Автоматическая сегментация фонем, используя искусственные нейронные сети // Электроника и электротехника. – Каунас: Технология, 2006. – № 8(72). – С. 39–42.

Автоматическая сегментация фонем часто используется в речевых технологиях. Цель этого исследования – установить способности искусственных нейронных сетей персептронов и обратного распространения ошибок (которые могут отличить линейные и нелинейные соотношения между предметами) отличить разные фонемы, используя разные векторы признаков, которые используются в распознавании речи и говорящего: коэффициенты линейного прогноза (LPC), кепстральные коэффициенты и коэффициенты Фурье трансформации (спектр плотности энергии). Искусственные нейронные сети могут быть использованы и в автоматическом определении начала и конца слова, потому что нейронные сети могут отличить не только вокализированные звуки от шума, но и невокализированные (которых энергия и спектр похожи). После проведенных экспериментов видно, что для сегментации фонем хорошо подходят все использованные векторы признаков, в то же время, пытаясь отличить разные фонемы от шума (для установления начала и конца слова), наилучше подходит спектр плотности энергии и не подходят кепстральные коэффициенты. Ил. 8, библ. 7 (на английском языке; рефераты на английском, русском и литовском яз.).

J. Kamarauskas. **Automatinis fonemų segmentavimas naudojant dirbtinių neuronų tinklus // Elektronika ir ektrotechnika.– Kaunas: Technologija, 2006. – Nr. 8(72). – P. 39–42.**

Automatinis fonemų segmentavimas – daug kur kalbos technologijoje taikomas uždavinys. Šio tyrimo tikslas ištirti perceptronų ir atgalinio sklidimo dirbtinių neuronų tinklų (galinčių įsisavinti tiesinius ir netiesinius pavyzdžių sąryšius) gebėjimą atskirti skirtingas fonemas panaudojant skirtingus kalbos signalų požymių vektorius, naudojamus kalbai ir kalbančiajam asmeniui atpažinti: tiesinės prognozės modelio (LPC) parametrus, kepstro koeficientus bei energijos tankio spektrą. Dirbtinių neuronų tinklus galima naudoti ir automatiniam žodžio pradžios ir pabaigos taškų nustatymui, kadangi dirbtinių neuronų tinklas sugeba atskirti ne tik vokalizuotus, bet ir nevokalizuotus garsus nuo triukšmo (kurių spektrai ir energija panašūs). Atlikus eksperimentus paaiškėjo, kad fonemoms segmentuoti gerai tinka visi panaudoti požymių vektoriai, tuo tarpu siekiant atskirti skirtingas fonemas nuo triukšmo (žodžio pradžios ir pabaigos taškams surasti), geriausiai tinka energijos tankio spektras ir netinka kepstro koeficientai. Il. 8, bibl. 7 (lietuvių kalba; santraukos anglų, rusų ir lietuvių k.).

DOI: 10.5755/j02.eie.10786