THE ALGORITHM FOR THE PHONEMIC LABELLING AND SEGMENTATION OF SPEECH WAVEFORMS USING FEATURE MAPS

VLADIMIR CHUCHUPAL

Computer Center of the Academy of Sciences of the USSR
Moscow, USSR, 117967

ABSTRACT

In this work the algorithm for the phonemic labelling and segmentation of speech waveforms is described. This algorithm is founded on the feature maps: the self-organized neural networks model. The model is able to form automatically a representation of distribution of speech signal parameters. The algorithm described below utilizes this ability in order to form criteria of phonemic labelling and segmentation. In the such manner we produce the representation for not only short-time signal parameters but also of the temporary trajectories of this parameters.

INTRODUCTION

One of the most succesful speech recognition methods is one, which founded on the use of statistical laws, which has been established in the speech signal parameters distribution. Therefore, it seems important to investigate methods which is able to accomulate the data about distribution of speech signal parameters, for example, to approximate the probability density function of mutial distribution of this parameters. Often this task can be solved satisfactory by means of self-organizing neural network models, in particulary, the model for the self-organized formation of structured feature maps /1/.

Let us $\mathcal{L}$ be a pattern space, the elements of $\mathcal{L}$ may be represented by vectors $\bar{x} \in R$ (pattern vector). The structured representation of is formed with the help of matrix $M_{L \times L}$ (feature map) with the elements $\bar{m}_{ij}$. Every m_{ij} is defined by it's time-variable weights $\bar{m}_{ij} = (\mu^k_{ij})_{k=1,n}$. Initially, the values of the $\bar{m}_{ij}$ choosed in the randomly manner. An algorithm creation of features map consist of two steps /1/. Let us, for the time moment t, t=0,1, ..., n, ... the input pattern vector would be x(t). Then, in the first step, we define the indexes i_0, j_0 of the element $m_{i_0 j_0} \in M$, such, that:

$$\| \bar{x}(t) - m_{i_0 j_0} \| = \min_{i,j} \| \bar{x}(t) - m_{ij} \| \quad (1)$$

In the second step the modifications of weights μ^k_{ij} is made. For $m_{i_0 j_0}$ and its neighbours (for example, if the radius r(t)=1, the neighbours for the $m_{i_0 j_0}$ will be $m_{i_0+1\, j_0}, m_{i_0-1\, j_0}, m_{i_0\, j_0-1}, m_{i_0\, j_0+1}$):

$$\bar{m}_{ij}(t) = \bar{m}_{ij}(t-1) + \alpha(t)(\bar{x}(t) - \bar{m}_{ij}(t-1)) \quad (2)$$

In equation (2) $\alpha(t)$ satisfy the conditions: $\sum_{t=0}^{\infty} \alpha(t) = +\infty$; $\sum_{t=0}^{\infty} \alpha^2(t) < +\infty$; $\alpha(t) > 0$

It was shown /1,2/ that for correct choise the values of the $\alpha(t)$ and r(t), described above process has the next properties. When $t \longrightarrow \infty$ the values of μ^k_{ij} change so, that adjacent elemehts of the matrix M respond to (in the sense of equation (2)) closed (in the sense of norm $\| \cdot \|$) vectors from space $\mathcal{L}$. The distribution of values μ^k_{ij} on the matrix M approximates the mutial distribution probability density function for patterns vectors.

The successful application of feature maps for fonemic labelling have been made in the work /3/. But the fonemic qualities of the sounds depend not only of it's short-time spectra, but also the context - phonemic qualities of the adjacent phonemes. In our investigation the method , was described, and feature maps, produced in the such manner, was used for creation the segments boundary criteria and accomulation the information about temporary traectories of spectral parameters. It's apparently, that this information may be useful for transeme segments analysis.

AN AUTOMATIC FORMATIONS THE CRITERIA FOR SETTING THE LABELS OF THE SEGMENTS BOUNDARY IN THE SPEECH SIGNAL

We assume, that the important role in the speech perception belongs to the stationary segments of speech and the silence segments. This segments may be viewed the adaptation's signals for our hearing system in the sense of adaptation to amplitude spectra of the sound. Therefore, the labels setting, in order to mark the stationary segments, may be useful on one hand, to produce the phonemic identification this segments, and on the other hand, to correctly identify the transition segments, which phoneme interpretation depend on long-time information. As the input patterns we used the short - time spectra $S(\omega,t)$ and the phonemic function /5/: $\Phi(\omega,t) = \ell g(|S(\omega,t)| - |S(\omega,t-\tau)|)$, where ω denote frequency, t - time, and τ - small time delay. We use the FFT algorithm in order to calculate the 252 - point amplitude spectra (divided into 21 frequency channel in the range 40 Hz-5 kHz) every 12.6 ms. Central frequency of each channel was equally spaced and the channel 22 contented the total energy of the segment. The values of fonemic function calculated from two adjacent short - time spectra. We used the synthetic sounds. Three sounds modelled the vowels. This formant frequency were spaced at 900 Hz, 1600 Hz, 2900 Hz. One sound was represented as an unvoiced fricative.

On the first step we formed two maps: map for short - time spectra and the second map for fonemic function values. The matrix M contained 6 6 elements in both cases. The process (1)-(2) contained T=20000 steps. The values and decreased linearly: $\alpha(t) = \alpha_0(1 - t/T)$, $r(t) = r_0(1 - t/T)$ where $\alpha_0 =$, $r_0 =$. We denoted sounds stimulus as A, B, C, D. The resulting maps are shown in the figures 1 and 2. In order to denote the elements of maps the next procedure was applied /1/. Approximately one hungreed of well known patterns of every sound were presented to input the algorithm (1)-(2). The element, mainly corresponded (in accordance with (1)) to patterns of the sound A was denoted A.

C - - D D D

- - - C - A

A - - A A A

A - A - - B

- - - - - -

A - - B - B

Figure 1. The feature map for the values of the short - time spectra. The symbol '-' denotes the nolabelled elements.

N - - - - -

- - - - - -

N - N - - -

N N N - - -

- - N - N -

S - N - - -

Figure 2. The feature map for the fonemic function values. The symbol 'S' corresponds the values of fonemic functions for stationary segments of the sounds A,B,C. The symbol 'N' denotes the same segments of the sound D. The symbol '-' denotes the transition segments.

In the segmentation and labelling algorithm we supposed, that elements denoted by the symbols 'N' and '-' would be correspond to nonstationary segments.

In the work /6/ it have been made suggestion about existance of special cells - detectors for phonemes boundary detection. The first question was: may the map of fonemic function values to be use as as the map of such detectors ? We tested this capability of the map using the continious signal, contained 140 above - mentioned sounds. On the map of fonemic function values we obtained the trajectory, consisted of the elements, that corresponded the sequence of input patterns. The algorithm produced the label of transition region (segments) when this corresponded element was belong to transition region of the map (or, another words, was denoted as transition element). The labels of stationary regions were produced in the such manner. In the case of stationary regions the algorithm made an attempt to interpret this segments in accordance with the map of short - time spectra. The analysis of the result shows that all stationary segments belonged vowels sounds have been labelled correctly. About 8% of the transition regions were omissed. All stationary segments were recognized (with respect to map of short - time spectra) right.

THE USE OF THE INFORMATION ABOUT TEMPORARY TRAJECTORIES OF THE PARAMETERS FOR THE FONEMIC LABELLING

In order to use the temporary trajectories of the parameters as a feature patterns, let us see the next feature map (denote it map III) formation process. The input vectors for this map consisted from the values of the outputs of the map of short - time spectra (map I). The dimensionality of the input vectors to map III is equal to the number of elements in the map I, and the values of the components of this elements are equal to output values (see equation (1)) of the correspond elements of the map I (these output values have been added during some times). It can be said, that each element of the map III is connected with each element of map I. In order to control the map III formation process we used the map of phonemic function values (map II). When the corresponded element of the mapII was the element, denoted as ststionary, the label of stationary segment was produced. Up to this moment the values have been summing up and the result was used as the input vector to map III. The produced label element of map II became nonactive for some time. For the formation of the map III we used the map I and map II, described above. The number of elements in the map III was 4x4. The process contains T=6000 steps. The values of the radius r(t) and parameter $\alpha(t)$ where chosen as it was shown below. The result is presented on the figure 3.

BB — — AA

— — — AA

BC AB — AA

CC — — AB

Figure 3. The feature map for temporary trajectories of parameters. Here AA, BB, CC, are corresponding to the stationary segments, BC and AB are corresponding to transition regions.

In the test signal for formation of the map III we used transition region beetween A and B, B and C, C and D, D and A only. It is clear, from the figure 3, that no exist elements, that correspond to transition regions CD, DA and sound D. We tried to label the test signal, described above, with the help of the map III. In this case the algorithm was the same as the algorithm for the creation of the map III. The only difference between then was that in the algorithm of the labelling, every input vectors was identified in according with map III. As it was expected, we received 100 of correct detection of transitions between B and C, A and B and stationary segments of A,B,C. But the detection of the sound D and transition regions CD and DA contained many mistakes.

CONCLUSION

It have been shown in our works, that use of model of the feature maps formation yields the possibility to form in the simple manner the labelling and segmentation rules founded on statistical properties of the signal. This rule uses the properties both stationary and transition segments of signal.

REFERENCE

1. Kohonen T. Self - Organization and Associative Memory, Springer, 1983
2. Cottrell M., Fort J.C. A Stochastic Model of Retinotopy: A Self-Organizing Process. Biol.Cybern., Vol.53, No 6, 1986.
3. Kohonen T., etc. Phonotopic maps - Insightful Representation of Phonological Features for Speech Recognition, Proc. of PRIP-84, Montreal, pp. 182-185, 1984.
4. Бондарко Л.В. Фонетическое описание языка и фонологическое описание речи. Л., Изд-во Ленинградского ун-та, 1981
5. Пирогов А.А. К вопросу о фонетическом кодировании речи. Электросвязь, 1967, №5, с. 24-31.
6. Чистович Л.А., Венцов А.В., Люблинская В.В. Слуховые уровни восприятия речи. Функциональное моделирование. В сб.: Акустика речи и слуха. Л.,"Наука", 1986.