SPEECH SOUNDS IN FREQUENCY FOLLOWING RESPONSES OF THE AUDITORY SYSTEM

ELENA A. RADIONOVA

Lab. of Hearing Physiology, I.P.Pavlov Institute
of Physiology, Leningrad, USSR, 199034

ABSTRACT

Speech sound signals are much better reproduced in the summed neuronal activity than in the activity of single neurons. The most complete reproduction is observed at the lower levels of the auditory system. At higher levels only information concerning the signal periodicity may be partly retained.

At present it is well known that sound signal parameters may be but poorly reflected in the impulse responses of single neurons of the auditory system. This is the case even for simple pure-tone signals. For instance, as it was found for the cat, at least about half of neurons from the cochlear nucleus (which is the first auditory brain level, receiving the whole auditory information from the ipsilateral cochlea of the inner ear via the auditory nerve fibers) show a pronounced nonlinear relation between the signal frequency and the impulse response value: at signal intensity of 50-80 dB this function has several (up to about ten) maxima separated by different frequency intervals over the frequency range from about 1 to 20 kHz. Besides, the time patterns of a single neuron responses often show very slight differences (if any) over a wide range of signal frequencies. At higher levels of the auditory system the correspondence between the sound signal parameters and the single neuron responses declines even to a greater extent. The above properties as well as some others, make each neuron, When alone, unable to indicate what kind of a sound signal comes to the ear.

Meanwhile the summed response of a number of neurons may give appropriate information about the signal presented to the ear. This possibility was first proposed as the so called "volley" principle /1/ and was then supported by the experiments with the "Frequency Following Response" (FFR) registered from the lower levels of the auditory system as the summed response of a group of neurons with the near spatial positions within a given brain region.

The FFR is the result of the activity of a number of neurons whose impulse responses are synchronized with a certain phase of the tonal signal. The upper frequency limit of this synchronization was reported as at least 5 kHz for the auditory nerve fibers, about 6-6.5 kHz for the cochlear nucleus level, with a pronounced diminution of this value at the higher auditory centers: to about 1.5 kHz at the midbrain level (inferior colliculus) and to about 1 kHz or even less resp. at the medial geniculate body and the auditory cortex levels /2/. Thus, the widest frequency range reproduced in the FFR is related to the lower levels of the auditory system.

It was found that FFR not only followed the tone frequency but could also reproduce the wave form of rather complicated sound stimuli. This was especially well observed at the cochlear nucleus level with complex harmonic signals containing 2 to 6 harmonics, as well as with the sound speech signals. For instance, when two-tone complex of the second and the third harmonics (or of some others) was presented while varying the signal waveform through variation of the phase of the higher harmonic, the FFR evoked by this complex signal usually reproduced rather precisely the waveform of the signal, with almost all the details: for each oscillation in the complex signal wave a cooresponding deflection in the FFR could be observed.

However, in some cases nonlinear phenomena took place, when one of the signal harmonics, usually the lower one, was fully suppressed in the FFR. It seems of interest that this suppression depended on the phase of the higher harmonic and could be observed within a certain phase range only.

Inspite of some nonlinear features, the FFR of the cochlear nucleus can reproduce rather well the sound speech presented to the ipsilateral ear. The speech reproduced in the FFR is well distinguished, the masculine and feminine voices can be well distinguished as well and even the person can be sometimes identified according to the voice reproduced in the FFR. The prosodic characteristics of the speech are also well reproduced in the FFR of the cochlear nucleus.

Thus, it may be concluded that at the cochlear nucleus level the sound speech signal can be rather fully described in the summed activity of the population of the neurons. The characteristic frequencies (CFs) of neurons forming such a population should be relatively similar, since these neurons are obvioslyy positioned rather near each other as their summed activity, the FFR,is recorded with the help of the same electrode. Besides, for the best reproduction of the speech signals, the CFs of the neurons whose activity forms FFR should lie at the upper frequency limit of the speech sound range (for instance, of about 4 kHz) or higher: such neurons would respond to the wide frequency range below the values slightly higher than the CFs. The neurons with the lower CFs would not respond to high components of the speech signals. Therefore only populations of neurons with sufficiently high characteristic frequencies would reproduce sound speech in their summed FFR.

At the higher level of the auditory system, namely, at the central nucleus of the inferior colliculus (IC) speech sounds may be reproduced in the FFR much more roughly than at the cochlear nucleus level. The sound speech, as it is reproduced in the FFR of the inferior colliculus, is not distinguishable now, though speech prosodic characteristics are well pronounced when the FFR is listened to through the audio reproduction system. These restrictions in the description of the sound speech in the FFR are obviously connected with the restricted frequency range (not higher than 1.5-1.7 kHz) reproduced in the FFR at the level of the inferior colliculus. This, in turn, may be connected with the properties of single neurons forming the FFR: the neuronal impulse activity is much lower than the activity at the cochlear nucleus level, its variability is greater, as well as nonlinear effects due in particular to neuronal inhibitory interconnections.

When analysing the inferior colliculus FFR to complex sound signals containing 2-6 harmonics it may be seen that, unlike the cochlear nucleus, the FFR from the IC does not describe the complex waveform in detail. Usually only the periodicity of the signal waveform is reproduced in the FFR quite well, especially the periodical sharp changes in the signal amplitude. Other, more delicate, though rather essential changes in the signal waveform produced for example by phase variations of the signal components may still be reflected in the FFR-waveform, but with many details lost.

As to the highest, cortical level of the auditory system, the FFR may be registered here only in a narrow low frequency range, for instance, of about 100 Hz /3/. Cortical FFR can reproduce only the low frequency envelope of complex sound signals, the speech sounds including, without any details concerning their waveform. The fact is, that neurons of the highest levels of the auditory system, the medial geniculate body and the auditory cortex, are greatly specialized concerning different sound parameters. Besides, their responses are variable and nonstationary to a great extent, with inhibitory effects well pronounced. These neuronal properties result in great restriction of both the FFR amplitude and frequency range, which makes the sound speech reproduction impossible in the summed neuronal activity at these auditory regions.

Thus, rather complete analogue description of speech signals in the neuronal FFR can be observed only at lower levels of the auditory system. At higher levels a restricted description based on definite signal features seems to substitute step-by-step the full description of the signals: now only pronounced changes in the signal envelope or transients may be reflected in the summed neuronal responses. Meanwhile it may be thought that at higher levels of the auditory system there would be a possibility to extract in a way the information from the lower levels concerning more complete description of complex signals, the speech signals including. It is not clear yet how it could be done. Still it may be supposed that the main function of the higher auditory centers (together with some other brain high centers) would be to form general ideas or at least sound images, i.e. mental pictures of the human voice, of the animal cry, of the step sound etc., which should be necessarily connected with a loss or diminution of the information relating to particular details of the real sound signals.

REFERENCES

/1/ E. G. Wever, "Theory of Hearing", Wiley, 1949.

/2/ E. A. Radionova, "Neurophysiological Studies of the Monaural Phase Sensitivity of the Auditory System". In: Sensory Systems, "Nayka", 1982, 72-86 (in Russian).

/3/ M. Steinschneider, J. Arezzo, H. G. Vaughan, Jr., "Phase-locked Cortical Responses to a Human Speech Sound and Low-frequency Tones in the Monkey". Brain Res., 1980, v. 198, 75-84.

THE ASSOCIATIVE BIONIC APPROACH TO THE DEVELOPMENT OF THE SPEECH SIGNAL PROCESSING CENTRAL MECHANISMS

A.A.Kharlamov

Moscow

The goal of the report is to show possibilities of the associative bionic approach to construction of the model of the speech signal processing central mechanisms. The basis of this approach is the data processing in the dinamic associative memory block (DAMB) and its usage for constructing hierarchical structure (HS) for phonetic, lexical and syntactical processing of speech.

INTRODUCTION

The analysis of the data representation methods that has been used in the speech recognition system shows that graphs (matrix) representation is the best. For example in the form of an evidently given graph or a hidden Markow model. The methods in question have some disadvantages: data representation inflexibility, graph labour consuming forming or need in large computer power. Hardware realisation of graph (matrix) data representation by the system of DAMB is free of these disadvantages.

The DAMB construction is based on some biological facts about structure and properties of neurons and their pulls. The uniform structure and a simple data processing algorithm allow to produce the DAMB using the microintegral technology as an integral system on the sylicon plate.

THE MODEL OF THE SPEECH SIGNAL PROCESSING CENTRAL MECHANISMS.

The main functions of DAMBs and HS formed of them are: storage of data with compact packing, reproducing it with the help of context association, and the statistical processing of the input data by picking out the number of different occurency frequency elements (i.e. vocabularies); extraction of vocabulary word relations in the input data - that allows to reconstruct the inherent input data structure. The above functions give us the possibility to model phonetic, lexical and syntactical levels of data processing by the HS of DAMB. It is supposed that the acoustical speech signal is preprocessed, optimal in each specific case, which is not discussed here.

The model in question consists of two data processing chanals: the coarse one and the precise one. When training the precise chanal performs compilation of a phonotype vocabulary $\{P\}$ and on its base the vocabularies of lexical level sublevels: i.e. the word vocabulary $\{L\}$ and the morph one $\{M\}$. When training the coarse chanal forms the syllable-phoneme vocabulary $\{SP\}$. The unit segmentation of the corresponding levels is performed by DAMBs as the natural feature of the data processing in them. The type of a unit is determined by the DAMB parameters (of a hiperqube dimension).

In the process of recognition the data in the input of the lexical level is represented in terms of syllable-phoneme vocabulary, i.e. syllable representation, as a number of syllable type sequences in the corresponding words, or morphs L^{sp}_j = = (SP_I, SP_2, ... , SP_i). The whole number of lexical level input is divided into the subvocabularies according to the equal syllable representation principle, these subvocabularies are indexed by this representation L^{sp}. If the vocabulary consists of only one lexical unit or there are high level constraints (contextual), which allow us to choose the necessary alternative, the recognition process is stopped. Let us assume $L_k \equiv L^{sp}_j$.

If it is necessary to choose a lexical unit from the subvocabulary $\{L\}$ L^{sp}_j, which is indexed by the given syllable representation L^{sp}_j, the precise chanal is used. In this subvocabulary the lexical units L_k and L_l (L_k = (P_I, P_2, ... , P_m, ...)) are divided by one or more phonetic element types P_m. To divide phonotypes the preprocessing form is used which is associatively related to the phonetic element type P_m. This division uniquelly determines the lexical level unit L_k.

The higher levels (from the lexical point of view), the syntactical, for instance, bring in additional (contextual) constraints on lexical level unit L_k choosing. In the process of training in the syntactical level the words and morphs relations (i.e. inflections) vocabulary $\{F\}$ and the type phrase vocabulary $\{Pr\}$ are compiled.

The above mentioned structure can be realized as a HS from DAMB.

THE DAMB FORMALIZATION.

The DAMB is a net of neuroliked elements (NE) with the input signal time summaring, which is accomplished by shift register of n cells - the model of the generalized dendrit . The DAMB consists of 2^n NEs and each of them models one of the n-dimensional unitary hiperqube node in the signal space.

The binary sequence $A = (\dots, a_{-I}, a_0, a_I, \dots, a_i, \dots ; a_i \in \{0,I\})$, the input sequence for the DAMB, is mapped into the hiperqube as a directed sequence of the nodes - trajectory $\hat{A} = F(A)$. Each n of the symbols from the sequence $(a_{i-n-I}, a_{i-n-2}, \dots, a_i)$ corresponds to the node $\hat{a}_i$ with the given coordinates. The initial sequence A can be restored from this trajectory $A = F^{-I}(\hat{A})$. F mapping has the property of associative addressing to the data with the help of context association (n sequential symbols).

The DAMB can operate in one of the following three modes: (I) training or perception; (2) reproduction; and (3)structural processing.

The training mode.

In the process of training NEs of the DAMB change their inner state under the influence of an input sequence. This changing means that the memory function H is inserted to the nodes of the hiperqube. That is why the trajectory is stored (i.e. in the current node the sequence next symbol is stored in case the record is autoassociative, or the informational sequence current symbol is stored in case the record is heteroassciative).

Let us provide function H with the thresholding properties. This allows to process the data that is stored in DAMB, statistically. The processing allows to compile the vocabulary of ivents $\{\hat{B}\}$, from the input of the DAMB as the number of sequences $\{A\} : \{\hat{B}\} = F(\{A\})$. Under the influence of the threshold value h of function H words of the vocabulary are either the union of input ivent trajectories (in that case the whole data is stored), or fuzzy or precize intersection (in this case more or less common part of data is stored). The identical parts of the sequences or events are mapped into the same chain, and different parts are mapped into the different ones. As a result a directed metrized graph or a graph-word is formed at the nodes of the hiperqubes. The trajectory attenuating models the forgetting process.

The reproduction mode.

The stored data reproduction using of F^{-I} mapping allows to recognize the input sequence by comparing it with the reproducing one according to a measure system. Only the pretrained DAMB can operate under reproducing. The n-member segment of the DAMB input addresses to one of the hiperqube nodes (to one of the NE) where some data is stored (the inner states were changed). The trained NE answer, added to the input n-member segment, determines a new address and thus a new node is addressed. And so on.

The structural processing mode.

Under the structural processing the input sequence is compared with the compiled in the DAMB vocabulary with the sequence segments changed by zero sequences, if these parts of sequences, corresponding to the parts of trajectories, coincide with the node sequences (chains of the vocabulary graph-words). Thus the special F_c^{-I} mapping allows to eliminate the vocabulary data from the input sequence, and to preserve only the relations of the vocabulary words. The abbreviation sequence (AS) $C = F_c^{-I}[F(A), \{\hat{B}\}]$ is formed. The mechanism of the AS forming allows to use the DAMBs for structural data processing within the DAMB hierarchical structure, as some rarefied parallel data flows can be united into one flow without losses.

The hierarchical structure of the DAMB.

There are some parallel processes $\{A\}^I \oplus \{A\}^2 \oplus \dots \oplus \{A\}^q = \{\mathfrak{A}\}$ - that is the situation - in the HS input. The vocabularies $\{\hat{B}\}$ of the most frequently occured situation $\{\mathfrak{A}\}$ events are formed in the first level DAMBs. After the vocabularies compilation, if sequences $\{A\}^q$ are given in the first level DAMB inputs, the AS are formed in their outputs. These AS converge in the second level DAMB inputs and compile the vocabularies $\{\hat{D}\}^r$, $(r = I, \dots, R;\ R < Q)$ in the DAMBs. Thus the input situation model is formed in the HS as a repeatedly enclosed directed metrized graph. In that graph the graph-words of the low level vocabularies are enclosed into the corresponding parts of the high level graph-words. The HS curtale the input data in the down-up direction and vice versa. That HS property allows to reproduce the stored situation with the help of association both from high level and low level. (Thus the HS can be used as an analyser and an effector as well).

CONCLUSION

The report is devoted to the development of a theoretical model of data processing at the phonetic, lexical and syntactical levels. That model allows to create a device for structural speech signal processing. That device automatically performs compilation of vocabularies of those level units, reconstruction of those level grammars, and recognition of the input events by matching them with the compiled vocabulary units.