KFKI-1985-42

I. BORBÉLY
B. LUKÁCS

THE DEVIATION FUNCTIONAL OF PHONEME RECOGNITION

Hungarian Academy of Sciences

CENTRAL
RESEARCH
INSTITUTE FOR
PHYSICS

BUDAPEST

KFKI-1985-42

THE DEVIATION FUNCTIONAL OF PHONEME RECOGNITION

I. BORBÉLY, B. LUKÁCS

Central Research Institute for Physics
H-1525 Budapest 114, P.O.B.49, Hungary

HU ISSN 0368 5330

ABSTRACT

Using general considerations and the standard methods of functional analysis we develop a formalism for the phoneme recognition problem. The conclusion is that a phoneme is characterized by not only its standard, but by the "directions" too, in which physically small changes cause maximal distortion of its character. In addition, the different phonemes are arranged into a system, whose basic organizing rules can be language-dependent.

АННОТАЦИЯ

Используя методы функционального анализа был развит общий формализм для распознавания фонем. Фонемы характеризуются не только стандартом, но также и "направлениями", в которых физически маленькие изменения дают наибольшие искажения. Основные правила, организующие фонемы в систему, могут зависеть от языка.

KIVONAT

Általános megfontolások és a funkcionálanalizis standard módszerei segitségével kidolgozunk a hangzófelismerés problémájára egy formalizmust. A végkövetkeztetés az, hogy egy hangzót nem csak standardja jellemez, hanem az olyan irányok is, melyben fizikailag kis változások maximális torzulást okoznak jellegében. Továbbá az egyes hangzók rendszert alkotnak, melynek alapvetõ szervezõ elvei nyelvenként különbözhetnek.

1. INTRODUCTION

Phonemes are the atoms of the speech. It means that they are shortest parts of speech remaining more or less unaltered when the fluent speech is built up from them [1]. In some languages other components (e.g. the stress or the intonation) may also be vital in word recognition; the location of the stress produces lexical differences in English or Russian, or it is well known that the intonation plays a similar role in Chinese, moreover, there are some insolated examples for this phenomenon even in Swedish [2], [3]. Nevertheless, even in such cases the phonemes remain important. The general success of alphabetic writing systems, adapted to a great variety of languages sometimes very far from the original Semitic languages for which their ancestors had been developed, indicates that some recognition pattern exists for phonemes. A simulation of such a pattern is the optimal solution for the automatic speech recognition. However, there is a quite general experience that one is completely lost at his first encounter with a foreign speech: it is not possible to recognize anything at all. Therefore in different languages these hypothetical phoneme recognition patterns can be different. Of course, for some languages they may be common in gross features. As an example we quote the usual and well known two formant distribution of vowels in Hungarian [4] and in American English [5], together with Lotz's data for the average formant frequencies in the Ottoman Turkish [6], (see *Fig.1*). In English the location of the different vowels seems to be more or less random, while the orientation of their regions are more or less radial, in the same time in Hungarian the structure is rather vertical, both for the location and for the orientation of the vowel regions, and such a vertical structure (for partly different wovels) is compatible with Lotz's Turkish data too. We shall return to this problem in Sect. 8, here *Fig.1* is merely a demonstration of the existence of different recognition systems.

In such a case the experience collected in a linguistic community is not necessarily convertible to an other language. It may be useful to build up a general formalism without references to specific languages, as far as it is possible. This is the goal of the present paper. Of course, one cannot develop a really general and language-independent formalism; the simplifying assumptions (which are always necessary) are characteristic for the authors'

preconceptions which are determined by their first language. E.g. here we restrict ourselves to the phoneme level, and treat them as separate indentities. It is a restriction, it cannot be too severe at least for Hungarian and for other languages with similar acoustic system, but there is no a priori way to decide, which are the languages for which this approximation is good enough.

For such a general treatment a formal scheme seems to be suitable. Clearly, no formal scheme can entirely solve any problem, nevertheless formalization is very useful since it gives the framework for the empirical information, and, what is more important, it organizes the research process and calls attention to crucial points. It is really astonishing how far reaching hypotheses can be at least formulated using very limited amount of actual information. So it is unwise not to take the advantages of an adequate formalization. We basicly use the standard methods of linear algebra and functional analysis inspired by some methods of pattern recognition [7].

2. GENERAL REMARKS

The phoneme structure and the recognition process may be different in different languages. Nevertheless, some elementary facts seem to be valid for all languages. Here we list three of them:

a/ The majority of the physiologically possible sounds does not correspond to any phoneme.

b/ A given sound of speech accepted as a realization of a phoneme belongs solely to that phoneme, i.e. practically there are no transition regions. Otherwise the mutual understanding would be impossible.

c/ Some realizations of a given phoneme can be clearly distinguished as sounds, therefore not the physiological limitation of hearing defines the domains of phonemes.

Points a/ and c/ suggest that the most important factor in the recognition is the analysing activity of the brain. The details of this activity may be quite complicated, nevertheless it is not hopeless to look for a relatively simple mathematical description. An encouraging example is the problem of color vision. There the final goal is to restore the reflection properties of the illuminated surface. The relative color constancy indicates that the brain is able, in fact, to achieve this goal by using some data of the illuminating and reflected lights. This seems to be a very complicated process, nevertheless the final result can easily be formulated by means of Riemannian geometry [8], [9].

3. THE DEVIATION FUNCTIONAL

Consider a linguistic community with complete mutual understanding. This mutual understanding implies that there exists a common probability of accepting a sound as the realization of a given phoneme. This acceptance probability is denoted here by p_α, where α labels the sounds; let us postpone the question which are the important characteristics of the sounds compressed here into a formal index. Obviously, there are minor individual variations in p_α even within a linguistic community, but one may neglect them in first approximation. The acceptance probability can be measured; for the scheme of an idealized measurement see Ref. 10. Since p_α is a probability, $0 \leq p_\alpha \leq 1$. Now, let us introduce the standard of a given phoneme: the standard is the sound for which p_α is maximal. We do not assume the unicity of the standard; a definite counterexample will be shown in Sect. 4. The acceptance probability for the other possible sounds is smaller. Therefore one can introduce a number Φ measuring the deviation from the standard via the acceptance probabilities $\Phi=\Phi(p)$. We want to reduce the recognition process to a minimum searching, therefore the following conditions are suitable for the actual gauge $\Phi(p)$:

$$\frac{d\Phi}{dp} < 0 \; ; \quad \Phi(0) = \infty ; \quad \Phi(p_{max}) = 0 \tag{3.1}$$

One can use a specific gauge

$$\Phi = -\ln(p/p_{max}) \tag{3.2}$$

which, obviously, satisfies the above conditions.

Eq. (3.2) gives

$$p = p_{max} e^{-\Phi_\alpha} \tag{3.3}$$

The mutual understanding requires

$$1 - p_{max} << 1 \tag{3.4}$$

otherwise communication would be difficult. For simplicity's sake here we use the approximation $p_{max}=1$.

If one knew p_α for all the possible sounds, then the human phoneme recognition could easily be simulated. Since this would need an infinitely long measurement, for practical purposes at first a guess is necessary for the form of the connection between the characteristics of the sound and p, and then a finite series of measurements can yield approximate values for the parameters in the formula for p.

In order to get a connection between the sound and α, one has to describe the sound mathematically. The sound is completely characterized by

the amplitude function F(t) of the oscillation; from obvious physical reasons we require

$$F(t < 0) = 0$$
$$\int_0^\infty F^2(t)dt < \infty \tag{3.5}$$

Instead of F(t) it is more convenient to use its Fourier transform z:

$$z(\nu) = f(\nu)e^{i\varphi(\nu)} = \int_0^\infty F(t)e^{i\nu t}dt \tag{3.6}$$

The real function $f(\nu)$ is the weight of the component with frequency ν, while $\varphi(\nu)$ is its phase. As F(t) is real, the following relations hold:

$$f(-\nu) = f(\nu)$$
$$\varphi(-\nu) = -\varphi(\nu) \tag{3.7}$$

that is, $\varphi(0) = 0$, and F(t) is completely determined by $f(\nu \geq 0)$ and $\varphi(\nu > 0)$.

Therefore the sound is completely characterized by a complex function $z(\nu)$, so

$$\Phi = \Phi(z(\nu)) \tag{3.8}$$

i.e. Φ is a functional.

The Ohm-Helmholtz Law [11] suggests that $\varphi(\nu)$ is irrelevant for sounds having no internal temporal structure. This may be the situation for phonemes whose realizations can be lengthened without limits, as vowels, nasals, fricatives and some liquidae, so it seems that for such phonemes $f(\nu)$ carries the basic information (nevertheless observe the short-long opposition for vowels in Hungarian, moreover the l-l' opposition in Slovakian). Therefore essential information could be obtained by investigating the restricted problem when the functional Φ is considered in the space of real functions [12]

$$\Phi = \Phi(f(\nu)) \tag{3.9}$$

In this paper we use this simplifying condition. We think that even in cases when the time structure of a phoneme is essential for its recognition, one can apply this approach by dividing the sound into characteristic parts and by analysing these parts separately.

By introducing the notation $\bar{f}(\nu)$ for the standard, eqs. (3.1) lead to

$$\Phi(\bar{f}) = 0$$
$$\Phi(f) > 0 \text{ if } f \neq \bar{f} \tag{3.10}$$

Therefore, if one is able to determine the deviation functional $\Phi(f)$, with this information the results of human phoneme recognition process can be simulated, since $f(\nu)$ can be measured by standard automatic methods. If one can guess a form for this functional with a finite set of unknown parameters, then the parameters can be fitted by measuring the acceptance probabilities $p(f)$ for a finite set of sounds, and then by further measurements the assumed form can be checked. Some technical difficulties can arise, nevertheless such an approach is well elaborated in experimental physics.

It is important to clearly distinguish here the acceptance probability $p(f)$ from occurrence probability $q(f)$, the second being the probability of the sound for the given phoneme in speech. They obviously differ, nevertheless, these quantities must be closely related, because speech and speech recognition are closely related processes too. Here we formulate a conjecture that they are approximately in functional dependence

$$p(f(\nu)) \approx p(q(f(\nu))) \tag{3.11}$$

(where $\approx$ indicates not simply an approximate equality but also some restrictions discussed in the next Section). Relation (3.11) will be verified in Sect. 6 in a simple model. If Rel. (3.11) were an exact equality, the deviation functional could be determined via $q(f)$ too, which is technically more easily measurable than $p(f)$.

4. NEUTRAL AMPLIFICATION

An elementary experience is that there are some parameters of the sound which, in a broad range of their values, are irrelevant in the recognition process; the most obvious of them is the total intensity. The advantage of this fact is obvious, otherwise one could communicate only with a given loudness fitted to the distance. Let us restrict ourselves to a single parameter λ. There are operations $f(\nu) \to T(\lambda)f(\nu) = f'(\nu)$ leaving Φ invariant:

$$\Phi(f') = \Phi(T(\lambda)f(\nu)) = \Phi(f) \tag{4.1}$$

Such operations will be called here symmetries. In many cases the operators $T(\lambda)$ can be parametrized in such a way that they form an Abelian group, i.e.

$$\begin{aligned} T(\lambda_1)T(\lambda_2) &= T(\lambda_1 + \lambda_2) \\ T(0) &= 1 \end{aligned} \tag{4.2}$$

Now let us try to guess the specific form of $T(\lambda)$ belonging to the neutral amplification of a sound. We can assume that in this case T is a multiplication by a function $t(\lambda;\nu)$:

$$(T(\lambda)f)(\nu) = t(\lambda;\nu)f(\nu) \tag{4.3}$$

There are even two hopeful candidates for $t(\lambda;\nu)$. The subjective intensity $r(I,\nu)$ of a monochromatic sound of frequency ν and physical intensity I is found to be [13]

$$r(I,\nu) \approx r_o(\nu)\ln[I/I_o(\nu)] \quad (\text{for } I > I_o) \tag{4.4}$$

where $I_o(\nu)$ is the sensitivity threshold. The logarithmic dependence is just the Weber-Fechner Law. Thus, by requiring that the neutral amplification add a constant λ to the subjective intensities of the constituents, one obtains

$$t(\lambda;\nu) = e^{\lambda/r_o(\nu)} \tag{4.5}$$

This seems, in fact, to be a subjectively neutral amplification for a sound as a sound. Nevertheless, the symmetry belonging to transformation (4.5) is not advantageous in the communication, since it is definitely not the transformation corresponding to the change of the distance between speaker and listener. By requiring <u>that</u> change to be a neutral amplification and symmetry, one gets

$$t(\lambda;\nu) = e^{\lambda} \tag{4.6}$$

Then eq. (4.1) means that Φ is a homogeneous functional of zero order [12].

If the neutral amplification is, in fact, the transformation (4.6) instead of (4.5), that is again an indication that the phoneme reconstruction is not simply a physiological process. Since elementary experiences seem to show that the change of distance does not influence the recognition until the intensity is not too low, here we accept eq. (4.6) as the correct symmetry.

Observe that $T(\lambda)\,\bar{f}$ is again a standard, because of eqs. (3.10) and (4.1). Thus $\bar{f}$ cannot be unique. Now, returning to eq. (3.11), it is clear that the existence of a symmetry means the existence of a parameter of which p is independent. On the other hand, q generally depends on this parameter when measured under usual circumstances, so the sign $\approx$ in eq. (3.11) stands for "approximately equal to, for special values of the symmetry parameters".

5. QUADRATIC APPROXIMATION

Eqs. (3.1-2), (3.9-10) and (4.1) fix only the most fundamental properties of the deviation functional Φ, the actual form should be determined from experiments. The standard way is to take a definite form with free parameters, which can be fitted. Nevertheless, the chosen form should be taken from experimental data too. This is a rather complicated procedure, for which a great amount of data and intuition is needed. Nevertheless, the most impor-

tant region for communication is where Φ is nearly O. For that region one may try to use a "power expansion" of Φ.

First we introduce a new variable $x=x(\nu)$

$$x(\nu) = \int_{O}^{\nu} (I_O(\nu'))^{-1} d\nu' \tag{5.1}$$

Then

$$f^2(x)dx = (f^2(\nu)d\nu)/I_O(\nu) = dI(\nu)/I_O(\nu) \tag{5.2}$$

Since for $\nu \to \infty$ $I_O(\nu)$ decreases very rapidly,

$$x(\infty) \equiv M < \infty \tag{5.3}$$

Therefore x remains in the [O,M] interval, which is very advantageous.

Eq. (3.10) shows that Φ (temporally denoted by Φ') starts quadratically in the difference from $\bar{f}$ at $f=\bar{f}$, i.e.

$$\Phi' = \int_{O}^{M}\int_{O}^{M} s(x,y)\varepsilon(x)\varepsilon(y)dxdy + \vartheta(\varepsilon^3) \tag{5.4}$$

where ε is some still undefined difference function, and $s(x,y)$ is a symmetric weight function; $\varepsilon=O$ is a true minimum if and only if

$$\int_{O}^{M}\int_{O}^{M} s(x,y)\varepsilon(x)\varepsilon(y)dxdy > O \tag{5.5}$$

for any $\varepsilon \neq O$. Then there remains the problem of proper definition of ε. It is convenient to incorporate the symmetry (4.6) (i.e. the homogeneous zero order nature of Φ) into the definition, which can be explicitely done by redefining Φ as

$$\begin{aligned} \varepsilon(\lambda,x) &= e^{\lambda} f(x) - \bar{f}(x) \\ \Phi(f) &= \min_{\lambda}\{\Phi'(\varepsilon(\lambda,x))\} \end{aligned} \tag{5.6}$$

where $\bar{f}$ is an arbitrary but fixed member of the set of standards, and Φ' is defined by eq. (5.4). The meaning of eq. (5.6) is that the symmetry connects a "ray" of sounds, and we compare $\bar{f}$ with the nearest member of the ray. By evaluating eqs. (5.4), (5.6) one obtains

$$\begin{aligned} \Phi = \Phi(f) = \{\int_{O}^{M}\int_{O}^{M} s(x,y)f(x)f(y)dxdy\}^{-1} * \\ * \{[\int_{O}^{M}\int_{O}^{M} s(x,y)f(x)f(y)dxdy][\int_{O}^{M}\int_{O}^{M} s(x,y)\bar{f}(x)\bar{f}(y)dxdy] - \\ - [\int_{O}^{M}\int_{O}^{M} s(x,y)f(x)\bar{f}(y)dxdy]^2\} + \vartheta((f-\bar{f})^3) \end{aligned} \tag{5.7}$$

which is, in fact, clearly homogeneous of zero order. This form is a direct consequence of the guessed type (4.6) of the neutral amplification; for any other form of the corresponding transformation the procedure could be repeated. If there are no other symmetries, then $f=e^{\lambda}\bar{f}$ is a true minimum of Φ; if other symmetries exist too, not incorporated into the definition of ε in eq. (5.7), then $\Phi(f) \geq \Phi(e^{\lambda}\bar{f})$ is valid.

It is not clear, which is the region for ε where the ε^3 terms are negligible in eq. (5.7); formally one can say that they remain negligible for $\Phi \lesssim 1$ (which is important for recognition) if Φ is slowly varying. According to the principle of Occam's razor, we assume this until counterevidences are not known.

Now we consider the functions $f(x)$

$$f(x < 0) = f(x > M) = 0 \tag{5.8}$$

to be the elements of a real Hilbert space $L^2[0,M]$ with a scalar product

$$(f,g) = \int_0^M f(x)g(x)dx \tag{5.9}$$

Then the operator $\hat{s}$

$$(\hat{s}f)(x) = \int_0^M s(x,y)f(y)dy \tag{5.10}$$

is completely continuous [12], and according to the Hilbert-Schmidt theorem its eigenfunctions form a complete orthonormal basis in the Hilbert space:

$$\begin{aligned} &\hat{s}f_i = k_i f_i \\ &(f_i,f_k) = \delta_{ik} \\ &\sum_{n=1}^{\infty} f_n(x)f_n(y) = \delta(x-y) \end{aligned} \tag{5.11}$$

with the following properties of the eigenvalues:

$$\begin{aligned} &\max|k_i| < C \\ &\lim_{i\to\infty} k_i = 0 \end{aligned} \tag{5.12}$$

On this basis the standard $\bar{f}$ and an arbitrary f belonging to the $L^2(0,M)$ space can be expressed as

$$\bar{f} = \sum_{n=1}^{\infty} \bar{\varphi}_n f_n$$

$$\bar{\varphi}_i = (\bar{f}, f_i)$$

$$f = \sum_{n=1}^{\infty} \varphi_n f_n \qquad (5.13)$$

$$\varphi_i = (f, f_i)$$

Then eq. (5.10) can be rewritten into the form

$$\hat{s}f = \sum_{n=1}^{\infty} k_n \varphi_n f_n(x)$$

$$s(x,y) = \sum_{n=1}^{\infty} k_n f_n(x) f_n(y) \qquad (5.14)$$

that is, s is separable. Therefore, instead of the function s(x,y) of two variables it is sufficient to use the infinite set of eigenfunctions $\{f_i(x)\}$ of one variable and the numbers $\{k_i\}$ and $\{\bar{\varphi}_i\}$. Thus the functional Φ, given by eq. (5.7), becomes a function $\tilde{\Phi}$ of the infinite set of variables φ_i:

$$\Phi(f) = \Big(\sum_{n=1}^{\infty} k_n \varphi_n^2\Big)^{-1} \Big\{ \Big(\sum_{n=1}^{\infty} k_n \varphi_n^2\Big)\Big(\sum_{n=1}^{\infty} k_n \bar{\varphi}_n^2\Big) -$$

$$- \Big(\sum_{n=1}^{\infty} k_n \varphi_n \bar{\varphi}_n\Big)^2 \Big\} \equiv \tilde{\Phi}(\varphi_i) \qquad (5.15)$$

Evaluating uneq. (5.5) one obtains

$$\sum_{n=1}^{\infty} k_n \varphi_n^2 > 0 \qquad (5.16)$$

Thus the eigenvalues are not negative; they form a monotonously decreasing series

$$k_i \geq k_j \geq 0; \quad i < j \qquad (5.17)$$

Since obviously the largest eigenvalues are the most important, uneq. (5.17) show how to truncate the infinite sums for approximation.

The form (5.15) given for the deviation functional seems to possess strange mathematical properties: it ceases to be a quadratic function of its argument, therefore no linear operator seems to correspond to it (as $\hat{s}$ of eq. (5.10) to eq. (5.4)). But, as it can be shown (cf. the Appendix), in second order its properties correspond to a quadratic form with one zero eigenvalue. This fact is very important, because it assures the existence of quadratic expressions we use later on in Sect. 7.

We have seen that further symmetries, unbuilt into the explicite form of Φ, violate the strict unequality in (5.5) and 0 eigenvalues are possible corresponding to symmetry directions. This phenomenon may disturb the actual

evaluation, thus it is useful to explore first the possible symmetries. This will be discussed in a subsequent paper.

6. CONNECTION BETWEEN ACCEPTANCE AND OCCURRENCE PROBABILITIES

At the end of Sect. 3 a conjecture was formulated about the approximate functional dependence between the acceptance and occurrence probabilities p(f) and q(f), since the pronounciation habits of the linguistic community form the individual's speech recognition via learning, and in its turn the fixed recognition pattern in the brain prevents serious changes in the pronounciation. It would be very desirable to exploit this connection, because the occurrence probability can easily be measured; here we verify the functional dependence in a simple model.

By definition q(f) is the occurrence probability of sounds formed as realizations of a given phoneme. There are other sounds which are not intended to be realizations of that phoneme at all and disturb the communication process; their occurrence probability is denoted by Q(f). Since they mainly are noises, their distribution is more or less uniform in the region where p(f) is substantial.

The recognition process of an individual is optimal if he accepts

a/ different representations of a given phoneme (pronounciated by different members of the community) with maximal probability; and, in the same time,

b/ other sounds, not belonging to that phoneme, with minimal probability.

Consider some set of sounds densely distributed in the space of all sounds, the index α labels them. Then Conds. a/ and b/ can be written as

$$\sum_{\alpha}(1-p_{\alpha})q_{\alpha} = \min. \tag{6.1}$$

$$\sum_{\alpha}p_{\alpha}Q_{\alpha} = \min \tag{6.2}$$

where $p_{\alpha} = p(f_{\alpha})$ and so on. Now, these conditions are inconsistent, namely, the solution of eq. (6.1) is $p_{\alpha} \equiv 1$, while that of eq. (6.2) is $p_{\alpha} \equiv 0$, independently of α. Therefore only a compromise can be achieved; the sum can be minimized for a function $\Psi = \Psi((1-p)q,pQ)$, instead of the two above functions $(1-p)q$ and pQ. A simple function of this form is

$$\Psi = ((1-p)q)^2 + \beta^2(pQ)^2 \tag{6.3}$$

where β is a constant expressing the relative weight of Conds. a/ and b/. Evaluating the minimum condition for Ψ

$$\sum_{\alpha}\{(1-p_{\alpha})q_{\alpha}\}^2 + \beta^2\sum_{\alpha}(p_{\alpha}Q_{\alpha})^2 = \min \tag{6.4}$$

one gets the solution for p as

$$p = \frac{q^2}{q^2 + \beta^2 Q^2} \tag{6.5}$$

Since Q is slowly varying in the neighbourhood of the standard, there p and q approximately fulfil Rel. (3.11). If there q dominates Q, and β is moderate, then $1-p(\bar{f}) \ll 1$, as we have assumed according to elementary experiences.

The result (6.5) does not reflect the fact that the functional dependence can be valid only up to symmetries; obviously the symmetries should be built into the form of Ψ. This problem will not be discussed here.

It is easy to see that the main reason of the approximate validity of a functional dependence is not the specific form (6.3) of the "compromise function" Ψ, but rather the approximate constancy of Q(f). Namely, for a general

$$\Psi = \Psi(p,q,Q) \tag{6.6}$$

the extramum condition (6.4) gives

$$\frac{\partial \Psi(p,q,Q)}{\partial p} = 0 \tag{6.7}$$

which is an implicite equation for p,

$$p(f) = p(q(f),Q(f)) \tag{6.8}$$

If $Q(f) \approx$ const., eq. (3.11) approximately holds.

7. THE PARAMETER SPACE

In some cases it is useful to use a basis different from that of (5.11) or from that of the extremal directions (cf. the Appendix). For an arbitrary basis $b_i(x)$

$$f = \sum_{r=1}^{\infty} c_r b_r(x) \tag{7.1}$$

and

$$\Phi(f(x)) = \Phi\left(\sum_{r=1}^{\infty} c_r b_r(x)\right) = \tilde{\Phi}(c_i) \tag{7.2}$$

Now, consider a symmetry of form (4.3); one can write

$$t(\lambda,x) = \sum_{r=1}^{\infty} t_r(\lambda) b_r(x) \tag{7.3}$$

and then the transformation (4.1) can be reformulated as

$$c_i \to c_i'(t_k(\lambda), c_k) \tag{7.4}$$

Since T is a symmetry, $\tilde{\Phi}$ can depend only on invariant combinations. For the special case (4.6) the invariant parameters can be

$$c_i^* = c_{i+1}/c_1; \tag{7.5}$$

$\tilde{\Phi}$ has a minimum at the standard parameters $\bar{c}_i^*$, therefore

$$\tilde{\Phi}(c_i) = \sum_{r,s=1}^{\infty} K_{rs}(c_r^* - \bar{c}_r^*)(c_s^* - \bar{c}_s^*) + \vartheta((c^*-\bar{c}^*)^3) \tag{7.6}$$

where the matrix K_{ik} is symmetric, and in the absence of other symmetries positive definite, otherwise semidefinite.

For practical purposes a finite basis is needed. Then one can improve the approximation if the basis functions $b_i(x)$ also contain some parameters to be fitted for optimally describing f(x):

$$f(x) \simeq \sum_{r=1}^{N} c_r b_r(x; p_{r\alpha}); \quad \alpha = 1...a \tag{7.7}$$

and thus

$$\Phi(f) \simeq \tilde{\Phi}(c_i, p_{i\alpha}) \tag{7.8}$$

i.e. $\tilde{\Phi}$ is a function of N(1+a) parameters, which belong to two groups, according to the different behaviour under neutral amplification. Thermodynamics offers some analogy for this: the coefficients c_i can be regarded as extensive parameters, while $p_{i\alpha}$'s are intensives [14]; the first group is multiplied in amplification, the second is not. Therefore one can reparametrize the problem as

$$\begin{gathered}(c_i, p_{i\alpha}) \to (c_1, c_i/c_1, p_{i\alpha}) \to (c_1, p_A) \\ \Phi = \tilde{\Phi}(p_A)\end{gathered} \tag{7.9}$$

The the N(1+a)-1 parameters p_A are coordinates in a parameter spece. The deviation is a scalar function, the sounds of equal acceptance are on surfaces which are closed without additive symmetries. Near the standard values p_A one can expand it as

$$\tilde{\phi}(p_A) \simeq \sum_{R,S=1}^{N(1+a)-1} \tilde{K}_{RS}(p_R-\bar{p}_R)(p_S-\bar{p}_S) + \vartheta((p_I-p)^3) \tag{7.10}$$

There again the matrix $\tilde{K}_{IK}$ is symmetric and positive (semi) definite, thus the surfaces ϕ=const. can be approximated by ellipsoids in the parameter space.

The problem of optimal parametrization is connected with the problem how to find the formants of phonemes. The standard carries much important and unimportant information about the phoneme. But only such parts are of crucial importance, whose small changes give essential changes in Φ. These parts are defined by the smallest axes of the ellipsoid in (7.10) or by the largest eigenvalues in eq. (5.14), and so on. The brain is sensitive for correlated changes in the regions characteristic for them. Therefore these sensitive parts form the basic formants of the standard.

It is not necessary that these parts correspond to peaks of the standard; however, one can guess that the peaks are amongst the characteristic parts. Namely, a sound is formed via a resonance process, in which the eigenfrequencies of the cavities of the sound channel appear. As these cavities are open, there is some damping, the eigenfrequencies are complex, manifested as peaks of various heights and widths. It suggests a parametric form

$$f(\nu) \simeq e^{-\alpha\nu}\{\sum_{n=1}^{N} \frac{a_n \Gamma_n}{(\nu-\nu_n)^2 + \Gamma_n^2} + \text{a smooth funct.}\} \qquad (7.11)$$

with $\alpha \simeq 6$ dB/octave. Then the parameters to be determined are the strengths, locations and widths of the resonance peaks, together with some parameters of the smooth background, which can be taken as a polynomial, for instance. The method of fitting by means of such functions is well elaborated [15], therefore objective results can be obtained.

8. THE STRUCTURE OF PHONEME SYSTEMS

Until now a chosen individual phoneme was considered. Nevertheless, generally the languages use several dozen phonemes which can be arranged into some structure characteristic for the language (or for a group of languages). The knowledge of this structure may give additional information on the functionals $\Phi_\Gamma(f(\nu))$ (where the capital Greek index stands for the different phonemes). In this Section we discuss the relations among the domains of the different phonemes.

Clearly, a trivial relation is a repulsion between phonemes. The model calculation of Sect. 6 indicates that the occurrence probabilities of different phonemes cannot have great values at the same place, otherwise the realizations of these phonemes would be confused. If (3.11) is approximately true, then the ellipsoids of different phonemes are disjunct. If measurements show doubly represented domains, then it is an artefact of the projection of the "true" infinite dimensional parameter space into a finite dimensional one, where originally different points can coincide.

The repulsion is a very important phenomenon, but it is almost trivial, and cannot generate a structure. Our guess is that Hungarian (possibly together with other related languages) yields an example for an additional relation, generating a definite structure. The necessary quantitative investigation will be done in a subsequent paper.

In Hungarian there are 8 short vowels [16]. A rule called vowel harmony places two restrictions to the words:

a/ A word can be built up from front or back vowels without mixing.

b/ There is a relation between the vowels of the root and the actual form of the suffix.

Rule a/ is rather a tendency, e.g. it does not hold for compound and borrowed words; Phoneme [e] (also e in Hungarian orthography, sometimes in linguistic texts written as ë) may occur also with back vowels; Phoneme [i] is a successor of both a front and a back vowel, etc. Nevertheless, Rule b/ is strict. It arranges 7 of the short vowels into 3 groups. For a suffix the group is lexically fixed, while the actual vowel is determined by the vowels of the root. The groups are as follow:

$$(a,e) \simeq ([\text{ɔ}]\ [\varepsilon]$$
$$(o,\ddot{o},e) \simeq ([o],[\phi],[e])$$
$$(u,\ddot{u}) \simeq ([u],[y])$$

Since these and only these short vowels can substitute each other without change in the meaning of the suffix, and the changes must not influence the recognition of the suffix, one can venture the hypothesis that the vowels belonging to a group are variants of some fictious vowel, i.e. that the functionals $\phi_\Gamma(f)$ show some group structure. In fact, as we mentioned in Sect.1, the available 2 formant measurements for Hungarian seem to indicate a vertical structure with the above groups, i.e. that $\bar{\nu}_1$ is common within a group.

Hungarian belongs to the Uralian family of languages [17]. In this family there is a general tendeny for some vowel harmony, although in some cases (as for e.g. Estonian and Vogulic) this tendency is rather weak. There is an other group of languages, the Altaic, whose connection with the Hungarian is not clear [18], but whose languages were in strong areal connection with Hungarian in the first millenium A.D. These languages also show vowel harmony which is the strongest in the Turkish and Mongolian subgroups. For the Ottoman Turkish J. Lotz recognized a vertical structure of 3 groups [6]:

$$(a,e) \simeq ([a],[\text{æ}])$$
$$(o,\ddot{o}) \simeq ([o],[\phi])$$
$$(u,\text{ı},\ddot{u},i) \simeq ([u],[\text{ɨ}],[y],[i])$$

the first and third group is the same for suffices, the second group cannot occur in them. Thus it seems that these cases are examples when the gramma-

tical structure "generates" a structure in the recognition. This hypothetical vertical structure cannot be a trivial consequence of human physiology, cf. the rather radial American English structure obtained by Peterson and Barney [5], discussed in Sect.1.

The acoustic structure generated by the vowel harmony may incorporate even some consonants. It is well known that [k] possesses "front" and "back" variants, regularly joined to front and back vowels in all European and Altaic languages, but not in the Arabic. Such an automatic correlation means that the orthography does not have to distinguish the different realizations, which may be, however, important for recognition. Of course, only measurements can show which consonants possess variants not reflected in orthography. For Hungarian, speech synthetization data [19] yield some suggestions for "front" or "back" versions of some consonant. *Figure 1* has demonstrated that first formants of the Hungarian vowels are characteristic for groups rather than for individual vowels; therefore let us consider a characteristic frequency of the consonant as a function of the subsequent vowel, as shown on *Figs. 2a-c* for the phonemes [k], [n] and [ɲ], on a double logarithmic coordinate system. Clearly, these three consonants show three completely different behaviours. For [k] the logarithms of the two frequencies seem to be proportional, by fitting a straight line on the logarithmic plot to the points one gets that

$$f_1(k) = 1.101(f_2(V))^{0.984} \tag{8.1}$$

where f_i denotes characteristic frequencies, V stands for "vowel". Ref. 19 was not intended to measure errors for frequency data, therefore it would be difficult to determine the error of the exponent in eq. (8.1), but it is reasonable to think that there is a linearity between $f_1(k)$ and $f_2(V)$ in the Hungarian [kV] syllables; the Hungarian [k] seems to be multiply represented.

The next example is [n]. One can see a saturation in $f_2(n)$ either for low and for high $f_2(V)$ values. Unfortunately, the middle region is unpopulated in $f_2(V)$, thus the functional dependence is by no means unique, nevertheless it can be well described by

$$f_2(n) = 1.425 + 0.175\text{th}(4.244f_2(V) - 6.131) \tag{8.2}$$

Thus [n] seems to be double-represented in Hungarian ($f_1(n)$ is constant).

The third example is [ɲ]: its formant frequencies seem to be constant according to Ref. 19; it has a single representation in Hungarian.

A possible interpretation of these curves is that some consonants participate in the vowel harmony rule of Hungarian: however the picture is still unclear to some extent, because not all the consonants belong to these three pure classes. Note that the vowel [ø] is "misplaced" on *Figs. 2a-c* among the back vowels; in the syntheses of Ref. 19 a peculiarly low second formant frequency was accepted for [ø], just at the lowest part of the boundary of

its region (cf. *Fig. 1*). This may have been a consequence of the empty region between [o] and [ø]; then a low f_2 means better discrimination between [o] and [e].

There existed a language with properly adapted orthography showing such versions for some consonants. It was the Old Turkish preserved mainly in the Orkhon inscriptions [20]. This alphabet distinguishes the front and back versions of some consonants, namely [20],[21]

b,d,g,j,k,l,n,r,s,t

while there are no two versions for the consonants

č,m,p,š,z,ŋ

As a first guess, we may regard the double-represented consonants (k has, in fact, not 2 but 5 versions) as suspects to be objects of the harmony rule. The data of the Orkhon alphabet for this question are roughly conformal to the acoustic structure suggested by the data of Ref. 19.

The investigation of such a structure of the phoneme system may shed some light on the reasons of changes in pronounciation, which sometimes lead to formation of new languages. This question will be discussed in a subsequent paper, here we only demonstrate that the reason is not simply a common feature of the physiology of the human speech channel. Namely, the fate of the intial [k-] phoneme was different in the Latin and Uralic groups. In the inheritors of Latin, [k-] remained unchanged before back vowels, excepting French, where [ka-] → [ša-], and became either affricate of fricative before front vowels, excepting the Logudorese in Sardinia, and the extinct Dalmatian before [e], where did not change; probably the first step was [c], conserved in Central European pronounciation of Latin and in the orthographies of Hungarian and Latin writing Slavic languages. The final result is [c],[θ] or [s]. This transition seems to be a permanent tendency among the Indo-European languages, since the most fundamental classification (into Western or kentum and Eastern or satem languages) is based on a (palatal k)→(s type fricative) transition [22]. On the other hand, the Uralic languages have fully opposite tendency. The phoneme [k-] remained unchanged before front vowels in the last few millenia in all the Finno-Ugrie languages, and before back vowels in the most ones [17], [23]. There is a change in the Ugric group, for (k+back vowel): in some dialects of the Ob-Ugric languages the final result has been [x], in the Hungarian [h].

Now, one may or may not think that the permanence of the Finno-Ugric [k] is a consequence of its multiple representation; obviously there is no great temptation to change the pronounciation of a consonant in the influence of the next vowel when the vowel already has influenced it, but first one should investigate the representations in the Indo-European counterexamples.

Nevertheless, in any case, one can arrive at the conclusion that the reasons determining the directions of some linguistic evolution processes are partly in the brain, not in the sound channel.

9. CONCLUSION

In this paper we introduced the deviation functional for mathematically describing the process of phoneme recognition. The argument of this functional is the Fourier transform of the amplitude function of the sound. Some approximations for the form of this functional have been discussed; it has been shown that if the functional Φ is slowly varying, then the region of a vowel is elliptical in the space of proper parameters. Some 2-formant measurements, in fact, indicate a roughly elliptic shape. An approximate functional dependence between the acceptance and occurrence probabilities has been verified in a simple model; this dependence could be checked in measurements.

It seems that the individual phonemes form some structure, which is language-dependent. Here we at least have demonstrated a difference between the vowel structures in American English and Hungarian, reflected on the 2-formant plots. If the vowels, in fact, form identifiable groups, then this structure has to appear in the forms of the functionals Φ belonging to the individual phonemes too. This seems to be suggested by the vertical locations and orientations of Hungarian vowels together with the fact that some consonants are not uniquely represented. These features, at least partially, seem to exist in Turkish languages too.

ACKNOWLEDGEMENTS

The authors would like to thank Prof. T. Tarnóczy, Klára Vicsi and A. Kaposi for illuminating discussions.

APPENDIX: THE EXTREMAL DIRECTIONS

Here, in order to study the properties of the deviation functional (5.15), we are going to look for the directions defined by the extrema of the change of the functional, i.e. for δf's which fulfil the following relations:

$$\begin{aligned} & f = \bar{f} + \delta f \\ & (\delta f, \delta f) \text{ is fixed} \\ & \delta\Phi = \Phi(\bar{f} + \delta f) - \Phi(f) = \text{extremum} \end{aligned} \tag{A.1}$$

Then, writing

$$\varphi_i = \bar{\varphi}_i + \varepsilon q_i ; \quad |\varepsilon| \ll 1 \tag{A.2}$$

eqs. (A.1) get the form

$$\sum_{r=1}^{\infty} q_r^2 = 1$$

$$\delta\Phi = \Big(\sum_{t=1}^{\infty} k_t \bar{\varphi}_t^{-2}\Big)^{-1} \Big\{ \Big(\sum_{s=1}^{\infty} k_s \bar{\varphi}_s^{-2}\Big)\Big(\sum_{r=1}^{\infty} k_r q_r^2\Big) - \tag{A.3}$$

$$- \Big(\sum_{r=1}^{\infty} k_r \bar{\varphi}_r q_r\Big)^2 \Big\} \varepsilon^2 + \vartheta(\varepsilon^3) = \text{extr.}$$

This is a variational problem with constraint, it can be solved by the Lagrange method, and the result is

$$q_i^{(o)} = \bar{\varphi}_i \Big/ \sqrt{\sum_{r=1}^{\infty} \bar{\varphi}_r^{-2}}$$

$$q_i^{(A)} = \frac{k_i \bar{\varphi}_i}{\alpha_A(\beta_A k_i - 1)} \cdot \frac{1}{\Big(\sum_{r=1}^{\infty} \bar{\varphi}_r^{-2}\Big)^{1/2}} \tag{A.4}$$

where

$$\alpha_A^2 = \frac{\sum_{r=1}^{\infty} \dfrac{k_r^{2-2}\varphi_r}{(\beta_A k_r - 1)^2}}{\sum_{r=1}^{\infty} \bar{\varphi}_r^{-2}} \tag{A.5}$$

and β_A is the Ath root of the equation

$$\lambda(\beta) = \sum_{r=1}^{\infty} \frac{k_r \bar{\varphi}_r^{-2}}{\beta k_r - 1} = 0 \tag{A.6}$$

The upper index of q_i labels the different solutions of the variational problem; for δf one gets

$$\delta f = \varepsilon \sum_{r=1}^{\infty} q_r f_r \tag{A.7}$$

Since $\delta^{(o)} f$ is an (infinitesimal) neutral amplification, it does not alter ϕ. One can verify that the functions

$$q^{(I)}(x) = \sum_{r=1}^{\infty} q_r^{(I)} f_r(x) \tag{A.8}$$

form an orthonormal basis. Expanding f on this basis

$$f(x) = \bar{f}(x) + \sum_{R=o}^{\infty} \varepsilon_R q^{(R)}(x) = \tag{A.9}$$

$$= \sum_{r=1}^{\infty} \{(1+\varepsilon_o)\bar{\varphi}_r + \sum_{R=1}^{\infty} \varepsilon_R \frac{\bar{\varphi}_r k_r}{\alpha_R(\beta_R k_r - 1)}\} f_r(x)$$

the deviation functional gets the form

$$\Phi(f(x)) = \sum_{R=1}^{\infty} \varepsilon_R^2 \beta_R^{-1} + \vartheta(\varepsilon^3) \tag{A.10}$$

With increasing I β_I is increasing and

$$\lim_{I\to\infty} \beta_I = \infty \tag{A.11}$$

This can be proven as follows. Let us differentiate the equation (A.6) defining β_I, then

$$\frac{d}{d\beta} \lambda(\beta) = - \sum_{r=1}^{\infty} \left(\frac{k_r \bar{\varphi}_r}{\beta k_r - 1}\right)^2 < 0 \tag{A.12}$$

Therefore $\lambda(\beta)$ is a decreasing function in the intervals where it is continuous; it possesses (infinite) jumps at $\beta = 1/k_i$. Now, let us start from a root of λ, β_{I-1}. Then there cannot exists a new root in the continuous region, the I^{th} root is behing $1/k_I$, thus

$$1/k_I < \beta_I < 1/k_{I+1} \tag{A.13}$$

Then eq. (5.17) leads to increasing β values, while eq. (5.12) excludes any finite upper bound.

Since the functions $q^{(I)}(x)$ are the solutions of a variational problem, and the series β_I^{-1} is decreasing, the expansion (A.9) possesses the minimal error when truncated.

REFERENCES

[1] Flanagan J.L., Speech Analysis, Synthesis and Reconstruction. Berlin, 1965

[2] Shapiro E., Amer. Anthr. 14, 226 (1912)

[3] Marnelius R., private communication

[4] Tarnóczy T., in: Általános nyelvészeti tanulmányok X. (ed. Telegdi Zs. and Szépe Gy.), Akadémiai Kiadó, Budapest, 1974. pp. 181 (in Hungarian)

[5] Peterson G.E. and Barney H.L., J. Acoust. Soc. Amer. 24, 175 (1952)

[6] Lotz J., in: Research in Altaic Languages (ed. Ligeti L.), Budapest, p. 135.

[7] Fukunaga K., Introduction to Statistical Pattern Recognition. Academic Press, New York-London, 1972

[8] Weinberg J.W., Gen. Rel. Grav. 7, 135 (1976)

[9] Lukács B., KFKI-1984-86

[10] Borbély I. and Lukács B., Proc. Symp. on Speech Acoustics, Budapest, 1980. p.25.

[11] Helmholtz H, Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik, Bruswick, 1863

[12] Kolmogorov A.N. and Fomin S.V., Elementy teorii funkcii i funkcional'nogo analiza. Nauka, Moscow, 1968

[13] Rzhevkin S.N., Sluh i rech'. Moscow-Leningrad, 1936

[14] Fényes I., Z. Phys. 134, S95 (1952)

[15] Borbély I. and Nichitiu F., Lett. Nuovo Cim. 16, 89 (1976)

[16] Lotz J., Ural-altaische Jahrbücher XXVI, 252 (1965)

[17] Harms R.T., The Uralic Languages, in: Encyclopedia Britannica, 1974

[18] Collinder B., Acta Univ. Appsalien. Acta Soc. Ling. Ups. Nova Ser. 1:4, 109 (1965)

[19] Olaszy G., Proc. 8th Colloq. on Acoust. Budapest, 1982, p. 204.

[20] Thomsen W., Inscriptions de l' Orkhon déschiffrées. Mem. de la Soc. Finno-Ougrienne 5. Helsingfors, 1896

[21] Vasilev D.E., Sov. Tjurk. 1976:1 p. 71

[22] Brugmann K., Kurze vergleichende Grammatik de indogermanischen Sprachen. Verlag v. K.J. Tübner, Strassburg, 1904

[23] Collinder B., Comparative Grammar of the Uralic Languages, Stockholm, 1960

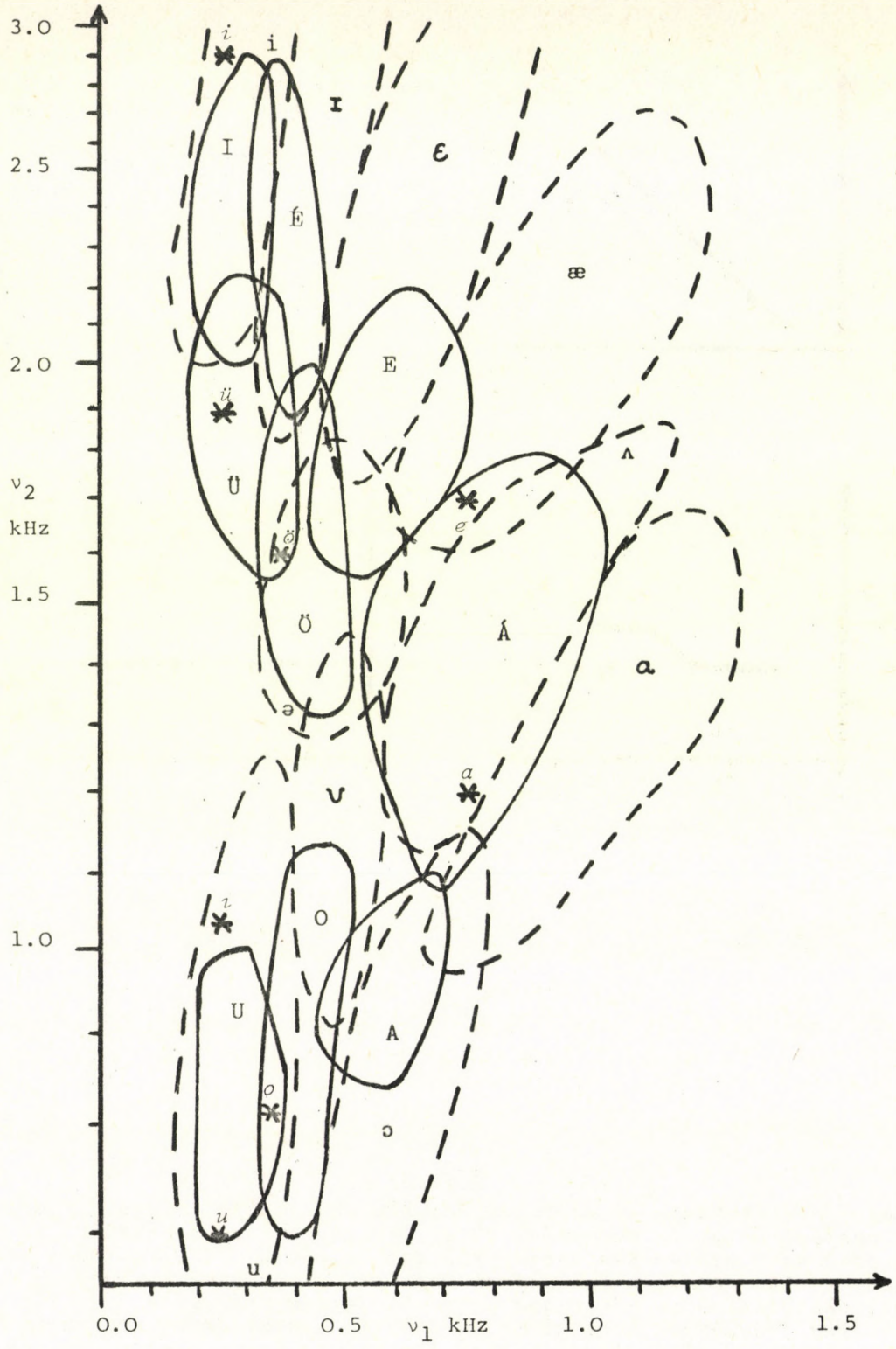


Fig. 1.

A comparison of the two formant pictures of American English, Hungarian and Ottoman Turkish vowel systems, according to Refs. 5, 4 and 6, respectively. For Turkish only mean frequency data are given.
American: dashed lines, small letters, IPhS orthography
Hungarian: continuous line, capitals, national orthography
Turkish: stars, italics, national orthography

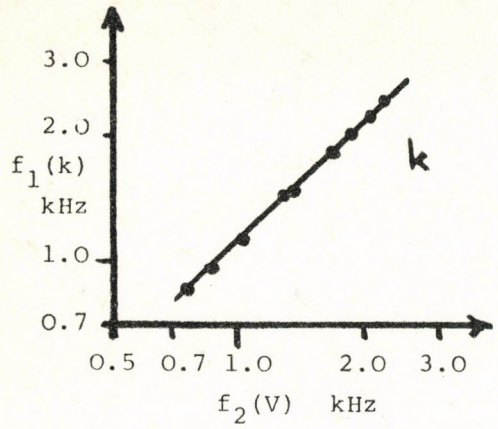


Fig. 2a.

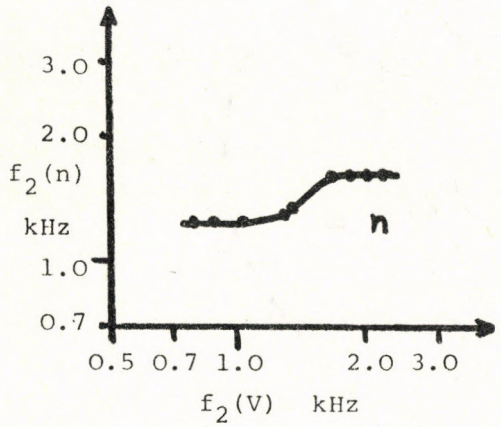


Fig. 2b.

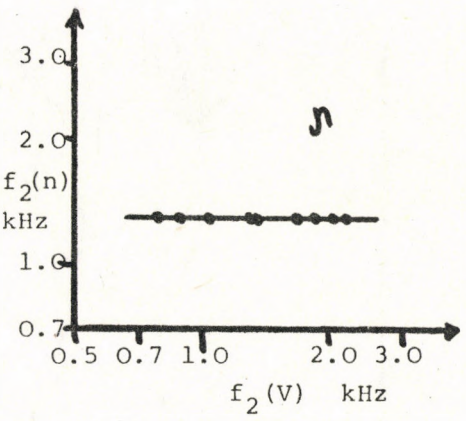


Fig. 2c.

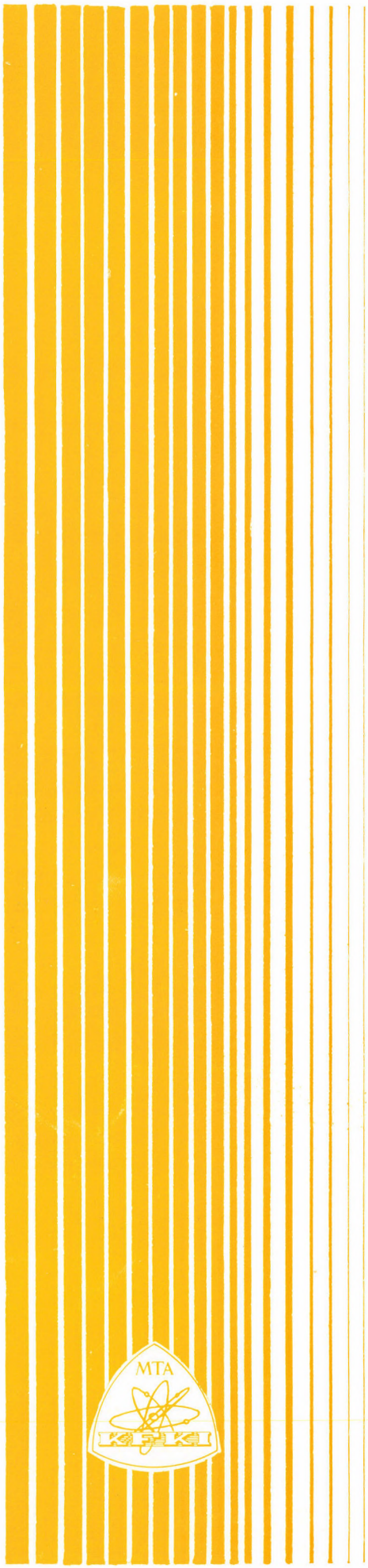
Fig. 2.

Examples for consonants of multiple, double and single acoustic representations, respectively, in Hungarian CV syllables; a characteristic frequency of the consonant versus the second formant of the vowel [19]. Both axes are logarithmic. The dots are taken from Ref. as frequencies of successful simulations, the continuous lines are given by fitting formulae, cf. eqs. (8.1-2). The frequencies are meant in kHz. The sequence of vowels is: [u], [o], [ɔ], [ø], [a], [y], [ɛ], [e] *and* [i].

Fig. 2a: The syllables [kV]; first noise maximum of [k]. The second one is constantly 4.5 kHz.

Fig. 2b: The syllables [nV]; the second formant frequency of [n]. The first one is always 0.250 kHz.

Fig. 2c: The syllables [ɲV]: the second formant frequency of [ɲ]. The first one is the same as for [n].



Kiadja a Központi Fizikai Kutató Intézet
Felelős kiadó: Bencze Gyula
Szakmai lektor: Kluge Gyula
Nyelvi lektor: Forgács Péter
Gépelte: Simándi Józsefné
Példányszám: 385 Törzsszám: 85-244
Készült a KFKI sokszorosító üzemében
Felelős vezető: Töreki Béláné
Budapest, 1985. április hó