EESTI NSV TEADUSTE AKADEEMIA TOIMETISED. 28. KÖIDE
FÜÜSIKA * MATEMAATIKA. 1979, NR. 3

ИЗВЕСТИЯ АКАДЕМИИ НАУК ЭСТОНСКОЙ ССР. ТОМ 28
ФИЗИКА * МАТЕМАТИКА. 1979, № 3

УДК 534 : 612—06

M. MIHKLA

A COMPUTER REPRESENTATION OF THE GEOMETRICAL CONFIGURATION OF THE VOCAL TRACT

(Presented by E. Lippmaa)

1. Introduction

One of the basic problems a designer of speech-oriented automatic control systems (ACS's) is faced with is finding a manner of coding speech units which would ensure their quickest reproduction without causing any damage to the speech and, at the same time, would occupy minimal memory space. Various ideas and methods have emerged; they are chiefly oriented to the acoustical representation of speech (segments of acoustical waves, spectra, formants).

The advantage of the acoustical approach is that all its necessary transformations are relatively uncomplicated (in a simpler case, it consists of just «patching up» the segments of an acoustical wave); its drawback, on the other hand, is a qualitative deterioration which accompanies the minimization of the coded information of the obtained speech. In this sense, there is more promise in using the geometrical configuration of the vocal tract which, while sufficiently economical, ensures a satisfactory quality of speech, and is versatile for purposes of various further applications.

Let us add that at the present level of computational mathematics carrying out the necessary transformations (first of all, that of «articulation — acoustics») is no serious problem. With regard to application, the ACS's where the coding of speech information is effected by representing the vocal tract configuration, have, despite a certain unwieldiness, two additional advantages over the systems realized so far: in case there is a change in requirements as to speech quality, the entire structure of representation will be retained, and, at the same time, it becomes possible to find an efficient application for the transformation «speech wave — vocal tract geometry». As the efficiency of a system which uses vocal tract geometrical configuration depends directly on dimensionality-reducing techniques, in order to obtain the optimal variant of representation the following techniques are compared: expansion into the Fourier series, cluster analysis, factor analysis, the method of principal components, and regression analysis.

2. Initial data

In the present work, the initial data are *X*-ray cineframes received from the midsagittal plain of the vocal tract. Several works on the geometrical configuration of the vocal tract are restricted to its sagittal model, i. e.

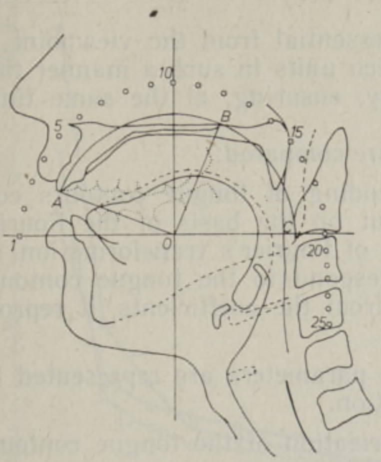

Fig. 1. The coordinate system and measuring points for the tongue contour.

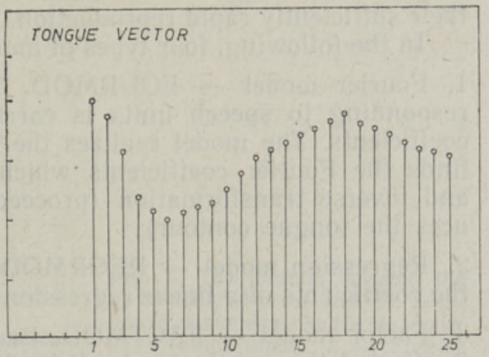

Fig. 2. A vector composed of 25 normalized variables which describe the tongue profile (shown in Fig. 1).

the third dimension in the region of the given vocal tract is regarded as constant [1-5]. We have adopted that simplification here as well. The sagittal model provides us with a basis for developing a dynamic model of the articulatory system.

Secondly, in our articulatory description of the vocal tract we have assumed that during the formation of the sounds the position of the lower jaw, velum, uvula and pharynx walls does not change. In other words, the assumption was that a sound is formed according to the position of the tongue (and lips). Thus the articulatory model of the vocal tract is reduced to the task of giving, in the case of each sound, as accurate and convenient a description of the position or movement of the tongue (and lips) as possible. The configurational description of the lips, being less complicated than that of the tongue, has been omitted.

In order to give a representation of the contour of the tongue one has to find a way of locating the system of co-ordinates in a certain established manner (Fig. 1) applicable in the case of all the speakers (in this case two different speakers were used) so as to avoid various troublesome deformations. The author uses a method developed by A. Eek and M. Remmel of the Institute of Language and Literature, Academy of Sciences of the Estonian SSR [6]. The position of the tongue at various points is determined not by perpendicular co-ordinates but by the pseudo-polar co-ordinates advocated by J. Liljencrants [4]. Fig. 1 shows that in the positions above the horizontal axis, the measurements are taken at intervals of ten degrees. In the lower part of the horizontal axis the tongue positions are measured as distances from the vertical axis at every 0.5 cm. Thus each tongue contour is characterized by a tongue vector (Fig. 2) composed of 23—25 variables.

3. Processing

For imitating the processes that take place in the vocal tract, one can use simple articulatory models. We shall compare some of the dimensionality-reducing techniques on the basis of those models in order to find a variant suitable for representing speech units. The criteria for comparison are represented by two aspects essential from the viewpoint of speech-oriented ACS's: the coding of speech units in such a manner that they would occupy the minimum memory, ensuring, at the same time, their sufficiently rapid reproduction.

In the following, four types of models are compared:

1. Fourier model — FOURMOD. The coding of tongue contours corresponding to speech units is carried out on the basis of the Fourier coefficients. The model realizes the steps of Fourier's transformation (it finds the Fourier coefficients which correspond to the tongue contour) and inverse transformation (proceeding from the coefficients, it reproduces the tongue contour).

2. Regression model — REGRMOD. The parameters are represented by the coefficients of a linear regression equation.

3. Factor model — FACTMOD. Parametrization of the tongue contours is effected through factor loadings and factor scores.

4. Component model — COMPMOD. The model is characterized by the matrix of weights and the principal components.

In addition to those models we also consider cluster analysis as an initial data dimensionality — reducing technique. Using cluster analysis we determine similar tongue contour groups and choose one representative out of each group which, with a certain degree of accuracy, describes all the elements of the group.

The main attention here is directed to the economical use of memory. The reproduction speed of speech units depends on the choice of a suitable algorithm, and the speed properties of the computer itself (in the present work, a computer of the type EC-1010 was used (operative memory 64K bytes, memory cycle 1.2 μsec)). With rapid computers the speed of reproduction becomes a factor of secondary importance, although one can never ignore it completely.

4. Results

Fig. 3 gives the tongue contour reproduction accuracy versus number of parameters for the factor model, Fourier's model and regression model. The reproduction accuracy is characterized by the mean-squared error $e(n)$:

$$e(n)=\sum_{i=1}^{M}(X_i(0)-X_i^n(s))^2/M, \qquad n=1, \ldots, N,$$

$X_i(0)$ — the i-th element of the original tongue vector; $X_i^n(s)$ — the i-th element of the tongue vector synthesized on the basis of n parameters; M — tongue vector measure; N — the maximum number of parameters.

One cannot construct an acceptable model on the basis of cluster analysis as in its case the procedures of grouping and minimizing the data are carried out without actually transforming the data. With regard to the parameters essential in the present work, the component and factor models are identical.

Fig. 4 depicts, for the three models being compared, the tongue contour reproduction time as a function of the number of parameters applied. We can observe that in the case of all the models the necessary

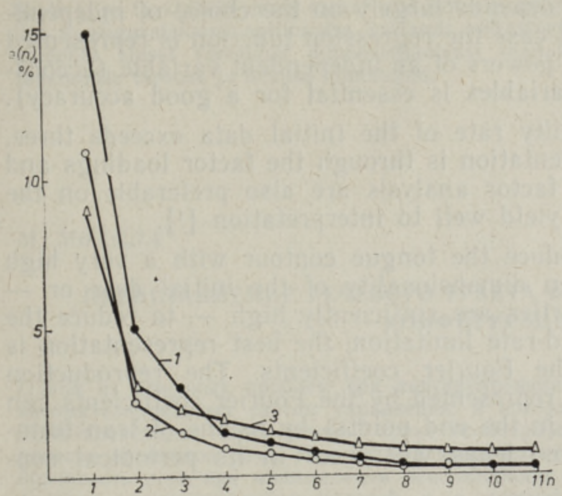

Fig. 3. The relation between the mean-squared error and the number of parameters in the analysis: *1* — factor model, *2* — Fourier model, *3* — regression model.

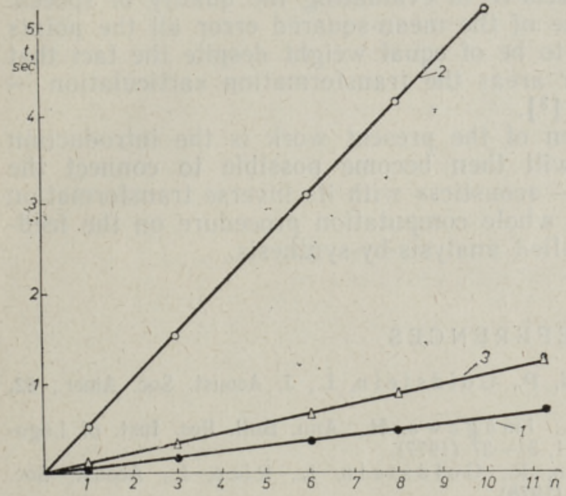

Fig. 4. The relation between the tongue contour reproduction time and the number of parameters (designations as in Fig. 3).

reproduction time *t* increases linearly with the increase in the number of parameters. The mean-square error $e(n)$ is inversely proportional to the number of parameters. In order to get the optimal variant of representation, one has to find a compromise between reproduction accuracy and reproduction speed.

Proceeding from the results, we offer the following variant of geometrical representation of the vocal tract in the computer:

1. In case of fairly strict limitations on the memory space to be used (tongue contours may be represented by only 1—3 parameters), the most suitable representation variant is via the regression equation coefficients. Fig. 2 shows that in the case of the REGRMOD reproduction accuracy is fairly high with a small number of parameters, yet above a certain point the reproduction curve becomes «saturated», i. e. a further increase in the number of parameters is not attended by any considerable increase in accuracy. On the other hand, we know that in the case of regression

models reproduction accuracy depends largely on the choice of independent variables — in the present case the regression function is represented by a polynomial of the integer powers of an independent variable (accordingly, the choice of correct variables is essential for a good accuracy).

2. If the allowed dimensionality rate of the initial data exceeds three, the optimum variant of representation is through the factor loadings and factor scores. The results of factor analysis are also preferable on the grounds that factor loadings yield well to interpretation [1].

3. If it is necessary to reproduce the tongue contour with a very high degree of accuracy at a given dimensionality of the initial data or — if the computer's speed properties are sufficiently high — to reduce the reproduction speed to a second-rate limitation, the best representation is obtained on the basis of the Fourier coefficients. The reproduction accuracy of a tongue contour represented by the Fourier coefficients can be improved upon (especially in the end points) by means of transforming the initial data so that no leaps will occur in its periodical continuation.

It should likewise be noted that the mean-squared error allows only a partial characterization of reproduction accuracy, and speech quality as the acoustical output is decisive in evaluating the quality of speech. The reason is that in the case of the mean-squared error all the points of the tongue are considered to be of equal weight despite the fact that not in all the tongue contour areas the transformation «articulation — acoustics» has a linear nature [3].

The immediate continuation of the present work is the introduction of the third dimension. It will then become possible to connect the transformation «articulation — acoustics» with its inverse transformation as well, and to construct the whole computation procedure on the feedback principle — as the so-called analysis-by-synthesis.

REFERENCES

1. Harshman, R., Ladefoged, P., Goldstein, L., J. Acoust. Soc. Amer., **62**, № 3, 693—707 (1977).
2. Kiritani, S., Sekimoto, S., Imagawa, H., Ann. Bull. Res. Inst. of Logopedics and Phoniatrics, № 11, 31—37 (1977).
3. Ladefoged, P., Harshman, R., Goldstein, L., Rice, L., Acoust. Soc. Amer., **64**, № 4, 1027—1035 (1978).
4. Liljencrants, J., Speech Transmission Laboratory QPSR (Stockholm), № 4, 9—18 (1974).
5. Maeda, S., C. N. E. T. Recherches/Acoustiques (Lannion), № 4, 131—143 (1977).
6. Eek, A., Remmel, M., Soviet Fenno-Ugric Studies, **5**, № 2, 141—145 (1969).

Academy of Sciences of the Estonian SSR,
Institute of Language and Literature

Received
March 29, 1979

M. MIHKLA

KÕNETRAKTI GEOMEETRILISE KONFIGURATSIOONI ESITUS ARVUTIS

Niisuguste kõnet kasutavate automatiseeritud juhtimissüsteemide loomisel, milles kõneinformatsioon esitatakse kõnetrakti geomeetrilise konfiguratsiooni kaudu, on olulisimaid probleeme kõneüksuste selline kodeerimine, mis kõnet kahjustamata tagaks nende võimalikult kiire taastamise ning samal ajal hõivaks kõige vähem mälu. Kõnetrakti geomeetrilise konfiguratsiooni optimaalse variandi väljaselgitamiseks on lihtsate artikulatoorsete mudelite alusel võrreldud dimensionaalsust madaldavaid tehnikaid. On esitatud kõne taastamise täpsuse ja kiiruse hinnangud.

М. МИХКЛА

ПРЕДСТАВЛЕНИЕ РЕЧЕВОГО ТРАКТА В ЭВМ В ГЕОМЕТРИЧЕСКОЙ КОНФИГУРАЦИИ

Одной из основных проблем при конструировании ориентированных на устную речь автоматизированных систем управления, в которых передача речевой информации в ЭВМ происходит через речевой тракт геометрической конфигурации, является кодирование элементов речи. Под оптимальным кодированием понимается требуемая разборчивость речи при минимальном использовании памяти ЭВМ. В работе на примере простых артикуляционных моделей, т. е. нескольких конфигураций речевого тракта, сравниваются алгоритмы, снижающие размерность данных, оценивается точность и оперативность воспроизведения речи.