THE METHOD FOR SOLVING INVERSE PROBLEM OF SPEECH PRODUCTION AND ARTICULATORY PORTRAY OF A SPEAKER

Yevgeni Vlasov, Natali Isayeva

Institute of Control Sciences
Academy of Sciences USSR
Moscow, USSR 117342

ABSTRACT

The accuracy of modern methods for determination of area function is not sufficient for practice. We present a numerical method for area function calculation with sufficient accuracy. Regulation continuum of this functions in finite region is called articulatory portray.

INTRODUCTION

Many year modelling of speech production processes has attracted investigators/1/, however its complexity up to the present does not lead to a wide introduction of such models into practice, despite the great efforts /2/ and intensively growing feasibilities. It has become clear, that the speech production processes are hierarchical and closely interacting /3/. In this case speech production model is expedient to be realized from bottom to top, using a lower level as a tool /4/. One of them is the articulatory model /5/, which synthesizes speech on the basis of solving the direct problem of speech production (vocal tract → acoustic) /6/.
The more accurate data of area functions have been obtained by G. Fant (1960) /7/ and up to now this work remains unique, because of its complexity. The LPC-method /8/ requires special measures (beforedistorting, etc.) for obtaining valid solutions. The tomography method /9/ enables us to determine the area in any section, however it requires multiple X-ray photographing of such sections along the axis for reconstructing only one area function. It was necessary to develop the method for an easier way of obtaining area functions without accuracy loss.
The paper presents the method for solving the inverse problem of speech production (acoustic → vocal tract), which allows the obtaining of "smooth" area functions and articulatory portray. The latter represents the region of permissible articulatory situations of a speaker. This method is based on the idea of "analysis from synthesis" and includes the algorithm /6/ with two-tier adaptive program complex (APC) /10/. The distinct features of the APC are automatical problem orientation to the class of the problems to be solved, supported by multidimensional optimization and associative information processing by a computer.

INPUT DATA AND ERRORS

Input data are easily-measured spectrum-time speech parameters: frequences, bandwidths and amplitudes of formants and also X-ray images of vocal tracts in sagital flatness of three speakers: two men /7,11/ and one woman /12/.
The formant frequences have been determined by sonagraph, the errors were 3-7 per cent. In future calculations the frequences vector $F^* = (F_i^*, i = 1,k)$ will be the standard and errors vector $\varepsilon^* = (\varepsilon^*, i=1,k)$ will be final accuracy.
From X-ray images we used the samples $H = (H_i, i = 1,M)$ of a heigt function $h(x)$, $0 \le x \le \ell$, where ℓ is the vocal tract length. Samples and length errors are respectively equal to 7 and 3 per cent.

THE METHOD

Taking into account the difficulty of obtaining an X-ray images, the method is realized by two variants: with X-ray images and without them.

The first variant (with an X-ray image). The area function $S(x)$ is represented as a product of the known height function $h(x)$ on a desired width function
The finite articulatory region determines

$$D_q: \left(q_i^{min}(x) \le q_i^{\circ}(x) \le q_i^{max}(x),\ 0 \le x \le \ell\right) \quad (1)$$

where $q_i^{\circ}(x)$ some initial approximation. Sampling of the all three function along the axis x in D_q gives respectively the vector of the lower boundary $W^{min} = (W_i^{min}, i=1,N)$, initial control vector $W^{\circ} = (W_i^{\circ}, i=1,N)$ and the upper boundary vector $W^{max} = (W_i^{max}, i=1,N)$ as shown in Fig. 1

$$D_w: \left(W_i^{min} \le W_i^{\circ} \le W_i^{max},\ i = 1,N\right). \quad (2)$$

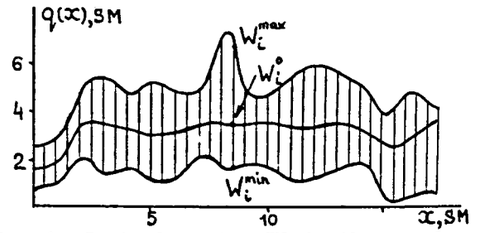

Fig. 1. Control vector W° is the samples of the function $q(x)$ in known boundaries (1)

Characteristics vector P, which characterizing the class of solved problems, includes F^*, ℓ , H:

$$D_p:(P=P_1,\dots,P_{K+m+1})=(F_1,\dots F_K,\ell,H_1,\dots H_m) \quad (3)$$

The mathematical statement of the speech production inverse problem has the following form:

$$L(F,F^*)\to\min,$$

$$L(F,F^*)=\Big(\sum_{i:\varepsilon_i\ge\varepsilon_j}(E_i-\varepsilon_j)^2\Big)^{1/2},\quad E_i=\frac{|F_i-F_i^*|}{F_i^*},\ i=1,K,$$

$$F=G(S(x)),\ S(x)=h(x)q(x),\ 0\le x\le\ell, \quad (4)$$

$$h(x)=R(H),\quad q(x)=R(W),$$

$$W\in\left(W_i^{min}\le W_i^{o}\le W_i^{max},\ i=1,N\right),$$

where L - functional, which depends on calculated frequencies F, G-an operator of the speech production direct problem, R - an operator transforming a given vector to a smooth function.
The method of solving problem is shown in Fig. 2 and consists of the following. Each problem is determined by concrete values of the vector P according to (3). Initial vector W^o gives the random width function q(x), which determines the random area function $S(x)=h(x)q(x)$. The operator G calculates F, which is compared with the standard F^* for determining L. The value L is analysed in the APC with the aim of optimizing the components for finding the minimum of L. When the final value L^* is achieved, the calculation process is finished and the decision vector (P,W^*,L^*) is stored in a computer memory. For a new problem P′ we take from the memory such an initial W^o in the set of earlier solved problems, whose vector P is closer to P′.

The second variant (without an X-ray image). The region is determined as follows:

$$D_S:\left(S^{min}(x)\le S(x)\le S^{max}(x),0\le x\le\ell,\ \ell\in[\ell^{min},\ell^{max}]\right), \quad (1')$$

sampling of which, gives the region $D_W(2)$. The vector P includes the formant frequencies F^*, bandwidths ΔF^* and amplitudes A^*. The two latter vectors may be not available.

$$D_p:\ (P=P_1,\dots,P_{3K})=(F_i^*,\Delta F_i^*,A_i^*,\ i=1,K). \quad (3')$$

In (4) the area function is formed directly from the control vector:

$$S(x)=R(W),\ 0\le x\le\ell,\ \ell\in[\ell^{min},\ell^{max}] \quad (4')$$

In other aspects this variant does not change and is illustrated in Fig. 3 (compare with Fig. 2).

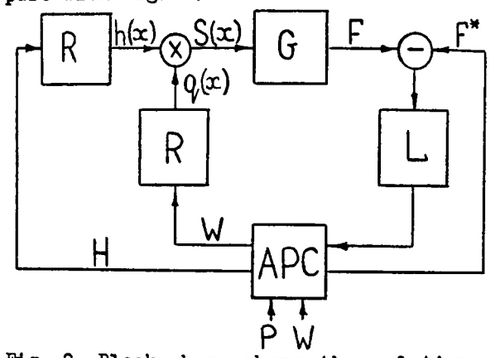

Fig. 2. Block-sheme shows the solution method for the first variant.

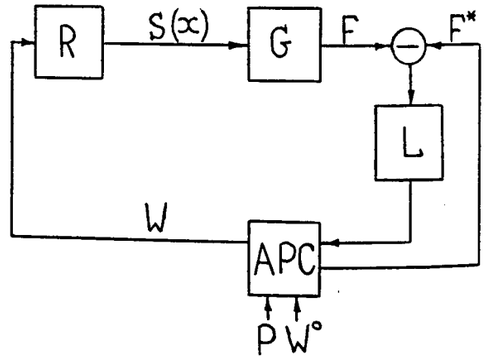

Fig. 3. Block-sheme shows the solution method for the second variant

REGULATION ALGORITHM

In general the inverse problems are mathematically noncorrect, i.e. they admit nonunique solution. Correctness of the considered problem is caused by constraining permissible solution region (1) and by the development of the special regulation algorithm.
This algorithm works along contour II in Fig. 4, while contour I denotes the functioning of the APC without it. At first

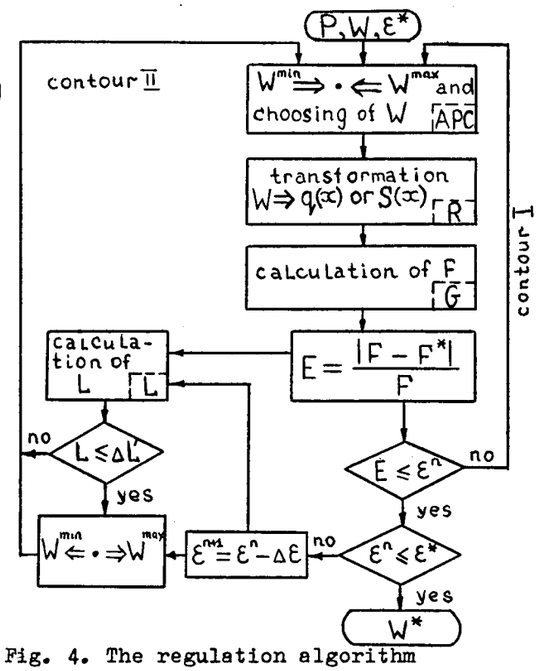

Fig. 4. The regulation algorithm

the rough stages value of accuracy ε^o is assigned. For a given P and initial W^o the error E is calculated and compared with ε^o. In contour I the ε^o is attained ($E\le\varepsilon^o$), then contour II gives the next accuracy $\varepsilon^1=\varepsilon^o-\Delta\varepsilon$, etc. It should be noted that in contour I the boundaries (2) are approached: $W^{min}\Rightarrow\cdot\Leftarrow W^{max}$, while in contour II they are expanded: $W^{min}\Leftarrow\cdot\Rightarrow W^{max}$ up to initial boundaries (2). The algorithm is also adapted to the change of L: contour II is switched over if the reduction velocity ΔL less than the threshold $\Delta L'$.
Thus, the reaching final accuracy ε^* is divided into a sequential stages, each of which gains a stage accuracy E^n

$$\varepsilon^o\ge\varepsilon^1\ge\dots\ge\varepsilon^n\ge\dots\ge\varepsilon^*$$

$$W^o\Rightarrow W^1\Rightarrow\dots\Rightarrow W^n\Rightarrow\dots\Rightarrow W^* \quad (5)$$

In this case sequence $S^n(x)$ tends to optimal $S^*(x)$.

RESULTS

The proximity criterion of functions S(x) $S^*(x)$, simular to /8/, is the mean square deviation, normalized by a maximum

$$\sigma=\Big(\frac{1}{M}\sum_{i<1}(S_i-S_i^*)^2\Big)^{1/2}\Big/\max_{1\le i\le M}S_i^* \quad (6)$$

Stability. The scatter of obtained solutions S (x) under the variations of the initial approximation W^o and boundaries W^{min} , W^{max} in (2) have been estimated. For similar phonems this scatter does not exceed 6.7 per cent with the deviation of control vectors from the initial values (Fig. 1) up to 120 per cent.

Convergence. The convergence to the accurate solution S (x) is guaranteed by the above regulation algorithm. Rejection of this algorithm leads to the interruption of convergence, as shown by the dotted line in Fig. 5. The continuous line shows the normal process of convergence: in points L_1 , L_2 , L_3 correction of stage accuracy ε^n takes place by contour II.

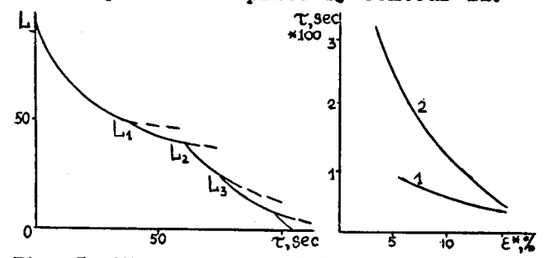

Fig. 5. Minimization of L shows the convergence of computer process.

Fig. 6. The illustration shows the benefit of regularization algorithm application.

Accuracy. Accuracy of $S^*(x)$ is estimated by (6) for the first speaker, since he has the exact area function /7/. Among the phonems the mean accuracy equals 8.3 per cent and it varies in the range of 4.1 - 12.9 per cent. With respect to the results /8/, where the LPC-method is used for the same speaker, the accuracy has increased by 2.7 per cent.

Computer time. The application of the requaltion algorithm provides not only required accuracy, but acceleration of the computer processes as well. Fig. 6 shows that the final accuracy ε^* is move benificially obtained making use of this algorithm since the solution time with the algorithm application (curve 2) is many times reduced as opposed to the one without algorithm application (curve 1). The quantity of benefit is increased with the increasing of the final accuracy, from 1.6 times at ε^* = 10 per cent to 4.5 times at ε^* = 5 per cent. Additionaly, owing to optimal fitting of the algorithm parameters the computer time is reduced by 2.5 -30 times. Among the phonems the average computer time is equal to 84 sec and varies in the range of 4.5-148 sec.

Comparison of two variants. Different input data application (with / without X-ray images) leads to the average error of 9.4 per cent in solutions. The average computer time in the second variant is greater by 18.4 per cent than the one in the first variant. Hence, decreasing of apriori information should be compensated

at the cost of increasing the computer time.

ARTICULATORY PORTRAY

Representation of the relationship of the solution $S(x)$ and input $h(x)$ as a functional dependence

$$S = S\left(h(x), x\right), \quad 0 \leqslant x \leqslant \ell \tag{7}$$

in three-dimensional space (S, h, x) leads to a complex surface as shown in Fig. 7.

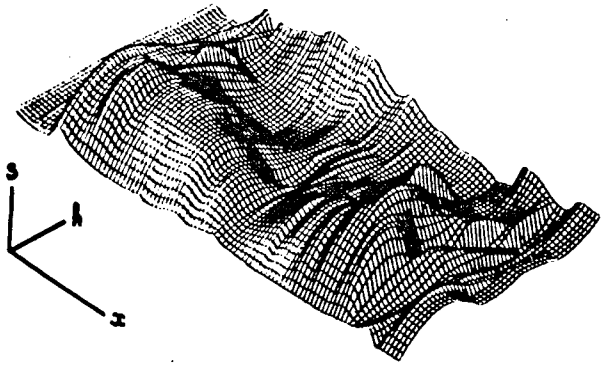

Fig. 7. Articulatory portray of a speaker

Such surface clearly presents the region of admissible articulatory situations of the speaker, therefore it is called an articulatory portray. Such protray vividly shows pharinx and oral (two saliences) and contraction of a larinx tube, a contraction caused by valum (pass between saliencea) and lips. The practical application of such a portray particularly consists in easy transformation of the flat X-ray image of vocal tract to area function with a given accuracy.

Comparison of portrays. The degree of individual distinction between the speakers is equal to 5.1 per cent for a diameter of a vocal tract and to 3.1 per cent for a longitudinal dimension, that may be interpreted as a value of articulation "quanta" /13/.

CONCLUSION

Vocal tract is usually approximated by a few cylinder sections, but in practice, a a more smooth area function is required. In proposed method there are no restrictions on a quantity of sections and computer time does not depend on the quantity of sections. The required accuracy, corresponding to input data errors, is also quaranteed. The method provides the reduction the compute time up to 4.5 sec, which equals one computer remembrance time. For articulatory synthesyzer /5/ smooth area function provides the improvement of the quality of a synthetic speech.

Computer time depends on the degree of APC-knowledge and the correct sequence of the problems to be solved. The better the APC is trained the shorter is the computer time. Due to the rational choice of problem sequences the training time decreases by 1.5-2.0 times with respect to random sequences.

In addition to it, this method may be used in speech analysis, medicine and in logophedia.

REFERENCES

/1/ J. Balazs, "In Memoriam Farkas Kempelen", Hungarian Papers in phonetics, ed. K. Bolla, No.13, 1984, p.11-21.

/2/ J. Allen, "A Perspective on Man-Machine Communication by Speech", Proc. IEEE 73 (11), 1985, p.1541-1550.

/3/ В.Н.Сорокин, "Теория речеобразования", Москва, Радио и связь, 1985.

/4/ P. Mermelstain, "Articulatory Model for the Study of Speech production", J. Acoust. Soc. Am. 50(4), 1973, p.1070-1082.

/5/ Е.В.Власов, "Акустический терминал для артикуляционного синтезатора речи", Тезисы I2-го Всесоюзного семинара по автоматическому распознаванию слуховых образов, Киев, Институт кибернетики АН УССР,1982,с.389-393.

/6/ Е.В.Власов,"Модификация метода Галеркина для расчета частотных параметров речевых сигналов", Проблемы построения систем понимания речи,Москва, Наука, 1980, с.136-142.

/7/ G. Fant, "Acoustic Theory of Speech Production", Gravenhage, Moution, 1960.

/8/ J.D. Markel, A.H. Gray, "Linear Prediction of Speech", N.Y., Springer-Verlag, 1976.

/9/ S.Kiritani, E. Takenaka, M. Sawashima, "Computer tomography of the vocal tract", Ann. Bull. Research Inst. Logopedics and Phoniatrics, Tokyo, No.12, 1978, p.1-4.

/10/ В.С.Широколава, Н.А.Исаева, "Двухъярусный обучающийся программный комплекс", Модели управления сложной программой, Москва, Институт проблем управления, 1986, с.3-10.

/11/ В.Н.Сорокин, "Механика движений языка", Описание и распознавание объектов в системах искуственного интеллекта, Москва, Наука, 1980, с.42-71.

/12/ K. Bolla, "A Phonetic Conspectus of Russian", Hungarian Papers in Phonetics, No.11, 1982.

/13/ K.N. Stevens, "The quantal nature of speech. Evidence from articulatory-acoustic data", Human com.:A unified view,McGrawHill,1972,p.51-66.