ARCHIVES OF ACOUSTICS
18, 4, 525–537 (1993)

ON THE MODEL OF VOCAL TRACT DYNAMICS

W. NOWAKOWSKA, R. GUBRYNOWICZ AND P. ŻARNECKI

Institute of Fundamental Technological Research
Polish Academy of Sciences
(00-049 Warsaw, Świętokrzyska 21)

An attempt to model a true vocal tract shape is presented. The base data were articulatory data taken from the cineradiographic recordings and the speech signal simultaneously registered (BOLLA, FÖLDI, 1987). Moreover, information concerning distances between the reference points, labiograms, photopalatograms and photolinguograms were used. A non-uniform three-dimensional vocal tract model was applied. The most difficult problem was how to evaluate the variations of the lateral dimensions along the longitudinal axis of the vocal tract. Only the lip opening dimensions for each speech sound were the objects of direct measurements. After the photo of the subject's hard palate one of the several plaster casts made for other male subjects was chosen as the most similar one. On its basis the dimensions in this region were reconstructed and assume to be constant. The lateral dimensions in other parts of the vocal tract were reconstructed after the formant frequencies measured for each Polish vowel spoken by the subject whose vocal tract was modelled. These dimensions were adjusted in such way that they were common for all vowels and the corresponding formant frequencies were obtained by varying vertical dimensions according to the X-ray data. This model was applied to study the relationship between the shape variation of the vocal tract and its acoustic output, especially in the case of study of transient sounds spoken with vocalic neighbourhood.

1. Introduction

Vocal tract model computation in the frequency domain is nowadays a well established procedure, especially for stationary speech sounds (FANT, 1960; MRAYATI, 1978; ATAL and al. 1978; BADIN and FANT, 1984; LIN, 1990). However, in most cases, the main object of vocal tract simulation is to obtain the best possible match of the calculated frequency responses with characteristics of the modelled speech sounds. Even if the vocal tract configuration taken for calculation is initially based on X-ray picture, it is next modified in order to achieve a higher degree of modelling accuracy in the frequency domain. It must be stressed that the continuous change of the tongue shape and position is hard to measure and to model adequately. Some basic facts are known, which describe certain stable articulatory mechanisms either in the steady state or in transitions; the rest is rather hypothetical. Another source of discrepancy between the vocal tract shape and its physical representation is its approximation by a number

of contignuous cylindrical tubes consisting, generally, of about 1 cm long, 17–20 segments. Although, in almost all systems, the vocal tract is represented by a cascade of cylindrical tubes of finite length, the form of the vocal tract cross-sections is far from circular, especially, in its pharyngeal and palatal parts. In the model calculation the cross-sectional shape has a certain influence on surface losses, but their effect on its frequency response is not considerable. The model composed of cylindrical lossy tubes was applied successfully in many researches and was the basis for formulation of the acoustic speech production theory (FANT, 1960).

Even an elementary articulatory model introduces an additional level for representation of speech phenomena such as coarticulation, reduction, assimilation and other context-dependent allophonic variations. This additional level appears to be more suited to human intuition in the manipulation of hypothesis than the lowest level of the exclusively acoustic signal description. Furthermore, new experimental possibilities are open to speech researcher at the acoustic level: some simple articulatory movements may induce very complex acoustic mechanisms that would be not recognized as basic in the speech production process.

However, the uniform cylindrical vocal tract (VT) model applied to reproduce some dynamic phenomena existing in natural speech does not always fit to describe them, especially in the output signal domain. Adopting linear form of the motion from the starting shape to a target one, the output signal calculated for VT of circular cross-section shape reveals stepwise variations non observed in natural signal. This signifies that the interpolated transitional VT configurations have no place in reality. In other words, the variations of the VT configuration (more exactly, of the area function) are not linear and the variations of the cross-section shapes should also be taken into account. This was the reason for applying a non-uniform vocal tract model to shape as precisely as possible, the contours of supraglottal organs obtained from cineradiography.

Application of an articulatory model instead of the physical one offers the advantage that the set of possible area functions is constrained to be compatible with anatomy, thus reducing ambiguities of solutions. In this work we tried to fit our articulatory model to a part of the accessible area function obtained by "direct" methods and fitted it at the same time to the acoustic data. The main criterion used to evaluate the quality of the vocal tract configuration approximation was the degree of discrepancy in formant space between the vocal tract transfer function and the frequency characteristic of the simulated speech sound.

We applied a non-uniform model in order to study dynamic phenomena in speech, in sequence of voiced sounds, like liquids and nasals with vowels, in particular. The main problem was to establish how the vocal tract configuration is changing from one position of the articulators corresponding to a preceding sound to the following one. An important part of the research was devoted to model the steady vocalic vocal tract shape suitable for the natural one in articulatory and formant frequency spaces. An example of mapping from the space of articulators to the frequency space in dynamic case was established for vocalic sequence of /u/ and /i/, with restriction to the time signal to be without jumps.

2. Acoustic model of the speech producing system

The human speech sound is characterized by the properties of the source of excitation and the acoustic transmission system. When assuming purely one-dimensional acoustical wave propagation in the vocal tract, it is known that the most accurate computational model is its representation as a transmission line analog. The vocal tract shape is usually modelled by a series of cylindrical tubes of finite length (Fig. 1).

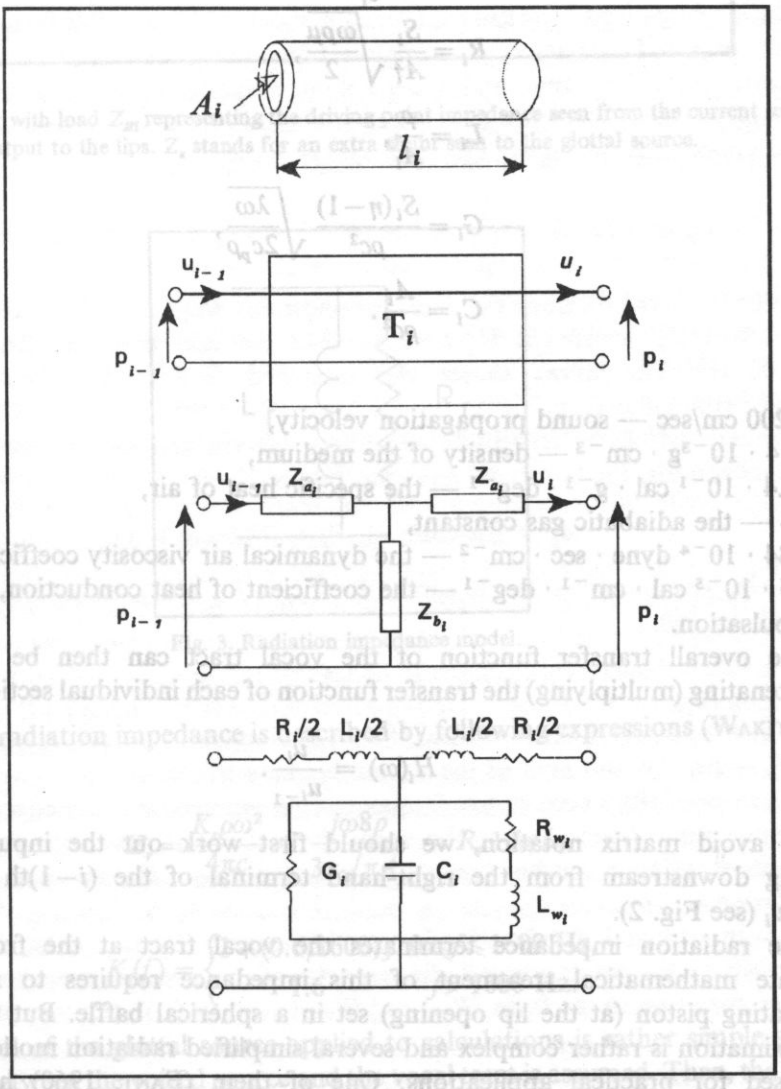

Fig. 1. T-network representation of a cylindrical section of length l_i.

The sound propagation in each cylindrical section is solved by applying the conventional quadrupole equations defining the relations between the input and output quantities (usually acoustic pressures and volume velocities)

$$\gamma_i = \alpha_i + j\beta_i = \sqrt{Z_{a_i} Z_{b_i}} = \sqrt{(R_i + j\omega_i L_i)\,(G_i + j\omega_i C_i + Y_{w_i})}, \tag{1}$$

with the condition that $|\gamma_i|\, l_i << 1$, and where

$$Z_{0_i} = \sqrt{\frac{Z_{a_i}}{Z_{b_i}}},$$

$$R_i = \frac{S_i}{A_i^2}\sqrt{\frac{\omega\rho\mu}{2}},$$

$$L_i = \frac{\rho}{A_i}, \tag{2}$$

$$G_i = \frac{S_i(\eta - 1)}{\rho c^2}\sqrt{\frac{\lambda\omega}{2c_p\rho}},$$

$$C_i = \frac{A_i}{\rho c^2}.$$

Here

$c = 35200$ cm/sec — sound propagation velocity,

$\rho = 1.14 \cdot 10^{-3}$ g · cm^{-3} — density of the medium,

$C_p = 2.4 \cdot 10^{-1}$ cal · g^{-1} · deg^{-1} — the specific heat of air,

$\eta = 1.4$ — the adiabatic gas constant,

$\mu = 1.84 \cdot 10^{-4}$ dyne · sec · cm^{-2} — the dynamical air viscosity coefficient,

$\lambda = 5.5 \cdot 10^{-5}$ cal · cm^{-1} · deg^{-1} — the coefficient of heat conduction,

ω — pulsation.

The overall transfer function of the vocal tract can then be obtained by concatenating (multiplying) the transfer function of each individual section defined as

$$H_i(\omega) = \frac{u_i}{u_{i-1}}. \tag{3}$$

To avoid matrix notation, we should first work out the input impedance looking downstream from the right-hand terminal of the $(i-1)$th section $Z_{B,i} = P_i/u_i$ (see Fig. 2).

The radiation impedance terminates the vocal tract at the front end. An accurate mathematical treatment of this impedance requires to regard it as a vibrating piston (at the lip opening) set in a spherical baffle. But this mode of approximation is rather complex and several simplified radiation models have been proposed for practical applications. One of them (FANT, 1960) has the form presented in Fig. 3.

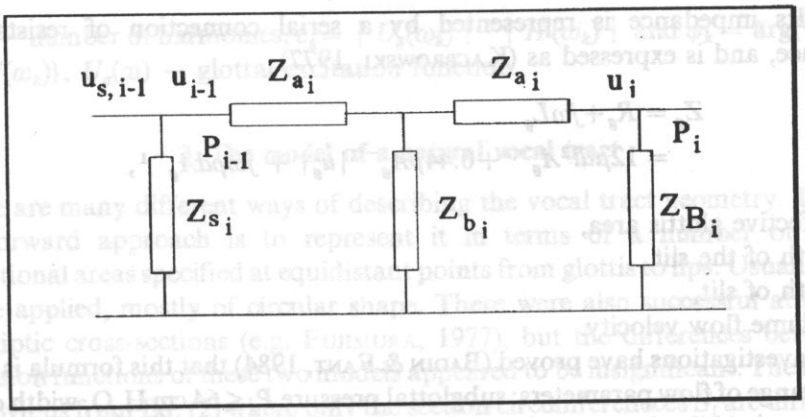

Fig. 2. T-network with load Z_{Bi} representing the driving point impedance seen from the current section's output to the lips. Z_s stands for an extra shunt seen to the glottal source.

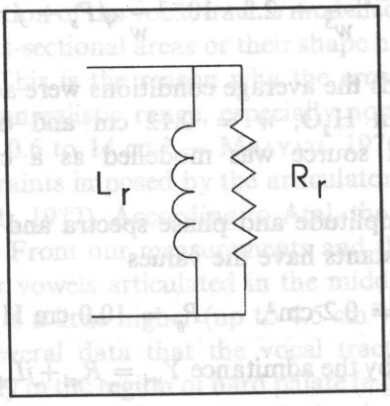

Fig. 3. Radiation impedance model.

Then the radiation impedance is described by following expressions (WAKITA and FANT, 1978):

$$Z_r = \frac{K_s \rho \omega^2}{4\pi c} + \frac{j\omega 8\rho}{3\pi\sqrt{\pi A_0}} = R_r + j\omega L_r, \tag{4}$$

with

$$K_s(f) = \begin{cases} 1+(0.6/1600)f & 0 \leqslant f \leqslant 1600 \text{ Hz}, \\ 1.6 & f > 1600 \text{ Hz}. \end{cases} \tag{5}$$

The model of the glottal source applied to calculations is rather simple and no interaction between the voiced source and the vocal tract is assumed. Then, the source of excitation for sonorants is defined as the volume velocity (u_g) of airflow through the

glottis. Its impedance is represented by a serial connection of resistance and inductance, and is expressed as (KACPROWSKI, 1977)

$$Z_g = R_g + j\omega L_g = 12\mu dl^2 A_g^{-3} + 0.44\rho A_g^{-2}|u_g| + j\omega\rho d A_g^{-1}, \tag{6}$$

where

A_g — effective glottis area,
l — length of the slit,
d — depth of slit,
u_g — volume flow velocity.

The investigations have proved (BADIN & FANT, 1984) that this formula is valid for a broad range of flow parameters: subglottal pressure $P_s \leqslant 64$ cm H_2O, width of the slit $0.1 \leqslant w(t) = A_g(t)/1 \leqslant 2$ mm, and mean volume velocity $|u_g| \leqslant 2000$ cm^3/s.

Assuming the geometrical glottal source dimensions to be constant, the formula for glottal impedance with $d = 0.3$ cm and $l = 1.8$ cm is

$$Z_g = 3.72 \cdot 10^{-4} \frac{l}{w^3} + 2.3 \cdot 10^{-2} \frac{1}{w}\sqrt{P_s} + j \cdot 1.95 \cdot 10^{-3} \frac{f}{w}. \tag{7}$$

For impedance calculations the average conditions were accepted, valid for medium voice effort $P_s \approx 10$ cm H_2O, $w = 0.12$ cm and over the frequency range $\leqslant 4500$ Hz. The larynx source was modelled as a current source generating pulses of triangular form.

For calculations of amplitude and phase spectra and the output signals of the modelled sounds, the constants have the values

$$A_g = 0.2 \text{ cm}^2, \qquad P_g = 10.0 \text{ cm } H_2O.$$

The losses represented by the admitance $Y_{wi} = R_{wi} + jL_{wi}$ are calculated separately for each section,

$$R_{wi} = S_i \cdot r_w \quad \text{and} \quad L_{wi} = S_i \cdot l_w,$$

where S_i is the circumference surface of the i-section and $r_w = 100.0$ [g/cm^{2}s], $l_w = 0.1$ [g/cm^2] — for the unit area of the circumference surface.

The slope of the voiced source frequency characteristic is assumed as -12 dB/oct. The connection of the nasal to the vocal channel is at the seventh section. All the constants mentioned can be changed in accordance with the assumed conditions of articulation.

The output signal of a given fundamental frequency F_0 was calculated by harmonic synthesis from the amplitude and phase spectra of the modelled speech sound. For its complex spectrum of the form $H(\omega) = |H(\omega)| \cdot \exp(j\phi(\omega))$, the output signal $f(t)$ is

$$f(t) = \sum_{k=1}^{N} c_k \cos(2\Pi k F_0 - \Phi_k), \tag{8}$$

where N — number of harmonics, $c_k = | U_g(\omega_k) | \cdot | H(\omega_k) |$ and $\phi_k = \arg\{(U_g(\omega_k)\} + \arg\{H(\omega_k)\}$, $U_g(\omega)$ — glottal excitation function.

3. The model of a natural vocal tract

There are many different ways of describing the vocal tract geometry. The most straightforward approach is to represent it in terms of a number of uniform cross-sectional areas specified at equidistant points from glottis to lips. Usually 17—20 areas are applied, mostly of circular shape. There were also successful attempts to apply elliptic cross-sections (e.g. FUJIMURA, 1977), but the differences between the transmission functions of these two models appeared to be insignificant. The reason of that is obvious from Eq. (2) where only the section circumfereneces S_i are different for the above two models. It is evident that augmentation of the circumference surface increases the losses related mainly to co-vibrating wall masses. This primarily has effect on q-factor of formant resonances, which is visible in modification of formant bandwiths.

When the transfer function of the vocal tract is modelled in the frequency domain, the absolute values of cross-sectional areas or their shape have a limited impact on its frequency characteristic. This is the reason why the cross-sectional areas in many models often vary in an unrealistic range, especially non an acceptable in case of continuous speech, (from 0.6 to 14 cm² — MRAYATI, 1976; up to 8 cm² — MAEDA, 1987), although the constraints imposed by the articulators on the vocal-tract shape are known (see ATAL et al., 1977). According to Atal, the cross-sectional areas vary between 0.1 and 3.5 cm². From our measurements and the X-ray data (BOLLA and FÖLDI, 1987) obtained for vowels articulated in the middle of the words spoken in isolation, the upper limit is a little higher (up to 4.5 cm²).

It is obvious from several data that the vocal tract is not unform along its longitudinal axis, especially in the region of hard palate (e.g. HIKI et al., 1986). To find the geometry of the upper oral cavities limited by the hard palate and the plane tangent to the upper teeth edge, ten subjects (8 male and 2 female) has been used to make dentist plaster casts of their palates. Then, each cast was cut at every 1 cm along its midline, in planes approximately perpendicular to the mid-sagittal line of the palate. In the Fig. 4, the shapes of the cross-sections in the hard palate region determined for two subjects are presented. It is evident that the shape approximation to the circular or elliptical form is not acceptable in this case, and the trapeziform shape is more adequate. As it will be seen next, the accepted form has a significant influence on the accuracy of modelling of dynamic vocal tract cross-sections shape variations during sounds sequence articulation.

The main object of our study was to reproduce the geometry of the vocal tract of a subject uttering a given vowel and to obtain the same spectral characteristic as those for the simultaneously registered sound. As basic data we applied the X-ray pictures showing the vocal tract mid-sagittal shapes (in the side plane) registered for Polish sounds (BOLLA and FÖLDI, 1987). Besides the pictures we used the data relating to

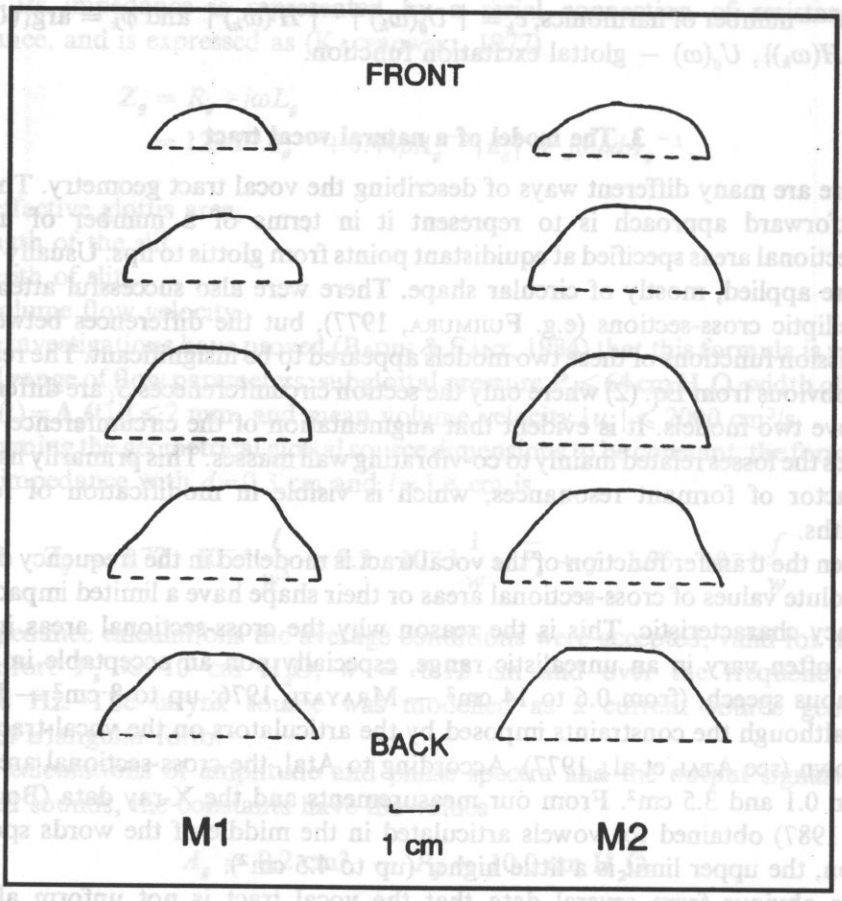

Fig. 4. Cross-sectional shapes of the oral cavity under hard palate taken for two male subjects.

distances between the referential points placed on the contours of the articulatory organs and photolabiograms with measurements.

The most difficult problem was to determine the third (lateral) dimension of the vocal tract. There were direct data only for its labial part, the cross-section shape of which was approximated by elliptical contours. The shape of the palatal cross-section of the vocal tract was assumed to be of trapeziform. In order to evaluate its dimensions, we chose (using the photo of the hard palate of the male informant participating in articulatory registration) the plaster cast of the similar dimensions and shape which fitted with reasonable accuracy (of the subject M1 from Fig. 4).

For this subject the configuration function has been determined using the measurement data concerning the distances between the reference points located on the midline of the palate and the tongue. The corresponding areas of cross-sections in palatal region are presented in the Fig. 4 (subject M1). One cm back from the rear edge of the palate a similar cross-section shape was accepted, and the shape of the

cross-sections of the lower oral and pharyngeal cavities of the vocal tract, up to the glottis, was modelled as elliptical. The vertical dimension was treated as a variable depending on its location and the articulated sound, and its value was taken from the mentioned conspectus. The dimension perpendicular to the mid-sagittal plane of the vocal tract was considered as a variable independent of the articulated sound, and its variation along the longitudinal axis of this part of the vocal tract is only due to anatomy. So, for all vowels the variation of this dimension is the same. In the palatal region only the trapezium upper side size is independent of the spoken sound and remains constant. The other side, e.g. the bottom one and the height of the trapezium are variable. The main object of modelling was to reproduce for each vowel, as exactly as possible, the full mid-sagittal shape and, at the same time, its frequency characteristic evaluated from the registered signal.

4. Results of dynamic modelling

After the X-ray data, labiograms, hard palate casts and the dimensions of the pharyngeal part evaluated on the basis of formant frequencies of all vowels, the spectral characteristics were calculated for each vocal tract configuration and compared with those obtained for real speech signal. This stage of modelling was performed in order to verify these dimensions of the vocal tract which do not vary from one vowel to another.

To model the time-dependent vocal tract configuration, in case of transient sounds, we have looked for rules of control of the area function changes. We have assumed that continuous spatial changes are modelled by linear variations of all variable dimensions, in the range determined by the starting and target configurations. It is obvious that the resulting cross-shapes obtained for intermediate states of articulation are varying in the non-linear manner.

To verify our approach we modelled transient articulation between two concatenated vowels /i/ and /u/, of extreme front-back opposition. The starging and target configurations were the same as those obtained for central part of the corresponding sounds spoken in the middle of the words and applied also to model their sustained phonation version. In this example of dynamic modelling the tongue movement is extreme. The duration of the transition from the first to the second vowel was equal to eleven pitch periods (about 73 msec), the same as in natural diphthong spoken by the subject under study. For each period an intermediate configuration with its corresponding transfer function and the output signal were calculated. The pitch period time synchronization imposed by the method of harmonic synthesis enables us to observe the transient output signal period-by-period and to detect easily any non-continuous, step-like transition.

In Fig. 5 the vocal tract characteristics (amplitude and phase) calculated for two concatenated vowels /i/ and /u/ and for eleven intermediary articulatory states are presented. The transitory output signal is on the top of the next figure. It is smoothly changing from state to state, without any abrupt jumps. Similar signal calculated for a cylindrical model is presented at the bottom of the figure. The irregular variations in

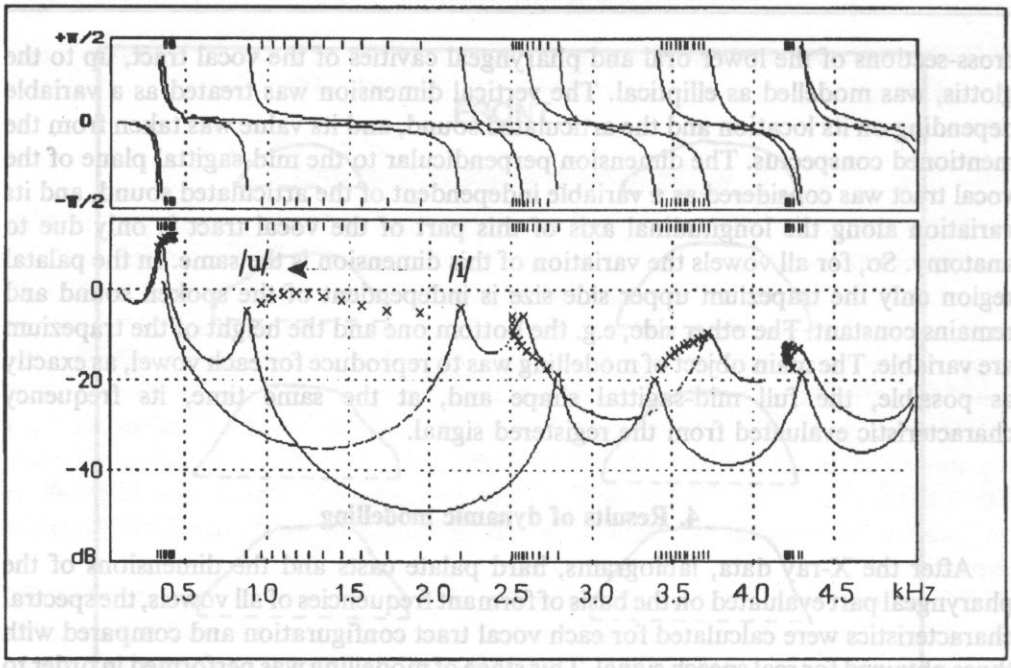

Fig. 5. Formant movements in transition from vowel /i/ to /u/ presented on phase (top) and amplitude (bottom) characteristics of vocal tract transmission function.

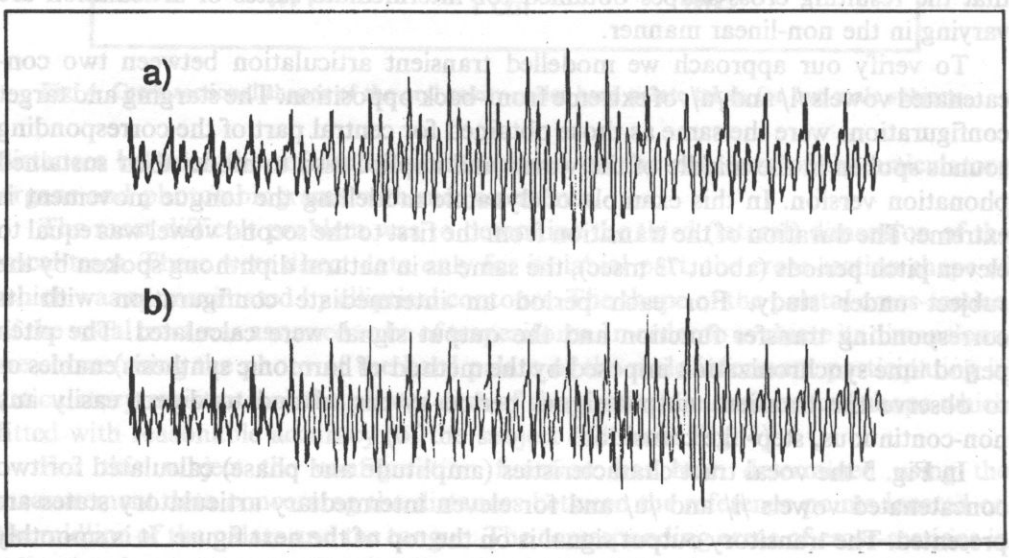

Fig. 6. Output signals for /i−u/ transition calculated for non-uniform vocal tract model (a) and circular model (b).

time are more accentuated than for the former signal. It seems that this model could not describe accurately the temporal changes in articulatory space, although the target configurations and characteristics were identical in both cases.

To compare, the formant trajectories were determined for non-uniform model and speech signal (Fig. 7). As a natural signal we analyzed a diphthong /ju/ in the Polish word "tiul" (tulle) spoken by the subject whose vocal tract was modeled. It seems that the movement of the place of articulation is rather correctly modelled. However, some

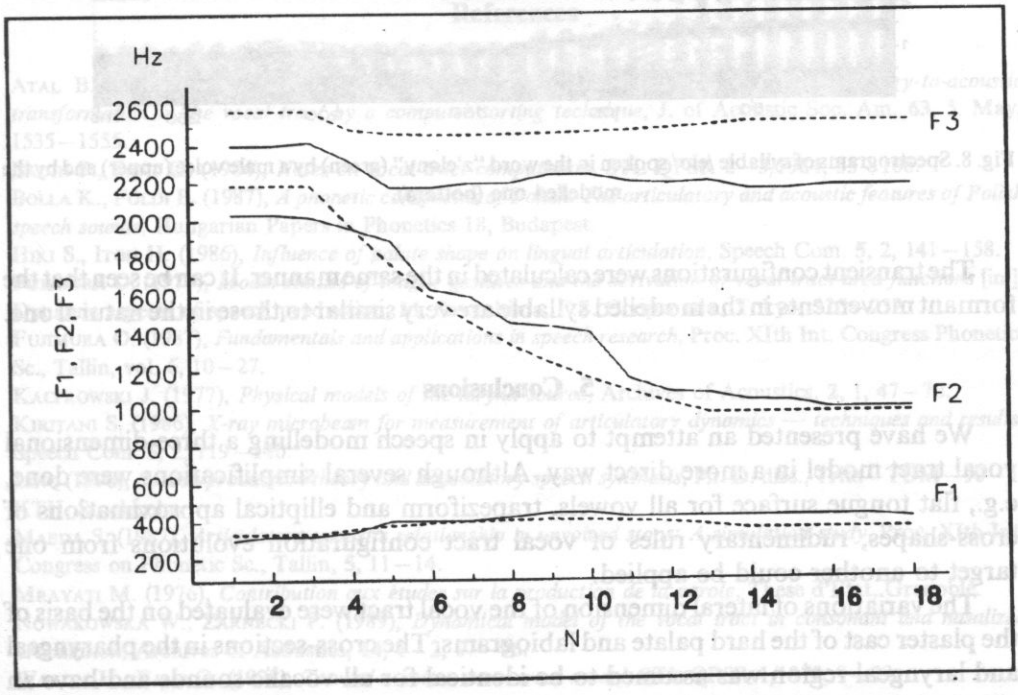

Fig. 7. Formant trajectories determined for natural (solid lines) and modelled (dashed lines) /i—u/ transition. *N* — pitch period continuous number.

discrepancies between the parameters of the modelled and the natural signal exist and this is probably due to the assumption of the movement synchronization of the whole tongue corpus. Nonetheless, for not so dramatic tongue movements the non-unform model was successful to model concatenations of vowels with different vowel-like segments (liquids, glides, for example). The example presented above could be also used to model the articulation of the syllable /ju/ with very short initial i-like segment and transition to vowel /u/.

Another example of results of dynamic modelling is presented in Fig. 8, where two spectrograms obtained for syllable /elo/ are presented. The upper spectrogram is obtained for male voice the vocal tract geometry of which was studied, the bottom picture is the result of modelling in which the vocalic target configurations (/e/, /o/ and /l/) were taken from the mentioned book (Bolla and Földi).

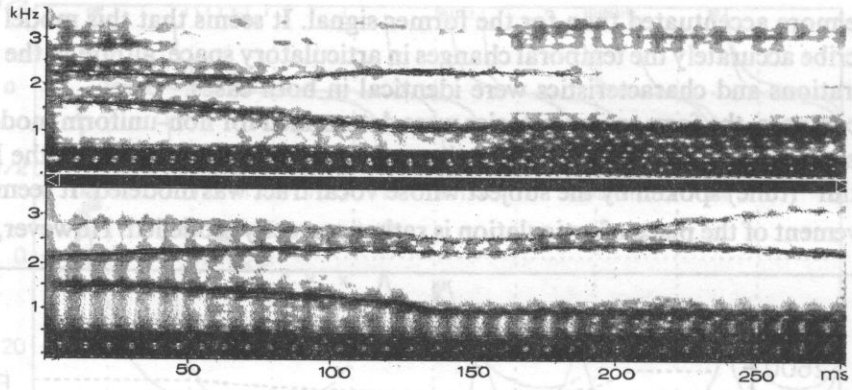

Fig. 8. Spectrograms of syllable /elo/ spoken in the word "z'elony" (green) by a male voice (upper) and by the modelled one (bottom).

The transient configurations were calculated in the same manner. It can be seen that the formant movements in the modelled syllable are very similar to those in the natural one.

5. Conclusions

We have presented an attempt to apply in speech modelling a three-dimensional vocal tract model in a more direct way. Although several simplifications were done, e.g., flat tongue surface for all vowels, trapeziform and elliptical approximations of cross-shapes, rudimentary rules of vocal tract configuration evolutions from one target to another could be applied.

The variations of lateral dimension of the vocal tract were evaluated on the basis of the plaster cast of the hard palate and labiograms. The cross-sections in the pharyngeal and laryngeal region was assumed to be identical for all vocalic sounds and have an elliptical shape with lateral dimension remaining constant for different configurations. These lateral dimensions were verified in the formant space for all six Polish vowels.

Our experiments in the development and use of three-dimensional (non-uniform) model have confirmed that it provides very effective means for the study of articulatory phenomena, especially of that connected with coarticulation pattern for different resonants. The model was also successfully applied to generate nasalized vowels and nasal consonants.

The timing characteristic of articulatory variations is linear, but it is also possible to adopt a non-linear one, e.g. of exponential form. The dynamic changes were modelled on the assumption that different articulatory dimensions vary synchronously during transitory speech waveform, from one to the following configuration. However, in the time signal a rising and sloping amplitude can be observed in the transitory part, not so accentuated in real signal. Its seems that the hypothesis of synchronous tongue movement should be verified as it was suggested by Fujimura (FUJIMURA, 1987), although the resulting modelled speech sounds are of very good quality.

Acknowledgment

The authors would like to express their gratitude to K. BOLLA and E. FÖLDI, researchers from the Phonetics Department at the University of Budapest, for providing us with the X-ray data and dynamic radiograms registered on videotape, without which we would not be able to carry out this research.

References

ATAL B.S., CHANG J.J., MATHEWS M.V., TUKEY J.W. (1978), *Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique*, J. of Acoustic Soc. Am. **63**, 5, May, 1535–1555.

BADIN P., FANT G. (1984), *Notes on vocal tract computation*, STL-QPSR 2–3/1984, 53–108.

BOLLA K., FÖLDI E. (1987), *A phonetic conspectus of Polish. The articulatory and acoustic features of Polish speech sounds*, Hungarian Papers in Phonetics 18, Budapest.

HIKI S., ITOH H. (1986), *Influence of palate shape on lingual articulation*, Speech Com. **5**, 2, 141–158.

FUJIMURA O. (1977), *Model studies of tongue gestures and the derivation of vocal tract area functions* [in:], Dynamic aspects of speech production, M. Sawashima, F.S. Cooper eds., Tokyo, 225–232.

FUJIMURA O. (1987), *Fundamentals and applications in speech research*, Proc. XIth Int. Congress Phonetic, Sc., Tallin, vol. **6**, 10–27.

KACPROWSKI J. (1977), *Physical models of the larynx source*, Archives of Acoustics, **2**, 1, 47–70.

KIRITANI S. (1986), *X-ray microbeam for measurement of articulatory dynamics — techniques and results*, Speech Com. **5**, 2, 119–140.

LINQ (1990), *Speech production theory and articulatory speech synthesis*, Ph. D. diss., Trita–TÖM–90–1, KTH, Stockholm.

MAEDA S. (1987), *Articulatory-acoustic relationship in unvoiced stops: A simulation study*, Proc. XIth Int. Congress on Phonetic Sc., Tallin, **5**, 11–14.

MRAYATI M. (1976), *Contribution aux études sur la production de la parole*, These d'Etat, Grenoble.

NOWAKOWSKA W., ŻARNECKI P. (1989), *Dynamical model of the vocal tract in consonant and nasalized articulation*, Archives of Acoustics, **14**, 1–2, 67–96.

WAKITA H., FANT G. (1978), *Toward a better vocal tract model*, STL-QPSR 1/1978, 9–29.