SPEECH INVERSION AND RE-SYNTHESIS

V.N. Sorokin,* & A.S. Leonov**, I.S. Makarov*, A.I. Tsyplikhin*

*Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia.
** Moscow Physical Engineering Institute, Moscow, Russia.

*vns@iitp.ru

Abstract

Inverse problems with respect to parameters of the articulatory model are solved for all types of sounds: vowels, semi-vowels, nasals, stops and fricatives in various contexts. Acoustical parameters of the speech signal and trajectories of some reference points inside the vocal tract serve as input data. 3.7%, 3.8% and 2.6% average approximation error for the first three formants, 8.5% for the specific frequencies of fricative spectra, 2.8% for the coordinates of reference points for all kinds of phonemes are obtained when both – acoustic and articulatory data are used. 1.8%, 1.6%, and 1.1% error for the first three formant frequencies, and 6% for the coordinates of reference points are obtained when only acoustic data are used. Original and re-synthesized utterances are found to be very similar in appearance, according to subjective assessment.

1. Introduction

The mathematical and physical aspects of speech inversion were studied in the fundamental works of Shirai, Sondhi, and their colleagues. In particular, Shirai introduced the norm of articulators excursion from the neutral position and the norm of articulator velocities into a criterion of optimality [1]. Sondhi has shown that only physically measurable acoustical parameters must be used, and applied a dynamic programming method to continuous speech inversion [2]. However, stability of their methods was not proven mathematically, and the accuracy of solutions was not studied. We lean upon the Tikhonov's regularization method [3] which guaranties the stability with respect to small perturbations, and the regularization coefficient depends on the correctness of speech production model and the accuracy of acoustical parameters measurement. We also found that the acoustical parameters and optimality criteria depend on speech segment type, while the codebook for the initial approximations must be formed by means of solving specific inverse problem using X-ray microbeam measurements. This task may be tackled by the use of additional information on trajectories of some reference flesh-points inside the vocal tract, recorded by the X-ray microbeam system or an articulograph synchronously with the speech signal. In the present study, such information for one male speaker was obtained from the database of Wisconsin University [4].

Hitherto, speech inverse problems have been solved mostly for isolated vowels or vowel sequences. This is because the inversion for stops, nasals and fricatives is highly difficult due to different and complicated acoustical processes. We present results of speech inversion for all kinds of phonemes - vowels, fricatives, stops and nasals. The adequate articulatory and acoustical models, the variational regularization of inversion, a code-book with support of articulatory data, and analysis of specific acoustical parameters for each kind of phonemes enable us to find suitable solutions.

2. Articulatory model

We use an 18-parametric articulatory model in inverse problem solving and speech synthesis. The model is mostly described in [5, 6], and some details are adopted from [7]. It contains the following articulatory parameters: area and height of the vocal slit, coordinates of the tongue root, degree of tongue rotation about its root, horizontal displacement of the jaw pivot, degree of jaw rotation, five coefficients for the eigenfunctions of elastic deformations of the tongue, height of the velum, degree of transversal deformation of the tongue, two coefficients for the eigenfunctions of transversal deformation of the pharynx, height of the lower lip, and lips protrusion. Besides, anatomic parameters for the speaker were obtained from the database. These parameters are: shape of the hard palate, shape and dimensions of the lower and upper jaw, pharyngeal width for the neutral position. A number of parameters used in the model (e.g. wall impedance and shape of the sinuïs piriformis) were found for all isolated vowels uttered by the speaker by means of minimizing discrepancy between measured and calculated formant frequencies. The articulatory model together with anatomic parameters computed for the given speaker enables us to calculate vocal tract shapes and distributions of cross-sectional area $S(x,t)$ along vocal tract mid-line.

3. Acoustical model

The acoustic processes in the vocal tract are bound up with energy losses. The main losses for fricatives are determined by viscous friction, which is proportional to oscillating velocity of air particles. At the same time, the losses for heat conductivity, which are in proportion to acoustic pressure, are relatively small. Therefore, we accept that the acoustic pressure $P(x,t)$ in the vocal tract obeys the following initial boundary problem for the one-dimensional wave equation with viscous losses:

$$\frac{\partial^2 P}{\partial t^2} - c_0^2 \frac{\partial}{\partial x} \left( \frac{\partial P}{\partial x} \right) = -2r \frac{\partial P}{\partial t} + F(x,t),$$

$$S(x,0) \left. \frac{\partial P}{\partial x} \right|_{x=0} = -\rho_0 \left( \frac{w}{S_{\omega}} \right)_0,$$

$$\left. \left( P(x,t) - \frac{Z_F}{\rho_0 S_{\omega}} \frac{\partial P}{\partial x} \right) \right|_{x=A} = 0, \quad 0 < x < l, \quad t > 0$$

(1)

Here $x$ is the coordinate along the mid-line of the vocal tract, $t$ is the time, $l$ is the length of the mid-line, $S$ is the cross-sectional area, $r$ is the coefficient of losses, $c_0$ is the sound velocity for the tract with yielding walls, $F$ are the distributed turbulent and lumped pulse sources of excitation, $w$ is the volume velocity of airflow through the vocal slit, $S_{\omega}$ is the...
vocal slit area, $Z$ is the radiation impedance, $\rho_0$ is the air density, and $\omega$ is a circular frequency.

The same wave equation is valid for the nasal cavity with the area function being time independent everywhere, except the velum region. When the velum is lowered the vocal tract is branched. The boundary conditions in the branching section are found from the continuity of the acoustic pressure and the acoustic volume velocity.

The resonance frequencies of the vocal tract can be computed for known area function $S(x,t)$ at each time moment by means of solving the spectral problem for the wave equation. We have found that the frequency of the first resonance is most affected by yielding walls while the frequencies of the second and third resonance are most affected by sinusis piriformis [8].

4. Aerodynamic models of excitation sources

Characteristics of excitation sources are determined by a system of non-linear differential equations of airflow dynamics in the vocal tract which is too cumbersome to put it here [9]. In particular, the airflow through the vocal slit can be described by the equation

$$
\rho_0 h_{\text{vs}} \left( 1 + \frac{S_{\text{vs}}}{2S_{\text{v}}} \right) \frac{dw}{dr} + \left( k_v - \rho_v \frac{S_{\text{v}}'}{2S_{\text{v}}'} h_{\text{vs}} \right) w + \frac{\rho_v c_v w^2}{2S_{\text{v}}} = \Delta p S_{\text{v}}.
$$

where $w$ is the volume velocity of the airflow, $h_{\text{vs}}$ is the depth of the vocal slit, $S_{\text{vs}}$ is the vocal slit area, $S_{\text{v}}$ is the area of the trachea at the entry to the larynx, $c_v$ is the coefficient of dynamic resistance, $\Delta p$ is the pressure difference between the upper and lower parts of the vocal folds, and $k_v$ is the coefficient of viscous friction. The vocal source of excitation of acoustical oscillations in the vocal tract is proportional to the time derivative of volume velocity $w$. The similar differential equation for $w$ describes airflow dynamics in tract constrictions formed for articulation of fricatives.

We accept the following model for generation of fricative noise by the use of calculated volume velocity $w$. First, the linear velocity of the airflow in the vocal tract is computed from the condition of airflow incompressibility $v(x,t) = w(t)/S(x,t)$. Then the airflow turbulence evolves as soon as Reynolds number $\text{Re}(x) = \rho_v(x)v(x,t)/\mu$, exceeds critical value $\text{Re}_{\text{crit}}$, where $\mu$ is the air viscosity and $h$ is some characteristic dimension of cross-section. The spectrum of the dipole source of turbulent excitation has maxima at the frequencies

$$
f_n = 0.2v(x_0)n/h(x_0), \quad n = 1,2.
$$

Here $x_0$ is the coordinate of lumped turbulent source. So, we can find special signals $\zeta_1(t), \zeta_2(t)$ obtained after filtering of white gaussian noise by the filters with poles $\omega_n = 2\pi f_n$ and damping coefficients $\gamma_n = 2\pi f_n$. Finally, the lumped turbulent source $F(x,t)$ in our acoustic model is described as follows:

$$
F(x,t) = (A_1 \zeta_1(t) + 0.1A_2 \zeta_2(t))\delta(x-x_0).
$$

The value $A_1 = k_1(Re_v^2(x_0) - Re_{\text{crit}}^2)$ presents here the noise amplitude. In more detail, we considered the mechanism of fricatives generating in [10].

5. Statement of inverse problem

The variational approach to the solution of our inverse problem consists in finding a time-dependent vector $u = u_k(t)$ of articulatory parameters such that it minimizes a criterion of optimality $\Omega[u]$ under constraints at each time moment, i.e.

$$
\Omega[u_k(t)] = \min \Omega[u(t)], \quad 0 \leq t \leq t_{\text{max}}
$$

subject to $\|A(u) - Z_0\|_2 \leq \delta$.

Here $A(u)$ denotes the operator of our “articulation-to-acoustics” model. In general, $z = A(u)$ is a procedure for calculating a vector $z$ of acoustical parameters by the use of an articulatory parameters vector $u$. The constraints in (5) define admissible values of articulatory parameters. In more detail, the functions $\Psi(z)$ determine restrictions for articulatory parameters and for forms of the vocal tract. The vector $z_k$ consists of acoustical parameters, experimentally measured with relative accuracy $\delta$. The inequality

$$
\|A(u) - z_0\|_2 \leq \delta
$$

means that we take into consideration only the admissible parameters with the relative discrepancy of order $\delta$ for experimental acoustical data and their calculated analogues. When we solve the problem (5) at a time $t$, we make an attempt to find among the admissible vectors $u$ such an articulatory vector $u_k(t)$, that minimizes the criterion $\Omega[u]$ and simulates experimental acoustical data with the same relative accuracy as they were measured. It is well known from the theory of non-linear ill-posed problems [8] that the minimization problem (5) has a unique solution $u_k(t)$, if the functional $\Omega[u]$ and the operator $A$ possess certain mathematical properties. The approximate solution $u_k(t)$ appears to be “sufficiently close” to exact one if the accuracy of the input data is “good enough”, i.e. the value $\delta$ is “small”. Thus, the articulatory vector $u_k(t)$ is a unique and mathematically stable approximate solution to our inverse problem.

Various forms of optimality criteria $\Omega[u]$ were under study in [11]. It was found that solutions of the problem (5) with so-called “instant criteria” differ essentially from solutions for “time averaged” or “integral” criteria. The integration time in the present study equals to phonetic segment duration. By physical motivation, we postulate that the optimality criterion in the problem (5) has a form $\Omega = \Omega_1 + \Omega_2$. Here the functional

$$
\Omega_1[u(t)] = \sum_{k=1}^N c_k \left( u_k(t) - u_k(0) \right)^2
$$

is proportional to the work of elastic forces when the articulators go from their neutral positions to the current ones, $c_k$ is the coefficient of elasticity. The term

$$
\Omega_2[u(t)] = \gamma \frac{\|u(t)-u(t-\Delta t)\|_2}{\|u(t-\Delta t)\|_2}.
$$

$\gamma = \text{const}$.
is proportional to a discrete analogue of the kinetic energy, calculated on a time grid with a step $\Delta t$. The symbol $\| \cdot \|_2$ is Euclidean norm of a vector. The residual

$$\Phi[u] = \|A(u) - z_0\|$$

(9)

is represented for the problem (5) in a form

$$\Phi[u] = b_1 \Phi_1[u] + \Phi_2[u]$$

(10)

with

$$\Phi_1[u] = \sum_{m=1}^{M} \left[ R(p_m, q_m) + \alpha \frac{dR(p_m, q_m)}{dt} \right].$$

(11)

Here the value $R(p_m, q_m)$ is Euclidean distance between measured reference points $p_m$ in the vocal tract and corresponding calculated points $q_m$. The coefficient $b_1=1$, if the solution to inverse problem is sought for acoustical and articulatory input data. If only acoustical data are used then $b_1=0$.

The summand $\Phi_2[u]$ determines the discrepancy of measured and calculated acoustical parameters for vowels, nasals, and fricatives

$$\Phi_2[u] = \beta \max \left| \frac{F_k}{F_k^*} - 1 \right|, \quad k = 1, 2, ..., K,$$

$$\beta = \text{const.}$$

(12)

The values $F_k^*$ have the meaning of measured characteristic frequencies for speech signal spectra, while $F_k$ are their analogues calculated by the use of our model in solving inverse problem. For vowels or nasals, $F_k$ should be interpreted as resonance frequencies. They can be found by solving the spectral problem for the wave equation. For fricatives, $F_k$ represent frequencies of the "center of gravity" in the high frequency region and frequencies of spectra crossing by its mean value. As a rule, $K=3$ for vowels and fricatives, while $K \leq 4$ for nasalized vowels and nasals.

In solving inverse problem, we applied an interactive procedure of segmentation for the speech signal. The procedure detects four basic types of segments i.e., vowel-like, nasal, fricative, closure. We selected the acoustical data $F_k^*$ for the functional $\Phi_2[u]$ in accordance with this segmentation procedure. Formant frequencies were found pitch synchronously.

The problem (5) is solved by use of quadratic approximation of the modified Lagrange function. The corresponding procedure of constrained optimization presented in MATLAB Software Package.

6. Numerical experiments

We solved inverse problem (5) for isolated vowels, diphthongs, syllables, words and sentences containing all sounds of English: vowels, semi-vowels, nasals, stops and fricatives. At first, the inverse problem was considered where the data were presented by acoustical parameters and coordinate tracks of reference points in the vocal tract as well. For such a problem, the mean error of the approximation of experimental formants by calculated ones comes to 3.7% for $F_1$, 3.8% for $F_2$ and 2.6% for $F_3$. Analogous mean error for fricative spectra was about 30.6%, whereas error for characteristic frequencies of fricative spectra amounts to 8.5%. The bigger error for spectra reflects their variability due to turbulent noise. The error of approximation for the coordinates of reference points by calculated ones was about 2.8% for all types of sounds. We estimate all these errors in the vector norm

$$\|z\|_\infty = \max \{ |z_k| : k = 1, ..., M \}.$$  

(13)

This estimation is more strict than the mean-square one. The solutions obtained for the problem of the first kind were collected in a codebook. This enabled us to solve the inverse problem of the second kind, i.e. only with acoustical parameters as input data. The initial guesses to start with the optimization in (5) were appropriately chosen from the codebook.

In the inverse problem of the second kind, the accuracy of formant frequencies was better while for the coordinates of the reference points approximation for vowels it was slightly worse in comparison with the inverse problem of the first kind. The first three formant frequencies were approximated with 1.8%, 1.6% and 1.1% error while the error for the coordinates of reference points increased to 6%. The better, in comparison with the problem of the first kind, accuracy of formant frequencies reconstruction is because of the absence of information on fleshpoints in the criteria of optimality. As a consequence of that, the optimizer aims only to acoustic data, making the solution more accurate for the account of the accuracy of fleshpoint coordinates approximation.

In solving inverse problems for fricatives with acoustical data only, we estimate approximation errors being 13.4% for characteristic frequencies and 3.2% for reference points. For the inverse problems of the first and second kind, the same calculated articulatory parameters differ from each other not more than 2%.

The vocal tract shapes, obtained in these inverse problems, resemble each other as well. It is clear from Fig. 2 where these shapes are represented for the consonant /ʃ/ of the syllable /æʃ/. When solving the inverse problem for nasals and stops, we made use of both acoustical and reference points data.
For nasals, the nasal formant and average spectrum at the nasal closure was used as input data. For stops, formant transitions in the vicinity of closure and the spectrum of burst was used. The approximation error for reference points comes to 3% in the solution of inverse problems for nasals and stops.

Thus, it turns out in our experiments, that the accuracy of found solutions is comparable with the accuracy of experimental input data both for reference points and acoustical parameters. Limited data of [4] does not allow for the verification of vocal tract shape computation. In principle, MRI technique might provide necessary information. However, in the contemporary technology, vocal tract shape registration and speech signal recording are separated in time which creates uncontrollable errors.

7. Speech synthesis

Our procedure for the speech synthesis is based on a convolution of impulse response of the vocal tract and an excitation source. Poles and eigenfunctions of the vocal tract are calculated from the area function. A transfer function of the vocal tract in parallel realization is computed for the poles and eigenfunctions, according to [10]. The Laplace-inverse of the transfer function produces an impulse response of the vocal tract.

The variational method of speech inverse problem solving provides accuracy which is sufficient to consider technical applications like articulatory synthesis, speech compression, recognition, and speaker verification. Satisfactory solution to the inverse problem for all sounds of English can be obtained if

- a codebook is constructed for real speech and articulatory data,
- a mathematical model of speech production is verified by high quality of speech synthesis,
- optimality criteria are physiologically justified,
- speech signal is segmented onto formant bearing segments, nasals, fricatives and stops.

The obtained high quality of re-synthesized speech and its resemblance to original utterance is an evidence of the adequacy for applied mathematical models of speech production and speech inversion technique.

8. Conclusions

The versional method of speech inverse problem solving provides accuracy which is sufficient to consider technical applications like articulatory synthesis, speech compression, recognition, and speaker verification. Satisfactory solution to the inverse problem for all sounds of English can be obtained if

- a codebook is constructed for real speech and articulatory data,
- a mathematical model of speech production is verified by high quality of speech synthesis,
- optimality criteria are physiologically justified,
- speech signal is segmented onto formant bearing segments, nasals, fricatives and stops.

The obtained high quality of re-synthesized speech and its resemblance to original utterance is an evidence of the adequacy for applied mathematical models of speech production and speech inversion technique.

9. References