A posteriori error estimates in voice source recovery

A S Leonov$^1$ and V N Sorokin$^2$

$^1$National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, 115409 Russia
$^2$Institute of Problems of Information Transfer, Russian Academy of Sciences, Bolshoi Karetnyi per., Moscow, 127994 Russia

E-mail: asleonov@mephi.ru vns@iitp.ru

Abstract. The inverse problem of voice source pulse recovery from a segment of a speech signal is under consideration. A special mathematical model is used for the solution that relates these quantities. A variational method of solving inverse problem of voice source recovery for a new parametric class of sources, that is for piecewise-linear sources (PWL-sources), is proposed. Also, a technique for a posteriori numerical error estimation for obtained solutions is presented. A computer study of the adequacy of adopted speech production model with PWL-sources is performed in solving the inverse problems for various types of voice signals, as well as corresponding study of a posteriori error estimates. Numerical experiments for speech signals show satisfactory properties of proposed a posteriori error estimates, which represent the upper bounds of possible errors in solving the inverse problem. The estimate of the most probable error in determining the source-pulse shapes is about 7-8% for the investigated speech material. It is noted that a posteriori error estimates can be used as a criterion of the quality for obtained voice source pulses in application to speaker recognition.

1. Introduction

The voice source pulse excites acoustic oscillations in the vocal tract. Mathematically, it is proportional to the time derivative of the volume velocity of the air flow in the glottis. Its characteristics are unique for each person. The individual features of voice source form give the basis for an approach to automated speaker recognition regardless of context, as well as for larynx pathology diagnostics and other applications. So, the inverse problem of finding parameters of voice source pulses from speech segments is of importance for applications.

Methods of solving this inverse problem are considered for a long time in many mathematical statements (see, e.g., a review in [1]). However, the question of error estimation for obtained pulse form remains poorly understood. The opinion about the accuracy of obtained solution is mainly formed intuitively. Therefore, it is not known quantitatively to what extent the found form of a particular source pulse can be considered reliable and whether it is possible to use this form, for example, in assessing pathology or in speaker recognition.

The reason is that the inverse problem of voice source recovery appears to be ill-posed and non-linear. So, its solution requires the use of special regularization methods (see, e.g., [2, 3]). Obtaining error estimates for solutions of such ill-posed inverse problems is a very complicated problem, which can be solved only by using significant a priori information about the desired solution. The information of this kind is not available in our task. Thus, the existing theory of a priori errors estimates (see, e.g. [3]) is practically not applicable to the inverse problem under
consideration. Therefore, we propose a method for voice source recovery with a quantitative \textit{a posteriori} error estimate for real speech signals, based on the theoretical results presented in [4, 5].

The initial in our study is a special mathematical model of speech generation by parametrical vocal sources, proposed in [6] (Sec.2). As in most known approaches, it is assumed in the model that the resonant frequencies (formants) of the vocal tract are the same at the intervals of open and closed vocal slit. The model takes into account the experimental dependence of the formants damping on the frequency, the radiation of speech signal from the lips etc. In processing, formants are calculated from the speech segment corresponding to the closed vocal slit. Signal segmentation, i.e., estimation of the instants of the glottis opening and closure, is performed according to the technique from [7]. For the adopted model, the inverse problem of determining shapes of the voice source pulse from registered segments of speech is posed in a form of parametrical operator equation.

In Sec.3, we give a variational technique for solving the inverse problem for each pitch period. To this end, we postulate a new parametric voice source, so-called piecewise linear source (PWL-source). Further, we apply presented method to the processing of speech signals. In fact, sources and processes in the vocal tract differ from their description in the mathematical model. In this regard, we demonstrate a computer study of the adequacy of the proposed model (Sec.4). The procedure of error estimate based on the work [4] is described in Sec.5. We apply it to different types of speech signals (Sec.6). In particular, we adduce the results of corresponding calculations for the speech signals from the database CMU ARCTIC [8] (2 speakers, about 22,000 analyzed impulses) and the base of Russian numerals (56 speakers, the total number of analyzed impulses is more than 322,000). It was found that, in general, PWL-sources adequately describe speech processes in the vocal tract. The most possible value of a posteriori error estimate in PWL-source recovery is about 7-8%. Sec.7 contains the conclusions.

A posteriori estimate gives upper bound for possible errors in the solution to the inverse problem for the selected segment of speech. If it turns out that this bound is sufficiently small, then the calculated form of the voice source preserves to a greater extent the individual properties of the speaker. Such form can be used in a procedure of speaker identification. Large upper bounds of the error serve as an indicator that the individual features of the speaker can be significantly distorted in found voice sources. Such source pulses should be excluded from the recognition procedure. In general, knowledge of the upper bound of errors makes it possible to more realistically assess the prospects for not only the speaker identification by voice source, but also other applied problems using such a source.

2. Mathematical model "from voice source to speech signal"

We use a model for the speech production of vowels based on the Webster horn equation with viscous-friction losses and with standard impedance boundary conditions (see [6]). The model relates the speech signal \( f(t) \) on a pitch period \( t \in [0, T_0] \) with the voice source function \( q(t) \) exciting vibrations in the vocal tract. The function \( q(t) \in C[0, T_0] \) is associated with the volume velocity \( v(t) \) of the air flow in the glottis by relation \( q(t) = q_0 v'(t), \quad q_0 = \text{const} > 0 \). We assume that there is no systematic distortion of the signal such as reverberation, and that the function \( f(t) \) recorded by the microphone is related to the pressure \( p(x, t) \) at the end \( x = l \) of the vocal tract by the formula

\[
f(t) = f_0 \left( \lambda p(l, t) + \frac{(1 - \lambda)}{T_0} \int_0^t p(l, \tau) d\tau \right), \quad f_0 = \text{const} > 0; \quad t \in [0, T_0].
\]

This assumption is based on the theoretical reasonings on the radiation of a speech signal from lips to a microphone. The value \( \lambda \in [0, 1] \) is determined by conditions for speech recording, the type of microphone, and other factors. When solving inverse problems, we will consider \( \lambda \) to be
determined. Then the connection between the source function $q(t)$ and the registered function $f(t)$ is postulated in the form [6, 9]:

$$f(t) = \int_0^t K(t - \tau) \left( \lambda q(\tau) + \frac{(1 - \lambda)}{T_0} \int_0^\tau q(\eta) d\eta \right) d\tau + \sum_{n=1}^{N_0} e^{-(\mu + \sigma(f_n))t} \left( A_n \cos (2\pi f_n t) + \frac{B_n}{f_n} \sin (2\pi f_n t) \right) + C_0$$

(1)

with

$$K(t) = K_0 \sum_{n=1}^{N_0} e^{-(\mu + \sigma(f_n))t} \frac{\sin \omega_m t}{\omega_m} \Delta(\omega) = (\omega - \omega_m) \prod_{m=2}^{N_0} \frac{\omega_m - \omega}{(m-1)^2}, \ K_0 = \text{const.}$$

(2)

Here $\{f_m\}_{m=1}^{N_0}$ are resonant frequencies of the vocal tract, $N_0$ is the number of such frequencies taken into account and $\omega_m = 2\pi f_m$. The function $\sigma(f)$ represents the dependence of the damping decrement on the frequency and is taken from practical measurements. Parameters of the model are the values $\mu$, $\lambda$, $A_n$, $B_n$, $C_0$, where $\mu > 0$ is the coefficient of viscous losses, the coefficients $A_n$, $B_n$, $C_0$ are determined by the initial conditions for the pressure in the vocal tract, and also by the numbers $n, \mu$. With known values $\mu$, $\lambda$, $A_n$, $B_n$, $C_0$ and for a given source function $q(t)$, formulas (1) and (2) make it possible to calculate the theoretical analogue of the recorded speech signal. The formula (1) can be written in the operator form

$$f = B(\mu, \lambda, A_n, B_n, C_0) [q(t)],$$

(3)

where the operator $B$, depending on the parameters, describes the mathematical model (1), (2).

We state our inverse problem of source recovery as follows. Given the fundamental tone period $T_0$, vocal slit closure $t_{cl} \in (0, T_0)$, frequencies $\{f_m\}_{m=1}^{N_0}$ and the speech segment $f(t)$, $t \in [0, T_0]$, we wish to find from the operator equation (1) the unknown source function $q(t)$ and associated source function $v(t)$, and the parameters $\mu > 0$, $\lambda \in [0, 1]$ and $A_n, B_n, C_0$.

3. Method for solving the inverse problem

It is common to use special parametric classes of functions $q(t) = q(t, a)$ to solve the inverse problem in question. Such functions reflect the approximate behavior of the source pulse and require physically and physiologically grounded a priori constraints on the parameters $a$. We use in this work a piecewise linear model of the voice source. A typical plot of such a source and its parameters $a = (t_1, t_2, t_3, t_4, t_5, A_1, A_2, A_3)$ are shown in Fig.1. The value $A_4$, presented in Fig.1, depends on other parameters and is determined from the condition $v(t_{cl}) = f_0^{t_{cl}} q(t)dt = 0$ as $A_4 = (A_1 t_2 + A_2(t_3 - t_1) + A_3(t_4 - t_2))/(t_{cl} - t_4)$. There are specific physiological constraints on the parameters $a$. For example, the inequality $0 < t_1 < t_2 < t_3 < t_4 < t_{cl}$ must hold. Without going into other details here, we denote the set of all these constraints as $M_p$. There are similar restrictions on the parameters $p = (\mu, \lambda, A_n, B_n, C_0)$. We shall denote them as $M_p$.

Now we formulate the method for solving our inverse problem in variational form. To this end, we introduce the relative residual of the operator equation (3):

$$\Phi(q(t, a), \mu, \lambda, A_n, B_n, C_0) = \frac{\|B(\mu, \lambda, A_n, B_n, C_0) [q(t, a)] - f(t)\|}{\|f(t)\|}.$$ 

Here we use the norm

$$\|f(t)\|^2 = \int_0^{t_{cl}} \left( f(t) e^{-\mu t} \right)^2 dt.$$ 

3
in the space $L_2[0, T_0]$. This corresponds to our desire to ensure the proximity of the problem data and their computed counterparts for the part of the pitch period where the source operates. The norm and the functional $\Phi$ actually specifies the closeness of the functions

$$U(t) = e^{-\mu t}f(t), \quad U_{\text{calc}}(t) = e^{-\mu t}B(\mu, \lambda, A_n, B_n, C_0)[q(t, a)].$$

Exactly these functions will be presented below as plots of $U(t), U_{\text{calc}}(t)$ (see Fig.2).

We reduce our inverse problem to the solution of an optimization problem, namely we wish to find a function $q = q(t, \bar{a}), \bar{a} \in M_q,$ and parameters $(\bar{\mu}, \bar{\lambda}, \bar{A}_n, \bar{B}_n, \bar{C}_0) \in M_p$ such that

$$\Phi(q(t, \bar{a}), \bar{\mu}, \bar{\lambda}, \bar{A}_n, \bar{B}_n, \bar{C}_0) = \min \{\Phi(q(t, a), \mu, \lambda, A_n, B_n, C_0) : a \in M_q, (\mu, \lambda, A_n, B_n, C_0) \in M_p\}.$$ (4)

In fact, problem (4) gives quasisolutions of the operator equation (3). Such problems can be ill-posed. The method for solving them are studied in detail in [10]. We note that the quantity $\Phi_{\text{min}} = \Phi(q(t, \bar{a}), \bar{\mu}, \bar{\lambda}, \bar{A}_n, \bar{B}_n, \bar{C}_0)$ found after solving the problem (4) shows to what extent the adopted mathematical model (1) is adequate to the data of the inverse problem. Problem (4) is a mathematical programming problem, and we solve it with the help of known stable numerical algorithms and program modules (see, for example, [2, 10]).

4. Model verification

For practical applications of the source recovery problem, it is important how our mathematical model (1) and forms of PWL-sources are adequate to the speech signals and experimental forms of voice source pulses. Fig.2 shows an example of solving our inverse problem for a speaker whose source functions $v_{\exp}(t)$ and their derivatives $q_{\exp}(t)$ can be obtained independently from additional experimental data presented in [11]. This allows us to compare the velocity pulses $v_{\exp}(t)$ and their derivatives $q_{\exp}(t)$ and functions $v_{\text{calc}}(t), q_{\text{calc}}(t)$ obtained by solving the optimization problem (4). All these functions are normalized to a maximum. One can see that the pulses of the volume velocity are recovered satisfactorily. Also, we note the high proximity of the functions $U_{\exp}(t)$ and $U_{\text{calc}}(t)$, which means the adequacy of the model used.

5. Method of a posteriori error estimation

The main difficulty in estimating the accuracy of voice source recovery is determined by numerical instability or ill-conditionality of the problem (4). It is for this reason that attempts...
to obtain theoretical a priori error estimates lead to estimates that can not be practically used. This is expressed in the fact that a priori estimates contain constants that either can not be found numerically, or can be estimated extremely roughly. As a result, attempts to express a priori error estimate as a number lead to overestimation of errors. In this connection, practical error estimates of the solutions to the problem (4) should be performed a posteriori, using, for example, the recently developed theory [4, 5]. Now we describe the technique of obtaining these a posteriori estimates for our problem.

After solving the problem (4), we have at our disposal quantities \( (q(t, \bar{a}), \mu, \bar{\lambda}, \bar{A}_n, \bar{B}_n, \bar{C}_0) \), which are approximations of exact values \( q(t, \tilde{a}), \mu, \tilde{\lambda}, \tilde{A}_n, \tilde{B}_n, \tilde{C}_0 \). Therefore, we can calculate the approximation \( \Phi^{(appr)}_{\min} = \Phi \left( q(t, \tilde{a}), \mu, \tilde{\lambda}, \tilde{A}_n, \tilde{B}_n, \tilde{C}_0 \right) \) for \( \Phi_{\min} \). We set the constant of the estimation algorithm, \( C > 1 \), close to unity (for example, \( C = 1.01 \)) and consider the extremum problem: find the quantities \( a^* \in M_q, (\mu^*, \lambda^*, A_n^*, B_n^*, C_n^*) \in M_p \), such that

\[
\frac{\|q(t, a^*) - q(t, \tilde{a})\|}{\|q(t, \tilde{a})\|} = \max \left\{ \frac{\|q(t, a) - q(t, \tilde{a})\|}{\|q(t, \tilde{a})\|} : a \in M_q, \Phi \left( q(t, a), \mu, \lambda, A_n, B_n, C_n \right) \leq C \Phi^{(appr)}_{\min} \right\}
\]

In this optimization problem, the first constraint means minimization over all admissible sources, and the second restriction corresponds to minimization over all sources and parameters that give residuals close to \( \Phi^{(appr)}_{\min} \). Thus, the value \( \varepsilon_q = \frac{\|q(t, a^*) - q(t, \tilde{a})\|}{\|q(t, \tilde{a})\|} \) represents the greatest relative deviation of the approximate solution to inverse problem (3) from all admissible source functions for which the values of the residual is close to optimal. The degree of this proximity is determined by the number \( C \). We take the value \( \varepsilon_q \) as a posteriori error estimate of found approximate solution \( q(t, \tilde{a}) \).

To calculate \( \varepsilon_q \), we need to solve the last extremum problem using known optimization methods. Similarly, one can find a posteriori error estimate for an volume velocity pulse \( v(t, a) = \int_0^t q(\tau, a) d\tau \):

\[
\varepsilon_v = \max \left\{ \frac{\|v(t, a) - v(t, \tilde{a})\|}{\|v(t, \tilde{a})\|} : a \in M_q, \Phi \left( q(t, a), \mu, \lambda, A_n, B_n, C_n \right) \leq C \Phi^{(appr)}_{\min} \right\}.
\]

In what follows, error estimates \( \varepsilon_v \) are presented, which were obtained numerically with the help of optimization modules of MATLAB software package.

6. Numerical experiments on the error estimation for volume velocity impulses

For model problems with known solutions \( q(t) \) and corresponding synthetic speech signals known with relative error \( \delta \), we can solve the problem (2) and find approximate PWL-solutions \( q(t, a_\delta) \) and corresponding function \( v(t, a_\delta) \). Then the accuracy of \( v(t, a_\delta) \) is given by the formula

\[
\Delta(\delta) = \frac{\|v(t) - v(t, a_\delta)\|_{L_2[0,T_0]}}{\|v(t)\|_{L_2[0,T_0]}}.
\]

Therefore, we can compare such a calculated accuracy with its estimate \( \varepsilon_v(\delta) \) found from (5). The results of this comparison are presented in Fig.3 for different \( \delta \).

Now we describe the results concerning a posteriori error estimates in voice source recovery with actual speech signals from the database CMU ARCTIC [8]. Speech signals of two speakers were analyzed: a man (BDL speaker) and a woman (SLT speaker). The total number of analyzed voice source pulses for each speaker exceeds 22,000. The plots on Fig.4A show the calculated standard distribution density \( p_v \) of the relative errors \( \varepsilon_v \), which were found using formula (5) for these speakers. Standard distribution density \( p_\Phi \) of relative residuals \( \Phi^{(appr)}_{\min} \) for these speakers are shown in Fig.4B.
Further numerical experiments were performed with speech signals from the database for Russian numerals. The speech was recorded under ordinary conditions in rooms with different dimensions and different natural-noise levels. Thus, the characteristics of speech signals were maximally close to the actual operating conditions for the speaker identification systems. Speech signals from 56 speakers were used in the experiments. The total number of analyzed voice source pulses is about 322,000. The corresponding results of estimating the upper bound of the errors, $\varepsilon_v$, and residuals, $\Phi_{\text{min}}^{(\text{appr})}$, for PWL-sources are shown in Fig.5A and Fig.5B correspondingly.

7. Conclusions
1) The rather simple model (1) of speech production and the piecewise linear model of the source make it possible to reproduce the studied speech material with good accuracy. The available deviations of the measured speech signals from those calculated by the model (1) are caused both by the simplification of the description of speech processes, and by experimental errors of various kinds that arise during registration of speech and its preliminary processing (segmentation). These errors, especially those that have a systematic character, can lead to the fact that the model signal calculated on the considered pitch period by means of found PWL-source is significantly different from registered. For this period, the model (1) is not adequate to the speech data used. The above graphs of the residual distribution (Fig.4B,5B) show the proportion of such speech segments in the used speech information.
2) The most probable value of the upper error limit $\varepsilon_v$ for found PWL voice source is estimated at approximately 7-8% for the speech material considered. The graphs in Fig.4A,5A allow us to analyze the frequency of the appearance of pulses with too large and too small error estimates, and make quantitative conclusions about the accuracy of the approximation of actual voice sources. The quantity 7-8% for source errors seems very good in comparison with rather large experimental data errors produced by actual reverberation and random errors of the order of 5-10%.

3) The values $\varepsilon_v$ can be used to estimate the quality of voice source impulses obtained in solving the inverse problem. The use of these quantities opens up new possibilities in the task of speaker recognition. Analyzing the obtained impulses, one can discard those of them that have large error estimates $\varepsilon_v$. Thus, it is possible to increase the reliability of recognition and accelerate the recognition procedure.

Acknowledgments
This work was supported by the Russian Foundation for Basic Research (project no.17-01-00159a) and the Program of Competitiveness Increase of the National Research Nuclear University MEPhI (Moscow Engineering Physics Institute); contract no. 02.a03.21.0005, 27.08.2013.

References
[1] Drugman T, Bozkurt B and Dutoit T 2012 Comput. Speech Lang. 26 20
[2] Tikhonov A N, Leonov A S and Yagola A G 1998 Non-Linear Ill-Posed Problems vol 1,2 (London: Chapmen & Hall)
[3] Engl H W, Hanke M and Neubauer A 1996 Regularization of Inverse Problems (Dordrecht: Kluwer)
[4] Leonov A S 2012 J. Inv. Ill-Posed Probl. 20 637
[5] Leonov A S 2012 J. Inv. Ill-Posed Probl. 22 713
[6] Leonov A S and Sorokin V N 2012 Doklady Mathematics 85 432
[7] Sorokin V N 2016 Acoust. physics 62 244
[8] CMU ARCTIC speech synthesis databases http://festvox.org/cmu arctic
[9] Leonov A S and Sorokin V N 2017 Acoust. physics 63 570
[10] Ivanov V K, Vasin V V and Tanana V P 2002 Theory of Linear Ill-posed Problems and Its Applications (Utrecht: VSP)
[11] Childers D G, Paige A and Moore A 1976 Archives of Otolaryngology 102 407