# A measure of differences in speech signals by the voice timbre

V. V. Savchenko

Published online: 11 March 2024

**Abstract**

This research relates to the field of speech technologies, where the key issue is the optimization of speech signal processing under conditions of a prior uncertainty of its fine structure. The problem of automatic (objective) analysis of the speaker's voice timbre using a speech signal of finite duration is considered. It is proposed to use a universal information-theoretic approach to solve it. Based on the Kullback-Leibler divergence, an expression was obtained to describe the asymptotically optimal decision statistic for differentiating speech signals by the voice timbre. The author highlights a serious obstacle during practical implementation of such statistics, namely: synchronization of the sequence of observations with the pitch of speech signals. To overcome the described obstacle, an objective measure of timbre-based differences in speech signals is proposed in terms of the acoustic theory of speech production and its "acoustic tube" type model of the speaker's vocal tract. The possibilities of practical implementation of a new measure based on an adaptive recursive filter are considered. A full-scale experiment was set up and carried out. The experimental results confirmed two main properties of the proposed measure: high sensitivity to differences in speech signals in terms of voice timbre and invariance with respect to the fundamental pitch frequency. The obtained results can be used when designing and studying digital speech processing systems tuned to the speaker's voice, for example, digital voice communication systems, biometric and biomedical systems, etc.

**Keywords** Acoustic measurements · Speech acoustics · Speech signal · (fundamental) pitch · Vocal tract · All-pole model · Berg's method

## Introduction

Voice timbre is among the primary acoustic characteristics of a speaker's vocal tract. As such, it has been capturing the attention of researchers and specialists across a wide range of fields for many years [1, 2]. Consequently, the voice timbre analysis is a classic problem in the field of acoustic measurements of speech signals [3–5], while the comparative analysis of speech signals based on the voice timbre is an important aspect of such problem. The latter is addressed when designing and studying modern automatic speech processing systems intended for a wide spectrum of purposes [5–8].

Translated from Izmeritel'naya Tekhnika, No. 10, pp. 63–69, October, 2023. Russian DOI: https://doi.org/10.32446/0368-1025it.2023-10-63-69.

Original article submitted September 18, 2023; approved after reviewing October 18, 2023; accepted for publication October 18, 2023.

https://doi.org/10.1007/s11018-024-02294-1

Despite years of research of the acoustic characteristics of the speaker's vocal tract, studies performed in this field show a clear tendency towards the development and expansion of this topic [9–11], because in the author's view, a number of unresolved theoretical problems still remain to date. One of the most important problems has to do with small observation samples [12]. In the studied case, the sample size is strictly limited by the duration of two to three periods of the fundamental pitch ($T_0$ = 5–10 ms), when the vocalized speech signal can be considered steady-state [13].

Since the voice timbre is defined by the fine structure of a speech signal within one such period, the sequence of observations $x(i)$, $i = 1, 2, \ldots$, should be synchronized with the vibrations of the speaker's vocal cords [14, 15]. However, under conditions of a prior uncertainty and small sample sizes, such synchronization presents a practically unresolvable problem. Therefore, the topic of this study is highly relevant.

The goal of this work is to develop an objective measure of differences in speech signals by the voice timbre, which does not require synchronization of observations with the fundamental pitch period. To achieve this goal, a universal information-theoretic approach and methodology of the acoustic theory of speech production were used.

This article was written to further advance the results of the previous studies performed by the author in collaboration with the personnel of the Laboratory of Algorithms and Technologies of Network Structure Analysis at the National Research University "Higher School of Economics" [16, 17].

## Problem statement

Let $x(t)$ represent a speech signal in discrete time $t = iT$, $i = 1, 2, \ldots, N$ with a period $T$ of sample $x(i) = x(iT)$ over the observation interval of a vocalized (vowel) speech sound having a duration of $T_{ob} = MT_0$, where $M \geq 1$. Assuming that the first sample $x(1)$ is co-located with the beginning of the observation interval, the sample size will be $N = nM$, where $n = [T_0 T^{-1}] = [F F_0^{-1}]$; $F_0 = T_0^{-1}$ is the fundamental pitch frequency; $F = T^{-1}$ is the speech signal sampling frequency; [·] denotes the integer part of a rational number. In such cases, we talk about synchronizing of observations (analysis [15]) with the fundamental pitch of the speech signal. We will now divide the $N$-sequence of samples $\{x(i)\}$ into $M$ partial sequences $\{x_m(i), m \leq M\}$ each having a dimensionality of $n = NM^{-1} \gg 1$. For example, at $F$ = 8 kHz and $F_0$ = 100 Hz (standard value of the fundamental pitch frequency for male voices [16]), there are $n$ = 8000/100 = 80 samples of $x_m(i) = x(mT_0 + iT)$ for $i \leq n$ within one (each individual) period of the fundamental pitch. According to the acoustic theory of speech production [18–20], these samples are the ones that determine the speaker's voice timbre. Therefore, formally, the voice timbre can be described using an intraperiodic (within the period of the fundamental pitch) function of autocorrelation of the sequence $\{x_m(i)\}$ of speech signal observations over a finite duration interval $T_0 = nT$ [10]. The statistical equivalent of this function is the empirical (sample) autocorrelation ($p$×$p$)-matrix [7]:

$$S_x \triangleq M^{-1} \sum_{m=1}^{M} \boldsymbol{x}_m \boldsymbol{x}_m^{\top}, \tag{1}$$

defined over a set $\{\mathbf{x}_m\}$ of $p$-dimensional (vector) observations $\boldsymbol{x}_m = \mathrm{col}_p\{x_m(i)\}$, synchronous with the fundamental pitch of the speech signal $x(t)$. Here, $\mathrm{col}_p\{\cdot\}$ is a column-vector having a dimensionality of $p \leq n$; and $\triangleq$ is equality by definition. Similarly, for any other speech signal $y(t)$, there is an empirical autocorrelation matrix:

$$S_y \triangleq M^{-1} \sum_{m=1}^{M} \boldsymbol{y}_m \boldsymbol{y}_m^{\top}, \tag{2}$$

where $\boldsymbol{y}_m = \mathrm{col}_n\{y_m(i)\}$ is the $p$-column-vector of synchronous observations $y_m(i)$ in discrete time $i = 1, 2, \ldots, n$.

Following the information theory of speech perception [6, 21], we will use matrices (1) and (2) as a basis of the information-theoretic approach to the automatic differentiation of speech signals $x(t)$ and $y(t)$ by the voice timbre.

## Kullback-Leibler divergence

We will now determine the Kullback-Leibler divergence [22] for two Gaussian laws of distribution of probabilities specified by their autocorrelation matrices $S_x$ and $S_y$ in a $p$-dimensional sample space[1]:

$$\rho_{x,y} \triangleq 0.5M\left[\mathrm{tr}\left(S_x S_y^{-1}\right) + \mathrm{tr}\left(S_y S_x^{-1}\right) - 2p\right] \geq 0 \tag{3}$$

where tr(·) denotes the trace (spur) of a square ($p$×$p$) matrix.

As shown in Ref. [21], Eq. 3 defines the asymptotically optimal (as $M \to \infty$) decision statistic in the problem of differentiating two speech signals $x(t)$ and $y(t)$ based on finite observation samples. However, the practical use of Eq. 3 as a measure of differences between such speech signals is greatly limited by the requirement for synchronization of their vector observations $\{\mathbf{x}_m\}$ and $\{\mathbf{y}_m\}$ with the fundamental pitch of the corresponding signal.

To circumvent the aforementioned issue, the problem at hand will be reduced to signal processing in the frequency domain, where there is fundamentally no need for synchronization of the observation sequence. In Ref. [23], the frequency equivalent of the information divergence (3) is justified using the formula for a scale-invariant modification of the COSH-distance[2]:

$$\rho_{x,y} = \sqrt{\left[F^{-1}\int_{-0.5F}^{0.5F} \hat{G}_x(f)\,\hat{G}_y^{-1}\;(f)\,df\right]\left[F^{-1}\int_{-0.5F}^{0.5F} \hat{G}_y(f)\,\hat{G}_x^{-1}(f)\,df\right]} - 1 \geq 0. \tag{4}$$

Bartlett's periodograms [24, 25] from Eq. 4:

$$\begin{cases} \hat{G}_x(f) \triangleq M^{-1}\sum_{m=1}^{M}(nT)^{-1}\left|T\sum_{i=1}^{n} x_m(i)\exp(-\mathrm{j}2\pi i f T)\right|^2; \\ \hat{G}_y(f) \triangleq M^{-1}\sum_{m=1}^{M}(nT)^{-1}\left|T\sum_{i=1}^{n} y_m(i)\exp(-\mathrm{j}2\pi i f T)\right|^2. \end{cases} \tag{5}$$

are used as statistical estimates of the intraperiodic spectra of power of speech signals $x(t)$ and $y(t)$ based on the discrete observation samples.

The scale invariance property of measure (4) can be easily confirmed by bringing the arbitrary gain coefficients for the signals $\{x_m(i)\}$ and $\{y_m(i)\}$ under the absolute value sign on the right-hand side of Eq. 5. The result will remain unchanged regardless [16]. However, this does not solve the main problem of automatic speech processing when analyzing the voice timbre, which is the synchronization of the observation sequence with the fundamental pitch of speech signals.

[1] The assumption of a Gaussian probability distribution does not limit the generality of the conclusions of this study, as this law is characterized by the maximum entropy for a given average power of the speech signal.

[2] COSH—cosine hyperbolic function.

## Method of asynchronous analysis of voice timbre

Considering that under the general assumptions [1], partial oscillations

$$x_m(i) = a_x h_{x,m}(i); \quad y_m(i) = a_y h_{y,m}(i), \ i = 1, 2, \dots, \tag{6}$$

(where $a_x$, $a_y$ = const) are determined by the dynamics of pulse response characteristics $h_{x,m}(i)$ and $h_{y,m}(i)$ of the linear (filter-based) "acoustic tube" type model of the vocal tract, which is inherently stable in terms of digital filtering [26], and therefore exhibit an attenuation behavior. We will rewrite Eq. 4 in an asymptotically equivalent form:

$$\begin{aligned}\rho_{x,y} = &\sqrt{F^{-1}\int_{-0.5F}^{0.5F}\frac{M^{-1}\sum_{m=1}^{M}\left|T\sum_{i=1}^{\infty}h_{x,m}(i)\exp(-\mathrm{j}2\pi ifT)\right|^2}{M^{-1}\sum_{m=1}^{M}\left|T\sum_{i=1}^{\infty}h_{y,m}(i)\exp(-\mathrm{j}2\pi ifT)\right|^2}df\times}\\ &\times\sqrt{F^{-1}\int_{-0.5F}^{0.5F}\frac{M^{-1}\sum_{m=1}^{M}\left|T\sum_{i=1}^{\infty}h_{y,m}(i)\exp(-\mathrm{j}2\pi ifT)\right|^2}{M^{-1}\sum_{m=1}^{M}\left|T\sum_{i=1}^{\infty}h_{x,m}(i)\exp(-\mathrm{j}2\pi ifT)\right|^2}df} - 1.\end{aligned} \tag{7}$$

The expressions under the absolute value sign from Eq. 7, through the Fourier transform of the corresponding pulse characteristics (6) in discrete time $i$, determine two complex transfer coefficients:

$$K_{x,m}(\mathrm{j}f) = T\sum_{i=1}^{\infty}h_{x,m}(i)\exp(-\mathrm{j}2\pi ifT);$$

$$K_{y,m}(\mathrm{j}f) = T\sum_{i=1}^{\infty}h_{y,m}(i)\exp(-\mathrm{j}2\pi ifT)$$

From Eq. 7 the following expression can be obtained:

$$\begin{aligned}\rho_{x,y} = &\sqrt{F^{-1}\int_{-0.5F}^{0.5F}\frac{M^{-1}\sum_{m=1}^{M}|K_{x,m}(\mathrm{j}f)|^2}{M^{-1}\sum_{m=1}^{M}\left|K_{y,m}(\mathrm{j}f)\right|^2}d\ f\times}\\ &\times\sqrt{F^{-1}\int_{-0.5F}^{0.5F}\frac{M^{-1}\sum_{m=1}^{M}\left|K_{y,m}(\mathrm{j}f)\right|^2}{M^{-1}\sum_{m=1}^{M}|K_{x,m}(\mathrm{j}f)|^2}df} - 1 =\\ &= F^{-1}\sqrt{\int_{-0.5F}^{0.5F}\frac{K_x^2(f)}{K_y^2(f)}df\int_{-0.5F}^{0.5F}\frac{K_y^2(f)}{K_x^2(f)}df} - 1.\end{aligned} \tag{8}$$

Thus, the problem comes down to determining the average statistical values of the squares of the amplitude-frequency characteristics (AFC) of the speaker's vocal tract:

$$K_x^2(f) \triangleq M^{-1}\sum_{m=1}^{M}|K_{x,m}(\mathrm{j}f)|^2; \ K_y^2(f) \triangleq M^{-1}\sum_{m=1}^{M}\left|K_{y,m}(\mathrm{j}f)\right|^2.$$

This is a typical problem of statistical analysis and speech modeling [6, 27]. A number of various theoretical approaches have been developed for solving this problem [18, 19], with the most relevant ones including the methods of parametric spectral analysis [24, 25], and specifically, the Berg's method[3].

## Example of practical implementation

According to the universal all-pole model of the speaker's vocal tract within short (10–20 ms) intervals of vocalized verbal speech, the desired amplitude-frequency characteristics can be determined using the formula for calculating the absolute value of the complex transfer coefficient of a recursive filter of the $p^{\text{th}}$ order [23]:

$$K_x(f) = b_x \left| 1 - \sum_{k=1}^{p} a_{x,p}(k) \exp(-\mathrm{j}2\pi k f T) \right|^{-1}; \tag{9}$$

$$K_y(f) = b_y \left| 1 - \sum_{k=1}^{p} a_{y,p}(k) \exp(-\mathrm{j}2\pi i f\ T) \right|^{-1}, \tag{10}$$

where $|f| \le 0.5F$; $b_x$ and $b_y$ are the gain factors of signals $x(t)$ and $y(t)$, respectively, in the speaker's vocal tract; $a_{x,p}(k)$ and $a_{y,p}(k)$ are the autoregression coefficients of the finite ($p^{\text{th}}$) order ($k$—coefficient number).

Considering Eqs. 9 and 10, we can rewrite Eq. 8 as follows:

$$\rho_{x,y} = \sqrt{F^{-1} \int_{-0.5F}^{0.5F} \left| \frac{1 - \sum_{k=1}^{p} a_{y,p}(k) \exp(-\mathrm{j}2\pi k f T)}{1 - \sum_{k=1}^{p} a_{x,p}(k) \exp(-\mathrm{j}2\pi k f T)} \right|^{2} df} \times$$
$$\times \sqrt{F^{-1} \int_{-0.5F}^{0.5F} \left| \frac{1 - \sum_{k=1}^{p} a_{y,p}(k) \exp(-\mathrm{j}2\pi k f T)}{1 - \sum_{k=1}^{p} a_{x,p}(k) \exp(-\mathrm{j}2\pi k f T)} \right|^{-2} df} - 1. \tag{11}$$

Written under the integral sign in Eq. 11 are the direct and inverse relationships of the squares of two normalized amplitude-frequency characteristics [9] and [10] (assuming that $b_x = b_y = 1$). In this case, gain coefficients $b_x$ and $b_y$ do not play a role. Autoregression coefficients $a_{x,p}(k)$ and $a_{y,p}(k)$ are adapted to the speech signals $x(t)$ and $y(t)$ for all $k \le p$ according to the samples of corresponding observations $\{x(i)\}$ and $\{y(i)\}$ obtained by using one of the known methods. For instance, this could be the Berg's method, which is based on the Levinson recursion [24]:

$$\begin{aligned}
&\forall q = \overline{1, p}\colon a_{x,q}(i) = a_{x,q-1}(i) + c_q a_{x,q-1}(q-i), \quad i = 1, 2, \ldots, q;\\
&c_q = -\frac{2\sum_{n=q}^{N-1} \eta_{q-1}(n)\, \nu_{q-1}(n-1)}{\sum_{n=q}^{N-1} \left[\eta_{q-1}^{2}(n) + \nu_{q-1}^{2}(n-1)\right]};\\
&\eta_q(n) = \eta_{q-1}(n) + c_q \nu_{q-1}(n-1);\\
&\sum \nu_q(n) = \nu_{q-1}(n-1) + c_q \eta_{q-1}(n)
\end{aligned} \tag{12}$$

[3] Researchers often prefer Berg's method over other parametric spectral analysis methods due to its well-known advantages in terms of computational speed and, most importantly, stability of the spectral estimates of the autoregressive type that are formed on its basis.

with the recursion initialization by a system of equalities $\nu_0(n) = \eta_0(n) = x(n)\backslash y(n)$, $n = 1, 2, \ldots, N$ (\—symbol of the choice function OR). The final values of recursion (12) (at $q=p$), taken with the opposite sign, determine two $p$-vectors of corresponding coefficients $\{a_{x,p}(k)\}$ and $\{a_{y,p}(k)\}$ on the right-hand side of Eqs. 9 and 10.

Thus, Eq. 11 together with recursion (12) defines a scale-invariant measure of differences in speech signals by the voice timbre of one or two different speakers. It does not require synchronization of observations with the fundamental pitch of speech signals. The potential of the proposed measure can be illustrated by the results of the experiment described below, in which the author's software Phoneme Training was used[4].

## Experimental procedure and results

The experimental program consisted of two stages.

*First stage.* During the first stage, the sensitivity of the new measure (11) to differences in the fine structure of speech signals was studied with the observations being asynchronous relative to the fundamental pitch. The study was focused on the long (approximately 1.5 to 2 s) signals in the form of the vowel phonemes of the reference speaker—the author of this article. Using the Phoneme Training software, each such signal was transformed into a sequence of homogeneous (monophonic) frames $x(t)\backslash y(t)$ of relatively short duration: $T_{ob} =$ 16 ms. In this case, it was assumed that all such frames are characterized by the same voice timbre. On the contrary, frames of different phonemes fundamentally differ from each other in terms of voice timbre.

The graphs shown in Fig. 1 represent timing diagrams of the phoneme "a" signal in case of synchronous (*a*) and asynchronous (*b*) discrete observations relative to the fundamental pitch. The sampling frequency (8 kHz) of the signal in both cases was consistent with the standard telephone bandwidth (4 kHz). In the first case, the signal covers two full periods of the fundamental pitch, while in the second case, it covers only one. As a result, the fine structures of the signals are very different, while the voice timbre in both cases is practically the same. This fact poses no contradictions, since a different fine structure of speech signals is considered. When analyzing the voice timbre, only the intraperiodic fine structure is considered, which severely limits its analysis by using classical periodogram estimates (5) in case of the asynchronous observations.

In accordance with the experimental procedure, measures (11) were calculated for different pairs of speech signals $x(t)$ and $y(t)$ within the set of prepared voice samples. For each such pair, two corresponding vectors $\boldsymbol{a}_x = \{a_{x,p}(k)\}$, and $\boldsymbol{a}_y = \{a_{y,p}(k)\}$ of autoregression coefficients of the order[5] of $p = 10$ were first calculated using algorithm (12). These vectors were then used to calculate the measure of differences $\rho_{x,y}$ according to Eq. 11. For example, for a pair of homonymous signals (see Fig. 1), the following two vectors were obtained: $\boldsymbol{a}_x =$ (1.364686; –1.08823; 0.532204; –0.80853; 0.906187; –0.43502; 0.107709; –0.17596; 0.40483; –0.12711) $\triangleq \boldsymbol{a}_x^*$; $\boldsymbol{a}_y =$ (1.368958; –1.01194; 0.457037; –0.84364; 1.049269; –0.55294; 0.145535; –0.27892; 0.529057; –0.18108), based on which measure $\rho_{x,y} = 0.009 \ll 1$ was calculated. Similar results were obtained for all other experimental pairs $\{x(t), y(t)\}$, composed of monophonemic voice samples from the reference speaker, within a range of $\rho_{x,y} =$ 0.005–0.025.

As can be seen from Table 1, which shows the average values of the measure of differences ($\rho_{x,y}$) for a set of 100 realizations of each separate variant of the experimental pair $\{x(t), y(t)\}$, the situation is sharply different for pairs of heterophonemic samples. The grayed out elements of the main diagonal of the Table correspond to the monophonemic variants of the experimental pair. All their values are considerably less than one, which indicates a high degree of similarity of the monophonemic speech signals in terms of the speaker's voice timbre. On

[4] The Phoneme Training phonetic analysis and speech training information system: [website]. URL: https://sites.google.com/site/frompldcreators/produkty-1/phonemetraining (access date: May 18, 2023).

[5] This order is intended for autoregressive simulation of 4–5 AFC resonances of a typical vocal tract when pronouncing vowels in the frequency bandwidth of 0 to 4 kHz.

**Table 1** Average values of measure $\rho_{x,y}$

| Phoneme | a | i | o | u | y | e |
|---|---|---|---|---|---|---|
| a | *0.012* | 11.807 | 8.568 | 72.994 | 6.876 | 2.202 |
| i | 11.807 | *0.021* | 17.825 | 33.372 | 2.567 | 5.314 |
| o | 8.568 | 17.825 | *0.010* | 5.898 | 14.300 | 8.858 |
| u | 72.994 | 33.372 | 5.898 | *0.023* | 22.808 | 41.118 |
| y | 6.876 | 2.567 | 14.300 | 22.808 | *0.014* | 2.169 |
| e | 2.202 | 5.314 | 8.858 | 41.118 | 2.169 | *0.011* |

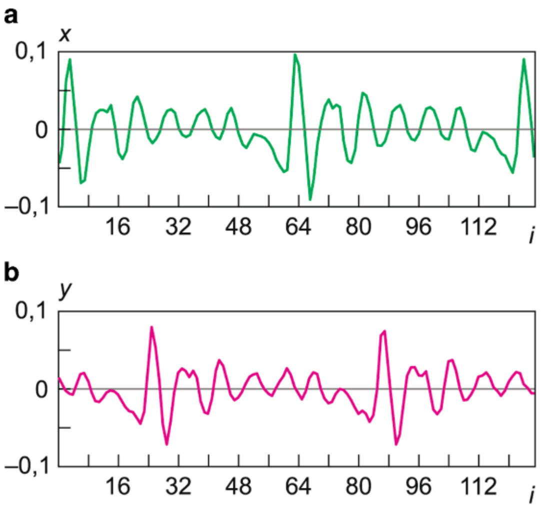


**Fig. 1** Timing diagrams of the phoneme “a” signal in two variants of discrete observations: synchronous (**a**) and asynchronous (**b**) with the fundamental pitch

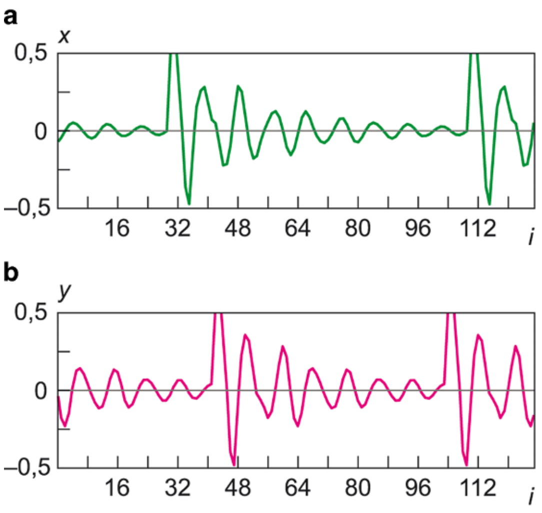


**Fig. 2** Timing diagrams of synthesized phoneme “a” signals with a pitch frequency $F_0$ = 100 Hz (**a**) and $F_0$ = 130 Hz (**b**)

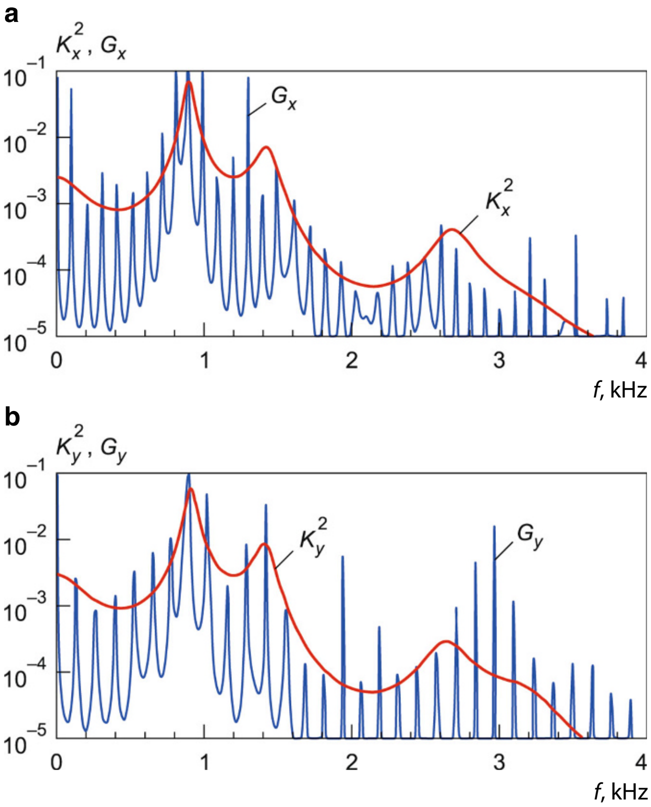
the contrary, all other elements of the Table consistently exceed one, indicating significant differences between heterophonemic signals. Therefore, a conclusion can be made about high sensitivity of measure (11) towards differences in speech signals in terms of the voice timbre.

*Second stage.* A pair of phoneme “a” signals were studied that were synthesized using a recursive filter scheme [23]:

$$\begin{cases} x\,(i) = -\sum_{k=1}^{p} a_{z,p}\,(k)\,x\,(i-k) + \eta_x\,(i)\,; \\ y\,(i) = -\sum_{k=1}^{p} a_{z,p}\,(k)\,y\,(i-k) + \eta_y\,(i)\,; \quad i = 0, 1, \ldots \end{cases} \tag{13}$$

The signals are characterized with pulse excitation $\{\eta_x(i)\}$ and $\{\eta_y(i)\}$ of different fundamental pitch frequencies: $F_0$ = 100; 130 Hz (130 Hz ≈ 1/7.7 ms). Here, the vector of autoregression coefficients $a_{z,p}(k) = a_{x,p}(k)$ was determined by the same vector $\boldsymbol{a}_x^*$, obtained during the first stage of the experiment in both cases. To maintain the small sample conditions, the duration of the synthesized signal samples in discrete time ($i$) was established as $N = 128$. The idea of the second stage was to ensure that the signals in each experimental pair $\{x(i), y(i)\}$ were similar in terms of the virtual speaker’s voice timbre, while being significantly different in terms of the fundamental pitch frequency ($F_0$) [27, 28]. These signals along with algorithms (11) and (12) were then used to calculate the vectors of autoregression coefficients: $\boldsymbol{a}_x$ = (1.366934; –1.11707; 0.595185; –0.85859; 0.900872; –0.40507; 0.008816; –0.05351; 0.328108; –0.0927) and $\boldsymbol{a}_y$ = (1.42998; –1.17923; 0.64873; –0.93677;

**Fig. 3** Graphs of the square of the normalized AFC $K^2_{x,y}$ (9) of the vocal tract linear model (13) at a pitch frequency $F_0 = 100$ Hz (**a**) and $F_0 = 130$ Hz (**b**) in comparison with autoregressive estimates of the discrete speech signal power spectrum $G_{x,y}$ (at high autoregression order)



1.072479; –0.59299; 0.211108; –0.26045; 0.482161; –0.1689) and the measure of their differences: $\rho_{x,y} = 0.0219 \ll 1$ based on the voice timbre. The obtained results are illustrated in Figs. 2 and 3.

The graphs shown in Fig. 2 represent timing diagrams of two synthesized phoneme "a" signals with different fundamental pitch frequencies ($F_0$), but practically identical intraperiodic fine structure. The graphs shown in Fig. 3 illustrate squared normalized AFC (order $p = 10$) of the linear model (13) of the vocal tract with pulse excitation for the fundamental pitch frequencies $F_0 = 100$ and 130 Hz. These characteristics are compared with the corresponding estimates $G_x(f)$ and $G_y(f)$ of the discrete speech signal power spectrum obtained using the Berg's method at a high value of autoregression order $p^* = 90$. (This order was established based on the requirement of $p \geq F/F_0 = 8000/100 = 80$ for the fine structure of the speech signal in the frequency domain [6]). In both cases, the AFC shape repeats the spectral envelope of the synthesized signals, which is practically independent of the fundamental pitch frequency. Therefore, it can be concluded that measure (11) is invariant with respect to the value of $F_0$, which is a key requirement for an objective measure of differences in speech signals when analyzing voice timbre [29, 30].

## Results and discussion

The theoretical justification of the measure of differences (11) is based on the principle of superposition of oscillations in linear systems [26]. According to this principle, a speech signal $\{x(i)\}$ with dominating vowel sounds is defined by the convolution of a periodic sequence of fundamental pitch pulses $\{\eta_x(i)\}$ with the pulse characteristic of the vocal tract $\{h_x(i)\}$. The square of the AFC (9) has a form of an envelope of the discrete speech signal power spectrum. Therefore, the acoustic theory of speech production considers the spectral envelope as the most comprehensive characteristic of the voice timbre in the frequency domain [18, 25].

The problem is that the concept of “spectral envelope” is not strictly defined in the theory. As a result, researchers still lack clarity on the issue of optimal estimation of the spectral envelope based on the speech signal [28–30]. Therefore, in the conducted study, this concept was not used for the synthesis of the measure of differences (11), but exclusively as an illustration of the synthesis results.

## Conclusion

The developed objective measure of differences in speech signals by voice timbre makes it possible to automatically assess the specifics of the fine structure of these signals under conditions of a prior uncertainty and small observation sample sizes. During practical implementation of the new measure based on a finite-order recursive filter with adaptive tuning of its parameters using the Berg’s method, it was established that there is no need to synchronize observations with the fundamental pitch of speech signals. The results of the conducted full-scale experiment confirmed the following two main properties of the proposed measure: high sensitivity to differences in speech signals by voice timbre and significant invariance to the fundamental pitch frequency.

The obtained results are intended for use when designing and studying digital speech processing systems tuned to the speaker’s voice, where the individual characteristics of the vocal tract are of primary importance [31–33]. Examples include digital voice communication systems, biometric and biomedical systems, etc. [4–9].

**Conflict of interest** The author declares no conflict of interest.

## References

1. Zhao, R., Erleke, E., Wang, L., Huang, J., Chen, Z.: The effects of timbre on voice interaction. In: Rau, P.-L.P. (ed.) Cross-Cultural Design: HCII 2023, Lecture Notes in Computer Science, vol. 14023. Springer, Cham (2023) https://doi.org/10.1007/978-3-031-35939-2_12
2. Ando, Y.: Temporal and spatial features of speech signals. In: Signal processing in auditory neuroscience, pp. 81–101. Academic Press, (2019) https://doi.org/10.1016/B978-0-12-815938-5.00009-1
3. Ternström, S.: Appl. Sci. **13**(6), 3514 (2023). https://doi.org/10.3390/app13063514
4. Song, W., Yue, Y., Zhang, Y., et al.: Multi-speaker multistyle speech synthesis with timbre and style disentanglement. In: Zhenhua, L., Jianqing, G., Kai, Y., Jia, J. (eds.) Man-machine speech communication: NCMMSC 2022, communications in computer and information science. Springer, Singapore (2022) https://doi.org/10.1007/978-981-99-2401-1_12
5. Jialu, L., Hasegawa-Johnson, M., McElwain, N.L.: Speech. Commun. **133**, 41–61 (2021). https://doi.org/10.1016/j.specom.2021.07.010
6. Savchenko, V.V.: Radioelectron. Commun. Syst. **64**(11), 592–603 (2021). https://doi.org/10.3103/S0735272721110030
7. Savchenko, A.V., Savchenko, V.V.: Meas. Tech. **64**(4), 928–935 (2022). https://doi.org/10.1007/s11018-022-02025-4
8. Wei, Y., Gan, L., Huang, X.: Front. Psychol. **13**, 869475 (2022). https://doi.org/10.3389/fpsyg.2022.869475
9. Xue, J., Zhou, H., Song, H., Wu, B., Shi, L.: Speech. Commun. **147**, 41–50 (2023). https://doi.org/10.1016/j.specom.2023.01.001
10. Li, J., Zhang, L., Qiu, Z.: 5th International Conference on Intelligent Control, Measurement and Signal Processing (ICMSP). Chengdu., pp. 833–837 (2023). https://doi.org/10.1109/ICMSP58539.2023.10171030
11. Igras-Cybulska, M., Hekiert, D., Cybulski, A., et al.: Work-in-Progress. In: 2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW) Shanghai. pp. 355–359. (2023) https://doi.org/10.1109/VRW58643.2023.00079
12. Cui, S., Li, E., Kang, X.: 2020 IEEE International Conference on Multimedia and Expo (ICME). London., pp. 1–6 (2020). https://doi.org/10.1109/ICME46284.2020.9102765
13. Gupta, S., Fahad, M.S., Deepak, A.: Multimed Tools Appl **79**, 23347–23365 (2020). https://doi.org/10.1007/s11042-020-09068-1
14. Dai, B., Zahorian, S.: J. Acoust. Soc. Am. **104**, 1805 (1998). https://doi.org/10.1121/1.423591
15. Zakhar’ev, V.A., Petrovskii, A.A.: Metody parametrizatsii rechevogo signala na osnove analiza, sinkhronizirovannogo s chastotoi osnovnogo tona v sistemakh konversii golosa. In: Proceedings of the 11th International Scientific and Technical Conference “Nauka – obrazovaniyu, proizvodstvu, ekonomike, vol. 1, pp. 203–204. BNTU, Minsk (2013). in Russian
16. Savchenko, V.V., Savchenko, L.V.: J. Commun. Technol. Electron. **68**(7), 757–764 (2023). https://doi.org/10.1134/S1064226923060128

17. Savchenko, A.V., Savchenko, V.V.: Radioelectron. Commun. Syst. **64**(6), 300–309 (2021). https://doi.org/10.3103/S0735272721060030
18. Gibson, J.: Information **10**(5), 179–189 (2019). https://doi.org/10.3390/info10050179
19. Herbst, Ch T., Elemans, C.P.H., Tokuda, I.T., Chatziioannou, V., Švec, J.G.: J. Voice (2023). https://doi.org/10.1016/j.jvoice.2022.10.004
20. Sadok, S., Leglaive, S., Girin, L., Alameda-Pineda, X., Séguier, R.: Speech. Commun. **148**, 53–65 (2023). https://doi.org/10.1016/j.specom.2023.02.005
21. Savchenko, V.V.: J. Commun. Technol. Electron. **64**(6), 590–596 (2019). https://doi.org/10.1134/S0033849419060093
22. Kullback, S.: Information theory and statistics. Dover, New York (1997)
23. Savchenko, V.V.: Meas. Tech. **66**(6), 430–438 (2023). https://doi.org/10.1007/s11018-023-02244-3
24. Marple Jr., S.L.: Digital spectral analysis, 2nd edn. Dover, New York (2019)
25. Savchenko, V.V.: Meas. Tech. **66**(3), 203–210 (2023). https://doi.org/10.1007/s11018-023-02211-y
26. Oppenheim, A., Schafer, R.: Discrete-time signal processing, 3rd edn. Pearson (2009)
27. Kathiresan, Th , Maurer, D., Suter, H., Dellwo, V.: J. Acoust. Soc. Am. **143**(3), 1919–1920 (2018). https://doi.org/10.1121/1.5036258
28. Kovela, S., Valle, R., Dantrey, A., Catanzaro, B.: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island., pp. 1–5 (2023). https://doi.org/10.1109/ICASSP49357.2023.10096220
29. Sun, P., Mahdi, A., Xu, J., Qin, J.: Speech. Commun. **101**, 57–69 (2018). https://doi.org/10.1016/j.specom.2018.05.006
30. Tohyama, M.: Spectral envelope and source signature analysis. In: Acoustic signals and hearing, pp. 89–110. Academic Press, (2020) https://doi.org/10.1016/B978-0-12-816391-7.00013-9
31. Savchenko, V.V.: Radioelectron. Commun. Syst. **63**, 42–54 (2020). https://doi.org/10.3103/S0735272720010045
32. Eggermont, J.J.: Brain responses to auditory mismatch and novelty detection. Academic Press, pp. 345–376 (2023). https://doi.org/10.1016/B978-0-443-15548-2.00011-9
33. Oganian, Y., Bhaya-Grossman, I., Johnson, K., Chang, E.: Neuron **111**(13), 2105–2118e4 (2023). https://doi.org/10.1016/j.neuron.2023.04.004

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Authors and Affiliations

✉ V. V. Savchenko
vvsavchenko@yandex.ru

**V. V. Savchenko**
National Research University Higher School of Economics, Nizhny Novgorod, Russian Federation