UDC 621.391
DOI: 10.15587/2312-8372.2018.134977

Omelchenko S.

DEVELOPMENT OF THE METHOD OF AUTOMATIC DETERMINATION OF THE SPEAKER GENDER ON THE BASIS OF JOINT EVALUATION OF FREQUENCY MOMENTS OF BASIC TONS AND FORMANT FREQUENCIES

Об'єктом дослідження є методи розпізнавання статі диктора по мовним сигналам. Одними з найбільш проблемних місць є недостатня вивченість вибору ознак і вирішальних правил. Це необхідно для підвищення ймовірності правильного розпізнавання і завадостійкості розпізнавання статі по мовним сигналам в умовах дії перешкод. Важливим також є простота реалізації алгоритмів розпізнавання статі дикторів.

Для розпізнавання статі диктора обрана нова сукупність класифікаційних ознак, що включають спільне використання оцінок середнього значення частоти основного тону, її коефіцієнта ексцесу, оцінок середніх значень формант і їх коефіцієнтів асиметрії. В ході дослідження використовувався метод статистичного випробування запропонованих алгоритмів на персональному комп'ютері. Експерименти проводилися з використанням реальних звукових сигналів, введених з мікрофона в персональний комп'ютер як для представників жіночої, так і чоловічої статі, і записаних у вигляді окремих файлів. Для цього було використано 10 еталоні 10 слів для кожного з 5 дикторів жінок та 5 дикторів чоловіків.

За результатами статистичних випробувань для алгоритму, що включає спільне використання оцінок середньої величини частоти основної тони, її коефіцієнта ефекту, оцінок середніх значень формантів та їх коефіцієнтів асиметрії, отримана оцінка середньої вірогідності правильного розпізнавання 1. При додатковій дії адитивної перешкоди типу гаусів білий шум і відношення сигналу/шум q=20, для такого алгоритму експериментально отримана вірогідність правильного розпізнавання – 0,8. Для алгоритму прийняття рішень, що використовує лише оцінки середньої величини частоти основної тони та її коефіцієнт ефекту, отримана оцінка середньої вірогідності правильного розпізнавання – 0,9. Це говорить про більшу завадостійкість таких алгоритмів.

В перспективі передбачається використання отриманих результатів не тільки для російської та української мов, але і для ряду іноземних мов.

Ключові слова: *розпізнавання статі диктора, формантного-смугові ознаки, коефіцієнт асиметрії, частота основного тону.*

1. Introduction

Algorithms for recognizing the speaker gender are necessary for solving a number of applied problems. The results of determining the speaker gender are used in systems of adaptive word recognition and speech phonemes, identification and verification of speakers, since recognition of the speaker gender allows significantly narrowing the range of values accepted by the signs.

Dimensions of the larynx, vocal folds and muscles that control their fluctuations, are different for men and women. This gives grounds for searching for distinctive features in the parameters of the voice excitation pulses and the digital filter of the speech formation model.

Therefore, it is important to investigate the methods of recognizing the speaker gender using speech signals.

2. The object of research and its technological audit

The object of research is the methods for recognizing the speaker gender by means of speech signals.

One of the most important steps, which ultimately determine the quality of classification, is the choice of classification characteristics.

Typically, as the information parameter, which is used to identify the speaker gender, use the pitch frequency. However, as practice shows, one tone frequency is not enough for reliable classification of the speaker gender.

Other characteristic disadvantages that are inherent in this object under the existing operating conditions are the complexity of implementation and low stability in the presence of high-level interference.

3. The aim and objectives of research

The aim of research is development of algorithms for automatic recognition of the speaker gender.

To achieve the aim, the following tasks are set:

1. To select new classification characteristics.

2. To develop the construction of a decisive rule (classifier), which are resistant to interference, Gaussian white noise.

3. To carry out experimental studies of the developed algorithms.

ISSN 2226-3780

4. Research of existing solutions of the problem

Previous studies on gender identification have offered many features and methods of classification. Isolation of the function is often performed using gender characteristics of speech, such as tone frequency, which is supplemented by cepstral signs [1, 2]. Other approaches are based on spectral features, such as linear prediction coefficients, reflection coefficients. Classification methods use hidden Markov models, Gaussian model mixtures [3, 4]. Also, multi-use approaches combining classification methods have been developed.

In work [5], features of the speaker gender recognition on 4th formant frequencies and 12 Mel-cepstral coefficients (MFCC), where the probability of correct recognition of 0.94 is received is considered.

In the existing system [6], fuzzy logic and neural networks are used. However, such system does not provide a high quality gender classification and is difficult to implement due to the complexity of network training.

In particular, the paper [7] is devoted to the consideration of the peculiarity of gender recognition by speech received from the phone. Various classification methods are considered, including the method of the k-nearest neighbor, the Bayesian approach, the multilayer perceptron using the Mel-cepstral coefficients (MFCC) as characteristics. The probability of correct recognition is 0.90.

In [8], the features of gender recognition by the frequency of the pitch and Mel-cepstral coefficients (MFCC) using logistic and linear regression are considered. The probability of correct recognition is 0.95.

In the system [9], Gaussian mixtures are constructed for Mel-cepstral coefficients (MFCC). Such system with 24 MFCC coefficients and 16 components of the Gaussian mixture of distributions has up to 100 % correct recognition. However, such a system is difficult to implement and training.

In an alternative solution to the problem described in [10], Gaussian mixtures are constructed for the Mel-frequency coefficients (MFCC). Such system has 92 % correct recognition.

The results of the analysis lead to the conclusion that the algorithms for recognizing the speaker gender are, as a rule, difficult to implement and do not satisfy the speaker's recognition quality.

5. Methods of research

It is assumed that the input of the recognition system receives a time sequence of samples of the speech signal $s(n)$, $n=\overline{0,N-1}$, taken with a sampling interval Δt.

It is necessary to build an algorithm that, according to the presented implementation of the speech, makes decisions about the belonging of the current structural speech units to the given types, classes and would provide the maximum of the average probability of correct gender recognition of the speakers P_{pr}.

Let's consider the work of the speaker gender recognizer. In order to obtain the dynamic features of a recognizable digital signal, words are divided into segments of the same length, which is usually 10–30 ms.

First, the segmentation of words, phonemes is performed to compile the stored standards of the speech units of the speaker. Such segmentation at the stage of recognition of speech units allows to exclude redundant decision making procedures for signals that do not carry verbal information or which are not integral speech units. The task of segmentation is to divide speech into structural units and to estimate their time boundaries. The segmentation algorithms are discussed in detail in [11, 12].

Assuming that within the sample the speech signal is stationary in a broad sense, the algorithm for filtering the speech signal in the frequency domain has the form:

$$\dot{x}(t)=\mathrm{Re}\left((N)^{-1/2}\sum_{m=0}^{N-1}C(m)H_{kor}(m)\exp\left(i\left(\frac{2\pi t}{N}\right)m\right)\right),$$

$$C(m)=(2N)^{-1/2}\sum_{\tau=0}^{2N-1}y_{\tau}^{j}\exp\left(-i\left(\frac{2\pi\tau m}{2N}\right)\right), \quad (1)$$

where input readings:

$$y_i^j=\begin{cases}s_i^j, & i=0,1,...,(N-1),\\ 0, & i=N,(N+1),...,(2N-1)\end{cases}$$

$H_{kop}(m)$ – frequency response of the filter.

One of the main parameters of oral speech is the frequency of repetition of vibrations of the vocal cords when pronouncing vocalized speech, called «Pitch». For recognition, it is possible to use the features of the pitch frequency distribution. Measurements on the speech signals made for the voices of five female speakers showed that the ranges of possible values of the pitch frequency from 135 to 522, and for five male speakers from 58 to 238 Hz. Although the ranges of estimates of pitch frequencies for men and women overlap, but differ in the average pitch frequencies of 128 Hz for men and 256 Hz for women.

To estimate the pitch frequency, it is better to use blocks that are voiced. There are many methods of calculating the sign of vocalization. For example, the sign of vocalization N_j^u is calculated by counting the number of zero-intersections for each sample of the j-th sample of the u-th word segment. The decision on vocalization is made in comparison with the threshold, calculated, for example, by the histogram method.

The histograms are asymmetrical – with respect to their mode: for female voices (Fig. 1), from the side of small periods the slope is steeper than for long periods, whereas in men the opposite picture is observed (Fig. 2). For such distributions, the gamma distribution is adequate. It is possible to use cumulants up to the 6th order, including odd ones, for recognition.

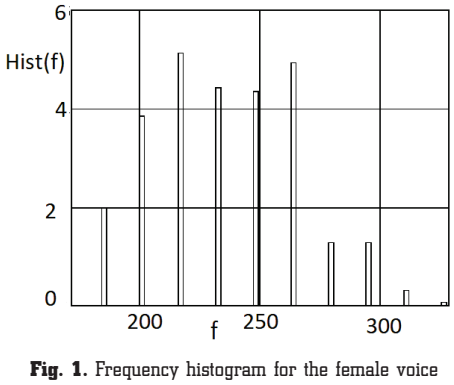

Fig. 1. Frequency histogram for the female voice

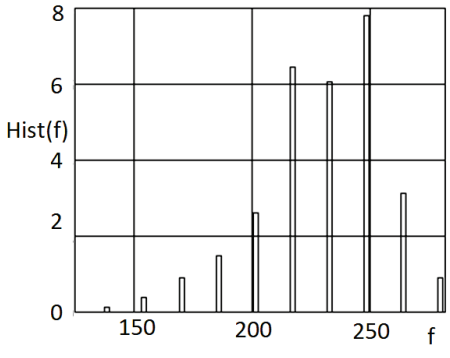

Fig. 2. Frequency histogram for the male voice

The cumulant coefficient γ_k is defined as follows:

$$\gamma_k = \frac{\mu_k}{\mu_2^{k/2}}, \quad (2)$$

where μ_k – cumulants of k order, uniquely related to the central moments.

The calculation of asymmetry and kurtosis makes it possible to establish the symmetry of the distribution of the random variable X with respect to the mathematical expectation $M(x)$. To do this, let's find the third central moment characterizing the asymmetry of the distribution law of a random variable. If it is zero $\mu_3 = 0$, then the random variable X is symmetrically distributed with respect to the mathematical expectation $M(X)$. Since μ_3 has dimension of the random variable in the cube, a dimensionless value is introduced: the asymmetry coefficient:

$$As = \frac{\mu_3}{\sigma^3}. \quad (3)$$

The central fourth-order moment is used to determine kurtosis, characterizes the flatness or sharpness of the probability density. The excess is calculated by the formula:

$$Es = \frac{\mu_4}{\sigma^4} - 3. \quad (4)$$

Estimate of the central moment by the pitch frequency f_o:

$$\mu_i = \sum_{k=1}^{N} \left(f_{ok} - M(f_o)\right)^i. \quad (5)$$

From Fig. 3 it can be seen that it is possible to draw a dividing line between the female and male classes.

The decision-making algorithm:

$$i = \operatorname{sgn}\left(mf_{ft} - k_1 \cdot As - fd\right), \quad (6)$$

where the sign function:

$$\operatorname{sgn}(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases}$$

mf_{ft} – estimate of the average value of the pitch frequency; As – estimate of the asymmetry coefficient; k_1, f_d – the coefficients of the decisive rule.

Considering the experimentally obtained concentration of the pairs of measurements of the mean pitch value f_{ft} and asymmetry A_s for female and male voices (Fig. 3), let's obtain approximate values of the decision rule coefficients $f_d = 170$ and $k_1 = 178$.

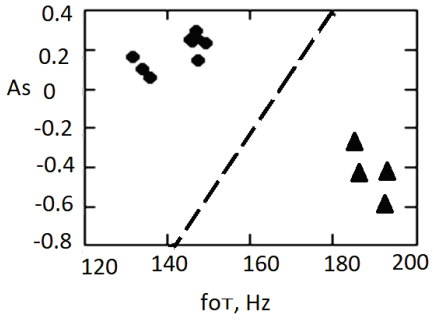

Fig. 3. Concentration of pairs of measurements of the mean value of the pitch frequency f_{ft} and asymmetry A_s for female (triangles) and male voices (dots)

Studies have shown that the use of variance $D(x) = \sigma^2$ and kurtosis Es estimates do not allow the use of a clear dividing line, although their use is possible.

It is also possible to use additional features that improve the recognition of the speaker gender and take into account the characteristics of the voice tract.

Analysis of the effect of changing the length of the speech-forming tract on the parameters of the voice showed that a decrease in the length of the speech-forming tract leads to a substantial increase in the frequencies of the formants. This explains the presence of higher frequencies formant in the female voice compared to the male.

There are methods of estimating the formant frequencies on the basis of estimates of the coefficients of autoregression, cepstral signs [13–15].

Estimates of the formant frequencies by the spectral-band method [16, 17] for each of the blocks can be calculated as the average effective frequencies from the corresponding output of the bandpass filter. In Table 1 for each m-th filter the boundary frequencies $f_h^{(m)}$ and $f_l^{(m)}$ are indicated.

Table 1

Boundary frequencies of filters

m	$F_l^{(m)}$, Hz	$F_h^{(m)}$, Hz
1	200	850
2	850	2200
3	2200	3000
4	3000	4000

Let's consider features of formation of formant-band signs. Blocks (samples) of speech are formed of the set of samples, which are taken with or without a 2–3-fold overlap. According to this method, the spectral-band signals corresponding to the probable arrangement of the formants are calculated, the bands of which are given in Table 1. The boundary frequencies $f_h^{(m)}$, $f_l^{(m)}$ correspond to the m-th formants at a sampling frequency of 8 kHz.

In this case, the estimates of the formant frequencies for a given sample are calculated by counting the number of zero-intersections of the speech signal from the corresponding output of the bandpass filter.

ISSN 2226-3780

The procedure for calculating the formant can be repeated, but the following are used as boundary frequency bands:

$$f_h^{(m)} = f^{(m)} + \Delta, \quad f_l^{(m)} = f^{(m)} - \Delta,$$

where $f^{(m)}$ – the formants computed at the previous stage; Δ – the range limits of the search formant. The simplest among the recurrent procedures is a two-stage one.

Studies have shown that the first and second formants and the measure of their excesses make the greatest contribution to recognition, where a linear division of classes is possible. Fig. 4 shows the obtained experimental concentration of pairs of measurements of the mean value of the second formant frequency and the kurtosis coefficient E_a for five female (triangles) and five male voices (squares). This allows the second formant to use linear separating boundaries between them. From the obtained experimental data, such linear separation is possible for the first formant and its kurtosis coefficient, but for the third and fourth formants it is difficult.

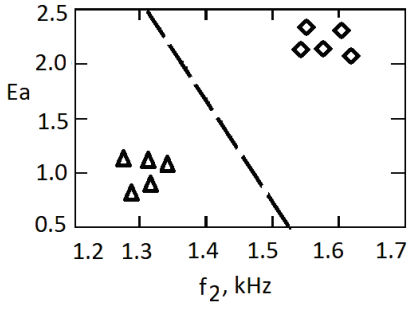

Fig. 4. Concentration of pairs of measurements of the mean value of the second formant frequency f_2 and the kurtosis coefficient E_a for female (triangles) and male voices (squares)

The decision-making algorithm:

$$i = \operatorname{sgn}\left(\begin{array}{l} mf_{ft} - k_0 \cdot As - fd + \sum_{k=1}^{4} k_k \left(mf_k - f_{dp_k} \right) + \\ + \sum_{k=1}^{2} ke\left(Es_k - Es_{dpk} \right) \end{array} \right), \qquad (7)$$

where mf_{ft} – estimate of the average value of the pitch frequency; As – asymmetry of the pitch frequency; mf_k – estimate of the frequency value of formants; Es_k – estimate of the kurtosis coefficient for the frequency of the k-th formant.

6. Research results

The tests of the above word recognition algorithms were performed on the basis of data entered into the computer from the microphone through the audio interface with a sampling frequency $F_s = 8$ kHz.

Tests were conducted on real samples of audio signals input to the computer from the microphone output.

Experimental studies of speech recognition algorithms with a one-stage determination of the number of zeros in the formant bands were carried out by the statistical test method using 10-signal samples for each of the 10 different male and female speakers. The parameters of the decision rule were estimated from samples, and the control samples of real signals were used to evaluate the quality of signal recognition.

The decision-making algorithm, taking into account the previously considered decision rule (7), has the form:

$$i = \operatorname{sgn}\left(\frac{f_{ft}}{178} - As - \frac{f_2}{1400} - \frac{Es_2}{1.5} + 1 \right), \qquad (8)$$

where f_{ft} – estimate of the average value of the pitch frequency; As – estimate of the asymmetry of the pitch frequency; f_2 – estimate of the frequency value of the 2nd formant; Es_2 – estimate of kurtosis for the frequency of the 2nd formant.

The decision-making algorithm for estimating the average value and the pitch frequency asymmetry coefficient:

$$i = \operatorname{sgn}\left(\frac{f_{ft}}{178} - As - 1 \right). \qquad (9)$$

The decision-making algorithm for estimating the average value and the frequency kurtosis of the 2nd formant:

$$i = \operatorname{sgn}\left(2 - \frac{f_2}{1400} - \frac{Es_2}{1.5} \right). \qquad (10)$$

According to the results of statistical tests, an average probability of correct recognition $P_{pr} = 1$ was obtained for each of the algorithms (8)–(10). Under the additional action of additive noise of the Gaussian type white noise and the signal-to-noise ratio $q = 20$, estimates of the mean probabilities of correct recognition were obtained experimentally P_{pr}. For the decision-making algorithm in accordance with formula (9), an estimate of the mean probabilities of correct recognition is experimentally obtained $P_{pr} = 0.9$. For the decision-making algorithm in accordance with the formula (8) $P_{pr} = 0.8$, and for the algorithm, the decision-making in accordance with the formula (7) $P_{pr} = 0.7$. Thus, under the influence of additive noise of the Gaussian type, high-level white noise is rational to use the decision-making algorithm in accordance with the formula (9), which takes into account the features of estimates of the mean value of the pitch frequency and the estimates of the pitch frequency asymmetry.

The conducted studies confirm the effectiveness of the applied algorithms.

7. SWOT analysis of research results

Strengths. Compared with analogues, the positive effect of the object of research in the form of constituent elements of the recognition system is optimization of the choice of decision-making features in order to increase the probability of correct recognition of the speaker gender, depending on the noise level. This includes modeling the recognition system on a personal computer.

Weaknesses. The weaknesses of the proposed effective parameters of the recognition system include the need for initial capital investments in the gender recognition system. It is also necessary to provide for the costs of their production at the place of use. Also, the weaknesses of the proposed solutions include their locality («point») with respect to the entire complex recognition system for different languages.

Opportunities. The proposed technical solutions to improve the quality of recognition helps to improve the quality of recognition of speakers, recognition of speech words, simplify the search in databases. This, in turn, will significantly reduce the cost of manufacturing systems. The expected profit is projected to be obtained in about 2–3 years, depending on the number of systems.

In the future, the use of the obtained results not only for Russian and Ukrainian languages, but also for a number of foreign languages is supposed.

Threats. The enterprise or the operating organization will require initial capital investments in the technical implementation of the recognition system. Also, the costs of their production are needed. Negative impact on the object of research of external factors in the form of the external environment and other operating conditions are due to the regulatory period of operation. It depends on the used developments. However, this period is not less than 5 years, which is more than sufficient for self-sufficiency of the developed organizational and technical solutions.

8. Conclusions

1. New classification features are selected, including the joint use of estimates of the mean value of the pitch frequency, its kurtosis coefficient, estimates of the mean values of the formants and their asymmetry coefficients.

2. Decisive rules for making decisions about the field of the speaker based on linear division by the weighted sum of the estimates of the proposed classification characteristics are constructed. Linear boundaries for separating the speaker gender are due to the compact arrangement of features for each type of speaker.

3. Based on the found performance characteristics, comparative studies of the speaker recognition algorithms are carried out.

Based on the results of statistical tests for the algorithm, including an estimate of the average value of the pitch frequency, its kurtosis coefficient, estimates of the mean values of the formants and their asymmetry coefficients, an estimate is obtained for the average probability of correct recognition 1. With the additional action of additive noise of the Gaussian type, white noise and the signal-to-noise ratio $q = 20$. For such algorithm, the probability of correct recognition is experimentally obtained – 0.8. For the decision algorithm, which uses only estimates of the average value of the pitch frequency and its kurtosis coefficient, an average probability of correct recognition is estimated – 0.9. This indicates more noise immunity of such algorithms.

The conducted studies of recognition algorithms confirm the possibility of obtaining an acceptable quality of recognition of the speaker gender on the basis of the use of:

– estimates of the average value of the pitch frequency and its asymmetry coefficients;
– estimates of the average frequency of the formants;
– estimates of kurtosis coefficients for formant frequencies.

References

1. Kalyuzhnyi A. Ya., Semenov V. Yu. Metod identifikatsii pola diktora na osnove modelirovaniya akusticheskikh parametrov golosa gaussovymi smesyami // Akustichniy visnik. 2009. Vol. 12, No. 2. P. 31–38.
2. Scheme E., Castillo-Guerra E., Englehart K., Kizhanatham A. Practical Considerations for Real-Time Implementation of Speech-Based Gender Detection // Lecture notes in computer science. 2006. Vol. 4225. P. 426–436. doi: http://doi.org/10.1007/11892755_44
3. Sorokin V. N., Makarov I. S. Opredelenie pola diktora po golosu // Akusticheskiy zhurnal. 2008. Vol. 54, No. 4. P. 659–668.
4. Robust GMM-based gender classification using pitch and RASTA-PLP parameters of speech / Zeng Y.-M. et al. // Proceedings of the Fifth International Conference on Machine Learning and Cybernetics. Dalian, 2006. P. 3376–3379. doi: http://doi.org/10.1109/icmlc.2006.258497
5. Faek F. Objective Gender and Age Recognition from Speech Sentences // Aro, The Scientific Journal of Koya University. 2015. Vol. 3, No. 2. P. 24–29. doi: http://doi.org/10.14500/aro.10072
6. Jayasankar T., Vinothkumar K., Vijayaselvi A. Automatic Gender Identification in Speech Recognition by Genetic Algorithm // Applied Mathematics & Information Sciences. 2017. Vol. 11, No. 3. P. 907–913. doi: http://doi.org/10.18576/amis/110331
7. Gender Identification using MFCC for Telephone Applications – A Comparative Study / Ahmad J. et al. // International Journal of Computer Science and Electronics Engineering. 2015. Vol. 3, No. 5. P. 351–355.
8. Levitan S. I., Mishra T., Bangalore S. Automatic identification of gender from speech // Proceeding of Speech Prosody. 2016. P. 84–88. doi: http://doi.org/10.21437/speechprosody.2016-18
9. Yucesoy E., Nabiyev V. V. Gender identification of a speaker using MFCC and GMM // 2013 8th International Conference on Electrical and Electronics Engineering (ELECO). Bursa, 2013. doi: http://doi.org/10.1109/eleco.2013.6713922
10. Harb H., Chen L. Gender identification using a general audio classifier // 2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698). Baltimore, 2003. doi: http://doi.org/10.1109/icme.2003.1221721
11. Presnyakov I. N., Omelchenko S. V. Pomekhoustoychivye algoritmy segmentatsii rechi v sistemakh obrabotki // Radiotekhnika. 2003. No. 131. P. 165–177.
12. Sorokin V. N., Tsyplikhin A. I. Segmentatsiya i raspoznavanie glasnykh // Informatsionnye protsessy. 2004. Vol. 4, No. 2. P. 202–220.
13. Presnyakov I. N., Omelchenko A. V., Omelchenko S. V. Avtomaticheskoe raspoznavanie rechi kanalakh peredachi // Radioelektronika i informatika nauchno-tekhnicheskiy zhurnal. 2002. No. 1. P. 26–31.
14. Rabiner L. R., Schafer R. W. Digital Processing of Speech Signals. Pearson; US edition, 1978. 962 p.
15. Marple S. L. Digital Spectral Analysis: With Applications/ Disk,Pc/MS Dos/IBM/Pc/at. Prentice Hall Signal Processing Series, 1987. 492 p.
16. Presnyakov I. N., Omelchenko S. V. Avtomaticheskoe raspoznavanie razdel'nykh slov i fonem rechi // Radioelektronika i informatika. 2003. No. 2. P. 41–47.
17. Presnyakov I. N., Omelchenko S. V. Algoritmy raspoznavaniya rechi // Avtomatizirovannye sistemy upravleniya i pribory avtomatiki. 2004. No. 126. P. 136–145.

Omelchenko Sergey, *PhD, Associate Professor, Department of Information Network Engineering, Kharkiv National University of Radio Electronics, Ukraine, e-mail: serhii.omelchenko@nure.ua, ORCID: http://orcid.org/0000-0002-3998-978X*