Speaker Verification Using Autoregressive Spectrum of Speech Signal in Composite Vector Stochastic Processes Model Representation

NatalijaV. Chmelarova (Kudriavtseva)[1],Vyacheslav A. Tykhonov[2], Valerij M. Bezruk[3], Pavel Chmelar[4], Lubos Rejfek[5]

[1,4,5] Department of Electrical Engineering, Faculty of Electrical Engineering and Informatics, University of Pardubice, Pardubice, Czech Republic

[2]Department of Radioelectronic Systems, National University of Radioelectronics, Kharkiv, Ukraine,

[3]Department of Radioelectronic Systems, National University of Radioelectronics, Kharkiv, Ukraine,

Email : [1]nataliia.kudriavtseva@gmail.com, [2]slavatihonoff@mail.ru, [3]valerii.bezruk@nure.ua

Abstract

This paper deals with the speaker verification system similar to a fingerprint or an eye scanner. For these purpose a long-term words' model and its spectral characteristics were used. The speaker verification method uses the word's sound parametric spectrum factorization in composite vector stochastic process representation based on the multiplicative autoregressive model. The developed method enables to receive the words' features with stable characteristics for the same speaker and differ for different speakers. During the training phase speaker's etalon frequencies has to be estimated for a pronounced word repeated several times. In the verification phase a speaker pronouncing the same word, word's frequencies are estimated and compared with the etalon frequencies database to find the best match or his deny. The results presented in the paper showed the high correct identification probability.

Keywords : Composite Vector Stochastic Processes Autoregressive Models, Power Spectrum Density, Speaker Verification

I. Introduction

Speaker recognition uses the acoustic features of speech to identify speakers. The research to develop robust speaker recognition systems is still ongoing since 40 years ago, such as modern applicable services includingbanking over telephone network, security control for confidential information, access to premises and forensics investigation. The task of speaker recognition comprises speaker verification (i.e. authenticating speaker's identity) and/or speaker identification (i.e. determining an unknown speaker's identity) (Kinnunen et al. 2010) and (Hansen et al. 2015).

In a difference from the classicalbiometrics where fixed parameters are used, the verification by voice has practically unlimited potential for reducing the error because of using more long speech messages. In recent times a considerable amount of research in automatic speaker verification focuses on the short-utterance issue, which is more challenging in practical scenarios (Poddar et al. 2018) and (Guo et al. 2016) .

In the case of the verification the user pronounces his identifier and it is required to confirm or denyhis voice. In the most of cases the user is interested in his identifier confirmation and he tries to not make changes in the speech password. During the identification the recognition system must confirm or deny the statement if the voice belongs to the one speaker from a speakers'database.Contemporary automatic speaker verification systems mostly use short-term spectral features (Chen et al. 2013), (Das et al. 2016) , (Li et al. 2015), (Li et al. 2016) and (Soldi et al. 2014). Mel-frequency cepstral coefficients, the perceptual linear prediction and the linear predictive are widely used feature extraction techniques due to their considerable performance and lower-computational complexity (Disken et al. 2017) , (Javanna et al. 2009) , (Salidullah et al. 2012) and (Salidullah et al. 2016) .

In this paper we present a new long term stochastic process model used for the speaker verification. The new autoregressive model generates distinctive frequency characteristics per speaker for each pronounced word. It leads to a high matching accuracy with low computation time.

Using the parametric power spectrum density frequency spikes of the speech signal as words' features, presented by composite vector stochastic processes is proposed in this paper. Such a representation by using the mostsubvector length enables to receive the stableword spectra in a bigger range. Therefore, we can observe the stable spectrum shape for each speaker and perceptibly differencespectra characteristics for different speakers. For the effective spectrum spikes' frequencies calculationwe are proposing to use the factorization multimode parametric spectra method (Kudriavtseva. 2014) for a speech signal created by the authors. In this casethe multimode parametric spectra it is presented by the set of single-mode spectra (Kudriavtseva et al. 2012) . For the spike frequencies calculation the autoregressive model coefficients describing a parametric spectrum of a single-mode spectra are used. Calculated spike differences in the recognition phase than allows to differentiate individual speakers.

The rest of this paper is organized as follows. In the Section 2 the long term stochastic process modelbasesare explained,theequation for parametric spectra estimation on this model base is shown and words long-term parametric spectra estimations are given. In the Section 3 the algorithm for the complex spectrum factorization into components and the method for spectral component peaks calculation, characterize a speaker in the verification process, are described. Here it is also presented the speakers' verification method including the etalon features library composition andthe decision rule. We summarized the speaker's verification results and our future research plan in the Section 4.

II. The Correlation Function and Spectra of Composite Vector Stochastic Processes Representation

In (Tykhonov et al. 2011) authors propose a new stationary stochastic process model called "Composite Vector Stochastic Processes" (CVSP)$x^n[t]$, whereby a stochastic process$x[t]$ can be presented as a "subvector" $x_i[t]$sequence, a segment of $x[t]$, having each the same length $n$and statistical characteristics. The CVSP presentation describes the long – term signal change without using the decimation process. In such representation the speech signal spectra contains information about the main tone vibration frequency, valuable for speech phonemes. The word's spectra using the CVSP presentation describes the most essential connections inside the phonemes and between the phonemes. Such a word analysis it is close to the human speech recognition, where it is perceiving not only phonemes' sounds, but also connections between the phonemes.

The equation for the signals' correlation function estimation using the CVSP representation is

$$R^n[k] = \frac{1}{M-k}\sum_{i=0}^{m-k}\sum_{j=0}^{n-1}(x[in+j]x[in+j+kn]), \qquad (1)$$

where M is the number of subvectors, n is the subvector's length. The expression (1) can be simplified to the following

$$R^n[k] = \frac{1}{N-kn}\sum_{i=0}^{N-kn}(x[i]x[i+kn]). \qquad (2)$$

If the vector count N is not divisible by thesubvector length n then M is integer part ofN/n. Notably,$R^1[k]$gives the autocorrelation function for the stationary stochastic process $x[t]$. After further arrangements of (2) we can receive the equation such as Yule-Walker for the AR CVSP model parameters calculation

$$R^n[j] = \sum_{i=1}^{p}\Phi^n[i]R^n[i-j], \quad j = 1,2,...p, \qquad (3)$$

where $\Phi^n[i]$are the AR CVSP model's coefficients. Obviously, that all features of autoregressive (AR) stochastic processes model are also valid for the AR CVSP model. Theequation (4) gives the our process power spectrum density (PSD) estimation expression (Tykhonov et al. 2011b)

$$P^n(\omega) = \frac{D_a^n}{\left|1-\sum_{i=1}^{p}\Phi^n[i]e^{-j\omega iT}\right|^2}, \qquad (4)$$

where $D_a^n$is the prediction error variance $a[t]$, T is the quantization process interval and$\omega = 2\pi f$is the cyclic frequency. Fig. 1-4 give AR PSDexamples using the CVSP representation for some words pronounced by one speaker. The resulting PSD have different spectra, thus can be used as the word's distinctive characteristic. The PSD distributions were estimated by using the AR model of 12-th order with the subvector length $n$equal to 20. The word's PSD analysis showed, that it is difficult to present the PSD as the word'scomponent phonemes set.

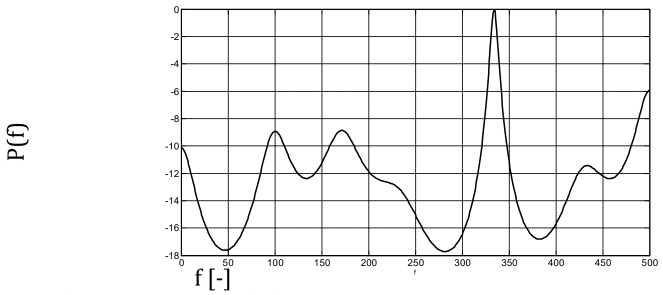


Fig. 1.PSD by AR (12) model using CVSP representation for Russian word “Езда” (driving)

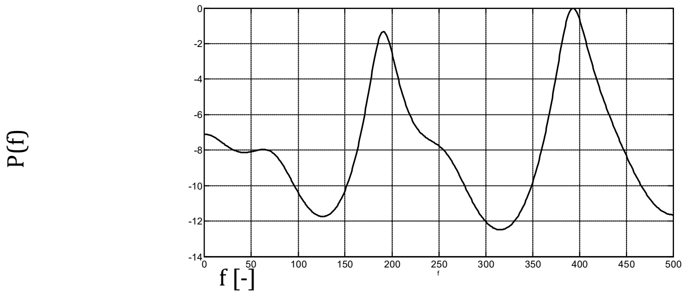


Fig. 2.PSD by AR (12) model using CVSP representation for Russian word “Литр” (litre)

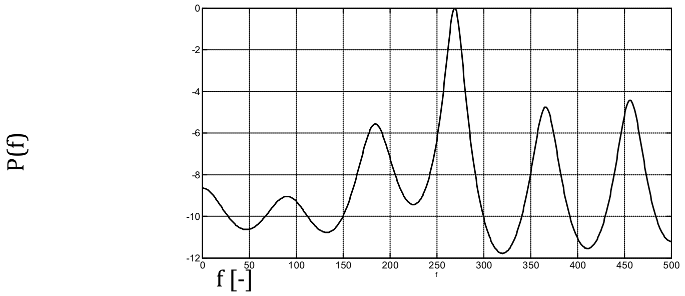


Fig. 3.PSD by AR (12) model using CVSP representation for Russian word “Село” (village)

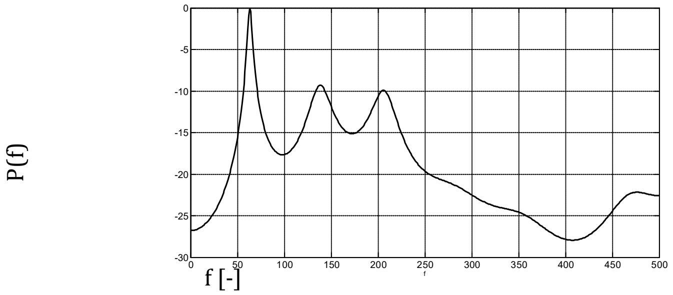


Fig. 4.PSD by AR (12) model using CVSP presentation for word "hello" for speaker 3, the word1

II. Speaker Verification using CVSP Model

In this research experiment the key word "hello" it is proposed to consider as a subvectors sequences set with the length equal to 20. The waveform results showed, thatthe AR multimode word spectra for each speaker are close by shape and have a salient difference for different speakers in the CVSP representation. The significant CVSP spectra representation advantage is a possibility to receive the main tone vibration frequencymodes for each speaker. This is an important feature for speaker verification. The multiplicative AR model $AR_1 \times AR_2 \times \ldots \times AR_k$ is created to use the spectral modes parameters as the features, but not as the spectra counts orthe shorten features set. The component models parameters$AR_i$are calculated by using the AR model characteristic equation roots. Then the multimode p-th order spectrum (4) is presented as a first and second order single-mode spectra set. Hence the multimode spectrum (4) can be presented as

$$P^n(\omega)=\frac{D_a^n}{\left|\sum_{i=0}^{p_1}\Phi_1^n[i]e^{-j\omega iT}\right|^2\left|\sum_{i=0}^{p_2}\Phi_2^n[i]e^{-j\omega iT}\right|^2\times\ldots\times\left|\sum_{i=0}^{p_k}\Phi_k^n[i]e^{-j\omega iT}\right|^2}, \tag{5}$$

where $\Phi_{\mathrm{m}}^n[i]$ are AR model coefficients that compose multiplicative model. For solving this verification task we proposing to factorize the multimode PSD into single mode components. It is easier to calculate spike's frequency and the spike's band width for single mode components. As the features we are proposing to use the spike's frequencies and the spike bandwidth with the level equal to 0.5. If the characteristic equation roots are complex, then the spike frequency verifies

$$f_s = \arccos\left(\Phi_m^n[1]/\sqrt{-\Phi_m^n[2]}\right)/\,2\pi T \tag{6}$$

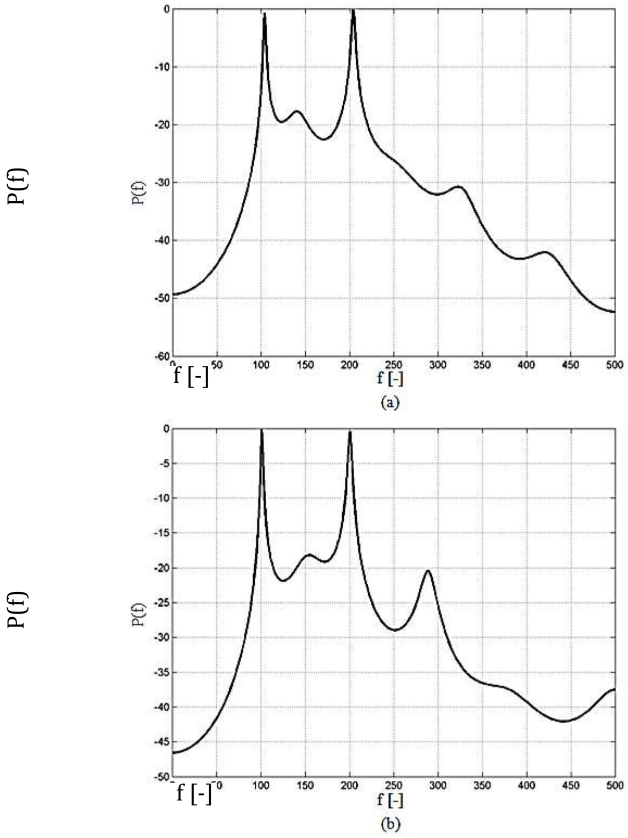
and the spike bandwidth is equal to

$$df = \ln(-\Phi_m^n[2])\,/\,2\pi T. \tag{7}$$

If the characteristic equation roots are rather real, then the spike frequency if $\Phi > 0$ or $f = Fd/2$, if $\Phi < 0$ the spike bandwidth in this case is equal to

$$df = \ln(\left|\Phi_m^n[1])\right|)/\pi T. \tag{8}$$

The ranges of the spectra distinctive parameters cited above for each speaker word utterance are determined at the stage of registration. During this stage, we save only the most stable spikes through all 10 word utterances per speaker. Table 1 gives the PSD spikes' frequencies of the words. These results were received after the factorization of multimode PSD and calculated using (6).



(a)

(b)

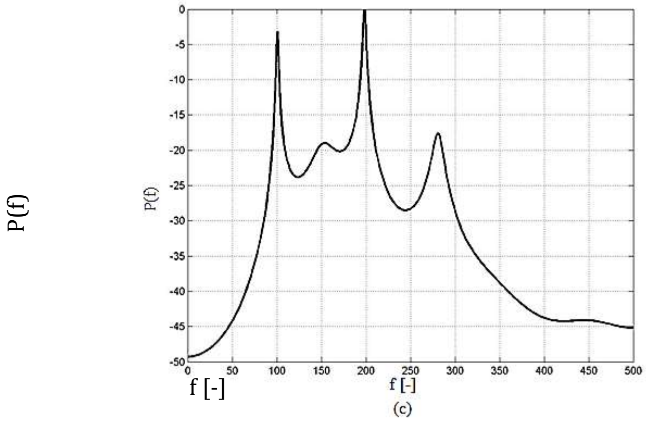


Fig. 5. The first speaker AR PSD examples for the word “hello”,
(a) – first word “hello”, (b) – second word “hello”, (c) – third word “hello”

The identifier word “hello”is used for speaker recognition. The AR PSD examples for the first speaker’s word “hello”, are shown in Fig.5. The multimode PSD factorization into single – mode spectra components of Fig. 5(c), are shown in Fig. 6.The well-defined narrowband spikes frequency f_S coincides with the single – mode components frequency. This effect confirms the PSD factorization method’s accuracy. The spikes’ frequencies and bandwidths calculated using equations (6) and (7) for diagrams in Fig. 6 are shown in the Table 1.

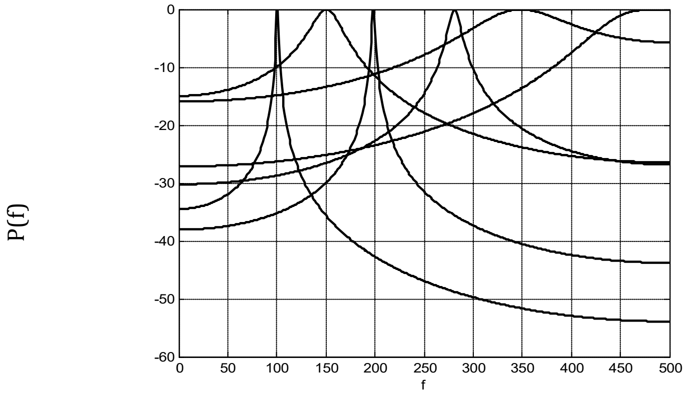


Fig. 6. The multimode PSD factorization into the single – mode spectra components of Fig. 5(c)

Table 1.The spikes’ frequencies and bandwidths for diagrams in Fig. 6

f [Hz]	**1**	**2**	**3**	**4**	**5**	**6**
f_s	453	198	151	101	281	343
df_s	93.9	2.76	29.5	1.72	11.94	96.2

For comparison the second speaker AR PSD examples for the word “hello” are shown in Fig.7. The multimode PSD factorization into the single – mode spectra components of Fig. 7(c) is shown in Fig. 8. The well-defined narrowband spikes frequency coincides wwiith the single – mode frequency componenttss. This also confirms the PSD facttoorization method’s accuracy. The spikes’ freqquencies and bandwidths calculated uussing equations (6) and (7) for diagrams in Fig. 8 are shown in Table 2.

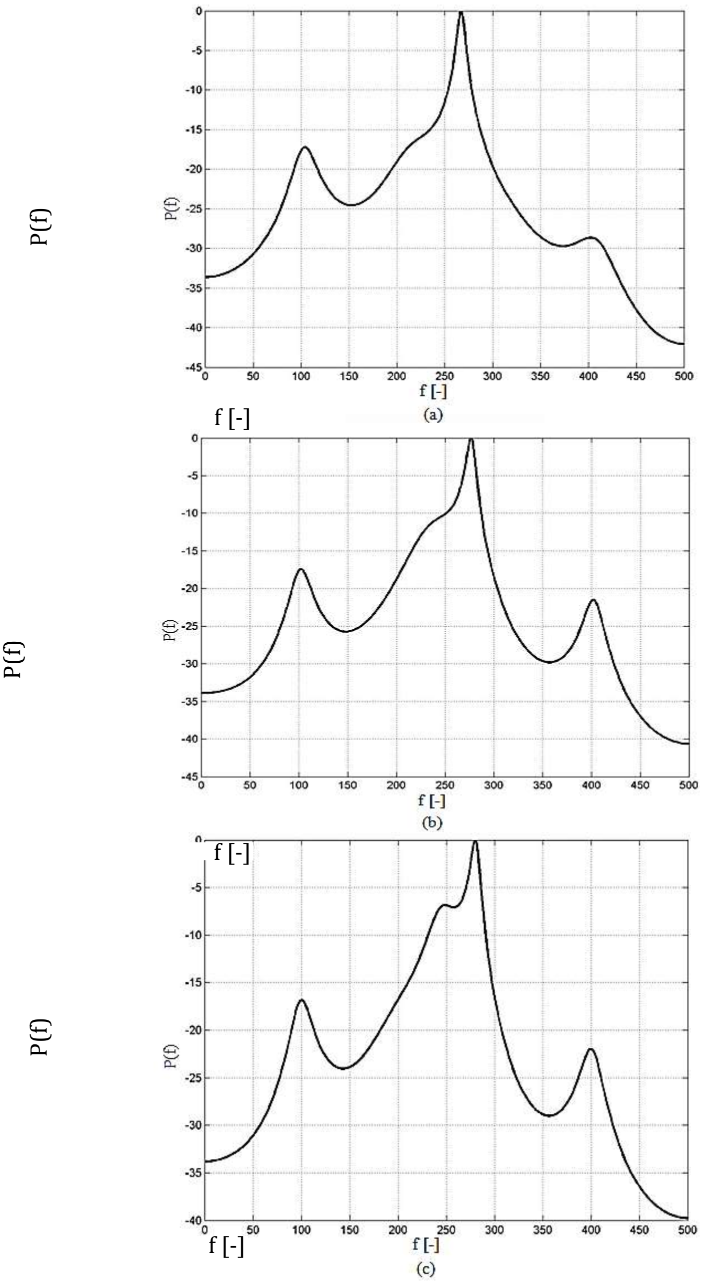


Fig. 7. The second speaker AR PSD examples for the word “hello”, (a) – first word “hello”, (b) – second word “hello”, (c) – third word “hello”

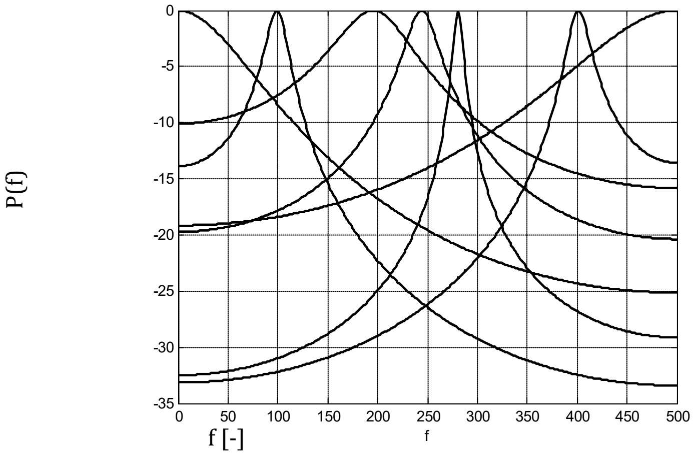


Fig. 8. The factorization of the multimode PSD to single – mode components of spectra that is shown in Fig. 7(c)

Table 2.The spikes' frequencies and bandwidths for Fig. 8

f [Hz]	1	2	3	4	5	6	7
f_s	500	0	100	244	197	281	400
df_s	219.7	152.7	21.0	31.8	73.1	9.1	21.7

The word "hello" was said by the 35 speakers 10 times for each speaker (12 women and 23 men). The three speakers changed slightly intonation during the words recording phase and the two speakers pronounced they words in noisy background in a classroom. This allows us to evaluate results of the proposed method in a different condition and test the system for its robustness.

When we are using the CVSP method we can consider the word like the one entire sample and not asfour samples as for filter method. The sample is presented by thesubvectors set with the length equal to 20 for each subvector. The CVSP method enables to find the long-term word spectral characteristics. The AR PSD were calculated for all 10 words of each speaker for 12-th and 20-th samples orders in the CVSP representation. The speakers' spectra have high modes and significant differences between each other using these AR models orders and subvector's length equal to 20. We suppose that this is the most applicable forthe model correct identification probability during the experiment.

Estimated frequencies of two speakers are shown in Fig. 5 and Fig. 7. As it is shown in figures, the PSD for each speaker are close among themselves and have differences for different speakers.This is the reason why to take the AR PSD frequency characteristics as the features of speakers.

For each word pronounced by speakers three, six or both the most stable spikes frequencies (f_s 3 andf_s6 in Table 3) were calculated during the training phase. The mean values of these stable frequencies calculated for each speaker are used as the speakers' features, stored in a database. The spike frequencies are calculated for the test speaker's word in verification phase and compared with all speakers' etalons in a database. The decision rule is to find the best match by the Euclidean distance calculation between the testing word and speakers' database. The result is the recognized speaker from database or his deny. In Table 3 are shown etalon frequencies for 10 selected speakers to present method properties.

Table 3. Example of etalon frequencies for 10 speakers

Spk.	f [kHz]	1	2	3	4	5	6
Spk.1	f_s 3	–	–	–	–	–	–
	f_s6	0.99	1.5	1.81	2.26	3.12	3.41
Spk. 2	f_s 3	0.64	1.23	1.93	–	–	–
	f_s6	0.49	1.1	1.73	2.06	2.79	3.59
Spk. 3	f_s 3	0.82	1.17	1.61	–	–	–
	f_s6	0.83	1.15	1.63	2.12	2.64	3.5
Spk. 4	f_s 3	–	–	–	–	–	–
	f_s6	0.52	1.11	1.67	2.19	2.3	3.64
Spk. 5	f_s 3	0.62	1.16	1.76	–	–	–
	f_s6	0.58	1.03	1.59	2.26	2.92	3.61
Spk. 6	f_s 3	0.82	1.57	1.97	–	–	–
	f_s6	–	–	–	–	–	–
Spk. 7	f_s 3	0.77	1.36	2.03	–	–	–
	f_s6	–	–	–	–	–	–
Spk. 8	f_s 3	–	–	–	–	–	–
	f_s6	5.13	1.26	1.49	2.03	2.79	3.55
Spk. 9	f_s 3	–	–	–	–	–	–
	f_s6	4.78	1.14	1.62	2.02	2.78	3.51
Spk.10	f_s 3	4.3	1.1	1.68	–	–	–
	f_s6	4.2	1.11	1.66	2.02	2.95	3.56

The correct identification probabilities for speakers from Table 3 and the all speakers' results are shown in the Table 4. This table also includes a comparison of 30speakerspronouncing words in a normal voice background with 5 speakers pronouncing words in different conditions mentioned above. The research results show the high correct identification probability. The mean probability value of the correct identification for all speakers is almost 91%. When we exclude the three speakers with slightly different intonation and the two speakers with noisy background the probability increases to 94%.

Table 4. Verification results

Spk.	Frequencies	Results "true"
Spk.1	f_s 3	–
	f_s6	100%
Spk. 2	f_s 3	96.43%
	f_s6	92.59%
Spk. 3	f_s 3	100%

	f_s6	100%
Spk. 4	f_s 3	–
	f_s6	100%
Spk. 5	f_s 3	85.71%
	f_s6	99.54%
Spk. 6	f_s 3	100%
	f_s6	–
Spk. 7	f_s 3	%71
	f_s6	–
Spk. 8	f_s 3	–
	f_s6	%99.17
Spk. 9	f_s 3	–
	f_s6	%75.17
Spk. 10	f_s 3	%93.8
	f_s6	%100
All 35 speakers		90.92%
30 speakers in normal conditions		94.15%

The Spk. 7 from Table 4 is an example of the speaker with slightly different word's pronunciation and Spk. 9 is a speaker who pronounced his words in noisy background.

Even when speakers pronouncing they words in worse condition the verification probability is higher than 70%.

IV. Conclusions

In this paper we presented the new speaker verification system using our novel stochastic process model CVSP.

This representation allows us to generate distinctive and stable frequency features for each speaker for each pronounced word. As it is shown in this research, the advantage of our model lies in the ability to generate strong PSD peaks which correspond to the pitch oscillation frequency for each speaker, crucial for the success of any speaker verification systems.In the CVSP representation a speech signal is divided into equally-size subvectors enable to analyze the speech signal in wider range.The stable parametric PSD peak frequencies were used to decrease the number of calculations. Our model leads to low verification time, because it is used only the stable mode frequencies of the resulting PSD during the matching.

The research results for 30speakers verification who pronounced the word "hello" ten times showed that the matching accuracy was high 94% and with 5 speakers we tested that this method recognize a speaker even in worse conditions.

This method is also possible to use for the connected speech recognition but this will need a future research. We verified the method ability and next we plan to test our model with a large established database e.g. RedDots and RSR2015.

V. Acknowledgements

The research was supported by the Internal Grant Agency of University of Pardubice, the project No. SGS_2019_021.

References

I. Chen Y. and Tang Z. M. (2013) ‘The speaker recognition of noisy short utterance. ICISBDE, pp. 666–671

II. Das R. K., and Mahadeva Prasanna S. R. (2016) Exploring different attributes of source information for speaker verification with limited test data, J. Acoust. Soc. Am., Vol. 1, pp. 184–190

III. Dişken G., Tüfekçi Z., and Saribulut L. (2017) A review on feature extraction for speaker recognition under degraded conditions’, IETE Tech. Rev., 2017, Vol 3, pp. 321–332

IV. Guo J., Yeung G., Muralidharan D., Arsikere H., Afshan A., Alwan A. (2016) Speaker Verification Using Short Utterances with DNN-Based Estimation of Subglottal Acoustic Features, Proc. Interspeech 2016, 2219-2222.

V. Hansen J. H. L., and Hasan T. (2015) Speaker recognition by machines and humans: a tutorial review, IEEE Signal Process. Mag., Vol. 6, pp. 74–99

VI. Jayanna H., and Prasanna S. M. (2009) Analysis, feature extraction, modeling and testing techniques for speaker recognition, IETE Tech. Rev. Vol 3, pp. 181–190

VII. KinnunenT., and LiH. (2010) An overview of text-independent speaker recognition: from features to supervectors, Speech Commun., Vol. 1, pp. 12–40

VIII. Kudriavtseva N. V. (2014) Factorization of processes parametric spectra on the base of multiplicative linear prediction polymodels, RADIOELEKTRONIKA 2014 24th International Conference, pp.1-4

IX. Kudriavtseva N. V. and Fil I. O. (2012) Factorization of Rhythmograms Parametric Spectra on the Base of Multiplicative Linear Prediction Models, IEEE East-West Design & Test Symposium (EWDTS’2012), Kharkiv, Ukraine, pp. 538-540

X. Li L., Wang D., Zhang C., and Zheng T. F. (2016) Improving short utterance speaker recognition by modeling speech unit classes, IEEE Trans. Audio Speech Lang. Process., Vol. 6, pp. 1129–1139

XI. LiZ.Y., ZhangW.Q. and LiuJ. (2015) Multi-resolution time frequency feature and complementary combination for short utterance speaker recognition, Multimedia Tools Appl., Vol. 3,pp. 937–953

XII. Poddar A., Sahidullah Md., and Saha G. (2018) Speaker Verification with Short Utterances: A Review of Challenges, Trends and Opportunities, IET Biometrics. 7 (2): 91–101.

XIII. Sahidullah M., and Saha G. (2012) Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition', Speech Commun., Vol. 4, pp. 543–565

XIV. Sahidullah Md., and Kinnunen T. (2016) Local spectral variability features for speaker verification". Digital Signal Processing, pp. 1–11

XV. SoldiG., BozonnetS., AlegreF., Beaugeant, Ch., and Evans N.(2014) Short-duration speaker modelling with phone adaptive training., Proc. Odyssey

XVI. Tykhonov V.A., and Fil I. O. (2011) Statistical modelling of composite vectorial stochastic processes,Radiotekhnika Journal, #165, KhNURE, pp. 173-176

XVII. Tykhonov V.A.,Kudriavtseva N. V., and Fil I. O. (2011b) Mathematical models of combining vectorial stochastic processes, Eastern-European Journal of Foremost Technologies, pp. 17-20