Simply Pattern Recognition as a Tool for Identity Verification

Karolina Kęsik
Institute of Mathematics
Silesian University of Technology
Kaszubska 23, 44-100 Gliwice, Poland
karola.ksk@gmail.com

***Abstract*—The increasing development of mobile and smart technologies caused that voice recognition and even analysis is becoming much more needed in recent years. For this reason, in this paper, the idea of voice recognition is presented. Proposed idea is based on classic approach called pattern matching. What distinguishes the technique is to present a sound sample in the form of a spectrogram ($2D$ image). Then, the features extraction is done not on the sound, but on the image, what allows first to build the pattern, and then the classification. In addition, the matching process is supported by the k-nearest neighbors technique. The entire process has been described, tested and discussed.**

I. Introduction

The Internet of Things is a concept that has become the driving force behind technological action. The increasing need to simplify life, as well its improvement, is mobilized not only by companies offering various types of equipment, or software with the *smart* note, but also by researchers. They focus on developing particular aspects that are components of any technique that is later assimilated by the industry, and hence distributed to our homes. The most important topic in this topic were widely described in [6], [10], [19]. The authors focused on emerging issues that are necessary from the industrial point of view.

It is hard to tell which components in large systems are important. Therefore, all are treated equally and developed at a similar level. Each software is installed on specific devices and is a link between the user and hardware. In the case of systems under the sign *smart*, various sensors are used to acquire knowledge about the environment. An example is a motion sensor or a camera that records an image and then serves to find some deviation from the norm or the appearance of some movement. One of video processing idea was presented in [17], where the authors described video tamper detection by the application of multi-scale mutual data. Another sensors are microphones that record the sound and voice. Sound recording devices allow to receive voice commands that will be important especially for people with disabilities. At first, the analog signal must be converted to discrete one (because of processing by computers), so processing of the signal is critical issue. In [2], [13], [16], the idea of using discrete and wavelet transformation to obtain audio signal in the form ready to analyzes was shown. Again in [11], [12] extraction technique for a specific parts of signals was shown. The method was tested on some popular voice distortion like cough. It is useful in authorization systems when a record is created and for verification process, only first/last name is required.

Of course, these systems to operate data obtained from many sensors need some algorithm to gather all these information and process them. If the system works in real time, a lot of data will come in every second. And this means that the software will not be able to process all at the same time, hence the idea is to use parallelization or give certain weights to incoming data. Queuing service is a stochastic model according to which it can direct the data handling from the sensors. An example of such a model is shown in [18]. Large amounts of data need fast sorting algorithms not only for sorting, but for searching a specific information in database, where all incoming data are stored. One of the latest achievement in these area are algorithms which are merged with multi-threaded processor [8], [9]. Many of new methods are based on artificial intelligence like neural networks or swarm intelligence [7], [20], [22]. All of these mentioned components are necessary in large systems, but it also need security against uncontrolled access to data or computational processes. Important work in these area is presented in [14], where almost all aspects and challenges in internet of things are described and discussed.

In this paper, the idea of identity verification process is described with background about interpreting audio signal to a form that allows analysis.

II. Signal theory

The processed signal should be given in a discrete form. Especially when the operations are performed by a computer. In practice, having an analog signal should be changed to a discrete equivalent. Unfortunately, even such a version is practically not useful. For this purpose, the signal must undergo a certain transformation, which will transform it to the form possible in the analysis. One of the most known transformation is Fourier's one. Suppose that $s(n) = (s_0, s_1, s_2, \ldots, s_{N-1})$ is a signal. Transformation of such a set will give $(S_0, S_1, S_2, \ldots, S_{N-1})$, where $S_i \in \mathbb{C}$ and it is done by

$$S_k = \sum_{n=0}^{N-1} s_n \exp\left(-\frac{2\pi ink}{N}\right) \qquad 0 \le k \le N-1. \tag{1}$$

While the discrete Fourier transform allowed for calculations on various machines, the operation time was still too long. In 1960s, two American scientists – James W. Cooley and John W. Tukey presented *Fast Fourier Transform* [3], [4], which is a technique of calculation transformation using recursion and *division and rule* method . Whole idea is based on the division of functions into even and odd indices in the following way

$$\begin{aligned} S_k &= \sum_{n=0}^{N-1} s_n \exp\left(-\frac{2i\pi nk}{N}\right) \\ &= \sum_{m=0}^{N/2-1} s_{2m} \exp\left(-\frac{2i\pi k(2m)}{N}\right) \\ &+ \sum_{m=0}^{N/2-1} s_{2m+1} \exp\left(-\frac{2i\pi k(2m+1)}{N}\right) \\ &= \sum_{m=0}^{N/2-1} s_{2m} \exp\left(-\frac{2i\pi km}{N/2}\right) \\ &+ \exp\left(-\frac{2i\pi k}{N}\right) \sum_{m=0}^{N/2-1} s_{2m+1} \exp\left(-\frac{2i\pi km}{N/2}\right), \end{aligned} \tag{2}$$

It is possible to analyze the sound in graphic form, but for this purpose the signal should be saved in the form of a so-called short-time transform as

$$S\{s[n]\}(m,f) \equiv S(m,f) = \sum_{n=-\infty}^{\infty} s[n]w[n-m]\exp(-jfn). \tag{3}$$

Using above equation, the signal can be presented as a graph of the amplitude spectrum, which is determined as

$$spectrogram\{s(t)\}(t,f) \equiv |S(t,f)|^2. \tag{4}$$

Presenting the calculated values from Eq. 4 on $2D$ graph, we have points and their values. There are two axes – OX which means time and OY representing the frequency. The value of a given point is represented by the shade of color which is understood as a intensity. Sample graphs are shown in Fig. 1.

III. Pattern recognition

Let us consider spectrogram as a set of point (x, y) with intensity in the range $\langle 0, 1\rangle$. On the spectrogram, the most important features will have the brightest shade, so the intensity value will have the smallest values.

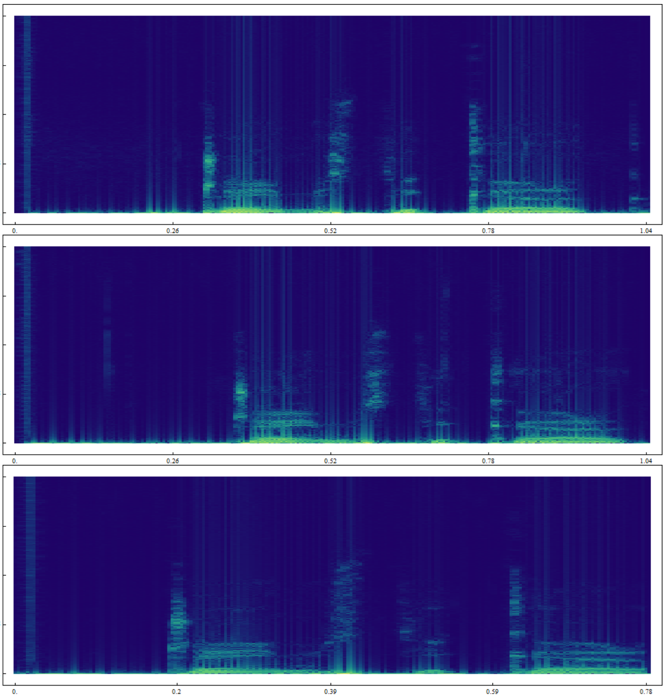

Figure 1: Spectrograms for three different samples belonging to one person pronouncing "*James Tiberius Kirk*".

At the beginning, let us focus on pattern creation process. The newly hired employee is asked to repeat his/her name at least 10 times. Each repetition is one recording. Then, 10 spectrograms are taken and used to create pattern based on these recordings. In ideal world, all samples should look similar. However, in practice it is not so easy because there can be worst quality of records, some noises and many other factors. For each sample, we find the value z_{max} with the lowest saturation. These values allows to define the μ which is a threshold value defined as

$$\mu = \begin{cases} 1.2 \cdot z_{max} & \text{if } \mu \neq 0 \\ 0.2 & \text{if } \mu = 0 \end{cases}. \tag{5}$$

Using μ value, it is possible to create a vector of the most characteristic points (with all points satisfied condition in Eq. (5)) in the following form

$$\xi^k = \left\{ \left(x_0^k, y_0^k, z_0^k\right), \left(x_1^k, y_1^k, z_1^k\right), \ldots, \left(x_m^k, y_m^k, z_m^k\right) \right\}, \tag{6}$$

where k is the number of a specific record, x_0^k and y_0^k are the point with the lowest saturation equal to z_0^k.

In this way, k sets will be created. All values are grouped by the k-nearest neighbors classifier to remove points at a short distance in each sets. A probability estimator is defined as

$$\hat{p}(k|x) = \frac{1}{K} \sum_{i=1}^{n} I(\rho(x, x_i) \le \rho(x, x^{(k)})) I(y_i = k) \qquad k = 1, \ldots, L, \tag{7}$$

where $\rho(\cdot)$ is metric, $x^{(k)}$ is k-th as to the distance to the point from the samples x. And using these, the classifier is formulated as

$$\hat{d}_{KNN}(x) = \arg\max_k \hat{p}(k|x). \tag{8}$$

After analyzing the points, there is a possible that sets will be have different numbers of elements. To fix it, sets will be pruned to the number describing the smallest set. Then, all sets ξ^k will create intervals for points in the pattern. Limits of these intervals are determined as

$$\begin{aligned}&\left\{\left(\left[\min\{x_0^k\},\max\{x_0^k\}\right],\left[\min\{y_0^k\},\max\{y_0^k\}\right],\right.\right.\\&\left.\left[\min\{z_0^k\},\max\{z_0^k\}\right]\right),\ldots,\left(\left[\min\{x_m^k\},\max\{x_m^k\}\right],\right.\\&\left.\left.\left[\min\{y_m^k\},\max\{y_m^k\}\right],\left[\min\{z_m^k\},\max\{z_m^k\}\right]\right)\right\},\\&\quad k\in\ \{1,2,\ldots,10\}\end{aligned} \tag{9}$$

IV. Experiments

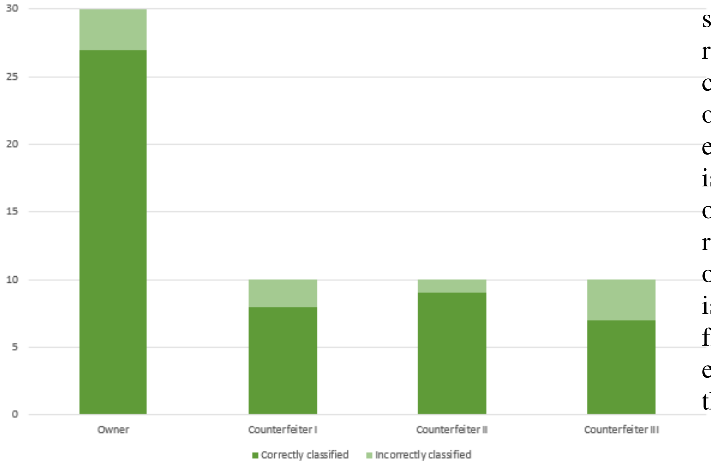

Figure 2: A graphical summary of the obtained results for all samples in the test database.

Proposed method was tested on a small dataset consisting only 60 samples, from which half of them contained the three words "*James Tiberius Kirk*" made by one person who is identified with this data (so called owner). The remaining 30 samples were created by three different people (so called counterfeiters), each of them has created 10 samples.

Using only 10 samples from the owner (selected randomly), pattern was modeled. Then, all samples in the database were checked for pattern match. If the compatibility was at least 80%, then it was marked as owner. Otherwise, the sample was marked as falsification.

The verification of the effectiveness of the proposed technique was examined by grouping the samples as TP (*true positive*), TN (*true negative*), FP (*false positive*), FN (*false negative*). For such divided results, accuracy was calculated as Γ, Dice's coefficient as Λ, overlap Ψ, sensitivity Υ and specificity Φ according to

$$\Gamma=\frac{TP+FN}{TP+TN+FP+FN}, \tag{10}$$

$$\Lambda=\frac{2TP}{2TP+FP+FN}, \tag{11}$$

$$\Psi=\frac{TP}{TP+FP+FN}, \tag{12}$$

$$\Upsilon=\frac{TP}{TP+FN}, \tag{13}$$

$$\Phi=\frac{TN}{TN+FP}. \tag{14}$$

Table I: Obtained solutions for voice recognition

TP	TN	FP	FN	Γ	Λ	Ψ	Υ	Ψ
27	3	6	24	0.85	0.64	0.47	0.53	0.33

The distribution of correctly and incorrectly classified samples is presented in Fig. 2, 3, 4, 5, 6. In the case of the owner, only 10% of correct samples were incorrectly classified. As the reason, some noise or recording time can be the issue. For samples made by three different counterfeiters, the average rate of fraud detection was 80% which is a good result considering the number of samples. A more detailed analysis of the measurements is shown in Tab. I, where the average effectiveness is 85%. Similarity coefficient reached 0.64 which is quite high value. However, it is worth noting that the obtained data should be contained within a fairly wide error range. Similarly with the other coefficients – the probability of obtaining a negative classification assuming that the sample is true is 0.33, and the probability of positive verification for fraud is 0.53. The obtained results indicate a high degree of effectiveness despite the number of sound samples as well as the extraction and classification technique itself.

V. Conclusions

In this paper, the idea of audio analysis based on the mechanism of pattern matching with k-nearest neighbors was presented. It is important to develop more different techniques for security due to the reduction in the number of calculations, simplifying the operation as well as increasing the precision of actions. This technique was implemented and tested on a small dataset consisting only 60 samples. Half of them belonged to the one person (called as owner), and the rest of them to three other people which were a forgery and used for verification purposes. Due to the noise and different recording times, the program incorrectly classified 10% of true records. However, it does not change the fact that the effectiveness of the proposed idea reached almost 85%. It is worth noting that it was tested for $k=4$, and increasing the number of neighbors resulted in a decrease in the correctness of classification, which may be due to the number of samples.

An important aspect of further research is increasing the database with a much larger number of recordings, increasing noises or problems with the voice of the recording person. It is particularly important to be able to bypass hoarseness or remove the cough. In the case of accuracy, the use of other, more complicated classification (like neural networks) methods may prove to be a much more favorable approach.

References

[1] F. Beritelli, G. Capizzi, G. L. Sciuto, C. Napoli, and F. Scaglione. Automatic heart activity diagnosis based on gram polynomials and probabilistic neural networks. *Biomedical Engineering Letters*, 8(1):77–85, 2018.

[2] D. Birvinskas, V. Jusas, I. Martisius, and R. Damasevicius. Eeg dataset reduction and feature extraction using discrete cosine transform. In *Computer Modeling and Simulation (EMS), 2012 Sixth UKSim/AMSS European Symposium on*, pages 199–204. IEEE, 2012.

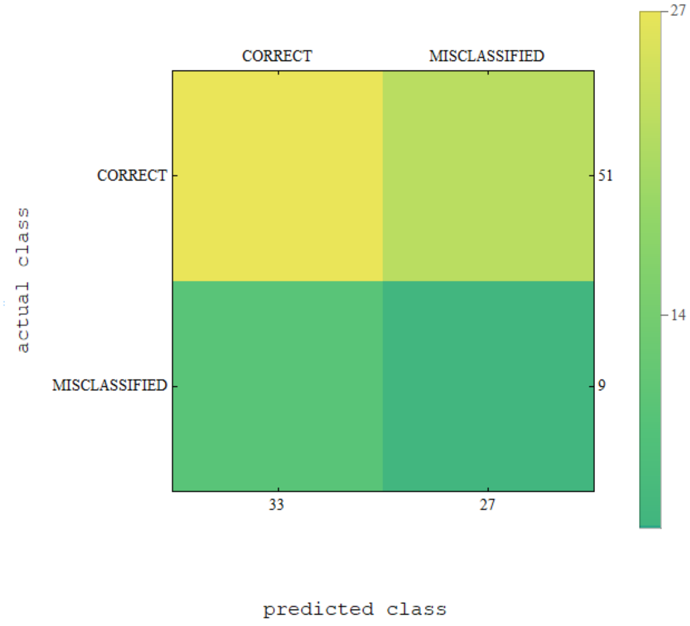

Figure 3: Confusion matrix for the owner's sample classification.

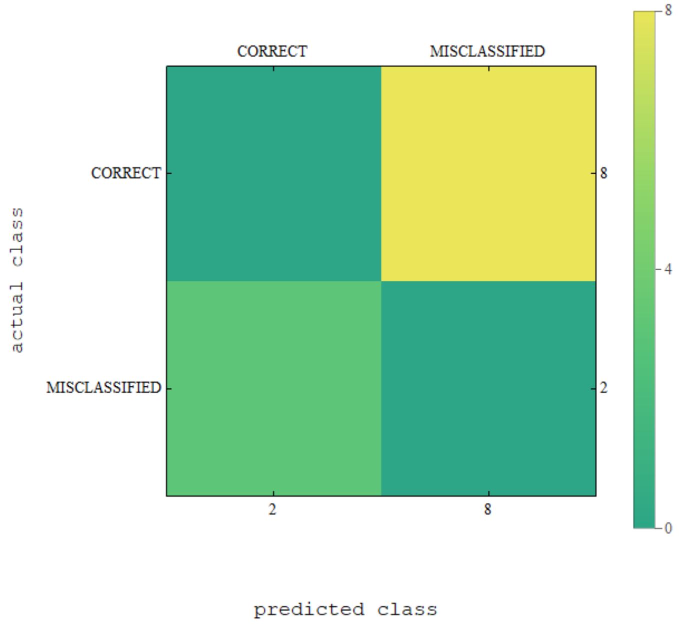

Figure 4: Confusion matrix for the first counterfeiter's sample classification.

[3] W. T. Cochran, J. W. Cooley, D. L. Favin, H. D. Helms, R. A. Kaenel, W. W. Lang, G. Maling, D. E. Nelson, C. M. Rader, and P. D. Welch. What is the fast fourier transform? *Proceedings of the IEEE*, 55(10):1664–1674, 1967.

[4] J. W. Cooley, P. A. Lewis, and P. D. Welch. The fast fourier transform and its applications. *IEEE Transactions on Education*, 12(1):27–34, 1969.

[5] R. Damaševičius, C. Napoli, T. Sidekerskienė, and M. Woźniak. Imf mode demixing in emd for jitter analysis. *Journal of Computational Science*, 22:240–252, 2017.

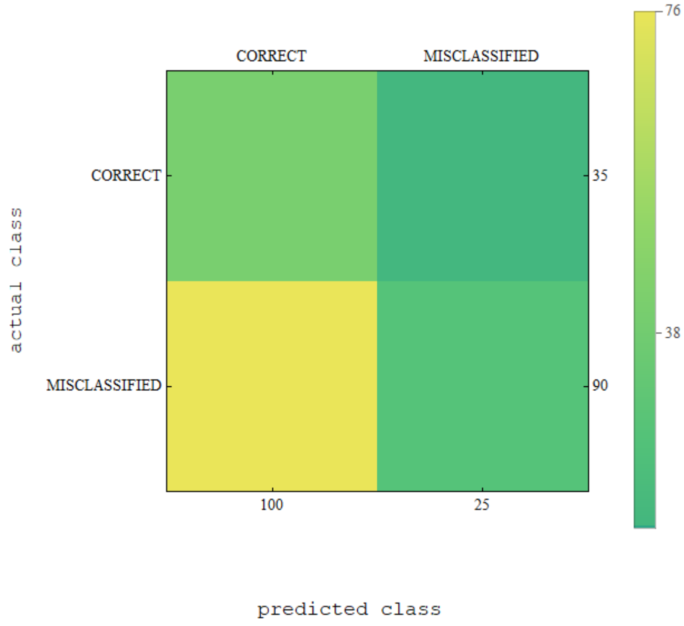

Figure 5: Confusion matrix for the second counterfeiter's sample classification.

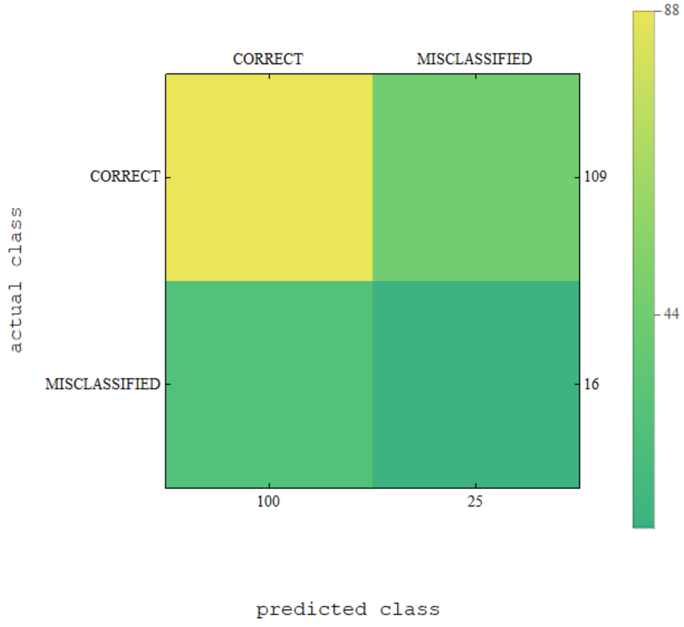

Figure 6: Confusion matrix for the third counterfeiter's sample classification.

[6] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami. Internet of things (iot): A vision, architectural elements, and future directions. *Future generation computer systems*, 29(7):1645–1660, 2013.

[7] P. N. Mahalle, P. A. Thakre, N. R. Prasad, and R. Prasad. A fuzzy approach to trust based access control in internet of things. In *Wireless Communications, Vehicular Technology, Information Theory and Aerospace & Electronic Systems (VITAE), 2013 3rd International Conference on*, pages 1–5. IEEE, 2013.

[8] Z. Marszałek. Parallelization of fast sort algorithm. In *International Conference on Information and Software Technologies*, pages 408–421.

Springer, 2017.

[9] Z. Marszałek. Parallelization of modified merge sort algorithm. *Symmetry*, 9(9):176, 2017.

[10] C. Perera, C. H. Liu, and S. Jayawardena. The emerging internet of things marketplace from an industrial perspective: A survey. *IEEE Transactions on Emerging Topics in Computing*, 3(4):585–598, 2015.

[11] D. Polap. Extraction of specific data from a sound sample by removing additional distortion. In *Computer Science and Information Systems (FedCSIS), 2017 Federated Conference on*, pages 353–356. IEEE, 2017.

[12] D. Połap and M. Woźniak. Extraction and analysis of voice samples based on short audio files. In *International Conference on Information and Software Technologies*, pages 422–431. Springer, 2017.

[13] N. Romano, A. Scivoletto, and D. Polap. A real-time audio compression technique based on fast wavelet filtering and encoding. In *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*, pages 497–502. IEEE, 2016.

[14] S. Sicari, A. Rizzardi, L. A. Grieco, and A. Coen-Porisini. Security, privacy and trust in internet of things: The road ahead. *Computer Networks*, 76:146–164, 2015.

[15] J. T. Starczewski, S. Pabiasz, N. Vladymyrska, A. Marvuglia, C. Napoli, and M. Woźniak. Self organizing maps for 3d face understanding. In *International Conference on Artificial Intelligence and Soft Computing*, pages 210–217. Springer, 2016.

[16] M. Vasiljevas, R. Turčinas, and R. Damaševičius. Development of emg-based speller. In *Proceedings of the XV International Conference on Human Computer Interaction*, page 7. ACM, 2014.

[17] W. Wei, X. Fan, H. Song, and H. Wang. Video tamper detection based on multi-scale mutual information. *Multimedia Tools and Applications*, pages 1–18, 2017.

[18] W. Wei, Q. Xu, L. Wang, X. Hei, P. Shen, W. Shi, and L. Shan. Gi/geom/1 queue based on communication model for mesh networks. *International Journal of Communication Systems*, 27(11):3013–3029, 2014.

[19] A. Whitmore, A. Agarwal, and L. Da Xu. The internet of things—a survey of topics and trends. *Information Systems Frontiers*, 17(2):261–274, 2015.

[20] M. Woźniak and D. Połap. Adaptive neuro-heuristic hybrid model for fruit peel defects detection. *Neural Networks*, 98:16–33, 2018.

[21] M. Wozniak, D. Polap, G. Borowik, and C. Napoli. A first attempt to cloud-based user verification in distributed system. In *Asia-Pacific Conference on Computer Aided System Engineering (APCASE)*, pages 226–231. IEEE, 2015.

[22] M. Woźniak, D. Połap, L. Kośmider, and T. Cłapa. Automated fluorescence microscopy image analysis of pseudomonas aeruginosa bacteria in alive and dead stadium. *Engineering Applications of Artificial Intelligence*, 67:100–110, 2018.