Journal of Universal Computer Science, vol. 24, no. 4 (2017), 444-459
submitted: 6/9/17, accepted: 15/4/18, appeared: 28/4/18 © J.UCS

How to Extract Interesting Information for Identity Verification Process from Spectrograms?

Kamil Książek, Karolina Kęsik and Zbigniew Marszałek
(Institute of Mathematics, Silesian University of Technology
Kaszubska 23, 44-100, Gliwice, Poland
KamilKsiazek95@gmail.com, Karola.Ksk@gmail.com
Zbigniew.Marszalek@polsl.pl)

Abstract: Nowadays, identity verification support systems are becoming more and more popular. Machine learning is one of the leading fields of research from all over the world. However, each classifier needs a large number of samples to be properly trained. Preparing such samples proves to be a big problem for several reasons. One of them is the quality of the recording, another is the problem of feature extraction. In this work, the idea of processing sound samples by using their graphical representation in the form of spectrograms is described. The process removes specific, redundant information from the samples and then performs feature extraction. The proposed technique has been tested for identity verification using convolutional neural networks. The performed tests and obtained results have been described and discussed to indicate numerous advantages and disadvantages of the proposed technique.

Key Words: identity verification, spectrogram, convolutional neural network, sample data

Category: F.2.0, F.3.1, H.5.1

1 Introduction

Speaker identification and verification systems are nowadays very important. In order to improve safety of users high efficiency of these methods is crucial. This is particularly important when the user confirms financial transactions. For this purpose both image and sound shall be used. Some biometric methods (e.g. iris recognition or fingerprint identification) are commonly applied. Ubiquitous electronics require well-functioning security systems [Jain et al. (2006)]. Almost every corporation needs employee identification system. Also during card transactions security and proper user verification is necessary. Another important way of verification is voice recognition. Many financial operations are done over the phone.

Identification consists in comparing the speaker with samples from database and choosing the best one (the most appropriate) [Kinnunen et. al. (2006)]. Verification is something another: it is ascertaining that given voice comes from a particular speaker. Verification is applied during Internet payments. Some biometric characteristic of a person could be additional security.

There exist some tools helpful in time-frequency analysis, i.e. spectrogram, periodogram or scalogram. In this paper we present extracting crucial information from spectrograms. Our novelty is proposition of removing noise and image approach in verification of sound samples. Spectrogram is a graphic representation of the spectrum of frequencies (dependent of time) of the given signal. In two-dimensional space X-coordinate (horizontal axe) represents time and Y-coordinate (vertical axe) frequency. During time-frequency analysis it is necessary to dump unwanted noises. After extracting key data, there will be carried out experiments by using convolutional neural networks. Neural networks are advancing field of artificial intelligence. They are used commonly during research concerning about pattern recognition. In this paper, identity verification technique based on image representation of audio file and the use of convolutional neural networks is presented, tested and discussed.

2 Related works

Sound identification and verification models are widely discussed by many researchers over the world. In [Melov et. al. (2017)] an online call recording diarisation system is presented. Its purpose is to identify a speaker by the call-centre operators. Some human characteristics are the basis of biometrics. In [Al-Kaltakchi et. al. (2017)] the authors present a speech biometric I-vector which is used in identification process. Extreme Learning Machine classifer ensures high throughput. Text-independent speaker identification system by using probabilistic method is presented in [Ma et. al. (2016)]. The authors transform histograms to estimate the probability density function (PDF). They called the system as super-mel-frequency cepstral coefficients (super-MFCCs). Results show that the efficiency of their novel approach is better than Gaussian mixture model. In [Yong et. al. (2016)] authors present speaker identification system performed on a large scale of voiceprint corpus (about 400 thousand people). An interesting application of identification is presented in [Woo et. al. (2016)]. The authors try to identify the speakers in TV news. Measurements demonstrate over 60% of effectiveness in case of speakers from CNN News. Authors of [Lukic et. al. (2016)] described convolutional neural networks and their design for speaker identification. The main instrument (input of neural networks) during experiments were spectrograms.

In [Viet et. al. (2017)] authors present a verification system which assesses if the speaker comes from known set. They employ also well-known MFCC and Gaussian mixture model (GMM). Also in [Boles et. al. (2017)] authors describe system based on MFCC and test on two databases. Moreover, in [Dhanush et. al. (2017)], discussion about decreasing efficiency of verification systems due to spoofing attacks is described. They propose a model useful in

spoof detection. Similarly in [Ergunay et. al. (2015)] authors describe risk connected with spoofing attacks in automatic speaker verification systems. They introduce the voice spoofing database and show impact of attacks on verification systems. Gender-dependent models and their advantages in comparison to gender-independent one are described in [Kanervisto et. al. (2017)]. The idea of using image approach to audio processing is described in [Polap (2016), Polap (2017)] where hybrid technique based on heuristic and neural methods are presented. Text-independent speaker verification by using probabilistic linear discriminant analysis (PLDA) is presented in [Chen et. al. (2017)]. Similarly, other approach about text-independent speaker verification and non-parallel voice conversion by using I-vector method and PLDA is concluded in [Kinnunen et. al. (2017)]. Also in [Snyder et. al. (2016)] authors discuss about text-independent speaker verification based on deep neural network. In [Ma et. al. (2017)] authors present technique normalizing the distribution mismatch. Duration mismatch compounds speaker verification in case of standard methods. An interesting approach by analyzing jitter and shimmer in speaker verification is inserted in [Polacky et. al. (2016)]. Authors of [Chen et. al. (2017)] engage very important topic: safety of smart-phones users. Their approach based on magnetic field coming from loudspeakers. Presented system achieves high precision. The widely held topic proves that speech recognition is an important subject of research in recent years.

In [Wlodarczyk et. al. (2017)] there was described system of mobile navigation for inline shipping by using Douglas-Peucker algorithm. The main task was simplification of lines and features of objects. A method helpful in determining depth of water (bathymetric data) is presented in [Wlodarczyk et. al. (2016)]. The main stages are clustering and generalization of data. There were compared some methods applied in clustering, among others neural networks. The authors of [Wei et. al. (2017)] take up the topic of video frame manipulation. Their paper presents a very effective method of detection position of video tampering (like copy or delete). In [Wei et. al. (2011)] was described a method based on multi-scale gradient descent to improve accuracy of navigation in wireless sensor networks (WSNs). Model of electric drive engine vehicle was shown in [Wozniak et. al. (2017)]. Simulation is based on intelligent calculations with application of neural networks. It ensures safety of transported load during driving over the terrain with obstacles.

3 Audio file processing

The recorded audio file for any decision support system must contain information indicating these particular person. In most cases, the type of such data will be the name of that person. Record a sound sample will be carried out when

the recording button will be turned on and stopped when it will be turned off. This will cause that the sample will contain more unnecessary information such as possible noise, other voices and even quiet intervals. Moreover, each of these factors may result in a lack of proper identification or even extend the calculation process performed by the machine. For this purpose, the sound sample is presented in the graphical form called spectrogram and then the extraction of only the first and last name is performed.

To be able to global analyze of the signal, some preparation needs to be done. The use of Fourier transform will not allow for a simple statement related to changes in frequency composition over time. It is a limitation, although a Short-Time Fourier Transform ($STFT$) can be used to move a window of a particular length and this will allow global analysis. For a given signal in discrete way $x[n]$, $STFT$ can be formulated as

$$\begin{aligned} STFT\{x[n]\}(m,\omega) &\equiv X(m,\omega) \\ &= \sum_{n=-\infty}^{\infty} x[n]w(n-m)\exp(-j\omega n), \end{aligned} \tag{1}$$

where $w(n)$ is a function described the Hann window in the following manner

$$w(n) = 0.5\left(1-\cos\left(\frac{2\pi n}{N-1}\right)\right). \tag{2}$$

$STFT$ allows to create a spectrogram, i.e. a plot of amplitude over the time as

$$spectogram\{x(t)\}(\theta,\omega) \equiv |X(\theta,\omega)|^2, \tag{3}$$

which can be seen in Fig. 1. A graphs show the amplitude spectrum for a given time point. Analysis of spectrogram is based on the location of the darkest points, that is, places where the value of the spectrum is the greatest.

Of course, recorded samples may contain some imperfections such as too long recording time, coughing or stuttering. To remove such a distortion, we propose the use of image processing. Suppose we have many samples for one person presented as spectrograms. On each graph, we will search for key points using the SURF (*Speeded Up Robust Features*) algorithm. It was proposed in [Bay et al. (2006)] as a tool to find important points on images. It is based on the calculation of the Hessian matrix as

$$H(x,\omega) = \begin{bmatrix} L_{xx}(x,\omega) & L_{xy}(x,\omega) \\ L_{xy}(x,\omega) & L_{yy}(x,\omega) \end{bmatrix}, \tag{4}$$

where $L_{xx}(x,\omega)$, $L_{xy}(x,\omega)$ and $L_{yy}(x,\omega)$ are the convolution with the second derivative of the Gaussian defined as

$$L_{xx}(x,\omega) = I(x)\frac{\partial^2}{\partial x^2}g(\omega), \tag{5}$$

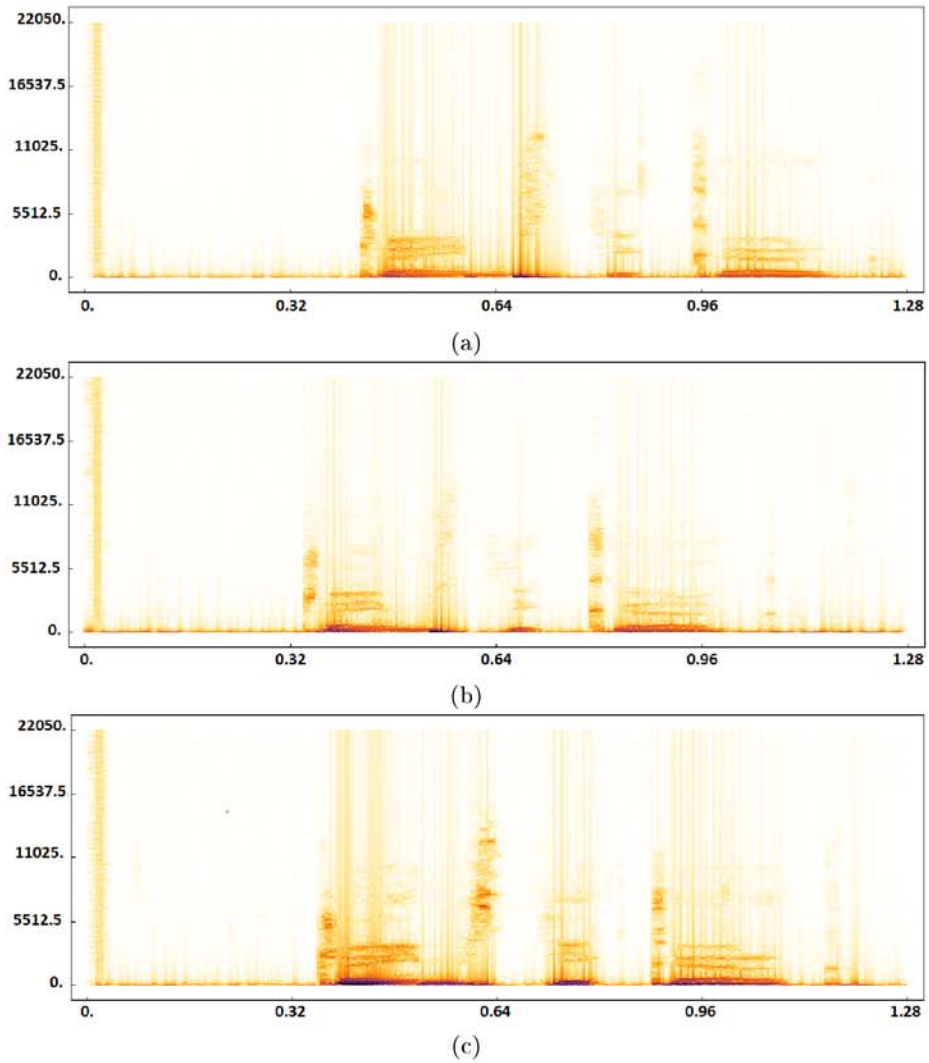

Figure 1: Sample spectrograms for sound samples with the sentence *James Tiberius Kirk*.

$$L_{yy}(x,\omega) = I(x)\frac{\partial^2}{\partial y^2}g(\omega), \tag{6}$$

$$L_{xy}(x,\omega) = I(x)\frac{\partial^2}{\partial x \partial y}g(\omega), \tag{7}$$

where $g(\omega)$ is the Gaussian kernel. And as $I(x)$, an integral image is understood, where x is the value described the sum of all pixel in the neighborhood calculated

as

$$I(x) = \sum_{i=0}^{i \leq x} \sum_{j=0}^{j \leq y} I(x, y). \tag{8}$$

The idea of SURF algorithm is based on non-maximal-suppression of determinant of Hessian matrix, what is described by

$$det(H_{approximate}) = L_{xx}L_{xy} - (wL_{xy})^2, \tag{9}$$

where w is the weight. This determinant helps to find the extremes values which are considered to be a key-points.

In this way, the key-points are found. Then, for each sample with the points, an empty bitmap is created. All key-points are transfered as a circle – the point is a center and r is a radius. All points in the circle are black ones, the rest is white. Based on such bitmaps, a general pattern for a particular person can be created. Let us define a minimum value ϕ that will be required to transfer a pixel to a pattern. The value is a count of black pixels in a given position on all bitmaps. Extraction of important information and removal the others is made by template matching technique. For a given sample, the pattern is moved in the search for the best fit. If at least 80% of the points are covered, the sample is truncated to the pattern size. This way, the information around the relevant are deleted.

4 Feature extraction

Described in previously section, method of preparing sample can be used for feature extraction. Let us describe a matrix defining a single feature of an image as

$$\begin{bmatrix} I_p(x-1,y-1) & I_p(x,y-1) & I_p(x+1,y-1) \\ I_p(x-1,y) & I_p(x,y) & I_p(x+1,y) \\ I_p(x-1,y+1) & I_p(x,y+1) & I_p(x+1,y+1) \end{bmatrix}, \tag{10}$$

where (x, y) is the key-point found on the sample by the use of SURF algorithm and $I_p(\cdot)$ is a pixel value with all color components.

So defined feature matrix is the size 3×3. Suppose, that SURF algorithm will returned at top k points, so k points make k feature matrix. It gives $9k$ pixels that can be combined into one, large image of the size $3k \times 3k$. This image will store all intensity values for each area, which can be used in the classification process.

5 Identity verification with convolutional neural network

Features saved in a fixed size image allow to apply the convolutional neural network (CNN) [Matsugu et al. (2003)] as classification tool in identity verification.

The model of CNN is inspired by the actions of the primary brain cortex. Unlike classic neural classifiers, the network has other types of layers. The first of layer is the convolution layer which performs feature extraction from image file. The layer is represented by 3–dimensional system of neurons stretched on 3 axes – width, height and depth understood as the average values of the filters. Suppose, ω is a matrix filter of size $m \times m$ that modified the image. The matrix is moved over the image with S steps. Layer's size is depend on the image (image size is $N \times N$), and it can be defined as

$$s_{output} = \frac{N-m}{S} + 1, \tag{11}$$

where s_{output} is the size of the output values from the whole network. In each layer l, for neuron x_{ij}, the sum of all inputs from the previous layer $l-1$ is calcualted and multiplied by the filter matrix as

$$x_{ij}^{l} = \sum_{a=0}^{m-1} \sum_{b=0}^{m-1} \omega_{ab} y_{(i+a)(j+b)}^{l-1}. \tag{12}$$

And the output value is calculated using

$$y_{ij}^{l} = \sigma(x_{ij}^{l}). \tag{13}$$

where $\sigma(x)$ is an activation function (for instance the sigmoid one).

Next type of layer is known as pooling one which reduce the size of the incoming image from previously layer. It works on the idea of maximum. For filter ω it takes only the maximum value from numbers covered by the filter and replaces the rest with the selected one. The third type of layer – called fully connected – is the representation of classic neural network. Each incoming pixel is treated as value for one neuron in the described layer called hidden. All these values are transferred to next layer, that is the output one.

Moreover, described network needs proper learning algoirthm. For this purpose, the backpropagation algorithm was created which is one of the most know technique of machine learning. As $f(\cdot)$ we define an error function, and $\frac{\partial f}{\partial y_{ij}^{l}}$ is the neuron output at position i, j in the layer l. The errors $\frac{\partial f}{\partial y_{ij}^{l}}$ are known (when we are in output layer). The whole algorithm is based on sharing the weight what is defined by a chain rule as

$$\frac{\partial f}{\partial \omega_{ab}} = \sum_{i=0}^{N-m} \sum_{j=0}^{N-m} \frac{\partial f}{\partial x_{ij}^{l}} \frac{\partial x_{ij}^{l}}{\partial \omega_{ab}} = \sum_{i=0}^{N-m} \sum_{j=0}^{N-m} \frac{\partial f}{\partial x_{ij}^{l}} y_{(i+1)(j+b)}^{l-1}. \tag{14}$$

To find the gradient value, an error $\frac{\partial f}{\partial x_{ij}^{l}}$ must be calculated using

$$\frac{\partial f}{\partial x_{ij}^{l}} = \frac{\partial f}{\partial y_{ij}^{l}} \frac{\partial y_{ij}^{l}}{\partial x_{ij}^{l}} = \frac{\partial f}{\partial y_{ij}^{l}} \frac{\partial \left(\sigma(x_{ij}^{l})\right)}{\partial x_{ij}^{l}} = \frac{\partial f}{\partial y_{ij}^{l}} \sigma'(x_{ij}^{l}), \tag{15}$$

where $\sigma(x)$ is the sigmoid function.

They were formulas for calculating error in current layer but the errors must be transferred to the previous one and it is done by

$$\frac{\partial f}{\partial y_{ij}^{l-1}} = \sum_{a=0}^{m-1}\sum_{b=0}^{m-1} \frac{\partial f}{\partial x_{(i-a)(j-b)}^{l}} \frac{\partial x_{(i-a)(j-b)}^{l}}{\partial y_{ij}^{l-1}} = \\ = \sum_{a=0}^{m-1}\sum_{b=0}^{m-1} \frac{\partial f}{\partial x_{(i-a)(j-b)}^{l}} \omega_{ab}. \tag{16}$$

Using Eq. (16), an error value may be defined as

$$\frac{\partial x_{(i-a)(j-b)}^{l}}{\partial y_{ij}^{l-1}} = \omega_{ab}. \tag{17}$$

Verification process is understood as the prepering dataset and create features matrices defined in Eq. (4), which are used to learn and test the classifier.

In our case, as a kernel (ω) we have used Gaussian blur 3 x 3 expressed as:

$$\frac{1}{16} \cdot \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix} \tag{18}$$

During classification by using convolutional neural networks we have applied following parameters: 4 convolutional layers, 4 pooling ones (during sampling), 1 hidden and 1 output layer in fully connected ones.

6 Experiments

For testing purposes, a database of 225 samples was created with the sentence *"James Tiberius Kirk"*. 200 samples came from one person and half of them were perfect samples without any imperfections, 25 contained coughing, 25 were shouted, 25 involved hoarseness and 25 stuttering. The remaining 25 samples were pronounced by others people – 5 files for each group. Samples from one person were used to validate the proposed technique of remove imperfections and then all samples were used to train CNN at 80% : 20% (learn to validate) and minimal error was set at the level of 0.001.

The obtained results from extraction technique are presented in Fig. 2. Proposed technique using SURF algorithm gave very good results and average efficiency was reached at 79%. For the stuttering samples, the worst results were obtained which were at level of 44%. For comparison, any other inaccuracies reached at least a level 84%. The same situation were achieved with the verification process. In Fig. 3 results for individual distortions in samples are presented.

Despite the low score for stuttering samples, average efficiency of neural classifier was approximately 82.7%.

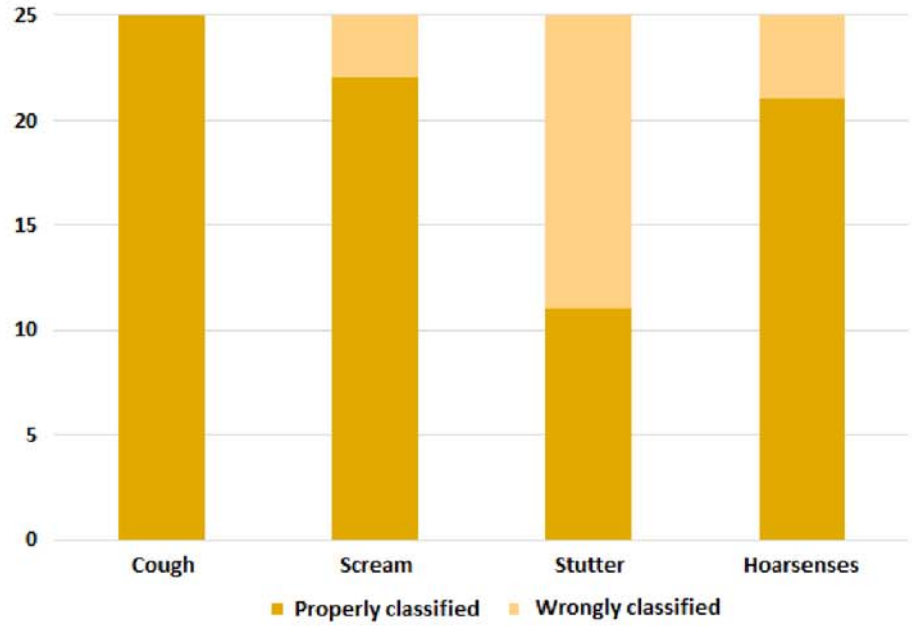

Figure 2: Data extraction results due to selected voice imperfections.

To assess quality of our test we propose some statistical measures. First of them: Γ is called accuracy – it is the proportion of the sum of correctly identified as positive (TP) and correctly classified as negative instances (TN) to the sum of the all instances in the set:

$$\Gamma = \frac{TP + TN}{TP + TN + FP + FN}. \tag{19}$$

False positive (FP) means that result of a measurement is false classified as positive. Inversely, false negative (FN) measure indicates that the occurrence is negative but really it is positive.

The another statistical metric is Dice's coefficient (Λ) – rate of similarity between sets. It is a kind of comparison between true cases classified by test and real true indicates.

$$\Lambda = \frac{2TP}{2TP + FP + FN}, \tag{20}$$

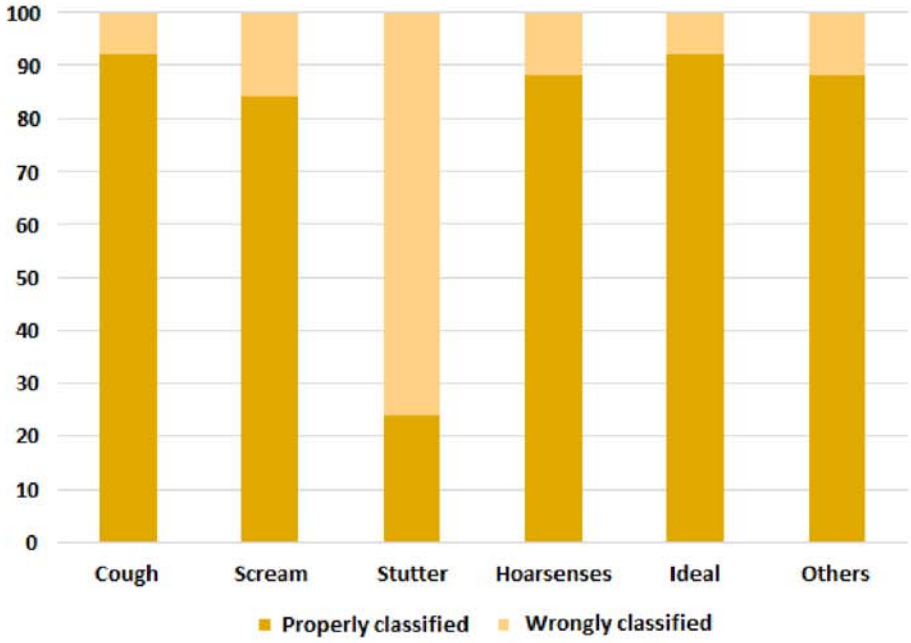

Figure 3: Classification results for all samples by CNN.

Ψ is very similar to Λ and is called the Jaccard index (JAC). JAC is always greater than Dice's coefficient (except 0 and 1). This index measures degree of overlap of two sets.

$$\Psi = \frac{TP}{TP + FP + FN}, \tag{21}$$

The next calculated metric (Υ) is called sensitivity or TPR (True Positive Rate). It represents proportion of true positive indicates to the combined sum of positive conditions.

$$\Upsilon = \frac{TP}{TP + FN}, \tag{22}$$

Φ, the last measure expresses specificity (called also True Negative Rate, TNR) that is the ratio of true negative indicates to the total number of negative ones.

$$\Phi = \frac{TN}{TN + FP}. \tag{23}$$

The results are presented in Table 1.

Table 1: The results.

	Γ	Λ	Ψ	Υ	Φ
cough	0.74	0.84	0.74	0.79	0
scream	0.7	0.82	0.69	0.79	0.69
stutter	0.3	0.36	0.22	0.69	0.16
hoarseness	0.74	0.84	0.73	0.81	0.25
ideal	0.89	0.94	0.88	0.96	0.11

One can see that accuracy was at least 0.7 (except stutter). The statistical tests confirmed that the most difficult obstacle for examined algorithm was stutter. Only sensitivity was acceptable. In other cases Dice's coefficient indicates that similiarity between scores given by the algorithm and real results is very high. Also Jaccard index shows high size of the intersection in four of five cases. Statistical tests claimed high rate of sensitivity - between 0.69 and 0.96. A bit worse were results specificity. Figures 4–8 show additional results - there are presented confusion matrices.

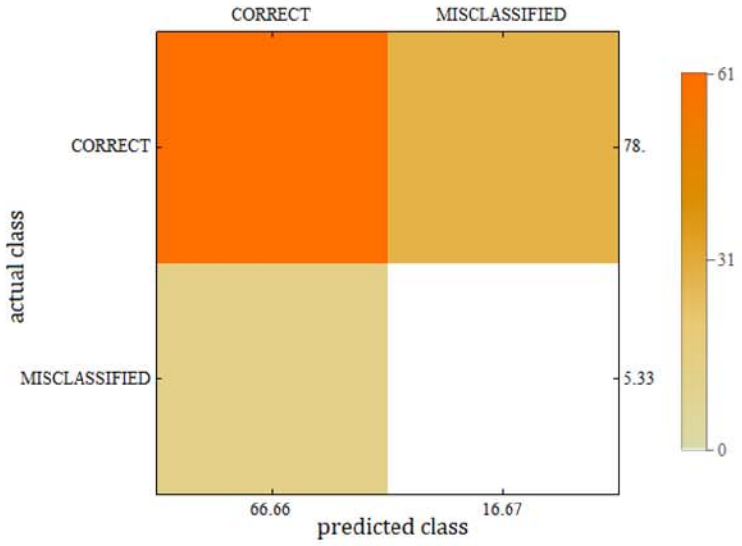

Figure 4: Confusion matrix for samples with cough

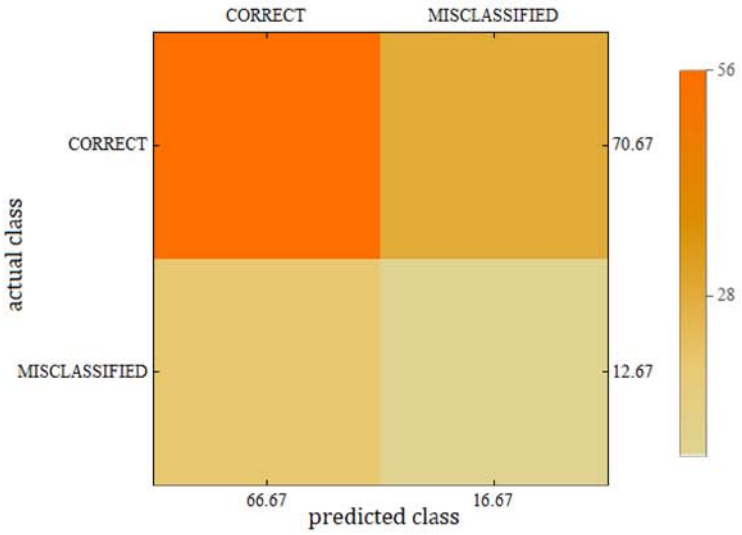

Figure 5: Confusion matrix for samples with scream

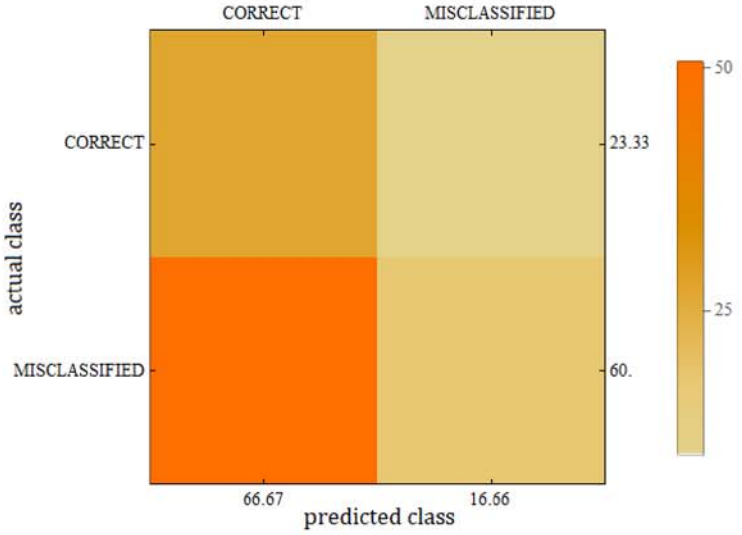

Figure 6: Confusion matrix for samples with stutter

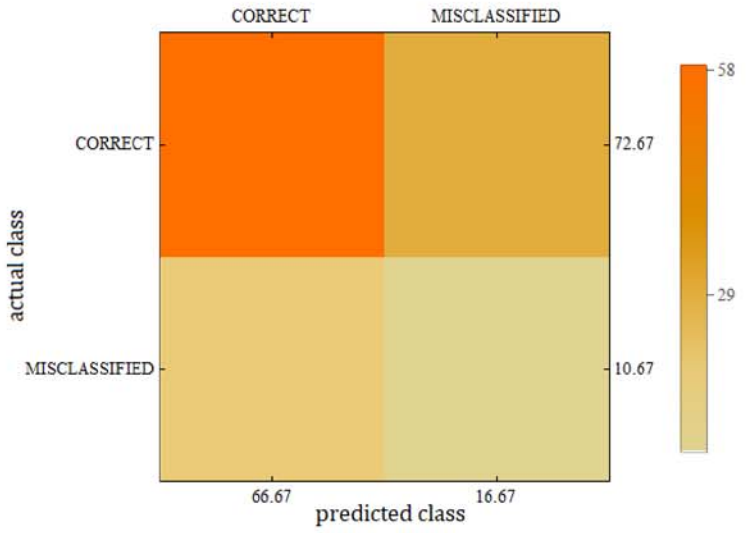

Figure 7: Confusion matrix for samples with hoarseness

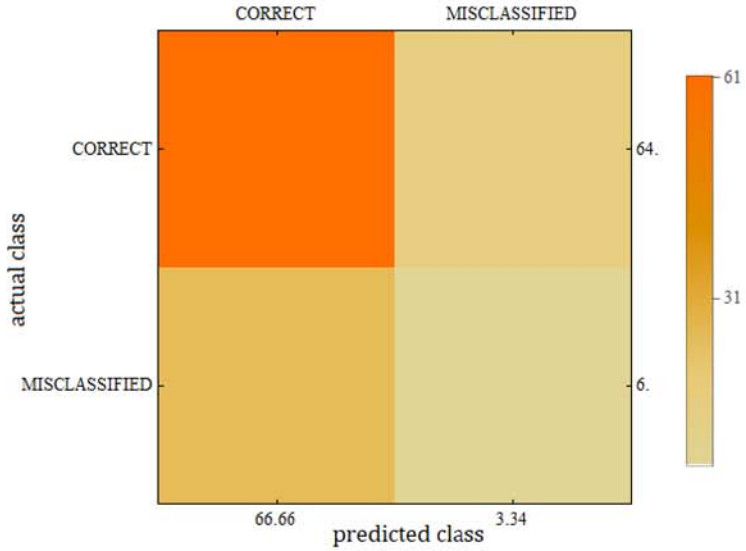

Figure 8: Confusion matrix for samples without imperfections

7 Conclusions

Presented technique of removal unnecessary data from audio sample based on spectrograms and image processing approach was presented. The obtained results were tested as the sample data for identity verification process using convolutional neural network. Despite the difficulty with stuttering, the obtained average results reached almost 83% of effectiveness. It is a high scores when we take into consideration the small number of samples in database. In future, we plan to use others key-point search algorithm to find the best one. We would like to also improve our method and compare described idea based on convolutional neural network with the other ones, typically used in verification process. Moreover, it seems be a good idea to use other image representation of sounds like scalograms and test the described method.

References

[Bay et al. (2006)] Bay, Herbert and Tuytelaars, Tinne and Van Gool, Luc, *Surf: Speeded up robust features*, Computer vision–ECCV 2006, Sprigner, 2006, pp. 404–417.

[Matsugu et al. (2003)] Matsugu, Masakazu and Mori, Katsuhiko and Mitari, Yusuke and Kaneda, Yuji, *Subject independent facial expression recognition with robust face detection using a convolutional neural network*, Neural Networks, vol. 16, no. 5, pp. 555–559.

[Jain et al. (2006)] Jain, Anil and Bolle, Ruud and Pankanti, Sharath, *Biometrics: personal identification in networked society*, Springer Science & Business Media, 2006, vol. 479.

[Kinnunen et. al. (2006)] Kinnunen, Tomi and Karpov, Evgeny and Franti, Pasi, *Real-time speaker identification and verification*, IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 1, pp. 277–288, 2006.

[Melov et. al. (2017)] Melov, Aleksandar and Gerazov, Branislav and Ivanovski, Zoran, *Delay based optimisation of an integrated online call recording speaker diarisation and identification system*, in: Smart Technologies, IEEE EUROCON 2017 - 17th International Conference, pp. 307–311, 2017.

[Al-Kaltakchi et. al. (2017)] Al-Kaltakchi, Musab TS and Woo, Wai L and Dlay, Satnam S and Chambers, Jonathon A, *Speaker identification evaluation based on the speech biometric and i-vector model using the TIMIT and NTIMIT databases*, Biometrics and Forensics (IWBF), 2017 5th International Workshop on, IEEE, pp. 1–6, 2017.

[Ma et. al. (2016)] Ma, Zhanyu and Yu, Hong and Tan, Zheng-Hua and Guo, Jun, *Text-Independent Speaker Identification Using the Histogram Transform Model*, IEEE Access vol. 4, pp. 9733–9739.

[Yong et. al. (2016)] Yong, Feng and Xinyuan, Cai and Ruifang, Ji, *Evaluation of the deep nonlinear metric learning based speaker identification on the large scale of voiceprint corpus*, Chinese Spoken Language Processing (ISCSLP), 2016 10th International Symposium, IEEE, pp. 1–4, 2016.

[Woo et. al. (2016)] Woo, Daniel N and Aygün, Ramazan S, *Unsupervised speaker identification for TV news*, IEEE MultiMedia, vol. 23, no. 4, pp. 50–58, 2016.

[Lukic et. al. (2016)] Lukic, Yanick and Vogt, Carlo and Dürr, Oliver and Stadelmann, Thilo, *Speaker identification and clustering using convolutional neural networks*, Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop, pp. 1–6, 2016.

[Viet et. al. (2017)] Viet, Quoc Nguyen and Tran, Bao Hung and Phuong, Bang Nguyen and Vu, Duc Lung, *A combination of Gaussian Mixture Model and Support Vector Machine for speaker verification*, Medical Measurements and Applications (MeMeA), 2017 IEEE International Symposium, pp. 432–436, 2017.

[Boles et. al. (2017)] Boles, Andrew and Rad, Paul, *Voice biometrics: Deep learning-based voiceprint authentication system*, System of Systems Engineering Conference (SoSE), IEEE, pp. 1–6, 2017.

[Dhanush et. al. (2017)] Dhanush, BK and Suparna, S and Aarthy, R and Likhita, C and Shashank, D and Harish, H and Ganapathy, Sriram, *Factor analysis methods for joint speaker verification and spoof detection*, Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference, pp. 5385–5389, 2017.

[Ergunay et. al. (2015)] Ergünay, Serife Kucur and Khoury, Elie and Lazaridis, Alexandros and Marcel, Sébastien, *On the vulnerability of speaker verification to realistic voice spoofing* Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference, pp. 1–6, 2015.

[Kanervisto et. al. (2017)] Kanervisto, Anssi and Vestman, Ville and Sahidullah, Md and Hautamäki, Ville and Kinnunen, Tomi, *Effects of gender information in text-independent and text-dependent speaker verification*, Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference, pp. 5360–5364, 2017.

[Chen et. al. (2017)] Chen, Liping and Lee, Kong Aik and Ma, Bin and Ma, Long and Li, Haizhou and Dai, Li-Rong, *Adaptation of PLDA for multi-source text-independent speaker verification*, Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference, pp. 5380–5384, 2017.

[Kinnunen et. al. (2017)] Kinnunen, Tomi and Juvela, Lauri and Alku, Paavo and Yamagishi, Junichi, *Non-parallel voice conversion using i-vector PLDA: towards unifying speaker verification and transformation*, Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference, pp. 5535–5539, 2017.

[Snyder et. al. (2016)] Snyder, David and Ghahremani, Pegah and Povey, Daniel and Garcia-Romero, Daniel and Carmiel, Yishay and Khudanpur, Sanjeev, *Deep neural network-based speaker embeddings for end-to-end speaker verification*, Spoken Language Technology Workshop (SLT), 2016 IEEE, pp. 165–170, 2016.

[Ma et. al. (2017)] Ma, Jianbo and Sethu, Vidhyasaharan and Ambikairajah, Eliathamby and Lee, Kong Aik, *Duration compensation of i-vectors for short duration speaker verification*, Electronics Letters, vol. 53, no. 6, IET, pp. 405–407, 2017.

[Polacky et. al. (2016)] Polacky, Jozef and Chmulik, Michal and Jarina, Roman, *Prosodic and voice quality features for speaker verification over coded channel*, Telecommunications and Signal Processing (TSP), 2016 39th International Conference, IEEE, pp. 327–330, 2016.

[Chen et. al. (2017)] Chen, Si and Ren, Kui and Piao, Sixu and Wang, Cong and Wang, Qian and Weng, Jian and Su, Lu and Mohaisen, Aziz, *You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones*, Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference, pp. 183–195, 2017.

[Polap (2016)] D. Połap, *Neuro-heuristic voice recognition*, Computer Science and Information Systems (FedCSIS), 2016 Federated Conference, IEEE, pp. 487–490, 2016.

[Polap (2017)] D. Połap, *Extraction of specific data from a sound sample by removing additional distortion*, Computer Science and Information Systems (FedCSIS), 2017 Federated Conference, IEEE, pp. 357–360, 2017.

[Wozniak et. al. (2017)] M. Wozniak, D. Połap, *Hybrid neuro-heuristic methodology for simulation and control of dynamic systems over time interval*, https://doi.org/10.1016/j.neunet.2017.04.013, Neural Networks, vol. 93, pp. 45–56, 2017.

[Wei et. al. (2017)] Wei Wei and Xunli Fan and Houbing Song and Huihui Wang, *Video tamper detection based on multi-scale mutual information*, Multimedia Tools and Applications, pp. 1–18, 2017.

[Wlodarczyk et. al. (2017)] Włodarczyk-Sielicka, Marta and Bodus-Olkowska, Izabela, *Simplification methods for line and polygon features in mobile navigation systems for inland waters*, 50 Scientific Journals of the Maritime University of Szczecin, Zeszyty Naukowe Akademia Morska w Szczecinie no. 50, pp. 105–111, 2017.

[Wlodarczyk et. al. (2016)] Włodarczyk–Sielicka, Marta and Stateczny, Andrzej *Clustering bathymetric data for electronic navigational charts*, The Journal of Navigation, vol. 69, no. 5, Cambridge University Press, pp. 1143–1153, 2016.

[Wei et. al. (2011)] , Wei, Wei and Qi, Yong, *Information Potential Fields Navigation in Wireless Ad-Hoc Sensor Networks*, Sensors, vol. 11, no. 5, ISSN 1424-8220, DOI: 10.3390/s110504794, pp. 4794–4807, 2011.