Centrum Rzeczoznawstwa Budowlanego

Inżynieria Bezpieczeństwa Obiektów Antropogenicznych 4 (2023) 26-34
https://inzynieriabezpieczenstwa.com.pl/

Inżynieria Bezpieczeństwa Obiektów Antropogenicznych

Behavioral features of the speech signal as part of improving the effectiveness of the automatic speaker recognition system

Dominik MAŁY[1]*, Andrzej P. DOBROWOLSKI[1]

[1] *Military University of Technology, Warsaw, Poland*

Abstract

The current reality is saturated with intelligent telecommunications solutions, and automatic speaker recognition systems are an integral part of many of them. They are widely used in sectors such as banking, telecommunications and forensics. The ease of performing automatic analysis and efficient extraction of the distinctive characteristics of the human voice makes it possible to identify, verify, as well as authorize the speaker under investigation. Currently, the vast majority of solutions in the field of speaker recognition systems are based on the distinctive features resulting from the structure of the speaker's vocal tract (laryngeal sound analysis), called physical features of the voice. Despite the high efficiency of such systems - oscillating at more than 95% - their further development is already very difficult, due to the fact that the possibilities of distinctive physical features have been exhausted. Further opportunities to increase the effectiveness of ASR systems based on physical features appear after additional consideration of the behavioral features of the speech signal in the system, which is the subject of this article.

This article was funded by the Military University of Technology as part of the UGB 866 project.

Keywords: automatic speaker recognition, automatic speaker recognition systems, physical features, behavioral features, speech signal

1 Introduction

Speech signal generation, processing and analysis are an everyday occurrence for everyone. What is judged by people as common and trivial only after closer analysis turns out to be an extremely complex process. Attempts to transfer human speech analysis capabilities to a computer show the scale of the complexity of this endeavor. The

* **Corresponding author:** E-mail address: (dominik.maly@wat.edu.pl) Dominik MAŁY

https://dx.doi.org/10.37105/iboa.187
Received 19 October 2023
Available online 5 December 2023
ISSN 2450-1859, eISSN 2450-8721
Published by Centrum Rzeczoznawstwa Budowlanego

possibility of extracting and interpreting the personal data contained in the speech signal (by processing it with the help of various technical solutions such as devices and software) is referred to as automatic speaker recognition - ASR (Automatic Speaker Recognition). The overwhelming number of solutions in the field of speaker recognition systems are based on distinctive characteristics derived from the structure of the speaker's vocal tract (laryngeal sound analysis), called physical characteristics of the voice. The widespread use of these characteristics has slowed down the possibility of further development of ASR systems. Overcoming these adversities is possible by analyzing and processing the behavioral characteristics of the speech signal. These include semantics, accent or pronunciation.

The article characterizes the most relevant behavioral features of the speech signal, used in the developed speaker recognition system. The designed algorithm for their extraction is described, as well as how to support the existing ASR system using physical features. The developed solution uses a proprietary speech signal preprocessing method that maximizes the distinguishing capabilities of behavioral features. The possibility of using the system as a standalone solution was also taken into account.

2 Behavioral features

Modern ASR systems are based mainly on the physical and prosodic characteristics of the voice, i.e. characteristics related to the physiological structure of the vocal tract, on the one hand, and intonation, the rhythm of speech and the manner of articulation, on the other hand (Dobrowolski, Majda, 2012). Features from the latter group - referred to as behavioral - are treated marginally in many speaker recognition systems. Naturally, speech also carries obvious information identifying the speaker's language, dialect or emotions, but also very subtle information identifying the region and environment of growing up, education and work. Figure 1 shows a general classification of speaker features extracted from the speech signal.

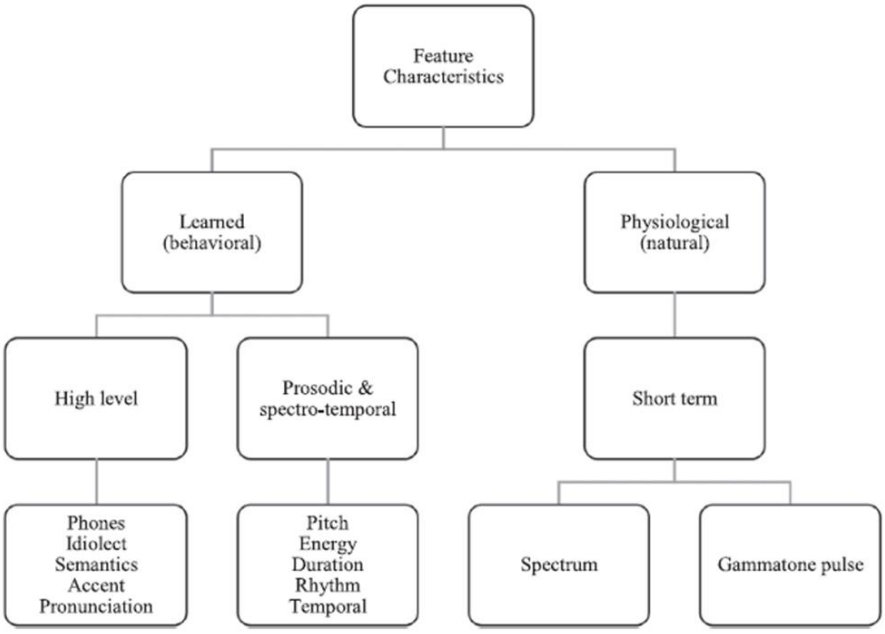

Figure 1. *Speech signal characteristics classification (Tirumala et al., 2017)*

A meticulous review of the current literature and a preliminary assessment of the discriminatory potential of the speaker's behavioral traits made it possible to define a set of 71 descriptors. Some of them are listed below:

- speech duration normalized against the duration of the entire recording,
- the average amplitude value from samples in thirds with the maximum of the sum of amplitude values,
- spectrum center of gravity,
- ratio of the maximum to the minimum pitch value,
- average spectral flatness measure.

In the final version of the algorithm, it was decided to use 27 behavioral characteristics from the original set of 71 features.

2.1 Speech duration normalized against the duration of the entire recording

The articulation process in each person proceeds at a different rate. Speech rate is defined as the number of phonemes uttered in a unit of time. The phonemes articulated in a given utterance are counted up, then this sum is divided by the number of seconds, omitting segments of silence (Woźniak, Soboń, 2015). The developed solution uses the *detectSpeech* command built into the Matlab environment. This function, in each loaded audio signal, detects and determines compartments, containing speech and silence. The total duration of speech, expressed in the number of samples of the signal, is divided by the total length of the recording also expressed by the number of samples. Since each training segment lasts for 3 seconds, it is possible to compare the descriptors obtained for different test subjects. Figure 2 illustrates the initial stage of determining the feature value, involving speech and silence detection.

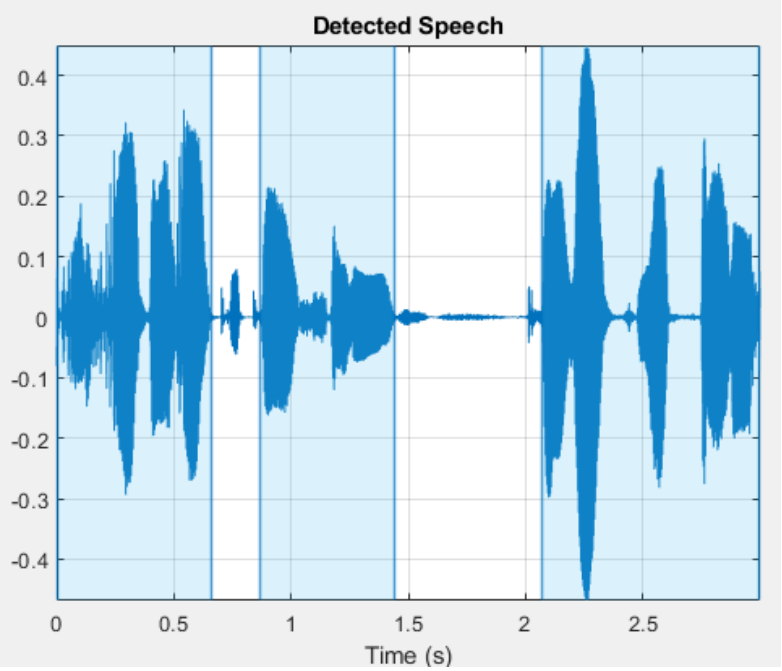

Figure 2. *An illustration of how the detectSpeech function works*

2.2 Average amplitude value from samples in thirds with the maximum of the sum of amplitude values

In acoustics, a third is a frequency band containing between two boundaries whose mutual ratio equals $\sqrt[3]{2}$. Another term for thirds is one-third-octave bandwidth. Energy is a signal parameter defined as the area bounded by the waveform of the square of the signal (Dobrowolski, 2018). In the developed solution, the signals are characterized by a sampling frequency of 8 kHz. For this frequency value, 33 thirds between 1 and 4 kHz are obtained. For each third, a spectrum is determined, and then the samples are summed to find the third with the largest value of the sum of the amplitude of the spectrum. The third characterized by the maximum amplitude is subjected to further analysis, where the average value of the amplitude per each sample included in it is determined. Thus obtained, the final parameter is another descriptor.

2.3 Spectrum center of gravity

Deterministic signals are characterized by various, usually general parameters. A more accurate description of the specific characteristics of a signal can be obtained by defining ordinary and central moments (Dobrowolski,

2018). The abscissa of the center of gravity of a signal is the point in time around which the signal is centered. In the case of frequency representation, we speak of the center of gravity of the spectrum. This is commonly referred to as the center frequency, the frequency for which the spectrum splits into two parts that are equal in energy. Figure 3 shows the center frequency of the spectrum, determined in the Matlab environment, corresponding to the normalized first-order moment.

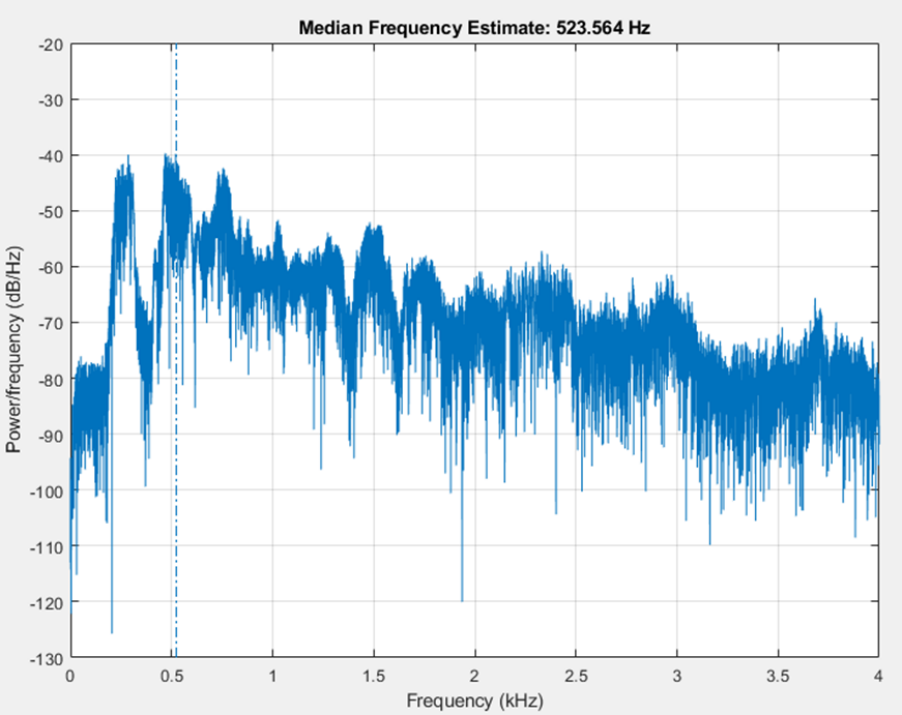

Figure 3. *Spectrum center of gravity*

2.4 Ratio of the maximum to the minimum pitch value

Human speech is an acoustic signal generated with the cooperation of many muscles, cartilage and ligaments. They are part of the natural "generator", which is the glottis. By changing the position of the various elements of the vocal tract, articulation is possible. The vocal cords (ligaments) can open and close, which changes the size of the trachea. This is associated with a change in the degree of airflow from the lungs. As air escapes from the lungs, the vocal folds make very rapid opening and closing movements, which is confusingly similar to vibration. This results in the sounding of speech, or phonation (Jaroszyk F., 2008) . The pitch of the tone is related with the rate of vibration, and thus corresponds to the frequency of the generated sound. In the developed solution, the maximum and minimum pitch are determined for each learning fragment. The two numbers are then compared, obtaining a descriptor that describes the dynamics of a person's voice.

2.5 Average spectral flatness measure

The spectrum, which is the result of Fourier analysis, represents the result of decomposing the signal into a sum of harmonic components. Interpretation and processing of the frequency representation broadens the spectrum of signal parameters that can be analyzed (Dobrowolski, 2018). One such characteristic is Spectral Flatness Measure – SFM. It determines the similarity of the analyzed acoustic signal to white noise. SFM ranges from 0 to 1, where the limits characterize "pure tone" or white noise, respectively. A flatness of 1 means that each frequency component in the spectrum is characterized by a similar power value, and the waveform of the spectrum will be smooth and even, resembling white noise. On the other hand, SFM of 0 says that the power is concentrated around a small number of harmonics, and thus the spectrum is more "pointed." Within each learning segment, its frequency representation is

determined. The algorithm then calculates the ratio of the parameters of the geometric mean and arithmetic mean of the spectrum, which is a numerical expression of the measure of flatness of the spectrum.

3 Developed ASR system

Most solutions in the area of automatic speaker recognition systems have a strictly similar structure. The structure is divided into several main modules (Reddy, Sumathi, 2021):

- acoustic signal acquisition module,
- signal pre-processing module,
- feature extraction and selection module,
- classification module.

Despite these similarities, it is the unique solutions used in the various segments of the system constitute the innovation of a particular solution. The algorithm developed by the authors also consists of the above modules, and its innovation comes from the novel method of splitting the speech signal and the way the system is used as a preselector system for the ASR system based on physical features (Kamiński , Dobrowolski, 2022).

3.1 Human speech signal acquisition and preprocessing

The first element of the described solution is signal acquisition. At this stage, the conditions for recording selected voice signals and the quality of the acoustic path used are of greatest importance. The next step is the extensive pre-processing of the speech signal. During this process, the decimation of the signal to a sampling frequency of 8kHz and amplitude standardization combined with speech dynamics compensation follow, respectively. In the next stage, long fragments of silence (exceeding 2 seconds) are removed, and then the signal is divided into overlapping three-second segments. The result of all the steps described above is the division of one 60-second recording into 58 three-second fragments. In this way, the distinctive nature of behavioral traits is highlighted which facilitates further analysis of the studied signal.

3.2 Extraction and selection of features

The key stage of the entire algorithm is the process of feature extraction and subsequent feature selection. From the recorded and preprocessed voice recordings, a set of features is extracted that should ensure the discrimination of the voices of different individuals, while maintaining the repeatability of these features for different utterances by the same speaker. In addition, a requirement is made for the features to retain the least possible variability over time, as well as to be robust to the health of the speakers and to possible attempts to imitate a given voice. The presented solution is based on the behavioral characteristics described earlier, which are obtained by temporal and spectral analysis of learning segments. For the most part, the behavioral characteristics are determined from three-second fragments of speech and then averaged. However, some are calculated directly from the entire analyzed voice signal. As a result, at the output of the speech signal algorithm, we get a 27-dimensional vector of features. These are already carefully selected descriptors that best describe the speaker. The very process of their selection was implemented using evolutionary methods, specifically a genetic algorithm. This step was repeated several times to be sure of the highest quality of the selected features.

3.3 Classification

The classification of the objects is based on the nearest average method. It is determined for each speaker based on a pre-selected set of 27 features. Through the previously described division of the training segment into a certain number of three-second fragments, it is possible to obtain less than sixty 27-dimensional feature vectors describing a single subject. Then, by averaging the corresponding features in the output, the final set of behavioral characteristics, which is the average value from the components, was obtained. The urban metric, also known as the

Manhattan metric, was used to determine the distance between the averages of different objects. It proved to be the most effective among other metrics, such as the Euclidean, correlation and cosine metrics. Urban distance is defined as the sum of differences measured along specific dimensions. Unlike the most popular Euclidean metric, Manhattan-type distance minimizes the impact of individual large differences by not compounding them.

4 Applications and results of the developed solution

The solution discussed in the article can work in two variants:

- Stand-alone ASR system,
- Preselector system to assist another previously developed ASR system.

In each of the solutions considered, the algorithm works very similarly. The differentiating aspect of the applications is the number of nearest centroids determined by the system. In the first case, the number of nearest neighbor centroids was narrowed down to one. As a result, the algorithm works like a standard automatic speaker recognition system matching a pre-learned closest object from the learning group to a test object. In the second case, on the other hand, the solution is extended to determine an additional number of consecutive nearest neighbors. Their number is top-down declared, but it is possible to adjust it to the ever-changing operating conditions of the system, as well as the growing voice base.

In the course of the research work, the free, publicly available LibriSpeech ASR corpus database included in the Open Speech and Language Resources – Open SLR. The authors decided to use voices marked as "clean" to be sure of high quality recordings. The database labeled SLR-12 contains more than 2,600 audio samples, where 1172 unique speakers can be highlighted.

The first step in testing the developed solution was to check the effectiveness of correct speaker identification in the set of 1172 speakers described above. For the tests, summed speech times of 30 seconds (25 seconds of the signal were used to teach the system, and the remaining 5 seconds for testing) and 60 seconds (analogously 35 and 25 seconds) were used, respectively. The efficiency of correct identification for the selected speech signal lengths were, respectively: 24.91% and 68.43%. At the same time, for the same base and the times of the training and testing segments, the system based on physical features achieved results equal to 86.26% and 98.21%. Such significant differences in the performance of the two systems prompted the authors to perform a solution fusion to maximize the identification performance of the physical feature-based solution (Kamiński , Dobrowolski, 2022) .

The next step of testing a behavioral ASR system is to use it as a selector that initially narrows down the number of cases closest to a given voice collected in the voice database. Then - within this limited set - the physical feature-based system searches for the right object using a classifier based on Gaussian mixtures (Kamiński , Dobrowolski, 2022). The results of these tests are shown in Tables 1 and 2.

	Metrics used			
Number of nearest neighbors	**Euclidean**	**Cityblock**	**Correlation**	**Cosine**
25	95.48	95.82	94.71	93.77
50	96.33	96.67	96.50	95.99
75	97.01	97.35	97.35	96.93
100	97.44	98.21	97.87	97.44
125	97.61	98.46	97.87	97.70
150	97.87	98.38	98.12	97.78
175	97.87	98.29	98.04	98.04
200	97.95	98.38	98.29	98.04
225	97.95	98.55	98.29	98.29
250	98.04	98.55	98.38	98.46
275	98.04	98.55	98.38	98.38
300	98.12	98.81	98.55	98.55
325	98.21	98.81	98.55	98.63
350	98.29	98.81	98.63	98.55
375	98.21	98.81	98.63	98.55
400	98.38	98.81	98.63	98.55
425	98.46	98.72	98.72	98.55
450	98.46	98.72	98.63	98.55
475	98.46	98.72	98.63	98.55
500	98.46	98.72	98.63	98.63
525	98.55	98.72	98.63	98.63
550	98.46	98.72	98.63	98.63
575	98.46	98.72	98.63	98.63
600	98.46	98.72	98.63	98.63

Table 1. *Efficiency values of ASR system based on physical features with applied behavioral preselector for a total speech time of 60 sec.*

	Metrics used			
Number of nearest neighbors	**Euclidean**	**Cityblock**	**Correlation**	**Cosine**
25	66.89	70.99	65.70	63.91
50	76.79	78.58	73.72	72.35
75	79.27	81.74	79.61	77.56
100	82.68	84.90	82.76	81.06
125	84.47	86.52	83.87	82.25
150	86.26	87.12	84.81	83.96
175	86.69	87.63	85.58	84.98
200	87.37	88.05	86.09	85.92
225	87.63	88.48	86.77	86.01
250	87.80	88.31	87.46	86.52
275	87.88	88.74	87.46	86.69
300	88.23	88.65	88.05	87.20
325	88.48	88.74	88.05	87.54
350	88.40	88.65	88.48	87.80
375	88.57	88.82	88.40	87.88
400	88.57	88.82	88.57	87.97
425	88.31	88.48	88.65	87.97
450	88.05	88.40	88.48	87.97
475	87.71	88.23	88.40	88.40
500	87.71	87.97	88.48	88.40
525	87.71	88.05	88.57	88.31
550	87.54	87.88	88.48	88.23
575	87.29	87.71	88.40	88.14
600	87.46	87.80	88.05	87.97

Table 2. *Efficiency values of ASR system based on physical features with applied behavioral preselector for a total speech time of 30 sec.*

5 Conclusions

The article reviews current solutions in the field of automatic speaker recognition systems. It was found that the vast majority of algorithms are systems that operate based on physical features. In response to this state of affairs, an approach based on behavioral distinctive features was proposed and implemented in the Matlab environment. After testing the algorithm operating as a standalone solution, it was decided to apply it in a novel way in the role of a preselector preliminarily narrowing down the number of potential matches for a given test voice. Supporting an automatic speaker recognition system based on physical features with a system based based on behavioral features is a unique solution. The authors, while analyzing available sources, did not find analogous or similar solutions. The proposed in this paper allowed reducing the number of errors from 161 to 132 in a set of 1172 speakers for a total speech time of 30 seconds (25 seconds of the signal were used to teach the system and the remaining 5 seconds for testing). For a 60-second speech signal (35 and 25 seconds, respectively), the number of errors was reduced from 21 to 14 in the same set of 1172 speakers.

Bibliography

1. Dobrowolski A. (2018), *Transformacje sygnałów: od teorii do praktyki*, Legionowo,
2. Dobrowolski A., Majda E. (2011), *Cepstral analysis in the speakers recognition systems*, 15th Conference on Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), pp. 85-90, Poznań,
3. Dobrowolski A., Majda E. (2012), Application of homomorphic methods of speech signal processing in speakers recognition system, *Przegląd Elektrotechniczny,* R. 88 NR 6/2012, pp. 12-16
4. Jaroszyk F. (2008), Biofizyka *Podręcznik dla studentów*, Warszawa, Wydawnictwo Lekarskie PZWL,
5. Kamiński K., Dobrowolski A. (2022), Automatic speaker recognition system based on gaussian mixture models, cepstral analysis and genetic selection of distinctive features, *Sensors*, 22(23), 9370, DOI: 10.3390/s22239370
6. Reddy Gade V. S. and Sumathi M. (2021), *A Comprehensive Study on Automatic Speaker Recognition by using Deep Learning Techniques*, 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, pp. 1591-1597,
7. Tirumala S. S., Shahamiri S. R., Garhwal A. S., Wang R. (2017), Speaker identification features extraction methods: A systematic review, *Expert Systems With Applications*, 90, pp. 250–271, DOI: 10.1016/j.eswa.2017.08.015
8. Woźniak T., Soboń J. (2015), Ocena płynności mówienia, *Nowa Audiofonologia*, 4(4), pp. 9–19, DOI: 10.17431/894809