

AUTOMATIC SPEAKER RECOGNITION SYSTEM SUPPORTED BY BEHAVIORAL FEATURES OF SPEECH SIGNAL

Dominik Mały¹⁾, Andrzej Dobrowolski²⁾, Kamil Kamiński³⁾

1) Military University of Technology, Faculty of Electronics, Institute of Radioelectronics, Sylwestra Kaliskiego 2, 00-908 Warsaw, Poland
(✉ dominik.maly@wat.edu.pl)

2) Military University of Technology, Faculty of Electronics, Institute of Electronic Systems, Sylwestra Kaliskiego 2, 00-908 Warsaw, Poland

3) Military University of Technology, Institute of Optoelectronics, Sylwestra Kaliskiego 2, 00-908 Warsaw, Poland

Abstract

The extraction and interpretation of personal data from speech signals, processed through various technical solutions, are key functions of Automatic Speaker Recognition (ASR) systems. Speech conveys information such as language, dialect, and emotions, making ASR systems increasingly essential due to the growing demand for human-computer interaction and biometric security applications in both military and civilian sectors. Voice, as a unique human characteristic, enables identification without additional attributes that can be lost or destroyed. However, while humans recognize voices naturally, machines face significant computational challenges. Despite advancements in automatic speaker recognition, many challenges remain. This article addresses the use of behavioral voice features in automatic speaker recognition (ASR) systems. The authors aimed to develop and implement a set of behavioral features in an existing ASR system that would increase the number of correct speaker identity identifications, particularly in the presence of various types of noise. By utilizing the publicly available LibriSpeech voice database, it was possible to compare the developed solution with other ASR systems. In addition, the authors developed a solution that can reduce the impact of external noise on speaker identity recognition accuracy. The key element proved to be the innovative data integration method, which leverages the advantages of various sources of distinctive feature sets. In the conducted experiments, the proposed ASR system demonstrated outstanding performance in automatic speaker recognition. Using the LibriSpeech database, it achieved identification rate exceeding 99% for the train-clean-100 subset and close to 99% for the train-clean-360 subset. Compared with traditional Gaussian Mixture Model (GMM) approaches, which typically achieve about 83% accuracy, the developed solution provides a substantial improvement, reaching identification rates above 99% and demonstrating performance comparable to, or even exceeding, that of modern deep learning-based techniques (approximately 98 to 99%).

Keywords: automatic speaker recognition, behavioral features, data fusion, distinct feature selection, genetic algorithm.

1. Introduction

The extraction and interpretation of individual data contained in speech signals using the capabilities of modern information technology is implemented in systems known as *Automatic Speaker Recognition* (ASR) systems. In addition to individual characteristics, speech also conveys various types of information, such as language, dialect, or emotions accompanying the speaker [1]. This article focuses on the extraction and interpretation of so-called behavioral distinctive features of the speaker, which have been used less frequently in speaker recognition systems compared to physical features. Behavioral characteristics are related to the origin, education, and general understanding of personality traits, while physical characteristics are directly related to the anatomical structure of the speaker, particularly the structure of his vocal tract. More broadly, behavioral speeches are verbal prompts where behavioral features reflect

the individual's manner of expression, speaking style, and articulation habits. Typical measurable examples include total speech time in the analyzed segment, the count of significant impulsive events, mean amplitude levels in octave and third bands, and the articulation speed. Automatic speaker recognition systems are increasingly being used in areas where a high level of security is required while maintaining a short identity verification time. However, these solutions are not completely immune to various types of fraud. The main threats they face include the following:

- impersonation attempts,
- highly similar voices present in large reference recording databases,
- pathological changes in vocal organs,
- insufficient data due to short voice recordings,
- negative impact of the entire acoustic transmission chain,
- disturbances occurring during the speech acquisition process,
- use of objects that modify voice properties, such as masks,
- voice manipulations, including deepfake-based voice synthesis.

Depending on the requirements that ASR systems must meet, their designers adopt various design approaches. Systems based on physical features achieve high accuracy even with relatively short training and testing recordings. In contrast, algorithms that rely on behavioral features require longer voice recordings to highlight their distinctive properties. Pathological changes in the larynx and upper respiratory tract strongly affect the values of physical characteristics while having a weak impact on behavioral biometrics. An additional advantage of high-level features, as opposed to those related to the vocal tract structure [2], is their resistance to noise and interference present in the transmission channel. Deliberate voice modifications, such as whispering, increasing pitch, or increasing articulation speed, affect the performance of ASR systems. The same applies to unintentional speech distortions caused by stress-related factors (e.g., jaw clenching) [3].

Taking into account only the vulnerability criterion to threats, systems based on behavioral speech signal features are more resistant than those relying on physical characteristics. However, when also the ease of speech processing and the long-term stability of the extracted features, solutions based on biometrics derived from the structure of the vocal tract structure appear more favorable. The choice of an appropriate speech parameterization method and, consequently, the entire ASR system design, depends on the specific requirements. It is not possible to create voice models that enable speaker identification with nearly 100% accuracy while simultaneously making them immune to external disturbances using only a single set of features. A compromise in this regard could be an algorithm that incorporates the fusion of physical and behavioral descriptors. With this approach, it is possible to improve the resistance of the ASR system to external factors while maintaining high operational accuracy [4].

The developed ASR system achieved very high identification rate, exceeding 99% on the *LibriSpeech* train-clean-100 dataset [5] and nearly 99% on the larger train-clean-360 dataset [5]. In particular, it reached 100% with feature-level fusion and 99.60% with the preselector, clearly outperforming classical methods such as *Gaussian Mixture Models* (GMM) (83.1%) [6] and surpassing or matching state-of-the-art solutions including *DeepSpeaker* [6], *Mel-Frequency Cepstral Coefficients* (MFCC) [7], and *SpeakerGAN* [7]. These results confirm the robustness and competitiveness of the proposed approach in automatic speaker recognition tasks.

2. State-of-the-art

This section presents a review of the scientific literature on ASR systems. To ensure comparability of the results, the selection of publications was restricted to studies utilizing various subsets of the *LibriSpeech* database [5]. This approach provides a common benchmark that allows the achievements of different methods to be evaluated under consistent conditions.

The authors of [6] used the train-clean-100 subset of the *LibriSpeech* database, which includes 251 speakers. For feature extraction, they employed MFCC [8, 9], *Linear Prediction Coding* (LPC), and the *Zero Crossing Rate* (ZCR) [10, 11] parameter. Speaker classification was then performed using GMMs. In contrast, the deep learning approach (*DeepSpeaker*) used 64 mel filters as input features. The authors achieved an identification accuracy of 83.1% for the GMMs and 99.8% for the *DeepSpeaker* approach.

In [7], the primary objective was to optimize the *SpeakerGAN* system in terms of resource efficiency and training time. The authors used the train-clean-100 subset of the *LibriSpeech* database. Feature descriptors were initially extracted using a mel filter bank and MFCCs. Each analyzed signal underwent a silence removal process. Ultimately, the researchers achieved an accuracy of 97.87% for the 64-mel filter variant. The extracted features were also used to train a *Convolutional Neural Network* (CNN), a modified version of the *SpeakerGAN* model. In this case, the *Identification Rate* (IR) was approximately 2% lower than that of the reference model. Notably, the authors also reported IR values as a function of the number of utterances per speaker.

Article [12] explores feature binarization for speaker recognition in embedded systems. This approach was motivated by the high computational complexity of traditional methods, such as MFCC and GMM, on hardware platforms. The authors employed linear prediction methods followed by MFCCs. As a result, they obtained first-order and second-order distinctive features (delta and delta-delta features) [13, 14], specifically their signs. The number of these signs was then summed to construct histograms of their frequency distribution. After appropriate normalization, the obtained results included an *Equal Error Rate* (EER) of 12.71% and an IR of 59.24%. The latter, referred to as Rank1 in the article, represents the frequency with which the correct speaker is ranked as the top match in a closed-set scenario. All experiments were conducted using the test-clean subset of the *LibriSpeech* database, which contains samples from 40 speakers.

The authors of [15] aimed to develop an ASR system based on CNN and LSTM neural networks. Their research was conducted using the TIMIT dataset (630 speakers) and the *LibriSpeech* dataset (251 speakers). Various feature extraction methods were explored, including MFCC, spectral centroids, spectral cutoff points, and mel spectrograms. Ultimately, the study employed MFCCs. The training segment lasted 12 to 15 seconds, while the test segment ranged from 2 to 6 seconds. In a closed speaker set based on [5], the authors achieved an identification accuracy of 97.85% for the CNN model.

Article [16] describes and compares two automatic speaker recognition systems based on deep learning techniques. The first is a reference system created using the Keras library [17], while the second is a custom modification. The authors proposed an innovative approach that employs a neural network with a simplified structure and Fast Fourier Transform (FFT) for feature extraction. The study was conducted on subsets of *LibriSpeech* (50 speakers) and *Multilingual Speech Diversity Dataset* (MSDD), the latter specifically created for this research, containing approximately 25 minutes of read text per speaker. The speaker identity recognition accuracy for both subsets was 91.27% and 94%, respectively.

In [18] a feature fusion method combined with a deep neural network is described. The approach integrates traditional MFCCs with their *Mel-Frequency Cepstral Coefficients Temporal* representation (MFCCT). Time-domain features were obtained by grouping the

initially extracted mel cepstral coefficients and then computing statistical properties such as minimum, maximum, standard deviation, and median for each subset. Ultimately, 12 temporal features were extracted. The final step was the fusion of characteristics from different domains into a single vector of distinctive features, which was then used to create speaker models using a deep neural network. The train-clean-100 subset of the *LibriSpeech* database [5] was used as the speech dataset. The speaker classification process was conducted in two stages: first, the gender of the subject was classified, followed by identity recognition within the specific gender group. The authors achieved speaker identification accuracies of 89% for women and 84% for men, with an EER of 0.11%.

In [19], an automatic speaker recognition system utilizing the spectral variant of MFCC, known as *Mel Frequency Spectral Coefficients* (MFSC), was described. Transformer neural networks were also employed. Features were obtained by omitting the final stage of the cepstral transformation, namely, the discrete cosine transform. In the course of their research, the authors compared three neural network models - *BiGRU*, *BiLSTM*, and *Transformer* - with two sets of features, resulting in six variants tested. The *train-clean-360* subset of the *LibriSpeech* database was used as the audio dataset. For each neural network, the spectral variant of mel-scale features yielded better results. The best identification accuracy, 96.08%, was obtained for the MFSC + *Transformer* combination. However, this performance improvement of approximately 0.2% over the second-best result came at the cost of a more than sixfold increase in testing time.

The authors of [20] investigated the use of the *Raw Audio Network* (RANet) neural network in conjunction with *Stochastic Gradient Descent with Momentum* (SGDM). This convolutional neural network processes raw (unprocessed) audio files, aiming to enhance robustness against interference compared to standard feature-based approaches. The study utilized all subsets of the *LibriSpeech* database, totaling nearly 22,000 voice samples from 2,484 speakers. The final accuracy of this approach reached 82.57%. Comparisons were also made with a conventional CNN, the *SincNet* model, and a variant of RANet using *Adaptive Moment Estimation* (ADAM).

In [21], the authors proposed the fusion of physical features (MFCC) with behavioral features (dependent on the speaker's pronunciation). Depending on the nature of the features, a convolutional network and a multilayer perceptron were used for extraction. Speech manner descriptors were extracted from recordings of Mandarin dialect speakers, including classifications of phoneme articulation types such as nasal, fricative, and oral sounds. Physical features were represented by *Mel-Frequency Energy Coefficients* (MFEC). An *Multilayer Perceptron* (MLP) network was also used for classification. One of the datasets used was the *train-clean-460* subset of *LibriSpeech*, containing 1,172 speakers. The results were presented by comparing the EER values of the proposed solution with *x*-vector and *d*-vector algorithms classified using the cosine metric and *Probabilistic Linear Discriminant Analysis* (PLDA). The lowest EER of 7.80% was achieved on the *LibriSpeech* database.

A review of the current literature in the field of ASR [1-18] shows that physical features, specifically MFCCs and their variations, remain dominant in automatic speaker recognition systems. Additionally, deep learning methods are increasingly popular for classification tasks. In contrast, behavioral characteristics remain underutilized. This article presents the results of implementing behavioral features as a standalone solution and in fusion with physical features, extending the existing ASR system described in [22]. The developed algorithm was initially trained on the *train-clean-460* dataset, which consists of the *train-clean-100* and *train-clean-360* subsets of the *LibriSpeech* database. The testing phase also included the NIST 2002 SRE dataset [23] and three proprietary voice datasets developed at the Faculty of Electronics, Military University of Technology, Warsaw, Poland.

3. Structure of the developed solution

3.1. Standardization of the speech signal

The designed ASR system is intended for use in commonly used telephone transmission conditions. For more than 100 years, telephones have operated within a limited frequency range of 300 to 3400 Hz. Although the frequency range of the human voice extends from approximately 50 to 8000 Hz, speech can still be intelligibly transmitted using a significantly narrowed bandwidth. It is widely accepted that human speech remains sufficiently comprehensible when confined to the 500–4000 Hz range [24, 25]. Considering these values, the developed system adopts a sampling rate of 8 kS/s.

Before distinctive features, the speech signal undergoes a standardization process, which involves adjusting the mean value of the speech signal to zero and the standard deviation to one. Another essential step in the preliminary processing of the speech signal is silence removal, silence naturally occurs in every utterance. From the perspective of behavioral speech characteristics, the distribution of silence segments carries a certain amount of distinctive information. For example, the relationship between speaking time and silence within a frame can be used to determine an individual's articulation speed. Therefore, in this system, only silence segments longer than 2 seconds are removed. Silence detection is performed using a procedure implemented in the Matlab environment, specifically the *detectSpeech* function [26], which is described in detail in [27].

3.2. Segmentation of the speech signal

Segmentation involves dividing the speech signal into short fragments, referred to as frames. Human speech is a nonstationary stochastic process characterized by significant temporal variations in signal properties. Dividing the speech signal into short fragments allows for its approximation as a quasi-stationary process [25]. Each frame is thus treated as a separate signal representing a specific speaker. The speech segmentation operation reduces to windowing, which involves multiplying signal samples by the samples of a rectangular window [28], characterized by an appropriately chosen width and shift.

Time-domain speech analysis involves significant redundancy in the information obtained. To reduce the amount of necessary data contained in the speech signal, frequency analysis is often used. The previously mentioned multiplication of signal samples by window samples in the time domain is equivalent to the convolution of their spectra in the frequency domain. Consequently, selecting the shape of the time window becomes an important aspect. The most commonly used window is the rectangular window, which has a narrow main lobe in its spectrum but unfortunately exhibits high side lobes. Other frequently used windows, such as Hann, Blackman, or Hamming, are functions composed of appropriate harmonic waveforms [29], offering a compromise between the width of the main lobe and the level of the side lobes.

Besides selecting the window type, its width and step size are crucial factors. Speech recognition systems typically employ variable window widths depending on the semantic content of the speech. This is due to the need to analyze various components of human speech, such as phonemes, syllables, or words. In the context of automatic speaker recognition systems, uniform segmentation with a fixed window length is generally used. Additionally, overlapping segments are often used to prevent data loss. In the developed system, a window length of 3 seconds with a 1-second shift was applied.

3.3. Behavioral features of the speech signal

The human voice, as analyzed by ASR systems, carries rich information useful in speaker identification. The extracted features can be classified into three main groups [30]:

- short-term spectral features,
- spectro-temporal features,
- high-level features (resulting from complex processing).

The first group consists of characteristics with a short duration (approximately 20–30 ms), whose properties depend on the structure of the vocal tract. Spectro-temporal features span hundreds of milliseconds and pertain to the intonation of speech, duration, or rhythm—often referred to as prosodic features, emphasizing their relationship with the sound of the voice. The last group includes so-called high-level features, which are influenced by socio-economic factors that have shaped the speaker, such as social status, place of birth, language used, or personality. In the literature, these features are called behavioral features. They reflect individual differences in articulation, resulting from learned habits related to the intonation of spoken phrases, speaking time, or loudness level. Apart from socioeconomic influences, one of the main factors modifying speech characteristics is a person's current emotional state. From both the speaker's and the listener's perspective, the impact of emotions on speech characteristics is noticeable.

Taking into account the components that influence the speech characteristics described above, it is quite challenging to define a universal set of behavioral features. In the ASR system described here, an initial set of 71 voice features was defined. These features were separately determined in two domains: the time domain and the frequency domain. Signal processing in both domains is based on the analysis of previously extracted 3-second frames. The initial set of 71 behavioral features includes such characteristics as:

- the ratio of speech duration to silence duration within a frame,
- speech duration within a frame normalized by the number of speech segments (average speech segment duration),
- the ratio of maximum to minimum fundamental frequency within a group of three frames,
- percentage of low-energy frames [10],
- number of significant impulse phenomena,
- number of voiced frames (relative to the entire signal),
- zero-crossing rate [10, 11],
- bandwidth occupancy containing 95% of the total power [31, 32],
- spectral roll-off values [33],
- spectral cut-off point values [11, 33],
- spectral kurtosis values [34],
- spectral skewness values [34],
- spectral spread values [11, 33, 34],
- sum of amplitudes in thirds and octaves and their derivatives, including mean values over a specified number of frames, standard deviations, maximum and minimum values.

In this study, some statistical descriptors of the signal, such as spectral skewness and kurtosis, are also included among behavioral features. Although these measures originate from the physical distribution of spectral values, they characterize properties such as asymmetry and peakiness of the spectrum, which in practice are shaped by the speaker's articulation habits and speaking style rather than by anatomical constraints.

3.4. Evolutionary methods applied to feature set selection

The feature extraction process in ASR systems typically generates a large set of descriptors

whose redundancy does not always translate into higher classification accuracy. Feature selection, by reducing the dimensionality of the vector, allows maintaining or improving identification accuracy while reducing the computation time [35, 36]. The most commonly used feature selection methods include Fisher's discriminant analysis, mutual correlation, principal component analysis, and genetic algorithms [37, 38, 39]. In this study, a genetic algorithm was applied. This procedure is inspired by mechanisms of biological evolution and operates on a population of individuals represented by binary characteristic vectors [1, 22, 40]. Each individual is evaluated using a fitness function defined as:

$$ff_i = 1 - Acc_i, \quad (1)$$

where Acc_i is the identification accuracy of the system for the i -th individual.

From the above relationship, it follows that the lower the value of ff_i , the higher the adaptation level of the i -th individual in the evolutionary process. Then, through selection operations, distributed crossover [2] and mutation (including Gaussian mutation [41]) the feature set is iteratively optimized.

The process continues until the maximum number of generations is reached or no further improvement in the fitness function is observed. As a result, the original set of 71 features was reduced to 27, which led to an approximately 5% improvement in speaker identification accuracy. In the following, the feature selection process is illustrated for successive generations of the evolutionary process. Figure 1 presents the feature selection process in the ASR system developed by the authors.

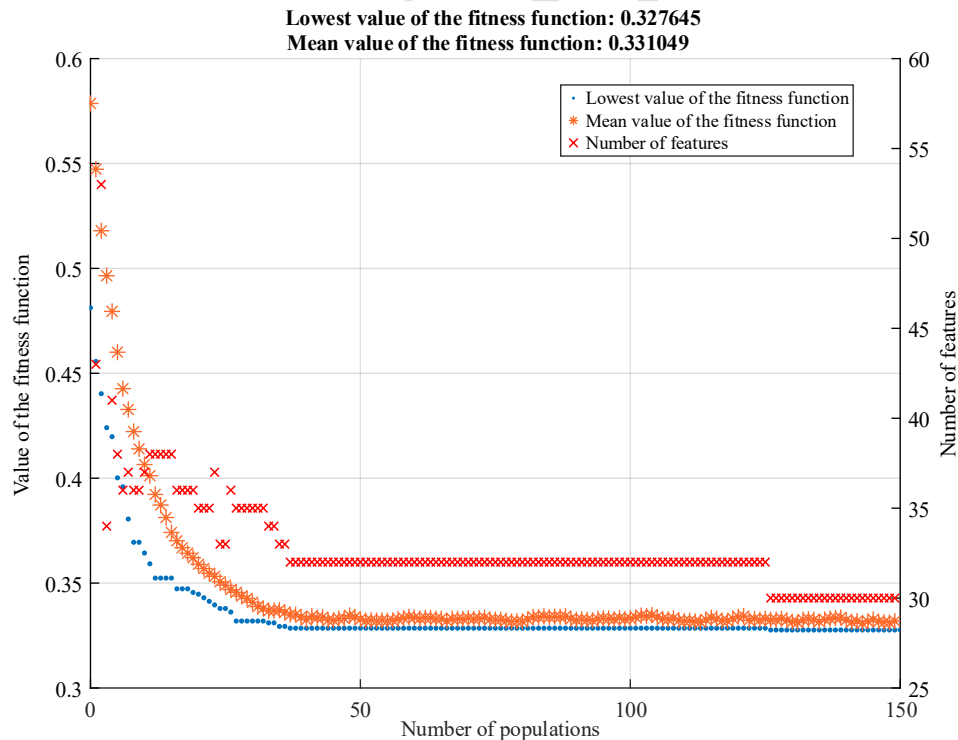


Fig. 1. Feature selection process using a genetic algorithm.

3.5. Applied classifier

The final stage of the speaker recognition process is classification, which involves assigning the signal source to the appropriate class [29]. The simplest classification method is based on measuring the distance between the analyzed feature vector and the vectors assigned to specific

classes. This method is called minimum distance classification, as it assumes that similar objects are located close to each other [29].

One of the fundamental classification methods based on the distances between feature sets is the nearest-neighbor method. Its principle is based on finding the closest vector from the training data relative to the test vector. An extension of this method is the *k*-nearest neighbors (*k*-NN) approach [42], which searches for the *k*-nearest neighboring vectors instead of just one. Unlike the classical approach, the *k*-NN classifier sums the distances to the *k* closest vectors rather than considering only the closest one. A key advantage of the nearest neighbors method is the ability to use different distance metrics. One of these is the Manhattan metric, *Cityblock* or *Taxicab* [43]. This distance is defined as the sum of absolute differences measured along specific dimensions. Its name originates from its similarity to cities with a square street grid, where a taxi can only move along designated streets [44]. Unlike the commonly used Euclidean metric, the Manhattan distance minimizes the influence of single large differences by avoiding exponentiation [45]. In the developed solution, various metrics (such as Euclidean, correlation, and cosine distance) were tested, but the Manhattan metric proved to be the most effective.

During the research, it was observed that the distinctive properties of behavioral features emerge when a relatively large amount of training data is available, compared to the requirements for extracting physical features. This, among other factors, led to the use of the previously described speech signal segmentation method, where the entire speech signal is divided into three-second frames with a one-second shift. In this case, a one-minute speech sample results in 58 analyzed fragments.

This segmentation approach means that a single voice sample is represented by multiple objects (feature vectors) belonging to the same class, creating a clustering effect [46]. In such cases, applying the *k*-NN classifier (and selecting an appropriate *k*-value) is not straightforward. A solution to this problem is the nearest mean classifier, which calculates the feature vector distance from the currently analyzed to the mean vectors representing each class [47]. Since the mean vectors serve as representatives of their respective classes in the training set, the amount of data required for storage during the training phase is significantly reduced.

4. Fusion of systems

To objectively assess the validity of the developed solution, it must be compared with another available system. According to the assumptions, the reference system chosen was *ARMiA* (*Automatyczne Rozpoznawanie Mówcy i Autoryzacja*) [22]. The *ARMiA* system, developed in 2018 by Maj. Dr. Eng. Kamil Kamiński, is an ASR solution designed for user identification based on voice samples. Its preprocessing stage includes signal normalization, silence removal using energy-based criteria, high-pass filtering, and frame segmentation with Hamming windowing. A multi-step frame selection process further refines voiced segments and reduces noise impact. Feature extraction relies on spectral, cepstral, and mel-cepstral analysis, complemented with weighted cepstral features and prosodic parameters. For classification, GMM are applied, enabling compact yet discriminative speaker models. Additionally, to ensure reliable comparison results, training and testing segment durations, as well as the same voice databases, were used. The study utilized the *LibriSpeech* database, which contains 1,172 speakers (*train-clean-460*). Tests were carried out for two variants of training and testing data lengths. In the first variant, a total of 30 seconds of voice recordings were used, divided into 25 seconds for training and 5 seconds for testing. The second option included a minute of speech, where 35 seconds were allocated for training and the remaining 25 seconds for testing. The authors of these solutions. The *ARMiA* system [22] was optimized for a total

segment length of 30 seconds, while the algorithm based on behavioral speech signal required a minimum of one minute of voice samples. Table 1 presents the results of the tests carried out.

Table 1. Comparison of ASR system accuracy based on training and testing segment lengths.

Training time / Testing time [s]	Behavioral feature-based system accuracy [%]	ARMiA system accuracy [%]
25 / 5	24.91	86.26
35 / 25	68.43	98.21

As the results show, the standalone behavioral system achieves an identification accuracy of just under 70% when analyzing one-minute speech samples. The reference *ARMiA* system, which is based on physical features, significantly outperforms the behavioral approach. Furthermore, a comparison was made regarding the number of correct identifications of the speaker identity between the behavioral system and the existing *ARMiA* system [22]. For 30 and 60 seconds of analyzed speech, the behavioral feature-based system performed approximately 70% and 30% worse, respectively, compared to the system based on physical voice characteristics. The obtained results clearly indicate that an ASR system based solely on behavioral features is not sufficiently effective in real-world conditions. Therefore, the next section of the article presents a solution that results from the fusion of these two biometric approaches.

4.1. Selection of the data fusion type

Automatic speaker recognition systems can be regarded as information measurement systems whose purpose is to collect as much data as possible to characterize the examined object (the speaker). Information from multiple such "measurement channels" can undergo integration processes, which can be carried out at various levels [48, 49]. This includes the hardware layer and the data acquisition process, information processing, and the synergistic use of data from different sensors. An important aspect is the proper definition of the fusion process as a component at each of these levels, where data is merged into a coherent, integral set [50, 51]. Selecting an appropriate data fusion mechanism is a complex task. The literature presents various classifications of the integration process [52, 53, 54]. One classification distinguishes three models based on:

- the level of abstraction in information integration [50],
- the level of abstraction of input and/or output data [55],
- the relationship between data sources [56].

The first model introduces a more detailed division into low-level, intermediate and high-level fusion. The low-level fusion approach is based on merging data from multiple sources into a single output signal. This method allows for the creation of data with greater informational significance compared to separately analyzed signals from integrated sources. It is often referred to as raw-data fusion. Feature-level fusion is another method of information integration, based on combining parameters extracted during the processing of input signals by different "sensors". The high-level fusion approach, often called decision fusion, involves combining final predictions (decisions) with respect to the classification of the same signal, but obtained from separate decision-making modules [57].

4.2. Nearest neighbor preselection method – decision-level fusion

The foundation of this data fusion method lies in the use of different classifiers in both ASR systems. The reference solution is based on GMM [40]. On the contrary, the algorithm

described in this article employs the nearest mean method for classifying the examined objects. The combination of these classifiers is illustrated in Fig. 2.

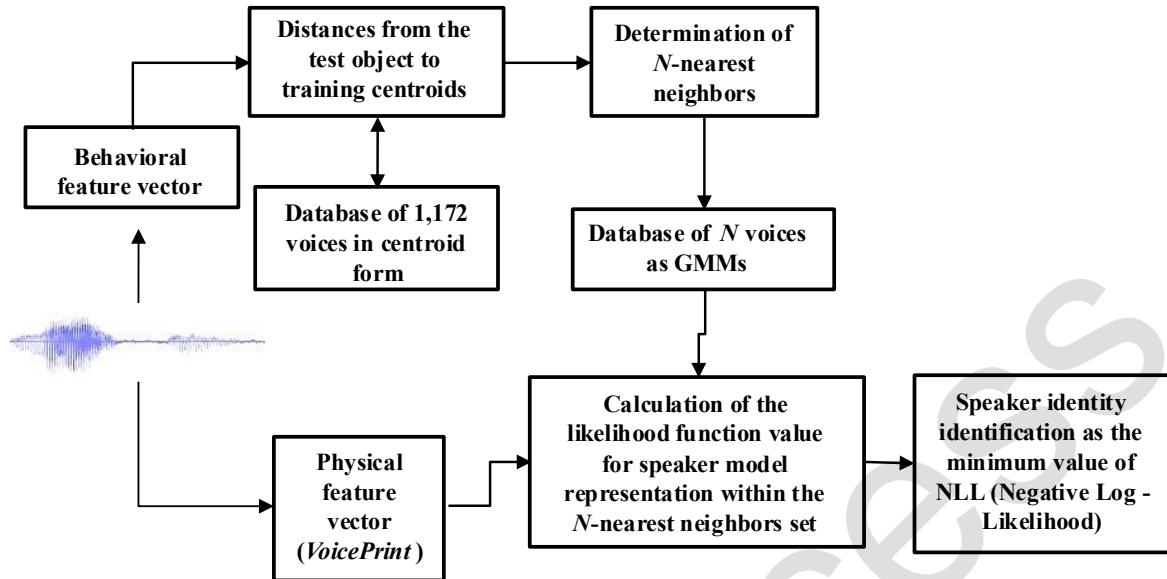


Fig. 2. Decision-level data fusion process.

According to the above scheme, the fusion process begins as soon as the physical and behavioral feature vectors of the object analyzed are determined. Next, a voice model of the object analyzed is created using GMM. At the same time, the algorithm based on behavioral speech signal features identifies the N nearest centroids (by calculating the distance from the analyzed object to the training data) from the full dataset. The next step is to reduce the full set of voice models (training data) to the N nearest neighbors. The classification process then follows the implementation described in [22].

During the research, various combinations of decision-level data integration methods were tested. Ultimately, however, the preselection method described above, which involves initially limiting the training dataset to the N samples most similar to the test object (based on a distance metric), proved to be the most effective. Table 2 presents the results of applying the "preselector" for 30- and 60-second speech samples. Green highlighted values indicate a higher number of correct identifications than in the baseline system described in [22] or each total duration (86.26% and 98.21%).

Table 2. ASR system accuracy based on physical features with an applied behavioral preselector.

	Speech duration [s]	Number of nearest neighbors	100	200	300	400	500	600
Applied distance metric	30	Euclidean	82.68%	87.37%	88.23%	88.57%	87.71%	87.46%
		Manhattan	84.90%	88.05%	88.65%	88.82%	87.97%	87.80%
	60	Euclidean	97.44%	97.95%	98.12%	98.38%	98.46%	98.46%
		Manhattan	98.21%	98.38%	98.81%	98.81%	98.72%	98.72%

4.3. Feature set combination method – feature-level fusion

The second method of data integration involves the fusion of two distinctive feature vectors to create a single set of resultant features. This approach is driven by the differing nature of physical and behavioral features. Such data consolidation allows for the creation of a more

information-rich distinctive feature vector that better characterizes the analyzed object. The process of feature-level data fusion is illustrated in Fig. 3. In this approach, all speech signal processing stages, including its parameterization, are performed separately in each of the "subsystems" based on behavioral and physical features. Then, the output vectors of distinctive characteristics are integrated to form a single, more comprehensive feature set that represents the analyzed object. The subsequent steps in the ASR system, such as the speaker's voice modeling and classification processes, are identical to those described in [1, 22].

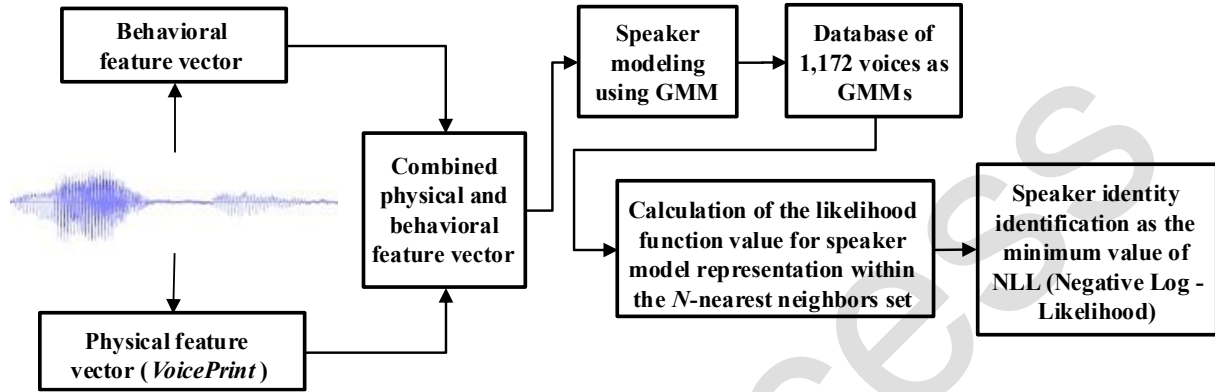


Fig. 3. Feature-level data fusion process.

As part of the conducted research, various post-fusion processing variants were tested, including different modeling and classification approaches. The study considered whether to apply techniques implemented in the system based on behavioral features, physical features, or a combination of both. The best results were achieved using the methods applied in the system described in [1, 22], specifically through the GMM classifier. However, an additional research aspect that needed to be considered was the method of consolidating the two distinct feature vectors. Since behavioral and physical feature sets differ not only in their nature, but also in their quantity, the selection of an appropriate integration method became a crucial aspect of the data fusion process. The set of physical characteristics can be described as a matrix $W_f(2)$ with dimensions $M \times N$, where M represents the number of frames analyzed and $N = 23$ is the number of features:

$$W_f = \begin{bmatrix} C_{f1_1} & C_{f2_1} & \cdots & C_{fN_1} \\ C_{f1_2} & C_{f2_2} & \cdots & C_{fN_2} \\ \vdots & \vdots & \ddots & \vdots \\ C_{f1_M} & C_{f2_M} & \cdots & C_{fN_M} \end{bmatrix}. \quad (2)$$

In contrast, behavioral features can be represented as a vector W_b consisting of $H = 27$ elements, each representing a specific behavioral feature:

$$W_b = [C_{b1}, C_{b2}, \cdots, C_{bH}]. \quad (3)$$

Due to the differences in the feature sets described above, two data fusion methods were tested. The first method involves combining the W_f matrix and the W_b vector in such a way that the behavioral feature vector, attached to the physical feature matrix, represents additional features of the analyzed object but only in the first frame. The second data fusion method simulated the determination of a behavioral feature vector for each frame processed in the ASR system [22]. This was achieved by multiplying a column matrix where each row – except for the first, which contained only ones – had a random value within a specified range (0.95÷1.05) with the W_b matrix. The resulting product was then appended to the W_f matrix.

After extensive testing, the authors finally decided to use the first feature-level fusion variant. Its practical implementation is carried out by multiplying a vertical vector W_{k_1} , whose number of rows corresponds to the number of frames processed in the solution described [22], with the horizontal vector W_b , and then appending the resulting output to the W_f matrix, forming the final matrix W_{fb_1} :

$$\begin{aligned}
 W_{k_1} &= [1, 0, \dots, 0]^T \\
 W_b &= [C_{b1}, C_{b2}, \dots, C_{bH}] \\
 W_{fb_1} &= \begin{bmatrix} C_{f1_1}, C_{f2_1}, \dots, C_{fN_1}, C_{b1}, C_{b2}, \dots, C_{bH}, \\ C_{f1_2}, C_{f2_2}, \dots, C_{fN_2}, 0, 0, \dots, 0, \\ \vdots, \quad \quad \quad \ddots, \quad \quad \quad \vdots, \quad \quad \quad \vdots, \quad \quad \quad \ddots, \quad \quad \quad \vdots, \\ C_{f1_M}, C_{f2_M}, \dots, C_{fN_M}, 0, 0, \dots, 0, \end{bmatrix} \quad (4)
 \end{aligned}$$

where: C_{fk} is the k -th physical feature and C_{bk} is the k -th behavioral feature.

Table 3 presents the results of the system described in [22], along with the feature fusion solution that uses 30 and 60 seconds of speech.

Table 3. Comparison of ASR (with feature fusion) accuracy based on training and testing segment lengths.

Training time / Testing time [s]	Behavioral Feature-based system accuracy [%]	ARMiA system accuracy [%]
25 / 5	86.26	57.25
35 / 25	98.21	97.25

From the above table, it is clear that despite creating a new set of descriptors for speaker characterization, it was not possible to achieve a higher accuracy than the original system based on physical features [22].

5. Selection of the data fusion type

The results presented in Section 4 on the fusion of a system based on physical features with a system based on behavioral features demonstrate that this process can improve the accuracy of speaker identification. To further investigate this process, this section presents studies conducted on recordings of artificially limited quality. The authors also decided to expand the voice dataset by adding more than 500 additional speech signals from various databases. Ultimately, the created voice dataset contains 1,688 samples, among which:

- 1,172 samples come from the SLR-12 database [5],
- 320 samples come from the NIST SRE 2002 database [23],
- 56 samples were created based on audiobooks,
- 85 samples come from the Ewelina Majda-Zdancewicz database [46],
- 24 samples come from the Daniel Posiadała database [58],
- 31 samples come from the Michał Bojsza database [59].

Also the training and testing times were different. Optimal ones were as shown in Table 3.

5.1. Method for presenting results

To evaluate the performance of the developed solutions in supporting the existing system [1], a series of experiments were conducted under various operating conditions. The time combinations of 25 seconds for training and 5 seconds for testing, as well as 35 seconds for training and 25 seconds for testing, were optimal for the ARMiA and behavioral feature-based systems individually. In the final fusion-based solution, however, another range of segment

durations was tested. Training durations ranging from 20 to 40 seconds (in 5-second increments) and corresponding testing durations of 10, 15, and 20 seconds. This configuration enabled a comprehensive evaluation of system performance across different segment lengths and facilitated the identification of the most effective setup for the final ASR implementation. The following systems were tested: the system described in [1], the standalone behavioral feature-based solution, the merging of both systems at the feature level and the fusion at the decision level for the 1,000 nearest neighbors. Additionally, before conducting the aforementioned tests, the authors of this article decided once again to use a *genetic algorithm* (GA) for feature selection in the feature-level data fusion variant. This operation resulted in a new set of 36 most distinctive features, including 19 physical features and 17 behavioral attributes. The experiments were conducted using the following types of recordings:

- uninterrupted (high-quality) recordings – shown in Fig. 4;
- recordings distorted by external noise – shown in Fig. 5;
- compressed recordings using commonly used codecs in telecommunication channels – shown in Tables 4 to 8.

In addition to testing ASR systems under noise-free conditions, experiments were also conducted in which analyzed speech signals were disrupted by external additive noise or interference [60]. A testing scenario was developed using three types of noise disturbances:

- white noise was generated separately for training and testing segments and added to the original voice dataset;
- crowd noise was introduced into each signal, simulating speech recording in an environment filled with other people conversing,
- heavy rain sound was used as the final type of disturbance.

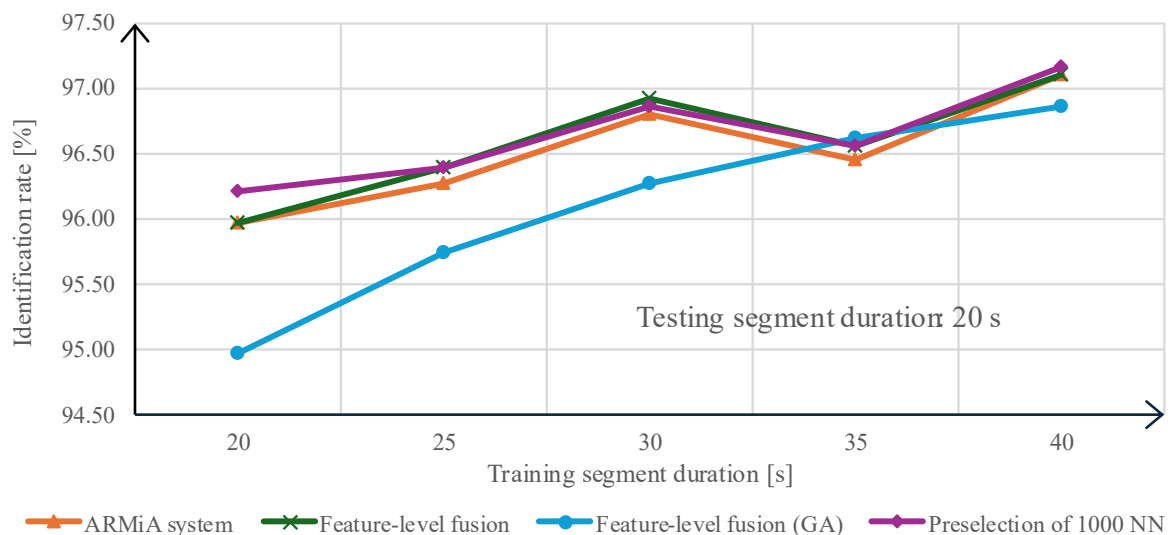


Fig. 4. Speaker identification rate of different ASR solutions for varying training segment durations.

This selection of noise types was driven by the need to evaluate the developed solutions in conditions as close to real world scenarios as possible. Additionally, each of these disturbances exhibits a different power distribution in the frequency spectrum. The tests were conducted for SNR levels of 20, 15, and 10 dB. A detailed description of the experimental setup and noise characteristics can be found in [4]. Figure 5 presents the results for the highest levels for each type of noise. A system variant was also tested where only the testing segments were affected by noise [61]. Figure 6 shows results obtained when only the test segment was degraded with white noise.

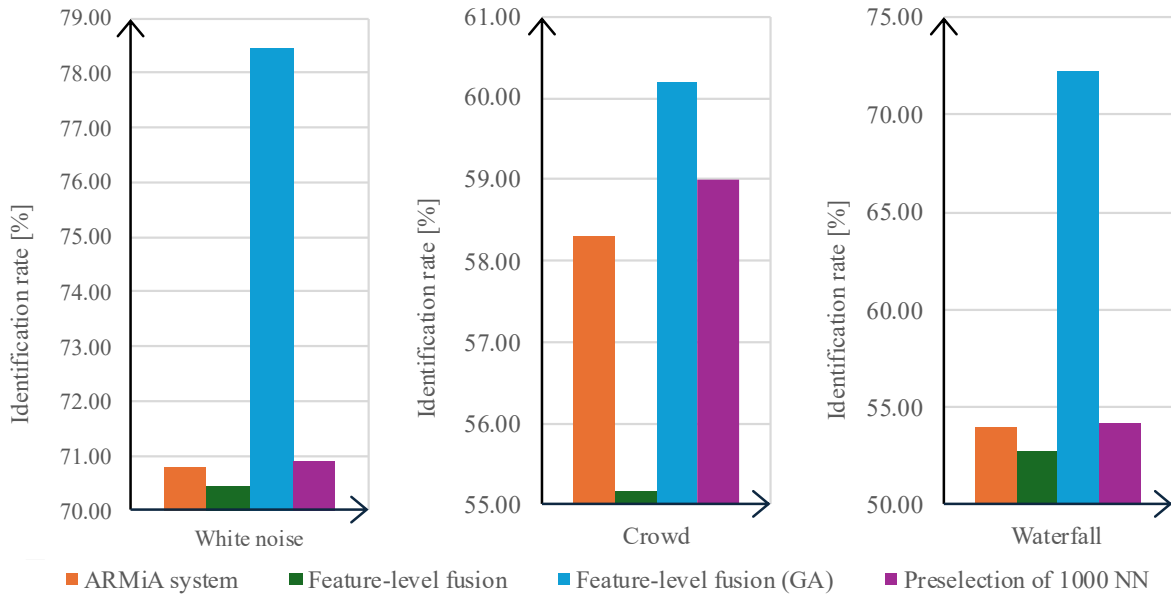


Fig. 5. Test results of different variants of the ASR system under degradation by various interferences, SNR = 10 dB.

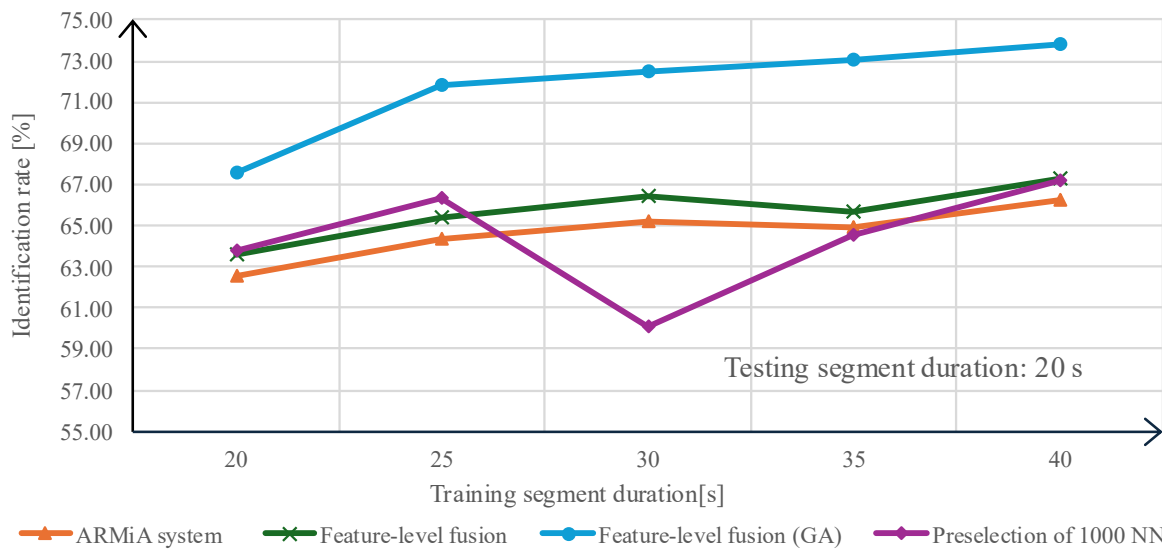


Fig. 6. Test results of different variants of the ASR system under degradation of the testing segment with white noise, SNR = 20 dB.

5.2. Evaluation of the developed solutions under different recording quality conditions

One of the important aspects to consider when using a selected ASR solution is the method of voice transmission through a transmission medium. The utilized telecommunication channel is a factor that modifies the transmitted speech signal due to the speech encoding process. This is a lossy voice compression process aimed at enabling its transmission within a telecommunication system [62]. The purpose of this experiment was to examine the number of correct speaker identity identifications under speech compression conditions. The following phonetic encoding standards were used: G-711, GSM 06.10, G.723.1, SPEEX. A description of individual speech encoding standards is provided in the article that describes the reference *ARMiA* system [22].

The results were presented only for the combination of 40 seconds of training and 20 seconds of testing. The green color indicates an IR higher than in the reference system, while the orange color represents the same rate.

Table 4. *ARMiA* system performance values for different encoding standards.

<i>Testing segment codec</i>	<i>Training segment codec</i>			
	G.711 - a-law	SPEEX	G.723.1	GSM 06.10
G.711 - a-law	96.80 %	94.85 %	91.47 %	92.89 %
SPEEX	94.61 %	96.15 %	93.19 %	91.82 %
G.723.1	90.70 %	93.13 %	95.38 %	89.04 %
GSM 06.10	92.36 %	91.71 %	89.16 %	96.33 %

Table 5. Performance values of the behavioral feature-based system for different encoding standards.

<i>Testing segment codec</i>	<i>Training segment codec</i>			
	G.711 - a-law	SPEEX	G.723.1	GSM 06.10
G.711 - a-law	57.82 %	27.55 %	34.42 %	37.50 %
SPEEX	26.13 %	44.02 %	28.91 %	23.58 %
G.723.1	30.45 %	28.97 %	44.14 %	36.55 %
GSM 06.10	35.60 %	24.70 %	37.74 %	50.65 %

Table 6. Performance values of the feature-level fusion variant for different encoding standards.

<i>Testing segment codec</i>	<i>Training segment codec</i>			
	G.711 - a-law	SPEEX	G.723.1	GSM 06.10
G.711 - a-law	96.80 %	94.91 %	91.88 %	93.01 %
SPEEX	94.25 %	96.39 %	92.77 %	91.53 %
G.723.1	90.82 %	93.07 %	95.38 %	89.69 %
GSM 06.10	92.36 %	91.35 %	88.74 %	96.39 %

Table 7. Performance values of the feature-level fusion variant after applying the genetic algorithm for different encoding standards.

<i>Testing segment codec</i>	<i>Training segment codec</i>			
	G.711 - a-law	SPEEX	G.723.1	GSM 06.10
G.711 - a-law	96.74 %	95.14 %	92.65 %	91.11 %
SPEEX	93.19 %	95.50 %	92.71 %	89.10 %
G.723.1	87.80 %	91.17 %	94.61 %	85.19 %
GSM 06.10	89.63 %	90.46 %	88.45 %	95.97 %

Table 8. Performance values of the decision-level fusion variant for different encoding standards.

<i>Testing segment codec</i>	<i>Training segment codec</i>			
	G.711 - a-law	SPEEX	G.723.1	GSM 06.10
G.711 - a-law	96.86 %	95.02 %	91.65 %	93.07 %
SPEEX	94.02 %	96.27 %	92.89 %	91.05 %
G.723.1	90.94 %	93.25 %	95.44 %	89.34 %
GSM 06.10	92.54 %	91.94 %	89.40 %	96.39 %

5.3. Computational efficiency

The final test conducted was the measurement of the average speaker identification time within a dataset of 10,000 voice models. The evaluated systems included the system from and the decision-level data fusion method. Each variant was tested 100 times and the obtained time measurements were averaged. The first ASR solution analyzed had an average speaker identification time of 5.78 seconds. In contrast, when using the behavioral preselector, the identification process required only 0.7 seconds. These results demonstrate that the decision-level fusion solution processes a single voice sample more than eight times faster. This significant reduction in processing time is due to the initial dataset of 10,000 voice models being reduced by the behavioral classifier to the 1,000 nearest neighbors relative to the analyzed (test) object. As a result, in the next stage, the module that uses physical features [1] compares the test model with a ten times smaller set of training models. Figure 7 illustrates the comparison of these results.

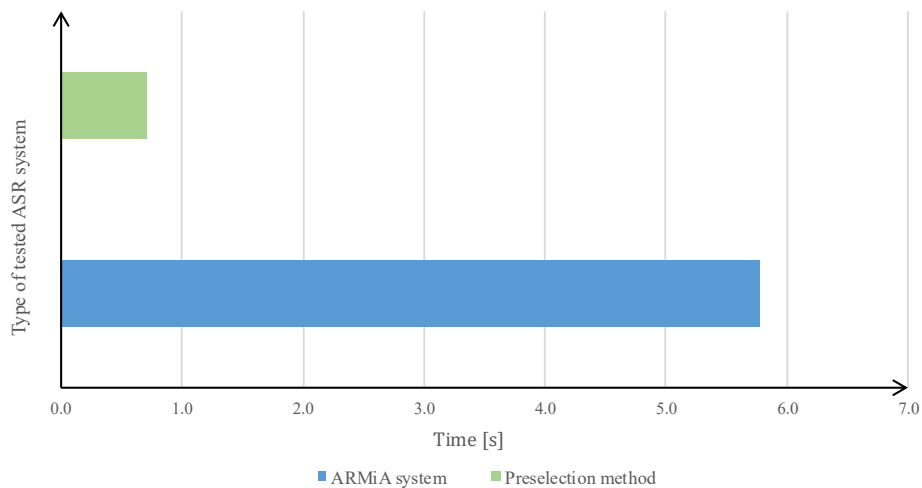


Fig. 7. Average speaker identification times in a dataset of 10,000 voice models.

The more than eight-fold reduction in analysis time comes with only a slight increase in disk storage size (for a dataset of 10,000 models), from approximately 32 to 34 MB. This change results from the need to store, in addition to GMMs, a set of behavioral feature vectors from the training data. However, significant improvement in speaker identification speed demonstrates that the use of a behavioral feature-based preselector significantly improves the scalability of the existing system [1].

5.4. Comparison with the literature

In Section 2, publications were cited in which the authors utilized different variants of the *LibriSpeech* dataset. Unfortunately, a fully objective comparison of the discussed solutions is not possible, as the speech signal datasets used are highly varied. Furthermore, authors do not always provide complete details of the described experiments (*e.g.*, the duration of training and testing segments). Therefore, the Table 9 below presents speaker identification accuracy results obtained under conditions as close as possible to those described in the cited publications. Due to the lack of complete information provided in these works, we decided to use a training and testing time combination that ensures the best operating conditions for our proposed solution. This configuration was validated through the tests shown in the previous sections and described in detail in [4].

Table 9. Comparison of the developed system with the literature

Dataset used	Training time / Testing time [s]	IR of the developed ASR system fusion	IR of comparative systems	Literature reference
<i>LibriSpeech</i> train-clean-100 (251 speakers: 125 women and 126 men)	40/20	100.00% – feature-level fusion 99.60% - preselector	83.1% - GMM 99.8% - <i>DeepSpeaker</i>	[6]
<i>LibriSpeech</i> train-clean-100 (251 speakers: 125 women and 126 men)	40/20	100.00% – feature-level fusion 99.60% - preselector	97.87% - 64 mel filter bank 95.25% - MFCC 97.31% - <i>SpeakerGAN</i>	[7]
<i>LibriSpeech</i> train-clean-360 (921 speakers)	40/20	98.59% – feature-level fusion 98.48% - preselector	96.08% - MFSC + <i>Transformer</i> 95.89% - MFSC + BILSTM 95.04% - MFSC + BIGRU	[19]

6. Conclusion

This article addresses the use of behavioral voice features in ASR systems. The authors aimed to develop and implement a set of behavioral features in an existing ASR system [1] that would increase the number of correct speaker identity identifications, particularly in the presence of various types of noise.

By utilizing the publicly available *LibriSpeech* voice database, it was possible to compare the developed solution with other ASR systems. However, this task is challenging due to the lack of standardized comparison procedures. However, the authors attempted to replicate the experimental conditions described in scientific publications as accurately as possible to ensure the highest reliability of the results obtained. On the basis of these results, they can be considered satisfactory, as a higher identification rate was achieved in all cases.

The results of experiments conducted on an expanded voice dataset, as well as in the presence of external noise, indicate that the authors have developed a solution that can reduce the impact of external noise on speaker identity recognition accuracy in ASR systems. The key element proved to be the innovative data integration method, which leverages the advantages of various sources of distinctive feature sets.

The developed ASR system still requires further refinement to meet the increasing challenges faced by voice biometrics. With the rapid advancement of artificial intelligence, the ability to fabricate voice samples (deepfake) and thus impersonate another person is becoming increasingly easier.

References

- [1] Kamiński, K. A., & Dobrowolski, A. P. (2022). Automatic speaker recognition system based on Gaussian mixture models, Cepstral analysis, and genetic selection of distinctive features. *Sensors*, 22(23), 9370. <https://doi.org/10.3390/s22239370>
- [2] Ming, J., Hazen, T. J., Glass, J. R., & Reynolds, D. A. (2007). Robust speaker recognition in noisy conditions. *IEEE Transactions on Audio Speech and Language Processing*, 15(5), 1711–1723. <https://doi.org/10.1109/tasl.2007.899278>
- [3] Staroniewicz, P. (2018). Influence of natural voice disguise techniques on automatic speaker recognition. In *2018 Joint Conference – Acoustics*, Ustka, Poland, <https://doi.org/10.1109/acoustics.2018.8502372>

- [4] Mały, D. (2025). Automatic speaker recognition system supported by behavioral features of speech signal [PhD Thesis]. Military University of Technology
- [5] Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia*, 5206–5210. <https://doi.org/10.1109/icassp.2015.7178964>
- [6] Natasya, C., Gunawan, A. a. S., & Komsiyah, S. (2024). Enhancing Two-Factor authentication with deep speaker recognition on Librispeech Dataset. In *2024 International Conference on Information Management and Technology (ICIMTech), Bali, Indonesia, 2024*, 777-782. <https://doi.org/10.1109/icimtech63123.2024.10780859>
- [7] Spisiak, M., Jakubec, M., Jarina, R., & Kasak, P. (2024). Efficient Low-Complexity Speaker identification based on a SpeakerGAN approach. In *2024 34th International Conference Radioelektronika (RADIOELEKTRONIKA)*, 27, 1–5. <https://doi.org/10.1109/radioelektronika61599.2024.10524050>
- [8] Kinnunen, T., Zhang, B., Zhu, J., & Wang, Y. (2007). Speaker Verification with Adaptive Spectral Subband Centroids. In *Lecture notes in computer science* (pp. 58–66). https://doi.org/10.1007/978-3-540-74549-5_7
- [9] Kinnunen, T., & Li, H. (2009). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52(1), 12–40. <https://doi.org/10.1016/j.specom.2009.08.009>
- [10] Scheirer, E., & Slaney, M. (2002). Construction and evaluation of a robust multifeature speech/music discriminator. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2, 1331–1334. <https://doi.org/10.1109/icassp.1997.596192>
- [11] Giannakopoulos, T., & Pirkakis, A. (2014). Introduction to Audio Analysis: a MATLAB approach. In *Elsevier eBooks*. <https://doi.org/10.1016/c2012-0-03524-7>
- [12] Laptik, R., & Sledevic, T. (2017). Fast binary features for speaker recognition in embedded systems. In *2017 Open Conference of Electrical, Electronic and Information Sciences (eStream)*, 1, 1–4. <https://doi.org/10.1109/estream.2017.7950317>
- [13] Kinnunen, T. (2006). Joint Acoustic-Modulation frequency for speaker recognition. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, 1, 665-668. <https://doi.org/10.1109/icassp.2006.1660108>
- [14] Kinnunen, T.H., LEE, K., & Li, H. (2008). Dimension reduction of the modulation spectrogram for speaker verification. *The Speaker and Language Recognition Workshop*. <https://cs.joensuu.fi/pages/tkinnu/webpage/pdf/odyssey2008modspec.pdf>
- [15] Prachi, N. N., Nahiyani, F. M., Habibullah, M., & Khan, R. (2022). Deep learning based speaker recognition system with CNN and LSTM techniques. In *2022 Interdisciplinary Research in Technology and Management (IRTM)*, 1–6. <https://doi.org/10.1109/irtm54583.2022.9791766>
- [16] Shirzad, A., Nasiri, A., Darshi, R., Safarpour, Z., & Abdollahipour, R. (2024). Text-independent speaker recognition: a deep learning approach. In *2024 10th International Conference on Control, Decision and Information Technologies (CoDIT)*, 1, 692–696. <https://doi.org/10.1109/codit62066.2024.10708578>
- [17] Keras: Deep Learning for humans. (n.d.). Retrieved January 26, 2025, from <https://keras.io>
- [18] Jahangir, R., TEh, Y. W., Memon, N. A., Mujtaba, G., Zareei, M., Ishtiaq, U., Akhtar, M. Z., & Ali, I. (2020). Text-Independent speaker identification through feature fusion and deep neural network. *IEEE Access*, 8, 32187–32202. <https://doi.org/10.1109/access.2020.2973541>
- [19] Bao, L., & Zuo, Y. (2023). Speaker identification based on MFSC Voice feature extraction using transformer. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, 1–7. <https://doi.org/10.1109/icdmw60847.2023.00008>
- [20] Saritha, B., Laskar, M. A., Laskar, R. H., & Choudhury, M. (2022). Raw waveform based speaker identification using deep neural networks. In *2022 IEEE Silchar Subsection Conference (SILCON)*, 1–4. <https://doi.org/10.1109/silcon55242.2022.10028890>
- [21] Hong, Q., Wu, C., Wang, H., & Huang, C. (2020). Combining deep embeddings of acoustic and articulatory features for speaker identification. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7589–7593. <https://doi.org/10.1109/icassp40776.2020.9053640>
- [22] Kamiński, K. (2018). Automatic Speaker system based on cepstral analysis of the speech signal Gaussian Mixture Models [PhD Thesis]. Military University of Technology.

- [23] Martin, A., & Przybocki, M. (2004). 2002 NIST Speaker Recognition Evaluation [Dataset]. In *Americanae (AECID Library)*. <https://doi.org/10.35111/axbp-bw85>
- [24] French, N. R., & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *The Journal of the Acoustical Society of America*, 19(1), 90–119. <https://doi.org/10.1121/1.1916407>
- [25] Makowski, R. (2011). Automatic speech recognition – selected problems. Publishing House of Wrocław University of Science and Technology.
- [26] detectSpeech. (2020). <https://www.mathworks.com/help/audio/ref/detectspeech.html>
- [27] Giannakopoulos, T. (2009). A method for silence removal and segmentation of speech signals, implemented in Matlab. www.di.uoa.gr
- [28] Zieliński, T. (2005). Digital signal processing. From theory to applications. WKŁ Publishing
- [29] Dobrowolski, A. (2018). Signal transformations – from theory to practice. BTC
- [30] Sreedharan, S., & Eswaran, C. (2018). An analysis of feature selection based on optimization algorithms for speaker verification. In *2018 Tenth International Conference on Advanced Computing (ICoAC)*, 105–111. <https://doi.org/10.1109/icoac44903.2018.8939099>
- [31] Dahlman, E., Parkvall, S., & Sköld, J. (2014). 4G: LTE/LTE-Advanced for mobile broadband. In *Elsevier eBooks*. <https://doi.org/10.1016/c2013-0-06829-6>
- [32] International Telecommunication Union. (2006). RECOMMENDATION ITU-R SM.328-11 * Spectra and bandwidth of emissions (Question ITU-R 222/1). https://www.itu.int/dms_pubrec/itu-r/rec/sm/R-REC-SM.328-11-200605-S!!PDF-E.pdf
- [33] Lerch, A. (2012). *An introduction to audio content analysis*. John Wiley & Sons. <https://doi.org/10.1002/9781118393550>
- [34] Peeters, G. (2004). *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*. http://recherche.ircam.fr/anasyn/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf
- [35] Loughran, R., Agapitos, A., Kattan, A., Brabazon, A., & O'Neill, M. (2017). Feature selection for speaker verification using genetic programming. *Evolutionary Intelligence*, 10(1–2), 1–21. <https://doi.org/10.1007/s12065-016-0150-5>
- [36] Quixtiano-Xicohtencatl, R., Flores-Pulido, L., & Reyes-Galaviz, O. (2006). Feature selection for a fast speaker detection system with neural networks and genetic algorithms. In *2006 15th International Conference on Computing*, 19, 126–134. <https://doi.org/10.1109/cic.2006.38>
- [37] Kamiński, K. (2015). Optymalizacja systemu automatycznego rozpoznawania mowy w warunkach zróżnicowanych torów akustycznych. *PRZEGLĄD ELEKTROTECHNICZNY*, 1(9), 91–94. <https://doi.org/10.15199/48.2015.09.23>
- [38] Barszcz, T., & Zabaryło, M. (2021). Automatic identification of malfunctions of large turbomachinery during transient states with genetic algorithm optimization. *Metrology and Measurement Systems*, 175–190. <https://doi.org/10.24425/mms.2022.138551>
- [39] Li, X., Jia, J., Yang, D., & Gu, Y. (2024). An integration method of a hybrid genetic algorithm and the Levenberg–Marquardt algorithm for ultrasonic testing. *Metrology and Measurement Systems*, 165–177. <https://doi.org/10.24425/mms.2024.148536>
- [40] Kamiński, K. A., Dobrowolski, A. P., Piotrowski, Z., & Ścibiorek, P. (2023). Enhancing web application security: Advanced biometric voice verification for Two-Factor authentication. *Electronics*, 12(18), 3791. <https://doi.org/10.3390/electronics12183791>
- [41] Bell, O. (2022). Applications of Gaussian mutation for self-adaptation in evolutionary genetic algorithms. *Journal of Machine Learning in Fundamental Sciences JMLFS-ID*. <https://arxiv.org/abs/2201.00285>
- [42] Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review*, 57(3), 238. <https://doi.org/10.2307/1403797>
- [43] Szabo, F. E. (2015). The Linear Algebra Survival Guide. In *Elsevier eBooks*. <https://doi.org/10.1016/c2012-0-06836-6>
- [44] Taxicab geometry. Retrieved February 10, 2025, from https://en.wikipedia.org/wiki/Taxicab_geometry

- [45] Dobrowolski, A., & Mały, D. (2023). Behavioral features of the speech signal as part of improving the effectiveness of the automatic speaker recognition system. *Inżynieria Bezpieczeństwa Obiektów Antropogenicznych*, 4, 26–34. <https://doi.org/10.37105/iboa.187>
- [46] Majda, E. (2013). Automatic system of reliable speaker recognition based on cepstral analysis of the speech signal [PhD Thesis]. Military University of Technology
- [47] Doddington, G. (2001). Speaker recognition based on idiolectal differences between speakers. In *7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, 2521–2524 <https://doi.org/10.21437/eurospeech.2001-417>
- [48] Henderson, T. C., Dekhil, M., Kessler, R. R., & Griss, M. L. (2005). Sensor fusion. In *Lecture notes in control and information sciences* (pp. 193–207). <https://doi.org/10.1007/bfb0015084>
- [49] Wójtowicz, B., Dobrowolski, A., & Tomczykiewicz, K. (2015). Fall detector using discrete Wavelet decomposition and SVM classifier. *Metrology and Measurement Systems*, 22(2), 303–314. <https://doi.org/10.1515/mms-2015-0026>
- [50] Luo, R., & Kay, M. (1989). Multisensor integration and fusion in intelligent systems. *IEEE Transactions on Systems Man and Cybernetics*, 19(5), 901–931. <https://doi.org/10.1109/21.44007>
- [51] Luo, R., Yih, C., & Su, K. L. (2002). Multisensor fusion and integration: approaches, applications, and future research directions. *IEEE Sensors Journal*, 2(2), 107–119. <https://doi.org/10.1109/jsen.2002.1000251>
- [52] Wójtowicz, B., Dobrowolski, A. P. (2013). Design of a sensor data integrator for uncontrolled fall detection. *Biuletyn WAT Vol. LXII(4)*, 229–240. <https://www.wojsko-polskie.pl/wat/u/d2/fa/d2fac4f4-fa1c-4161-af8a-462df5696ad4/2013-04.pdf>
- [53] Rothman, P. L., & Denton, R. V. (1991). Fusion or confusion: knowledge or nonsense? *Proceedings of SPIE, the International Society for Optical Engineering/Proceedings of SPIE*, 1470, 2–12. <https://doi.org/10.1117/12.44835>
- [54] Elmenreich, W. (2002). Sensor Fusion in Time-Triggered Systems [PhD Thesis]. <https://www.researchgate.net/publication/215499135>
- [55] Dasarthy, B. (1997). Sensor fusion potential exploitation-innovative architectures and illustrative applications. *Proceedings of the IEEE*, 85(1), 24–38. <https://doi.org/10.1109/5.554206>
- [56] Durrant-Whyte, H. F. (1988). Sensor Models and Multisensor Integration. *The International Journal of Robotics Research*, 7(6), 97–113. <https://doi.org/10.1177/027836498800700608>
- [57] Wójtowicz, B. (2015). Fall detector using wavelet-based generation of distinctive features and an SVM classifier [PhD Thesis]. Military University of Technology
- [58] Posiadała, D. (2015). Study of the effectiveness of a speaker recognition system in a diverse acoustic environment [B.Sc. Thesis]. Military University of Technology
- [59] Bojsza, M. (2021). Assessment of the impact of noise and disturbances on the effectiveness of an Automatic Speaker Recognition system under conditions of linguistic diversity [M.Sc. Thesis]. Military University of Technology.
- [60] Kamiński, K., Dobrowolski, A. P., & Tatoń, R. (2019). The assessment of efficiency of the automatic speaker recognition system for voices registered using a throat microphone. In *Proc. SPIE 11055, XII Conference on Reconnaissance and Electronic Warfare Systems, 11055*. <https://doi.org/10.1117/12.2524591>
- [61] Kamiński, K., Dobrowolski, A. P., & Majda-Zdancewicz, E. (2014). Ocena funkcjonalności systemu rozpoznawania mowy dla zdegradowanej jakości sygnału głosowego. *Przegląd Elektrotechniczny*, 90, Article 8. <https://doi.org/10.12915/pe.2014.08.38>
- [62] G, N. B., Anees, M., & G, T. Y. (2023). Speech coding techniques and challenges: a comprehensive literature survey. *Multimedia Tools and Applications*, 83(10), 29859–29879. <https://doi.org/10.1007/s11042-023-16665-3>



Dominik Mały received the M.Sc. degree in electronics and telecommunications from the Faculty of Electronics, Military University of Technology, Warsaw, Poland, in 2021. He is currently pursuing a Ph.D. degree at the Military University of Technology in the field of automation, electronics, electrical engineering, and space technologies. He is an engineer at the Institute of Radioelectronics,

Faculty of Electronics, Military University of Technology. His research interests include automatic speaker recognition systems and optimization methods.



Andrzej Dobrowolski received the D.Sc. degree in electronics in 2010 from the Military University of Technology, Warsaw, Poland. In 2017, the President of Poland awarded him the title of Professor of Technical Sciences. From 2012 to 2016, he served as Deputy Dean for Scientific Issues at the Faculty of Electronics, Military University of Technology, where he also managed doctoral programs. From 2016 to 2020,

he served as Dean of the Faculty of Electronics, and in the following four years, he held the position of Vice Rector for Scientific Issues at the Military University of Technology. He is the author or coauthor of four books and more than 200 journal and conference papers. His research interests include digital signal processing, artificial intelligence, medical diagnosis, and biometrics. He has accumulated over 9,000 hours of teaching experience and has supervised eight Ph.D. dissertations.



Kamil Kamiński received the Ph.D. degree in technical sciences in the field of electronics from the Faculty of Electronics, Military University of Technology, Warsaw, Poland, in 2018. Since 2011, he has been conducting research on the digital processing of speech signals. He has authored more than 20 scientific papers. His research interests include voice signal processing, biometric data analysis, and

AI-based methods. He is currently an Assistant Professor at the Institute of Optoelectronics, Military University of Technology. In addition, he is the co-founder of BITRES, a company that specializes in the implementation of voice biometrics systems.