Changes in the results of voice biometric software using different methods (GMM-UBM, i-vector) in the case of different speech tasks and voice sample durations

Attila Fejes[1,2], Dávid Sztahó[3]

[1]*Special Service for National Security Institute for Expert Services*
[2]*Doctoral School of Law and Political Sciences, Széchenyi István University*
[3]*Budapest University of Technology and Economics*

Abstract

During forensic speaker comparison, the audio forensics expert appointed to perform the investigation works with audio recordings of different types and durations. Distinct speech samples and durations affect the probability data. In order to evaluate biometric identification results, the probability value of the data obtained must be determined so that the expert's report can be accurate and interpreted by other actors in the public proceedings. In the present study, the speech samples of 78 speakers from the forensic voice sample database were compared within the framework of the FORENSICSpeech research project (Beke et al., 2020). The samples include three different types of speech: spontaneous, read, and narration speech. The recording of the samples was repeated after an average of two weeks, and then the audio files were cut into 20, 40, 60, 80, 100, and 120 seconds in duration using automatic editing. The aim of this study is to show how different speech styles and durations affect voice biometric identification results.

Results show that EER (Equal Error Rate) and FRR (False Reject Rate), Cllr (Log likelihood ratio cost) values decrease with increasing duration; however, in the 20–120-second range, the change is not continuous. Similarly, the lowest EER, FRR, Cllr, and Cllr- min values occur in the case of spontaneous speech, followed by narration, while the speech samples of information exchange give the highest Cllr values. The data as a whole is characterized by the fact that the more advanced i-vector method tends to provide more efficient, lower error-rate person identification results.

1. Introduction

The application of biometric methodology is an important element in speech-based speaker identification in forensic science. Biometrics determines the likelihood of the sample owner's identity using certain biological or behavioral char-

Email addresses: fejes.attila@nbsz.gov.hu (Attila Fejes), sztaho.david@vik.bme.hu (Dávid Sztahó)

Beszédtudomány – Speech Science 2025/2. 132–153. DOI: 10.15775/Besztud.2025.2.132-153

acteristics. In the case of voice biometrics, voice is the feature whose uniqueness makes it possible to use as a biometric feature. This is due to the fact that every human being has different biological (physical) dimensions; no two human bodies are exactly alike, including all organs involved in speech production. Speech is also influenced by the individual's personality, sociocultural environment, education, emotional and intellectual intelligence, and a number of factors that may not appear together in the case of another person (Anil et al., 2011; Gráczi et al., 2022; Leemann et al., 2025).

The advantage of voice biometrics is that the result is independent of the expert performing the test; its validity and error rate can be accurately measured, and it can be well-automatized, so it is suitable for mass data processing. Technology provides a probabilistic result, so it does not define a categorical identity or difference. The probability of identity is determined in different forms by the high-tech systems available today, calculated from Score data (Morrison, 2013; Kelly et al., 2019), Likelihood Ratio (LR) (Van der Vloed, 2016; Zhang. & Tang, 2018), and its decimal-based logarithmic value (LLR) (Jessen et al., 2019). In our research, we performed measurements with the Batvox software, working with two distinct version numbers and different biometric identification engines. Both versions' output is LR data, and their inputs are the voice samples to be compared. Version 3.1 is based on GMM-UBM (Gaussian Mixture Models – Universal Background Model) (Zhang. & Tang, 2018), while version 4.1 uses the PLDA (Probabilistic Linear Discriminant Analysis) method with i-vector extraction (Van der Vloed, 2016).

In the detection and proof of criminal offenses, it is common for the audio of an unknown person to be compared with the speech sample of a known speaker recorded by an expert working on the case. During sampling, the expert records several speech samples of different types of the known person and compares each with the unknown speaker's voice recording (Fejes, 2022). The type of the speech sample influences the result of the identification; however, the extent of this can be determined by implementing performance tests. The probability value is also

affected by the duration of the research material, for which different biometric methods set dissimilar threshold levels (Meuwly, 2009).

In the present study, we aimed to show the effect of changing the type and duration of the speech sample on the biometric identification results. These factors are important because in forensic science, the expert often only has short recordings at his disposal, so we need to understand the relationship between duration and performance for a given method. On the other hand, in forensic audio sampling, the expert working on a given case uses several types of samples, so we need to know the characteristics of different types of samples in terms of identification results. In our study, we focused on speech duration and speech type. For the comparison, we created the same test conditions for both versions, and in this way, we produced 36 identification matrices per software version using samples from speakers of both genders. Data were converted into LLR format for evaluation, and performance metrics and other data were evaluated using the Bio-Metrics 1.8 software (Kelly et al., 2019).

1.1. Methodology of Forensic Voice Comparison in Hungary

The purpose of Forensic Voice Comparison (FVC) is to determine the probability of the compared speaker's identity. In typical cases, there is a sample of a recorded unknown person via wiretapping and a suspect speaker whose identity is known. In other cases, samples of unknown speakers need to be compared to determine the probability of identity.

The FVC methodology includes auditory phonetic-linguistic analysis, acoustic measurements, and the use of voice biometrics technology. In the auditory analysis, the expert examines the features of articulation, language and speech, dialect, idiolect, hesitation phenomena, speech pathology, etc. With special expert software, one can measure, for example, the similarity of formants in matching sounds, fundamental frequency (f0), and formant frequencies (European Network of Forensic Science Institutes, 2022). After the phonetic-linguistic analysis and acoustic measurements, the audio forensics expert assesses the

similarities and differences in these features and determines the probability of identity utilizing available scientific background information.

Voice biometric measurements are the final stage of the speech analysis because the results can influence the expert, as they may cause cognitive bias (Kovács, 2017). The auditory phonetic-linguistic analysis and the acoustic measurements depend on the expert's judgement. In contrast, voice biometric measurements are objective and reproducible, and the results are independent of the person conducting the analysis. Automatic speaker recognition systems are the state-of-the-art technologies in voice comparison nowadays, and an efficient and powerful tool (Ramos, 2007:9–10). The voice biometric methodology is used in Forensic Automatic Speaker Recognition (FASR) systems, which are used by audio forensic experts. Of these, Batvox is not the latest technology, but it is still in use in Hungarian forensic science.

2. Methods

2.1. Measurements

Speech samples were selected from the FORENSICSpeech (Beke et al., 2020; Sztahó et al., 2021) project database. We used Spontaneous, Read, and Narration speech samples. Recordings were conducted in a quiet office room, similar to those found in expert sampling, using a laptop, an external sound card, and a condenser microphone. There was no background noise at the recording site, and the recording was made at a sampling rate of 44,100 Hz with a depth of 16 bits. The age of the speakers ranged from 16 to 48 years, and 39 female and 39 male speech samples were measured. Samples were recorded during two separate sessions on different days, which will be referred to as sessions 1 and 2. Spontaneous conversations recorded at the first session served as models (i.e., the known speaker's speech sample). These were compared with the Spontaneous, Read, and Narration-like monologue (the recollection of the events of the speaker's previous day) audio recordings, comparing the first session with the second session of audios. Samples were cut into clips of six different dura-

tions with 20-second increments, ranging from 20 seconds to 120 seconds. The methods of audio expert sampling were applied to the recordings, similarly to a real forensic condition. All recordings were automatically edited. First, pauses longer than 500 milliseconds were removed, and then the samples were divided into chunks with the desired durations. No phonological boundaries were considered during the chunking. Each chunk has the exact target duration. Accordingly, during the measurements, we created 36 identification matrices per software version with speech samples of male and female speakers, as all three different styles were compared for each of the six different durations. A typical matrix contains 39 x 39 = 1,521 probability values, as the software compares all speakers against each other. There are LLR data of 39 same (SS-Same Source <same speakers>) and 1,482 different (DS-Different Source <different speakers>) speakers. On the x and y axes are voice samples from the same speakers, recorded at different times and speech styles. The samples from the first recordings are plotted on the y-axis, and the samples from the second recordings are plotted on the x-axis. The samples from the first recordings were always the same, while in each matrix, the length and speech style of the samples from the second recordings were adjusted.

2.2. The Bayesian framework and the method of evaluating the results

In the methodology of speech-based personal identification, we analyze (through perceptual and acoustic-phonetic studies) and measure (using voice biometrics) various sound parameters (Drygajlo et al., 2015). Depending on the methodology, the characteristics are evaluated in text, measured manually, or the biometric software calculates the probability of identity from the data using mathematical and statistical methods (Craig, 2010). Since we do not know the characteristics of speech that uniquely represent the speaker, we do not look for matching data; instead, we compare and infer probability. In addition to that, the even greater difficulty is that there is within-speaker variation in voice characteristics. Thus, in a speaker identification study, two hypotheses must be considered and calculated by the voice biometrics software: the probability of evidence

(Morrison, 2009). Accordingly, the null hypothesis (H0) is that the speakers on the two audio recordings being compared are the same, and the alternative hypothesis (H1) is that the speakers on the two recordings are different. The software then provides the probabilities for each of these hypotheses to be true. The Bayesian framework (Meester & Slooten, 2021) is suitable for determining the strength of the evidence and the probability of identity by considering and calculating both probabilities. In the Bayesian approach, the competence of the audio engineering expert conducting the study is to determine and evaluate LR. The relationship between probabilities is shown by the formula in Figure 1 below.

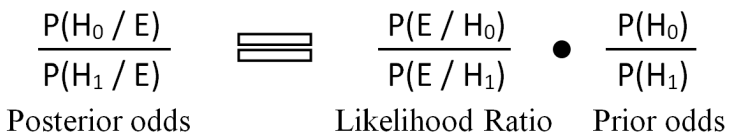

Figure 1: Bayesian framework formula.

H0 is the hypothesis that the speaker on the two recordings is the same, and $P(E|H0)$ is the probability of observing the evidence given that H0 is true. H1 is the hypothesis that the speaker on the two recordings is different, and $P(E|H1)$ is the probability of observing the evidence given that H1 is true. E (Evidence) denotes the sound recording as evidence, and P (Probability) expresses probability. It can be seen that a priori knowledge is weighted by the strength of the evidence as determined by the expert, taking into account both hypotheses. The statement about the strength of evidence is the Likelihood Ratio (LR), which is the ratio of the probability of the two hypotheses, denoted by H0 and H1 in the formula.

Biometric measurements were performed with Batvox 3.1 and 4.1. For both versions, the user must create a reference population database, which should consist of audio files with the same characteristics (gender, language, channel, and speech type) as the audio recording group used as a model. The same

reference population database, comprising 80 female and male audio files, was used for both software versions in our study. The audio files of the reference population databases were randomly selected from the Hungarian Spontaneous Speech Database (Gósy et al., 2012). The first step in biometric identification is feature extraction in the case of both software versions. Due to the uniqueness of the vocal tract and other organs involved in speech, the forms of speech sounds are unique characteristics of the speaker, so their acoustic parameters can be measured and subtracted by the software as described below.

Voice Biometrics technology, as applied in our research, utilizes the envelope of the audio signal spectrum to extract its characteristics. To do this, both Batvox versions split the voice into 20-millisecond windows with 50% overlap. Then, they extract the individual characteristics from the spectrum using the Mel-Frequency Cepstrum Coefficient (MFCC) method. In addition to spectral characteristics, the subtraction of phonetic and prosodic characteristics, fundamental frequency, and energy conditions of voice provide additional data on the speaker's speech. After feature extraction, the system sets up feature vectors and performs speech modeling. The two Batvox versions used in the study have the same interface, and the measurement sessions are also configured in the same way, with one exception: in the more modern version (4.1), the known speaker's audio recording has a minimum duration of 30 seconds, while version 3.1 defines a minimum of 40 seconds as an input requirement.

The results were obtained in LR format, which were converted to a decimal-based logarithmic format (LLR) for full-scale evaluation, as only a Tippet Plot graph can be created for LR format data. "The Tippett plot is a cumulative probability distribution plot expressing the proportion of likelihood ratios (LRs) greater than a given value, i.e., $P(LR(H) > LR)$, for cases corresponding to the H0 hypothesis (biometric samples are from the same source) and the H1 hypothesis (biometric samples are from different sources)" (Oxford Wave Research, 2025). In this study, H0 and H1 in the above Bayesian approach formula are not the same as the H0 and H1 denoted below. Outside the formula, H0 data is equal to the probability values of same speakers, and H1 data is equal to the

probability values of different speakers. The evaluation was performed using Oxford Wave Research Bio-Metrics 1.8 (Oxford Wave Research, 2025) using the following outputs to analyze the data:

- Mean of H0 and H1: arithmetic mean of LLR data from hypotheses H0 and H1;
- Standard Deviation of H0 and H1: standard deviation of LLR data from hypotheses H0 and H1;
- Log likelihood ratio cost (Cllr): the degree of calibration, the accuracy of the system;
- False acceptance rate (FAR) is the rate at which the comparison between two different individuals' samples is erroneously accepted by the system as a true match. In other words, FAR is the percentage of impostor scores that are higher than the decision threshold;
- False rejection rate (FRR) is the percentage of times when an individual is not matched to his/her own existing reference templates. In other words, FRR is the percentage of the genuine scores that are lower than the decision threshold;
- Equal error rate (EER) is the rate at which both acceptance and rejection errors are equal (i.e., FAR=FRR). Generally, the lower the EER value, the higher the accuracy of the software;
- Detection Error Trade-off (DET) plot: represents FAR and FRR values.

3. Results

The values of Same Source (SS) LLR data, which indicate the identity of speakers, tend to increase with longer, higher-quality audio materials, and the variance also decreases as the system's robustness improves. The values of Different Source (DS) LLR data, which indicate the likelihood that the compared

audio samples come from different individuals, typically decrease as the signal-to-noise ratio (SNR) and duration increase. This is because a low SNR reduces the performance of the analysis, while a longer duration enhances it. In general, the effectiveness of feature extraction is reduced by noise and shorter audio length. The degree of standard deviation of SS and DS LLR data, interpreted separately, suggests discriminatory power: a high-performance voice biometrics software identifies matching speakers with low standard deviation and high probability results, while also distinguishing between different samplers with low standard deviation and values closer to the minimum.

The smaller the Cllr value, the more accurate the system is, and the $0 <$ Cllr < 1 relationship is a feature of well-calibrated (Sztahó & Fejes, 2023) applications.

The speech style notations used in Tables 1 through 6 and Figures 2 through 6 are as follows:

- 2.1: spontaneous (sess2_task1) audio sample,
- 2.2 : read (sess2_task2) audio sample,
- 2.3 : narration (sess2_task3) audio sample.

A decimal logarithmic transformation was applied to the LR values. The primary advantage of the LLR data obtained in this manner is the symmetric scale. LLR values less than 0 suggest different speakers (H1) and ones greater than 0 suggest identical ones (H0). The evaluation was based on the trends observed in Tables 1 through 6 and Figures 2 through 6, as well as on the DET curves generated using Bio-Metrics software.

3.1. Mean values of H0 and H1

In the case of the two different biometric identification software versions, as the speech duration increases, SS values also show an increase in both genders. However, in certain cases, different phenomena can be observed. The LLR results are shown in Tables 1 and 2 and Figure 2.

Table 1: H0 mean values of LLR data for audio samples of female speakers.

mean of H0, female speakers						
duration (s)	**Batvox 3**			**Batvox 4**		
	Spo	**Read**	**Narr**	**Spo**	**Read**	**Narr**
20	9.28753	7.20152	9.06353	5.5569	3.15539	4.52925
40	9.51955	8.53179	9.30864	6.45635	4.13735	5.06021
60	9.62084	9.02421	9.40291	7.11815	4.37704	5.77591
80	9.82359	9.0712	9.50403	7.4881	4.81714	6.27967
100	9.86962	9.49392	9.64835	7.75766	4.70182	6.51493
120	9.92827	9.33723	9.66373	7.73667	4.99168	6.47422

Table 2: H0 mean values of LLR data for audio samples of male speakers.

mean of H0, male speakers						
duration (s)	**Batvox 3**			**Batvox 4**		
	Spo	**Read**	**Narr**	**Spo**	**Read**	**Narr**
20	9.45068	7.45508	9.32333	5.74415	3.73451	5.08057
40	9.81685	8.27898	9.95291	6.60618	3.93672	6.45817
60	9.98095	8.60391	9.97565	7.79713	4.2863	6.99744
80	9.99395	9.02827	9.96294	7.62297	4.40198	7.36554
100	9.99173	8.90239	9.58461	7.90105	4.38097	6.35662
120	9.99994	9.17295	9.99726	8.20106	4.38394	7.57775

In the case of the Narration (Narr) type audio samples of male speakers, a slight decrease is observed for both Batvox versions at a duration of 100 seconds, after which it jumps to the local maximum of the average SS for the samples with a duration of 120 seconds. The highest LLR values were measured in Spontaneous (Spo) speech, followed by Narration, then Read speech. Note that data should be interpreted separately for each system, as the different biometric methods of the software versions affect the order of magnitude of the LLR values. For samples 2.1 and 2.3, it can be seen that even at 20 seconds, the LLR is above

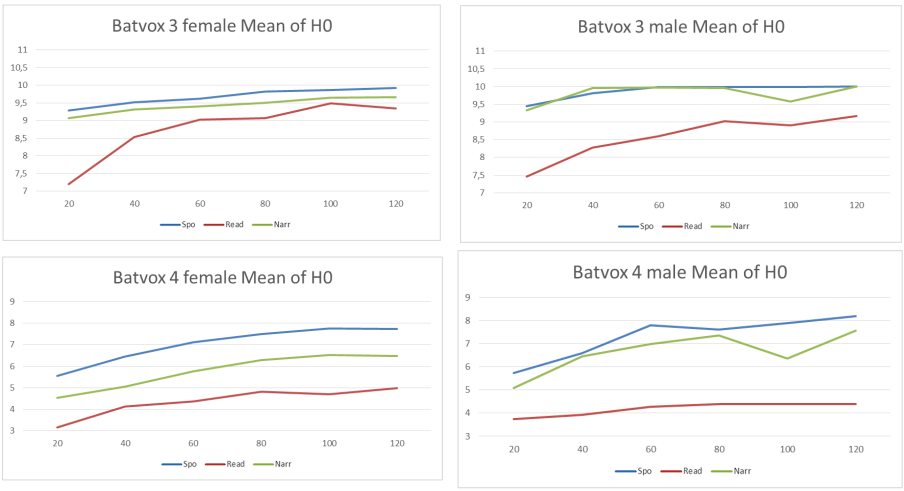

Figure 2: Graphs of the mean of LLR data for Hypothesis H0.

9 for Batvox 3.1, while Batvox 4.1 is more sensitive to speech duration based on the SS averages.

Contradictory results were obtained for the mean of the H1 DS data (see Tables 3 and 4), and our hypothesis – the mean decreases with increasing speech duration – cannot be supported in either case. For all three speech styles and both software versions, the mean of the DS data tends to increase, indicating that the longer the speech duration, the less likely the system is to differentiate speakers. However, it does not indicate a malfunction of the system, but rather reveals that the average of DS values cannot be used as a measure of performance. This statement is supported by the relations discussed in the following subsections.

3.2. Standard Deviation of SS and DS values

For Standard Deviation (SD), we assumed that longer speech duration results in lower SD, thus increasing the discriminating power of the applied method, but this was only partially confirmed by the data shown in Tables 55 and **??** and Figure 3.

Table 3: DS data mean values of LLR data in the speech samples of female speakers.

mean of H1, female speakers						
duration	**Batvox 3**			**Batvox 4**		
(s)	**Spo**	**Read**	**Narr**	**Spo**	**Read**	**Narr**
20	−0.4067	−0.4665	−0.5536	−2.2132	−2.3277	−2.3132
40	−0.235	−0.3643	−0.4232	−2.1648	−2.2746	−2.3075
60	−0.1089	−0.2745	−0.3603	−2.1117	−2.2828	−2.2945
80	−0.0763	−0.251	−0.2648	−2.1115	−2.2485	−2.2681
100	0.01711	−0.1147	−0.203	−2.054	−2.2508	−2.2709
120	0.05172	−0.0751	−0.2005	−2.0531	−2.2432	−2.2514

Table 4: DS data mean values of LLR data in the speech samples of male speakers.

mean of H1, male speakers						
duration	**Batvox 3**			**Batvox 4**		
(s)	**Spo**	**Read**	**Narr**	**Spo**	**Read**	**Narr**
20	−0.3803	−0.6292	−0.2814	−1.7706	−1.8261	−1.8127
40	−0.3156	−0.5153	−0.2109	−1.7646	−1.8184	−1.762
60	−0.2259	−0.4623	−0.1506	−1.6615	−1.8013	−1.7183
80	−0.2729	−0.4894	−0.0934	−1.6659	−1.8268	−1.6934
100	−0.1643	−0.4638	−0.1912	−1.6305	−1.7941	−1.7092
120	−0.1663	−0.4585	−0.0756	−1.5791	−1.8282	−1.664

From the above data, it can be seen that in Batvox 3.1, the Standard Deviation of the SS data values decreases with increasing speech duration in all three speech styles; however, a notable jump can be observed in the Narration and Read style samples. The SD of DS data shows an increasing or fluctuating trend depending on the software version and the gender of the speakers. Nevertheless, it should be noted that the absolute range of the values is smaller than in the case of SS data. We also found that the two software versions are characterized by a lower Equal Error Rate for the same speakers and a higher Equal Error Rate for different speakers. This statement is supported by the histograms

Table 5: H0 SD of LLR data in the speech samples of female speakers.

Standard Deviation of H0, female speakers						
duration	**Batvox 3**			**Batvox 4**		
(s)	**Spo**	**Read**	**Narr**	**Spo**	**Read**	**Narr**
20	1.47925	2.87787	1.76351	2.77122	2.27976	2.83024
40	1.47628	2.18109	1.54118	2.72329	2.98783	2.80453
60	1.12278	1.91644	1.38321	2.79839	3.03602	2.60051
80	0.56944	1.78398	1.2422	2.63205	3.07059	2.77972
100	0.55016	1.29549	1.21685	2.56946	3.04013	2.83464
120	0.44792	1.52514	1.2573	2.43123	2.96467	2.77677

Table 6: H0 SD of LLR data in the speech samples of male speakers.

Standard Deviation of H0, male speakers						
duration	**Batvox 3**			**Batvox 4**		
(s)	**Spo**	**Read**	**Narr**	**Spo**	**Read**	**Narr**
20	1.17316	2.84889	1.43012	2.37577	2.57552	2.61177
40	0.82076	2.53241	0.1818	2.34037	2.64598	2.55186
60	0.10194	2.32753	0.1301	2.18984	2.86471	2.41174
80	0.03495	2.29878	0.21188	2.34126	2.79255	2.47985
100	0.02988	2.25904	1.34614	2.0958	2.851	2.69386
120	0.00035	1.90806	0.01453	1.9704	2.80516	2.38392

shown in Figure 4, where you can see the distribution of the values generated during the comparison of spontaneous speech samples of female speakers with Batvox 3.1. The y-axis shows the distribution rate, and the x-axis shows the LLR probability values.

3.3. EER and Cllr values

EER and Cllr values are measures of the performance of the biometric speaker identification software. EER is a measure of discrimination that shows how well a software can distinguish between same and different (SS-DS) speakers. In forensics, we typically compare the speech samples of two speakers to

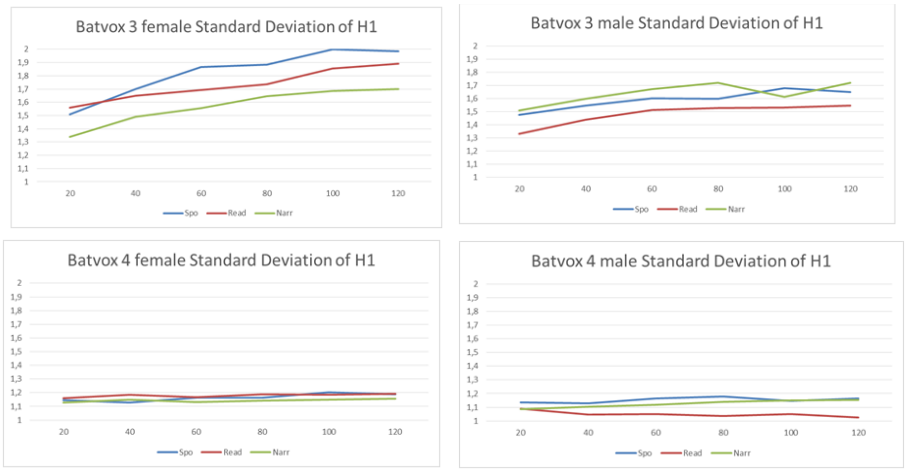

Figure 3: Graphs of the Standard Deviation data for DS.

determine the likelihood of identity, making it particularly important for the results to be robust enough to identify the same speaker and differentiate between different individuals. Another important criterion is to keep the error rates as low as possible (FAR, FRR, EER) in order to prevent erroneous expert reports. When determining performance, it is essential to conduct measurements in the same test environment (speech sample database), thereby comparing different software versions and performing tests with the same speech samples. Tables 7 and 8 and Figure 5 below show the EER data for the two Batvox systems for speakers of both genders.

For both systems, we obtained low EER results in the vast majority of both female and male samples, indicating that the system reliably separates the same and different individuals based on their voice patterns, even at short speech durations. More so in the case of female speakers, and to a lesser extent for male speech samples, it can be seen that the newer, more modern Batvox 4.1 software achieves a lower EER value. However, for both genders, it can be observed that the values for the 20-second recordings contradict the trend: they exhibit smaller or larger values compared to the subsequent 40-second measurement runs in several cases.

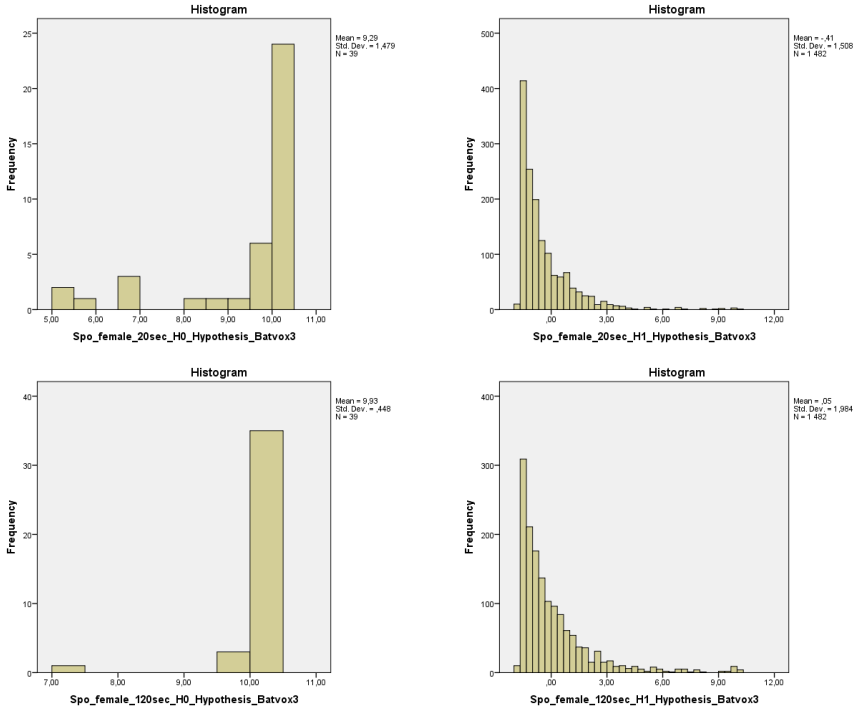

Figure 4: Histograms of SS and DS data for 20- and 120-second audio samples of the same speech style (The y-axis shows the distribution rate, the x-axis shows the LLR probability values).

Table 7: EER values for speech samples of female speakers.

EER, female speakers						
duration	**Batvox 3**			**Batvox 4**		
(s)	**Spo**	**Read**	**Narr**	**Spo**	**Read**	**Narr**
20	0.5735	5.1282	2.193	0.4386	5.0607	2.2942
40	2.5641	2.8677	2.5641	0.5735	4.9595	2.0243
60	2.5641	2.6653	2.5641	0.3036	5.1619	0.1687
80	2.0429	2.5978	2.5304	0.5398	2.5641	0.2024
100	2.1592	2.5641	2.5641	0.2699	0.6073	0.2362
120	2.1255	2.5978	2.5641	0.1012	1.9568	0.1687

Table 8: EER values for speech samples of male speakers.

EER, male speakers						
duration	**Batvox 3**			**Batvox 4**		
(s)	**Spo**	**Read**	**Narr**	**Spo**	**Read**	**Narr**
20	0.4723	4.5547	2.2605	2.5641	2.8003	2.5978
40	0.5398	2.5978	0.135	0.5735	7.4224	0.3374
60	0.135	2.8003	0.2024	0.1687	5.1282	0.1012
80	0.1012	3.0702	0.2699	0.1687	5.1282	0.135
100	0.0375	5.1282	2.5641	0.0337	7.6586	2.5978
120	0.1012	2.5641	0.1687	0.135	4.892	0.2362

Figure 5: Graphs of EER values.

Cllr refers to the accuracy of a biometric speaker identification software, with lower values being more favorable. Cllr measures the discrimination error (how much overlap between H0 and H1 LRs there is) and calibration error (whether the LRs are too large or too small). A Cllr of 0 is a perfect system, and a Cllr of 1 is a system that is completely worthless (performs at chance level). Tables 9 and 10 show the Cllr data for the two systems and speakers of both genders.

Table 9: Cllr values for speech samples of female speakers.

Cllr, female speakers						
duration	**Batvox 3**			**Batvox 4**		
(s)	**Spo**	**Read**	**Narr**	**Spo**	**Read**	**Narr**
20	0.5136	0.5202	0.4496	0.1523	0.2365	0.1722
40	0.5952	0.5489	0.512	0.1462	0.2228	0.1591
60	0.6604	0.584	0.5401	0.1476	0.2114	0.1383
80	0.6744	0.595	0.5814	0.1472	0.1846	0.1377
100	0.7264	0.659	0.6064	0.157	0.1752	0.1352
120	0.7347	0.677	0.6087	0.1524	0.1716	0.1368

Table 10: Cllr values for speech samples of male speakers

Cllr, male speakers						
duration	**Batvox 3**			**Batvox 4**		
(s)	**Spo**	**Read**	**Narr**	**Spo**	**Read**	**Narr**
20	0.52341	0.44963	0.56021	0.19851	0.25501	0.21117
40	0.55134	0.49186	0.59174	0.18901	0.25449	0.18765
60	0.55134	0.51943	0.62482	0.20237	0.25218	0.19076
80	0.56838	0.50988	0.65201	0.20331	0.23497	0.19612
100	0.61989	0.51981	0.60192	0.20291	0.26457	0.20612
120	0.61216	0.51705	0.6584	0.21286	0.24123	0.2006

During our measurements, we also observed fluctuating trends in Cllr values in the case of Batvox 4.1. In the case of Batvox 3.1, Cllr increases with increasing speech duration. To obtain an accurate picture of the characteristics of the operation of both software versions, LLR values were plotted on histograms. Paired t-tests were performed between the trials of same and different speakers to show the separation power between the two hypotheses.

The histograms in Figure 6 show that the results of the same and different speakers are well separated by both software versions. It can be seen that the more advanced Batvox 4.1 is more likely to identify the match and differentiate

between different speakers. It is more sensitive to speech duration compared to version 3.1, yet it identifies matching speakers with a high LLR at 60 seconds. The broader histogram of the same speaker LLR values suggests that this system is more sensitive to similarities/differences in speaker voice characteristics and can measure this similarity better.

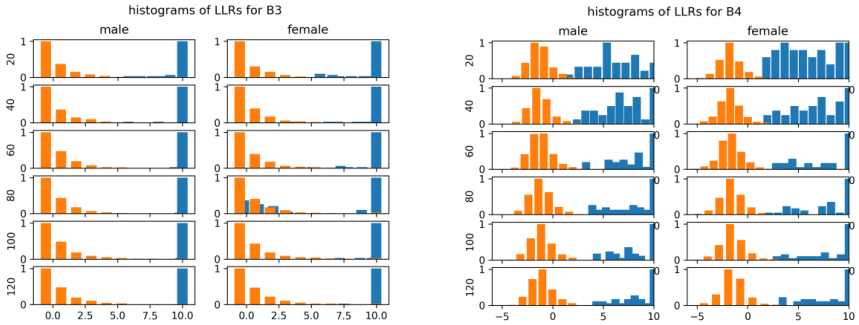

Figure 6: Histograms of LLR data. Yellow: different speaker trials, blue: same speaker trials. B3 and B4 represent Batvox 3.1 and 4.1 systems, respectively.

Although p-values do not reflect the distinctive power between the two groups (same speaker versus different speaker), the tendency in their value suggests that there is a real effect of sample duration. For better visualization, the logarithm of the p-value is shown in Figure 7. Indeed, this is not a standard way of representing the significance of differences between groups, and the p-value cannot be considered a "measure" of the difference, but it does give us a general idea of the trend in the magnitude of differences between groups as a function of recording length.

4. Conclusion

In forensic identification tests for forensic purposes, the expert often only has a short duration of speech samples available. In such cases, it is possible to determine the probability of speaker identity with high accuracy using voice biometrics. In our research, we have demonstrated that even for voice recordings with a gross duration of 20 seconds, in which the net, uninterrupted speech

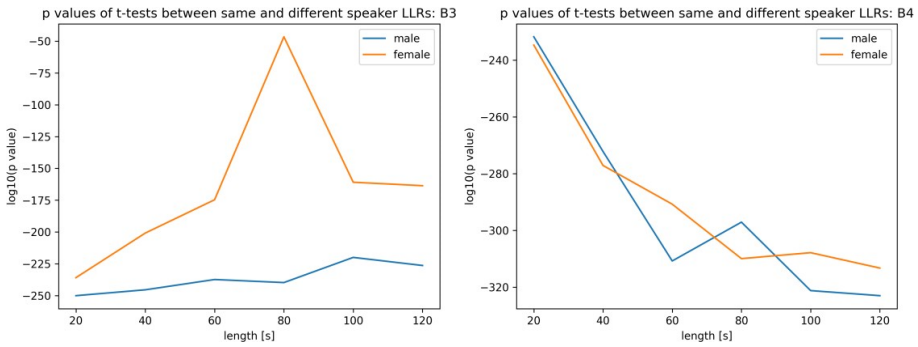

Figure 7: The p-values of the t-tests.

duration is even lower, the automatic identification software is likely to identify or distinguish between different speakers. Overall, the more advanced Batvox 4.1 performs better than the previous version, Batvox 3.1; the Cllr and EER values are mostly lower for Batvox 4.1. In general, the higher the SS value and the lower the DS data is, the better performance we can expect from the system. However, the older software version also produced good results, with a low EER error rate for shorter recordings.

Three different types of speech were used as model test recordings (spontaneous, read, narrative style) that modeled the "known speaker" speech sample of the common forensic case, and compared this with the spontaneous sound sample of the "unknown person" two weeks apart. By evaluating the measurement results, we obtained better results with the spontaneous and narrative-type speech samples compared to the samples of the "read" speech style. This is promising in terms of expert voice sampling methodology: in the future, the use of read voice sampling in biometric speech identification is of limited use; forensic voice comparison methodology should therefore adapt these results. The other conclusion is that the error rate is significantly reduced at 60 seconds, so voice biometric measurements can be made reliably at or above this audio duration. The forensic audio sample database created as part of the FORENSICSpeech research project provides an excellent research base for forensic biometric speech identification studies. A basic requirement for our research is to have more than

one speaker-style speech sample recorded at different times. Thus, in the future, the methodology of identification tests to be performed on sound recordings in Hungarian can be developed using new research results.

The Hungarian-language database of forensic speech samples will also provide significant support for speech recognition research (Kamath et al., 2019), which requires a large corpus of Hungarian-language data. A speech sample database modeling a typical forensic case is a prerequisite for both speech recognition and speaker identification research. It can be used for performance studies and to support research on speech and speaker recognition. Using the above results allows the development of systems with higher accuracy in the future.

References

Anil, K., Arun, A., & Karthik, N. (2011). *Introduction to Biometrics*. New York, Dordrecht Heidelberg London, DOI: Springer. doi:10.1007/978-0-387-77326-1.

Beke, A., Szaszák, G., & Sztahó, D. (2020). Forvoice120+ magyar nyelvű utánkövetéses adatbázis kriminalisztikai célú hangösszehasonlításra. In G. Berend, G. Gosztolya, & V. Vincze (Eds.), *XVI. Magyar Számítógépes Nyelvészeti Konferencia* (pp. 95–101). Szeged.

Craig, A. (2010). *Mathematics and Statistics for Forensic Science*. Chichester: John Wiley & Sons Ltd.

Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J., & Niemi, T. (2015). *Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition*. European Network of Forensic Science Institutes. Agreement Number: HOME/2011/ISEC/MO/4000002384.

European Network of Forensic Science Institutes (2022). Best practice manual for the methodology of forensic speaker comparison.

Fejes, A. (2022). Hangazonosítás. In C. Fenyvesi, C. Herke, & F. Tremmel (Eds.), *Kriminalisztika*. Budapest: Ludovika Egyetemi Kiadó.

Gráczi, T., Fejes, A., Krepsz, V., & Huszár, A. (2022). Speaker recognition over the course of 10 years and across speech style. *Alkalmazott Nyelvtudomány, 22: Különszám*, 94–109.

Gósy, M., Gyarmathy, D., Horváth, V., Gráczi, T., Beke, A., Neuberger, T., & P, N. (2012). Bea: Beszélt nyelvi adatbázis. In M. Gósy (Ed.), *Beszéd, adatbázis, kutatások* (pp. 9–25). Budapest: Akadémiai Kiadó.

Jessen, M., Bortlík, J., P., S., & A, S. Y. (2019). Evaluation of phonexia automatic speaker recognition software under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01. *Speech Communication*, *111*, 22–28. doi:10.1016/j.specom.2019.05.002.

Kamath, U., Liu, J., & Whitaker, J. (2019). *Deep Learning for NLP and Speech Recognition*. Switzerland AG: Springer Nature. doi:10.1007/978-3-030-14596-5.

Kelly, F., Fröhlich, A., Dellwo, V., Forth, O., Kent, S., & Alexander, A. (2019). Evaluation of vocalise under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01. *Speech Communication*, *112*, 30–36. doi:10.1016/j.specom.2019.06.005.

Kovács, G. A. (2017). Egyes kognitív, emberi tényezők szerepe a szakértőivélemény-alkotásban. *Belügyi Szemle*, *65*, 89–103. doi:10.38146/BSZ.2014.10.7.

Leemann, A., Perkins, R., Buker, S., & Foulkes, P. (2025). *An Introduction to Forensic Phinetics and Forensic Linguistic*. New York: Routledge. doi:10.4324/9780367616595.

Meester, R., & Slooten, K. (2021). *Probability and Forensic Evidence*. Cambridge: Cambridge University Press. doi:10.1017/9781108596176.

Meuwly, D. (2009). Speaker recognition. In A. Jamieson, & A. Moenssens (Eds.), *Wiley Encyclopedia of Forensic Science Volume 5*. West Sussex: John Wiley & Sons Ltd.

Morrison, G. (2009). Forensic voice comparison and the paradigm shift. *Science and Justice*, *49*, 298–308. doi:10.1016/j.scijus.2009.09.002.

Morrison, G. (2013). Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences*, *45*, 173–197. doi:10.1080/00450618.2012.733025.

Oxford Wave Research (2025). Bio-metrics. URL: https://oxfordwaveresearch.com/products/bio-metrics/#:~:text=The%20Tippett%20plot%20is%20a%20cumulative%20probability%20distribution,H1%20hypothesis%20%28biometric%20samples%20are%20from%20different%20sources%29.).

Ramos, D. (2007). Forensic evaluation of the evidence using automatic speaker recognition system.

Sztahó, D., Beke, A., & Szaszák, G. (2021). Forvoice 120+: Statisztikai vizsgálatok és automatikus beszélő verifikációs kísérletek időben eltérő felvételek és különböző beszéd feladatok szerint. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, 2021. január 28–29.

Sztahó, D., & Fejes, A. (2023). Effects of language mismatch in automatic forensic voice comparison using deep learning embeddings. *Journal of Forensic Sciences*, *68*, 871–883.

Van der Vloed, D. (2016). Evaluation of batvox 3.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01. *Speech Communication*, *100*, 13–17. doi:10.1016/j.specom.2018.04.008.

Zhang., C., & Tang, C. (2018). Evaluation of batvox 4.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01. *Speech Communication*, *85*, 127–130. doi:10.1016/j.specom.2016.10.001.