© by the Institute
of Forensic Research

ISSN 1230-7483
e-ISSN 2720-5983

Problems of Forensic Sciences 2025, vol. 142, 143, 143–155

DOI: 10.4467/12307483PFS.25.007.22914

CHALLENGES FOR AUDITORY AND AUTOMATIC SPEAKER RECOGNITION: EVALUATING CASES OF HIGHLY SIMILAR VOICES

Jacek KUDERA[1,2], Naima Islam NODI[3], Julia ROEGNER[3]

[1] *Trier University, Department of Phonetics, Trier, Germany*
[2] *WSB Merito University in Wrocław, Poland*
[3] *Trier University, Computational Linguistics, Trier, Germany*

Abstract

Identical twins present a difficult case for both auditory and machine speaker recognition. This paper addresses this challenge and presents the findings of two studies: an auditory speaker discrimination test and a machine-based task using forensic automatic speaker recognition (ASR) system. The outcomes of the perceptual judgement task were compared with the log-likelihood ratios (LLRs) yielded by an x-vector-based speaker recognition system. Although the task was given to lay listeners, as opposed to forensic phonetic experts, the results appear to be congruent with the scores yielded by a state-of-the-art automatic system. The human raters were more accurate in judging same-speaker pairs than different-speaker pairs. The machine approach showed better performance in both conditions tested as compared to human listeners. Overall, the voices that were difficult for human listeners were different from those that the ASR system struggled with.

Keywords

Forensic phonetics; Automatic speaker recognition; Auditory assessment; Machine-based approach; Similar voices; Identical twins.

Received 30 June 2025; accepted 23 September 2025

1. Introduction

Natural inter-speaker variability of speech characteristics plays a key role in speaker recognition. Idiosyncratic properties of speech help to tell speakers apart in both auditory and machine-based processing. Such a discrimination task, however, becomes more complex if the variability of speech is limited, e.g., due to a similar anatomy of the vocal tract and the articulators [1–3]. Identical, monozygotic twins present an extreme case of voice similarity due to the genetic information that determines the shape of their vocal tracts [4]. Similarity increases even further when twins are socialized in the same environment and acquire the same language variety [5]. Therefore, the speech of monozygotic twins presents valuable and challenging material for testing the performance of automatic speaker recognition systems. Previous studies addressing extremely low inter-speaker variability of speech have focused on temporal features, such as pause length [6], speaking tempo [7], as well as spectral characteristics of the speech signal [8], including fundamental frequency estimates [9], formant dynamics [10], and glottal source features [11]. Studies comparing speech features of identical twins have covered a variety of linguistic units including selected phonemes [12], phoneme sequences [13], words [14], or degree of coarticulation [15].

Currently, speaker recognition studies seem to be dominated by two approaches. The first involves machine-based analysis of the representation of the speech signal, while the second relies on human perceptual judgement based on auditory-acoustic evaluation. The approach presented in this paper involves a combination of perceptual and machine-based methods for speaker discrimination. When it comes to the analysis of twin voices, the results of automatic speaker comparison studies suggest that ASR systems are able to discriminate monozygotic twins on the basis of speech samples, but report gender effects with superior performance of the ASR system for male voices as compared to female voices [16]. State-of-the-art systems – even if fine tuned for cases of extreme voice similarity such as twin speech – often show variable performance that *inter alia* depends on sample length [17]. Therefore, the successful automatic discrimination of speech samples with extremely low inter-speaker variability seems to be a consequence of the system applied. These findings invite hybrid approaches that involve both human listeners and machines [18]. Previous studies comparing the performance of machines and human listeners showed that human assessments positively correlated with automatic estimates of voice similarity [19]. In this study, we employed a hybrid approach to test the case of extremely low inter-speaker variability and asked the following research questions: How does the performance of human listeners compare to that of ASR in a discrimination task on highly similar voices of identical twins? Is there a gender effect in human listeners? What are the parameters used by listeners to make their consistency?

2. Methods

2.1. Recordings

The recordings of the identical twins used in our experiments were made for a television reality show called *Die Zwillingsshow* (*The Twin Show*), which was broadcasted by RTL in 2004. Throughout the show, the identical twins were given different tasks, one of which was to record read speech used in the experiment and spontaneous speech. Recordings were made at 48 kHz sampling rate and 16-bit depth. The read speech was a German version of *The North Wind and the Sun*. An acoustically treated environment was used for the recording sessions. A total of 55 pairs of twins (30 female and 25 male, group mean age 34 years) were recorded.

2.2. Experiment 1: Human assessment

50 native speakers of German (30 males and 20 females) served as listeners. None of them reported any speech or hearing disorders. The participants were aged between 21 and 68 years (mean = 36). None of the participants had any training in phonetics. Participants were paid for the completion of the study. Participants were presented with pairs of speech samples. They had to decide in a forced choice format whether the two speech samples originated from the same speaker or from different speakers. Participants were instructed to use headphones. Subjects were exposed to 55 pairs of samples (30 female twin pairs and 25 male twin pairs). Sample lengths ranged from 9 to 13 seconds. We aimed at keeping the sample durations constant, but some variation resulted from preserving pauses and intonation contours at phrase level. The order of the audio files was randomised such that there was one identical twin sample for every four different-speaker pairs. The samples included two parts of the German version of *The North Wind and the Sun*. One recording included the second and third sentence of the fable (see Appendix, Part 1), whereas the other recording included the remaining sentences (Part 2). For each pair of recordings, listeners were asked whether they thought both audio recordings were from one person or two different people. When the ground truth is ‘same’, both audio samples were from one of the twins. Participants could repeat the audio samples if they wished.

The experimental session was preceded by 3 trial questions to familiarise the participants with the procedure and, if necessary, to adjust the volume to a comfortable level. The participants were also asked to indicate their confidence on a scale from 1 (not at all confident) to 5 (very confident). Once they had completed the discrimination task, we asked them about the features they had used in assessing the speakers. They could name as many as they wanted. The experimental trials were followed by a debriefing session in which we explained the purpose of the study. The entire session lasted approximately 35 minutes.

2.3. Experiment 2: Machine-based approach

In this study, we used an established forensic ASR system (Phonexia Voice Inspector, version 5.1.0) [20]. We performed the analyses with known ground truth. The target scores were calculated using a population set that consisted of 40 male native speakers of the German language. Each native speaker in the population set was represented by two files of read speech.

The content, mode and style of the read speech were similar to that of the analysed samples, which made our population dataset tailored to the tested case [21]. We used a so-called one-to-one mode in which both target and non-target scores are calculated on a selected population set [22]. As the recordings were made in an acoustically controlled environment, the signal-to-noise ratio (SNR) threshold was set to 3 dB. The automatically assigned voice activity detection labels were compared with the waveforms and spectrograms of the recordings and visually checked for correct boundaries by a trained phonetician.

The results of speaker comparison are typically presented as a numerical log-likelihood ratio (LLRs) [23], which is currently the accepted method in reporting the outcomes of forensic speaker comparisons [24, 25]. This consensual ratio rooted in Bayesian reasoning in our case reflects the score under two competing hypotheses: that the tested speech samples originate from the same speaker (H0) and that the speech samples originate from two different speakers (H1) [26, 27]. Then following the ENFSI (The European Network of Forensic Science Institutes) guidelines [28] for evaluative reporting in forensic science, these scores are associated with verbal correlates of the yielded ratios for ease of interpretation.

2.4. Data analysis

For Experiment 1, we analysed 5500 observations (50 participants × 55 pairs × 2 responses: binary same-speaker vs. different-speaker judgement and scalar confidence score). We calculated the inter-rater agreement for assessing same-speaker and different-speaker pairs with Krippendorff's alpha, performed a Levene's test for equality of variances, and conducted a Mann-Whitney test for evaluation of the differences in confidence ratings.

For Experiment 2, we analysed the log-likelihood ratios of evidence given same-speaker and different-speaker hypotheses as yielded by Phonexia Voice Inspector computed with a recommended mode of generalized logistic probability density function (version 5.1.0), in which the maximum likelihood estimates are obtained for type IV generalized logistic distribution with location and scale parameter [22]. We report the verbal ratios that correspond to the LLRs obtained.

3. Results

3.1. Experiment 1: Human assessment

The analysis revealed that participants were more successful in recognising audio samples from same-speaker pairs compared to different-speaker pairs. The overall accuracy reached 68.6%, where the percentage of correct recognition of the audio samples from the same speakers was 93.3%, and the performance in recognizing the audio samples from different speakers was 62.4% (for detailed pairwise analyses see Section 6). The inter-rater agreement, computed as Krippendorff's alpha ($\alpha = 0.348$), indicates low inter-rater agreement and may suggest considerable individual differences in performance or high level of difficulty of the experimental procedure.

A Levene's test proved that confidence scores significantly differ between the tested conditions. The results indicate that the variances between the two groups were not homogeneous ($p < 0.05$). Given the unequal variances, Welch's t-test was performed to compare the confidence ratings. The results revealed a significant difference ($t = 9.65$, $p < 0.05$), suggesting that participants were more confident when recognising audio samples of same-speaker pairs compared to different-speaker pairs. The significant results of a Mann-Whitney test ($p < 0.05$) support this finding. This also implies that listeners were more confident when they were correct than when they were wrong. Figure 1a illustrates higher confidence when recognising audio samples of same-speaker pairs compared to different-speaker pairs.

The analysis of the dataset confirms earlier results that female voices are more difficult to assess than male ones. However, the difference was not statistically significant. For male twin pairs, the overall accuracy found was 70.46% (same-speaker pairs 94.80% and different-speaker pairs 64.67%). For female twin pairs, the overall accuracy was 66.90% (same-speaker pairs 92.00% and different-speaker pairs 60.35%).

Listeners stated that dialect, intonation and articulation of /r/ were the most salient features that they used to compare speakers. Since the twin pairs came from all over Germany and siblings had grown up together, it seems hard to conceive how dialect would have played a role, but this is what was indicated by the listeners.

3.2. Experiment 2: Machine-based approach

System validation was performed using the log-likelihood-ratio-cost (Cllr), a commonly used metric to evaluate the efficiency of forensic ASR systems [29]. Cllr values range from 0 to 1. The values close to 0 suggest a well calibrated system [30]. The overall ASR performance, based on the tested population set is EER = 1.25% (EER threshold = 4.58 and Cllr = 0.102), which suggests that the system has been calibrated well.

The results yielded by the forensic ASR system were both LLRs and verbal ratios associated with likelihood intervals. The verbal ratios rely on gradual modifiers of evidence strength that pose twelve degrees of certainty, i.e., limited (1), moderate (2), moderately strong (3), strong (4) , very strong (5), extremely strong evidence (6), either for or against prosecution hypothesis (the numbers in the parentheses correspond to the intervals of LLR yielded by the forensic ASR system). Positive LLRs are interpreted as favouring the same-speaker hypothesis, while negative LLR values support the different-speaker hypothesis. The system calibrated for German read speech has achieved an accuracy of 100% on the dataset with known ground truth. The automatic system performed perfectly in that it always correctly identified whether the result was the same or different. However, the degree (probability rating) of these consistency differs.

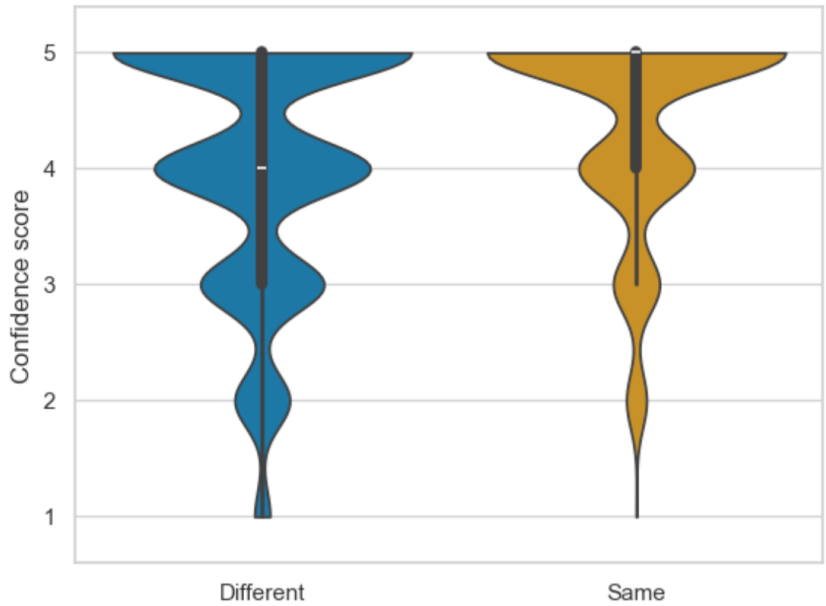

Fig. 1a. Distribution of confidence scores of auditory assessment (y-axis: absolute values, x-axis: speaker pair type).

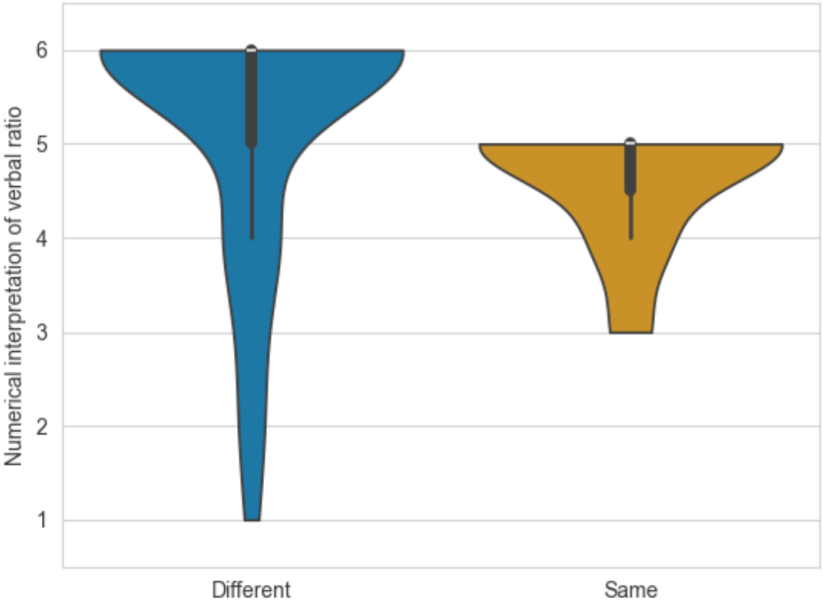

Fig. 1b. Distribution of LLR-based verbal ratios (y-axis: absolute values, x-axis: speaker pair type)

The extreme LLR values in favour of same-speaker hypothesis ranged from 2.95 (Pair 5: Moderately strong evidence to support prosecution hypothesis) to 4.64 (Pair 28: Very strong evidence to support prosecution hypothesis), whereas the LLR scores supporting the different-speaker hypothesis ranged from −0.24 (Pair 15: Limited evidence against prosecution hypothesis) to −20.45 (Pair 58: Extremely strong evidence against prosecution hypothesis). The results revealed that 75.0% of ratings for different-speaker pairs reached higher absolute LLRs than the mean scores for same-speaker pairs. These observations are illustrated in the density plot in Fig. 1b.

Levene's test ($F = 13.3$, $p < 0.05$), Welch's t-test ($t(31.66) = -5.26$, $p < 0.05$), and Mann-Whitney ($U = 108.0$, $p < 0.05$) with $\alpha = 0.05$ indicate significant differences in variance, mean, and ranks of LLR values for both tested conditions, i.e., same-speaker and different-speaker pairs. However, when the datasets are contrasted by gender, the same tests show no significant differences in LLRs for female and male pairs.

3.3. Comparison between human listeners and ASR results

Although in general confidence scores do not always correlate with performance in auditory tasks, there is no other empirical scale that can serve as basis for comparison of certainty in auditory speaker discrimination and machine-based approach. To compare the outcomes of the two methods as presented in Fig. 2, we relied on human same-speaker vs. different-speaker judgements, the confidence scores and respective verbal interpretation of LLRs (pink bars). The average confidence score calculated over all responses (blue triangles) refers to the average of the ratings given, with negative values for responses in favour of different speakers.

When analysing only individual correct human judgements (blue bars), the mean confidence score for 11 same-speaker pairs as well as 10 different-speaker pairs (11F, 15F, 18F, 30F, 33M, 43F, 47M, 55F, 59M, 60M) is higher than the LLR yielded by Phonexia Voice Inspector. Even though male participants tend to show a higher degree of confidence regardless of accuracy [30], our study revealed that the differences in confidence scores between human listeners were minimal (mean confidence scores: females = 4.07, males = 4.15).

4. Discussion

ASR has been shown to be highly effective for distinguishing the speech of identical twins, i.e., in cases of extreme voice similarity. The overall performance of the forensic ASR system surpasses the results obtained by the lay listeners. The ASR system correctly assessed all 16 speaker pairs that were misjudged by human lay listeners, that is, the ASR system outperformed human listeners in assessing 16 pairs. They were all different-speaker pairs, 10 of them female

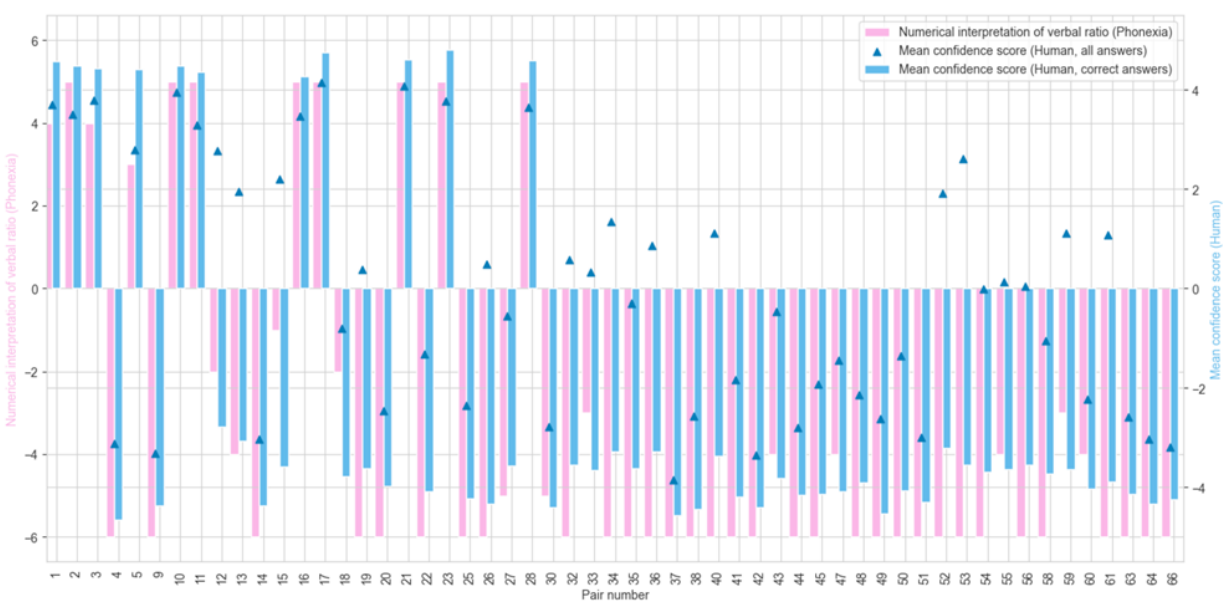

Fig. 2. Results of human auditory assessment (triangles: all answers; bars: correct judgements only) and machine-based speaker comparison (left y-axis: verbal ratio corresponding to LLRs; right y-axis: mean confidence score; x-axis: pair number).

and 6 male. These findings resonate with earlier observations that female voices are more challenging to interpret than male voices [16]. The correspondence between human and machine results is generally low in terms of degree (operationalised as confidence in human listeners and LLR in automatic comparison).

No statistically significant differences were observed between the male and female twin subsets with either the human listeners or the ASR. These findings diverge from previous studies, which suggested that female voices are more challenging to recognise than male voices. However, we observed that the forensic ASR system provided higher degrees of certainty for different-speaker pairs, whereas human lay listeners showed higher confidence ratings for same-speaker pairs. Nevertheless, human errors are all related to identity. This, in combination with the higher confidence ratings for the same-speaker pairs, shows that lay listeners tend to have more confidence in incorrect results than in correct ones. This is a question that should be explored further.

The limitations of our exploratory contribution relate to the population set, which can always be expanded to include more representative speakers. Another restriction is the low potential for explaining the results yielded by a machine-based approach. Although in our study the population set has not been ideally gender-balanced, we have observed a satisfactory performance of the model with acceptable calibration thresholds [30]. We also recognise the risk of using confidence scales as a basis for comparing human and machine results.

5. Conclusions

The rapid development of speaker recognition technology has made forensic ASR systems increasingly used by law enforcement and investigative agencies. The automatic approach to speaker discrimination tasks, however, requires a large population dataset that represents the demographic stratum of the speakers under investigation and reflects the quality of the recordings being compared [31]. The human-led approach, on the other hand, requires expertise and training. In this paper, we employed a hybrid approach to test the human auditory ability and performance of the machine to tell speakers apart under the condition of extreme voice similarity.

We showed that discrimination of similar voices is possible even for untrained native speakers exposed to speech samples of identical twins. If the experimental procedure had been repeated with expert phoneticians and forensic experts, the results would have probably been even more robust. Furthermore, our participants were able to point to the features that seemed useful in the speaker discrimination task. Whereas the cues that appear informative to human listeners are interpretable, the cues that are informative to the forensic ASR system are inexplicable to humans. The results of the study show that ASR can distinguish between the voices of even identical twins under ideal circumstances. In fact, it outperforms phonetically untrained humans and has achieved a perfect score. Other commonly used speaker comparison systems, such as VOCALISE [32], could also be used to compare performance with the fifth generation of Phonexia Voice Inspector model.

The results of the experiment are not directly relevant to forensic casework. The main reason is that untrained listeners were used as subjects as opposed to experts with extensive training in forensic phonetics. Furthermore, the materials did not conform to forensic real-world conditions, which usually means that only small quantities of degraded speech are available for analysis. Finally, cases involving identical twins are extremely rare. What does happen frequently, though, is that there are one or more siblings whose voice is often confused with that of the suspect, and whose anatomy can be assumed to resemble it more closely than that of an unrelated person. In these cases, which generally prove to be difficult to cope with even for the trained expert, ASR systems may be of help. What this study does demonstrate, though, is that ASR has come a long way when it comes to discriminating voices of speakers whose vocal tracts are almost anatomically identical (cf. Künzel, who in 2010 found a number of errors when using a GMM-UBM system [16]). However, there remain some intrinsic sources of error in the ASR systems such as the accuracy of VAD labelling, arbitrarily set SNR thresholds, and the quality and quantity of the population data set. Furthermore, there is no reason to assume that two voices that sound very similar to human observers, but which have not necessarily been produced by similarly shaped vocal tracts (so-called voice twins) [18], will be more difficult for ASR systems to judge than any other two voices.

Overall, lay listeners clearly showed a bias towards identity – they made more errors in the different-speaker condition than in the same-speaker condition. The ASR system, on the other hand, made more errors in terms of degree of certainty in the same-speaker condition. In a way, it was thus more conservative in its judgement than the human listeners. Interestingly, we have not observed any significant differences between the gender of speakers. Finally, this contribution may

also be useful for comparing the performance of different forensic ASR systems. Analysing the results of the tests with known ground truth is essential for system benchmarking.

Data availability

The experimental data and code are available in the following OSF repository: https://osf.io/kxu6v/.

Declaration of interests

The authors declare no competing interests.

Appendix

Part 1

Sie wurden einig, dass derjenige für den Stärkeren gelten sollte, der den Wanderer zwingen würde, seinen Mantel abzunehmen. Der Nordwind blies mit aller Macht, aber je mehr er blies, desto fester hüllte sich der Wanderer in seinen Mantel ein.

Part 2

Endlich gab der Nordwind den Kampf auf. Nun erwärmte die Sonne die Luft mit ihren freundlichen Strahlen, und schon nach wenigen Augenblicken zog der Wanderer seinen Mantel aus. Da musste der Nordwind zugeben, dass die Sonne von ihnen beiden der Stärkere war.

References

1. Jessen M. Forensic phonetics. Lang Linguist Compass. 2008;2(4):671–711. Available from: https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-818X.2008.00066.x
2. French P. An overview of forensic phonetics with particular reference to speaker identification. Int J Speech Lang La. 1994;1(2):169–81.
3. Nolan F. Auditory and acoustic analysis in speaker recognition. In: Gibbons J, editor. Language and the Law. London: Longman; 1994. p. 326–45.
4. Beck JM. Organic variation of the vocal apparatus. In: Hardcastle W, Laver J, Gibbon F, editors. The Handbook of Phonetic Sciences. Oxford: Blackwell Publishers; 1997. p. 256–97.
5. Mayr R, Price S, Mennen I. First language attrition in the speech of Dutch–English bilinguals: the case of monozygotic twin sisters. Biling-Lang Cogn. 2012;15(4):687–700.
6. San Segundo E, Tsanas A, Gómez-Vilda P. Euclidean distances as measures of speaker similarity including identical twin pairs: A forensic investigation using source and filter voice characteristics. Forensic Sci Int. 2017;270: 25–38. Available from: https://www.sciencedirect.com/science/article/pii/S0379073816304960
7. Whiteside SP, Rixon E. Speech tempo and fundamental frequency patterns: a case study of male monozygotic twins and an age- and sex-matched sibling. Logopedics Phoniatrics Vocology. 2013;38(4):173–81.
8. Van Gysel W, Vercammen J, Debruyne F. Voice similarity in identical twins. Acta Otorhinolaryngol. Belg. 2001;55(1):49–55. Available from: http://europepmc.org/abstract/MED/11256192
9. Cavalcanti JC, Eriksson A, Barbosa PA. Multiparametric analysis of speaking fundamental frequency in genetically related speakers using different speech materials: Some forensic implications. J. Voice. 2024;38(1):243.e11–29. Available from: https://www.sciencedirect.com/science/article/pii/S0892199721002927
10. Zuo D, Mok PPK. Formant dynamics of bilingual identical twins. J Phonetics. 2015;52:1–12. Available from: https://www.sciencedirect.com/science/article/pii/S0095447015000182
11. San Segundo E, Gómez-Vilda P. Evaluating the forensic importance of glottal source features through the voice analysis of twins and non-twin siblings. Lang Law/Linguagem e Direito. 2017;1(2):22–41.
12. Nolan F, Oh T. Identical twins, different voices. Int J Speech Lang La. 1996;3(1):39–49. Available from: https://journal.equinoxpub.com/IJSLL/article/view/10105
13. Loakes D. A forensic phonetic investigation into the speech patterns of identical and non-identical twins [PhD dissertation]. University of Melbourne, School of Languages; 2006.
14. Whiteside S, Rixon E. Speech patterns of monozygotic twins: An acoustic case study of monosyllabic words. The Phonetician. 2001;84:9–22.
15. Whiteside SP, Rixon E. Speech characteristics of monozygotic twins and a same-sex sibling: An acoustic case study of coarticulation patterns in read speech. Phonetica. 2003;60(4):273–97.
16. Künzel HJ. Automatic speaker recognition of identical twins. Int J Speech Lang La. 2010;17(2):251–277.
17. Cavalcanti JC, da Silva RR, Eriksson A, Barbosa PA. Exploring the performance of automatic speaker recognition using twin speech and deep learning-based artificial neural networks. Front Artif Intell. 2024;7:1287877.
18. Gerlach L, McDougall K, Kelly F, Alexander A. Voice twins: discovering extremely similar-sounding, unrelated speakers. In: Interspeech 2023; Dublin: ISCA. p. 2553–2557. doi: 10.21437/Interspeech.2023-2134

19. Gerlach L, McDougall K, Kelly F, Alexander A, Nolan F. Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features. Speech Commun. 2020;124:85–95. Available from: https://www.sciencedirect.com/science/article/pii/S016763932030251X
20. Phonexia Voice Inspector version 5.1.0. Available from: https://www.phonexia.com/use-case/audio-forensics-software/ Accessed on 05.02.2024.
21. Hughes V. Sample size and the multivariate kernel density likelihood ratio: how many speakers are enough? Speech Commun. 2017;94:15–29.
22. Phonexia Voice Inspector – User Manual. Phonexia s.r.o. Manual version 2023-12-07.
23. Brümmer N, Swart A. Bayesian calibration for forensic evidence reporting. In: Li H, Meng HM, Ma B, Chng ES, Xie L, editors. Interspeech 2014. 15th Annual Conference of the International Speech Communication Association. Singapore: ISCA; 2014. p. 388–92.
24. Evett IW. Towards a uniform framework for reporting opinions in forensic science casework. Sci. Justice. 1998;3(38):198–202.
25. Morrison GS, Zhang C, Rose P. An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. Forensic Sci Int. 2011;208(1–3):59–65.
26. Drygajlo A, Jessen M, Gfroerer S, Wagner I, Vermeulen J, Niemi T, et al. Methodological guidelines for best practice in forensic semi-automatic and automatic speaker recognition. Frankfurt: Verlag für Polizeiwissenschaft; 2015.
27. Morrison GS, Enzinger E, Hughes V, Jessen M, Meuwly D, Neumann C, et al. Consensus on validation of forensic voice comparison. Science & Justice. 2021;61(3):299–309. Available from: https://www.sciencedirect.com/science/article/pii/S1355030621000083
28. McKenna L, McDermott S, O'Donell G, Barrett A, Rasmusson B, Nordgaard A, et al. ENFSI Guideline for Evaluative Reporting in Forensic Science: Strengthening the evaluation of forensic results across Europe (STEOFRAE). Wiesbaden, Germany: European Network of Forensic Science Institutes; 2015. Available from: http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf
29. Brümmer N, Du Preez J. Application-independent evaluation of speaker detection. Computer Speech & Language. 2006;20(2–3):230–75.
30. Reilly D, Neumann DL, Andrews G. Gender differences in self-estimated intelligence: Exploring the male hubris, female humility problem. Front. Psychol. 2022;13:812483.
31. Kudera J, Coccia M, Fadaeijouybari S, Preidt T, Ranjan A, Braun A. Voice cloning and mismatch conditions in forensic automatic speaker recognition. In: Karpov A, Delić V, editors. Speech and Computer. Cham: Springer Nature Switzerland; 2025. p. 171–84. Available from: https://doi.org/10.1007/978-3-031-78014-1_13
32. Kelly F, Forth O, Kent S, Gerlach L, Alexander A. Deep neural network based forensic automatic speaker recognition in vocalise using x-vectors. In: Audio Engineering Society Conference: 2019 AES International Conference on Audio Forensics. Porto, Portugal: Audio Engineering Society; 2019. p. 151–57.

ORCID
Jacek Kudera 0000-0003-3678-1067
Naima Islam Nodi 0009-0009-9456-4045
Julia Roegner 0009-0005-4553-1483

Corresponding author
Dr. Jacek Kudera
Trier University
Universitätsring 15
D-54296 Trier
e-mail: kudera@uni-trier.de

TRUDNOŚCI ZWIĄZANE ZE SŁUCHOWYM I AUTOMATYCZNYM ROZPOZNAWANIEM OSÓB NA PODSTAWIE GŁOSU: ANALIZA PRZYPADKÓW GŁOSÓW BARDZO PODOBNYCH

1. Wprowadzenie

Kluczową rolę w rozpoznawaniu osób na podstawie głosu odgrywa naturalne zróżnicowanie cech mowy. Te właśnie niepowtarzalne właściwości głosu umożliwiają odróżnianie osób zarówno przez ludzi, jak i w trybie maszynowym. Pojawiają się jednak trudności w sytuacji niewielkiego zróżnicowania mowy, wynikającego np. z podobnej struktury anatomicznej narządów fonacyjnych i artykulacyjnych [1–3]. Monozygotyczne bliźnięta jednojajowe stanowią skrajny przypadek podobieństwa głosu, spowodowanego informacjami genetycznymi określającymi kształt toru głosowego [4]. Podobieństwo to jest jeszcze większe, gdy bliźnięta są wychowywane w tym samym środowisku i uczą się tej samej odmiany języka [5]. Dlatego mowa monozygotycznych bliźniąt jest cennym i skomplikowanym materiałem do analizy skuteczności automatycznych systemów rozpoznawania mowy (ASRM). Wcześniejsze badania nad skrajnie małym zróżnicowaniem mowy skupiały się na parametrach czasowych, takich jak długość pauz [6] czy tempo mówienia [7], oraz na parametrach widmowych sygnału mowy [8], w tym na ocenie składowej głównej [9], dynamiki formantów [10] i cechach źródła krtaniowego [11]. Badania porównujące cechy mowy bliźniąt jednojajowych obejmowały różne jednostki językowe, w tym wybrane fonemy [12], sekwencje fonemów [13], słowa [14] czy stopień koartykulacji [15].

Obecne w badaniach nad rozpoznawaniem osoby mówiącej dominują dwa podejścia. Pierwsze z nich polega na analizie maszynowej reprezentacji sygnału mowy, drugie – na słuchowo-akustycznej ocenie dokonywanej przez człowieka. W niniejszej pracy zastosowano podejście łączące metody słuchowe i maszynowe rozróżniania głosu. W odniesieniu do analizy głosów bliźniąt wyniki badań sugerują, że ASRM są w stanie rozróżnić bliźnięta jednojajowe na podstawie próbek mowy, ale badacze stwierdzają wpływ płci, tj. ASRM skuteczniej rozpoznają głosy męskie niż żeńskie [16]. Trafność nowoczesnych systemów, nawet tych dostosowanych do głosów skrajnie podobnych, w tym głosów bliźniąt, często waha się w zależności od długości próbki i innych czynników [17]. Z tego powodu do skutecznego rozpoznawania próbek mowy o skrajnie niskim zróżnicowaniu pomiędzy osobami mówiącymi wydaje się przyczyniać stosowany system. Wyniki badań skłaniają do poszukiwania podejść hybrydowych, łączących ocenę przez człowieka i maszynową [18]. Wcześniejsze publikacje porównujące skuteczność metod maszynowych i oceny przez człowieka w odniesieniu do podobnych głosów stwierdziły dodatnią korelację pomiędzy tymi dwiema metodami [19]. W niniejszej pracy zastosowano podejście hybrydowe w celu analizy przypadku skrajnie małego zróżnicowania pomiędzy osobami mówiącymi. Zadano następujące pytania badawcze: Jak skuteczna jest ocena przez człowieka w porównaniu z ASRM do celów rozróżniania bardzo podobnych głosów bliźniąt jednojajowych? Czy płeć wpływa na słuchaczy? Na podstawie jakich parametrów słuchacze dokonują oceny?

2. Metody

2.1. Nagrania

Próbki głosu bliźniąt jednojajowych analizowane w ramach niniejszej pracy nagrano na potrzeby programu telewizyjnego typu reality show pt. „Die Zwillingsshow" („Bliźnięta"), emitowanego przez stację RTL w 2004 r. W ramach serialu bliźnięta jednojajowe miały do wykonania różne zadania, m.in. nagrać czytany na głos tekst – nagrania te wykorzystano w doświadczeniach – oraz nagrać mowę spontaniczną. Nagrania wykonano z częstotliwością próbkowania 48 kHz i głębią 16 bitów. Czytanym tekstem było niemieckie tłumaczenie bajki Ezopa „Wiatr Północny i Słońce". Nagrania prowadzono w otoczeniu dostosowanym akustycznie. Łącznie nagrano głosy 55 par bliźniąt (30 par kobiet i 25 par mężczyzn, średnia wieku 34 lata).

2.2. Doświadczenie 1: Ocena przez człowieka

Słuchaczami było pięćdziesięciu rodzimych użytkowników języka niemieckiego (20 kobiet i 30 mężczyzn). Żaden słuchacz nie zgłaszał zaburzeń mowy lub słuchu. Wiek uczestników mieścił się w przedziale 21–68 lat (średnia wieku 36 lat). Żaden uczestnik nie był szkolony w zakresie fonetyki. Uczestnikom zapłacono za udział w badaniu. Odtwarzano im parami próbki mowy i proszono ich o stwierdzenie na zasadzie wymuszonego wyboru, czy poszczególne pary próbek pochodziły od tej samej osoby, czy od różnych osób. Uczestnicy odsłuchiwali nagrania w słuchawkach. Słuchaczom odtworzono 55 par próbek (30 par bliźniaczek i 25 par bliźniaków) o długości od 9 do 13 sekund. Starano się stosować próbki o jednakowej długości. Różnice w długości wynikały

z zachowywania pauz i konturów intonacji na poziomie wyrażenia. Nagrania odtwarzano w losowej kolejności, tak by na każde cztery pary różnych osób przypadała jedna próbka bliźniąt jednojajowych. Próbki zawierały dwie części niemieckiego tłumaczenia „Wiatru Północnego i Słońca". Jedno nagranie obejmowało drugie i trzecie zdanie bajki (zob. Załącznik, cz. 1), a drugie nagranie – pozostałe zdania (Załącznik, cz. 2). Słuchaczy pytano, czy ich zdaniem poszczególne pary nagrań pochodziły od tej samej osoby, czy od dwóch różnych osób. W przypadkach gdy prawdą podstawową była „ta sama osoba", obie próbki głosowe pochodziły od któregoś z bliźniąt. Na życzenie uczestników nagrania odtwarzano ponownie.

Przed fazą doświadczalną uczestnikom stawiano trzy pytania próbne, żeby zaznajomić ich z procedurą badawczą i w razie konieczności zmienić głośność do poziomu, który im odpowiadał. Uczestników poproszono również o ocenę, jak bardzo byli pewni swoich odpowiedzi w skali od 1 (zupełny brak pewności) do 5 (wysoka pewność). Po etapie rozróżniania próbek słuchaczy pytano, na podstawie jakich cech oceniali osoby mówiące. Słuchacze mogli wymienić dowolną liczbę cech. Po próbach doświadczalnych przeprowadzono część podsumowującą, w trakcie której uczestnikom objaśniono cel badania. Cała sesja trwała ok. 35 minut.

2.3. Doświadczenie 2: Podejście maszynowe

W niniejszej pracy wykorzystano stosowany w praktyce automatyczny system rozpoznawania mowy do celów kryminalistycznych (Phonexia Voice Inspector, wersja 5.1.0) [20]. Dane analizowano na podstawie znanej prawdy podstawowej. Docelowe wartości punktowe obliczono na podstawie próby populacyjnej złożonej z 40 mężczyzn będących rodzimymi użytkownikami języka niemieckiego. Każdy natywny użytkownik w próbie populacyjnej był reprezentowany przez dwa pliki zawierające mowę czytaną. Mowa czytana była zbliżona pod względem treści, trybu oraz stylu do analizowanych próbek, dzięki czemu użyty w pracy zestaw danych populacyjnych był dostosowany do badanego przypadku [21]. Zastosowano tzw. tryb jeden-do-jednego, w którym wyniki zarówno docelowe, jak i niedocelowe oblicza się na podstawie wybranej próby populacyjnej [22]. Ponieważ próbki nagrano w otoczeniu akustycznie kontrolowanym, próg stosunku sygnału do szumu (SNR) ustawiono na 3 dB. Automatycznie przypisane etykiety wykrywania aktywności głosowej zostały porównane z przebiegami fal i spektrogramami nagrań, a następnie sprawdzone wizualnie pod kątem poprawności granic przez wykwalifikowanego fonetyka.

Wyniki porównania mówców zazwyczaj przedstawia się jako numeryczny współczynnik logarytmicznej wiarygodności (*log-likelihood ratio*, LLR) [23], który jest obecnie przyjmowaną metodą raportowania wyników kryminalistycznych porównań mówców [24, 25]. Ten oparty na rozumowaniu bayesowskim, uzgodniony współczynnik w tym przypadku odzwierciedla wyniki dwóch konkurencyjnych hipotez: hipotezy, że dane próbki mowy pochodzą od tej samej osoby (H0) oraz hipotezy, że próbki mowy pochodzą od dwóch różnych osób (H1) [26, 27]. Następnie, zgodnie z wytycznymi Europejskiej Sieci Instytutów Nauk Sądowych (European Network of Forensic Science Institutes, ENFSI) [28] dotyczącymi sporządzania raportów oceniających w naukach sądowych, wyniki te powiązano z werbalnymi korelatami uzyskanych stosunków dla ułatwienia interpretacji.

2.4. Analiza danych

W ramach Doświadczenia 1 zanalizowano 5500 obserwacji (50 uczestników × 55 par × 2 odpowiedzi: zero-jedynkowe stwierdzenie tego samego mówcy bądź różnych mówców oraz skalarna ocena punktowa pewności). Obliczono zgodność pomiędzy uczestnikami w odniesieniu do stwierdzania tego samego mówcy bądź różnych mówców za pomocą współczynnika α Krippendorffa oraz przeprowadzono test Levene'a dla równości wariancji i test Manna-Whitneya do sprawdzenia różnic w ocenie pewności.

W ramach Doświadczenia 2 analizowano współczynnik logarytmicznej wiarygodności danych na podstawie hipotez o tym samym mówcy oraz różnych mówcach, uzyskany za pomocą systemu Phonexia Voice Inspector i obliczony przy użyciu zalecanego trybu uogólnionej logistycznej funkcji gęstości prawdopodobieństwa (wersja 5.1.0), który polega na tym, że maksymalne oszacowania prawdopodobieństwa odpowiadają uogólnionemu rozkładowi logistycznemu typu IV z parametrem lokalizacji i skali [22]. W pracy przedstawiono współczynniki werbalne odpowiadające otrzymanym LLR.

3. Wyniki

3.1. Doświadczenie 1: Ocena przez człowieka

Analiza ujawniła, że uczestnicy trafniej rozpoznawali pary próbek mowy pochodzących od tego samego mówcy niż te od różnych mówców. Ogólna dokładność wyniosła 68,6%, przy czym udział poprawnie rozpoznanych próbek od tego samego mówcy wynosił 93,3%, a próbek od różnych mówców – 62,4% (szczegółowe analizy względem par przedstawia ustęp 6). Obliczony współczynnik α Krippendorffa (α = 0,348) wskazuje na niewielką zgodność między słuchaczami, co z kolei sugeruje, że występowały znaczne indywidualne różnice w trafności oceny lub że procedura doświadczalna miała duży stopień trudności.

Test Levene'a dowiódł, że wartości pewności różniły się znacząco w zależności od warunków doświadczenia. Ponieważ wariancje okazały się nie być jednolite pomiędzy obiema grupami ($p < 0{,}05$), przeprowadzono test t Welcha w celu porównania ocen pewności. Wyniki testu pokazały istotną różnicę ($t = 9{,}65$, $p < 0{,}05$), co sugeruje, że uczestnicy byli pewniejsi swoich ocen przy rozpoznawaniu próbek mowy bliźniąt niż próbek mowy różnych osób. Obserwację tę potwierdzają istotne statystycznie wyniki testu Manna-Whitneya ($p < 0{,}05$), co sugeruje również, że słuchacze czuli się pewniej w odniesieniu do próbek ocenianych prawidłowo niż próbek ocenianych błędnie. Z danych na Ryc. 1a wynika, że słuchacze deklarowali większą pewność przy rozróżnianiu próbek mowy pochodzących od par tego samego mówcy niż od par różnych mówców.

Analiza zbioru danych potwierdziła wcześniejszą obserwację, że głosy żeńskie sprawiają oceniającym większe trudności niż męskie. Różnica ta nie była jednak istotna statystycznie. Ogólna dokładność oceny dla par bliźniaków wyniosła 70,46% (pary bliźniąt 94,80%, pary różnych osób 64,67%). Ogólna dokładność oceny dla par bliźniaczek wyniosła 66,90% (pary bliźniąt 92,00%, pary różnych osób 60,35%).

Słuchacze stwierdzili, że dialekt i intonacja mówców oraz wymowa przez nich głoski [r] były najbardziej wyrazistymi cechami, na podstawie których oceniali próbki. Ponieważ pary bliźniąt pochodziły z całych Niemiec i dorastały one razem, trudno jest uzasadnić, w jaki sposób dialekt mógł przyczynić się tu do wyników, ale takie uzyskano odpowiedzi od słuchaczy.

3.2. Doświadczenie 2: Podejście maszynowe

System walidowano za pomocą współczynnika logarytmicznej wiarygodności i kosztu (*cost log-likelihood ratio*, CLLR), który jest miarą powszechnie stosowaną do oceny wydajności kryminalistycznych ASRM [29]. Współczynnik ten przyjmuje wartość od 0 do 1. Wartość zbliżona do 0 świadczy o dobrej kalibracji danego systemu [30]. Ogólna skuteczność ASRM, stwierdzona na podstawie badanej populacji, wyniosła EER = 1,25% (próg EER = 4,58, CLLR = 0,102), co wskazuje, że użyty system był dobrze skalibrowany.

Wyniki generowane przez system stanowiły zarówno LLR, jak i współczynniki werbalne powiązane z przedziałami prawdopodobieństwa. Współczynniki werbalne opierają się na stopniowych modyfikatorach mocy dowodów, które informują o dwunastu stopniach pewności, tj. pewności ograniczonej (1), umiarkowanej (2), umiarkowanie silnej (3), silnej (4), bardzo silnej (5) i ekstremalnie silnej (6), na korzyść bądź niekorzyść hipotezy strony oskarżającej (liczby w nawiasach odpowiadają przedziałom LLR uzyskanym za pomocą kryminalistycznego ASRM). Dodatnie wartości LLR interpretuje się jako przemawiające za hipotezą o tym samym mówcy, ujemne – za hipotezą o różnych mówcach. Automatyczny system skalibrowany do mowy czytanej w języku niemieckim osiągnął stuprocentową dokładność w odniesieniu do zestawu danych o znanej prawdzie podstawowej. Stuprocentowa dokładność oznacza w tym przypadku, że system zawsze poprawnie stwierdzał identyczność bądź różność wyników. Stopień (ocena prawdopodobieństwa) tych stwierdzeń był jednak różny.

Skrajne LLR wspierające hipotezę o tym samym mówcy przyjmowały wartości od 2,95 (para 5: umiarkowanie mocne dowody wspierające hipotezę strony oskarżającej) do 4,64 (para 28: bardzo mocne dowody wspierające hipotezę oskarżenia), podczas gdy LLR przemawiające za hipotezą o różnych mówcach wahały się od −0,24 (para 15: ograniczone dowody przeciwko hipotezie strony oskarżającej) do −20,45 (para 58: skrajnie mocne dowody przeciwko hipotezie oskarżenia). Okazało się, że 75,0% ocen par różnych mówców osiągnęło wyższe bezwzględne wartości LLR niż średnie wyniki dla par tego samego mówcy. Powyższe obserwacje ilustruje wykres gęstości na Ryc. 1b.

Test Levene'a ($F = 13{,}3$, $p < 0{,}05$), test t Welcha ($t(31{,}66) = -5{,}26$, $p < 0{,}05$) oraz test Manna-Whitneya ($U = 108{,}0$, $p < 0{,}05$) przy $\alpha = 0{,}05$ wskazują na istotne różnice w wariancji, średniej i przedziałach wartości LLR dla obu badanych warunków, tj. par tego samego mówcy i par różnych mówców. Porównując jednak zestawy danych według płci, te same testy nie wykazują istotnych różnic w LLR dla par kobiet i par mężczyzn.

3.3. Porównanie wyników oceny przez człowieka i za pomocą ASRM

Chociaż ogólnie rzecz biorąc, wartości punktowe pewności nie zawsze korelują ze skutecznością wykonania zadań słuchowych, nie ma innej empirycznej skali, która mogłaby posłużyć do porównywania pewności w słuchowym rozróżnianiu mówców przez ludzi i w podejściu maszynowym. Jak pokazuje Ryc. 2., w niniejszej pracy porównywano wyniki uzyskane obiema metodami na podstawie odpowiedzi uczestników, którzy oceniali, czy próbka pochodziła od tego samego mówcy, czy od różnych, oraz wartości punktowych pewności i odpowiednich werbalnych interpretacji LLR (różowe słupki). Przeciętna pewność obliczana dla wszystkich odpowiedzi (niebieskie trójkąty) odnosi się do przeciętnych przypisanych wartości, przy czym wartości ujemne dotyczą stwierdzeń o różnych mówcach.

Podczas analizy wyłącznie poprawnych ocen dokonanych przez człowieka (niebieskie słupki) okazało się, że średnia pewność dla 11 par tego samego mówcy i 10 par różnych mówców (11F, 15F, 18F, 30F, 33M, 43F, 47M, 55F, 59M, 60M) była wyższa niż LLR uzyskane za pomocą systemu Phonexia Voice Inspector. O ile

słuchacze płci męskiej wykazują zazwyczaj wyższy poziom pewności niezależnie od trafności oceny [30], o tyle w niniejszym badaniu stwierdzono minimalne różnice w pewności między słuchaczami (średnia pewność: kobiety = 4,07, mężczyźni = 4,15).

4. Dyskusja

Kryminalistyczny ASRM okazał się bardzo skutecznie rozróżniać przypadki głosów skrajnie podobnych, czyli mowę bliźniąt jednojajowych. Ogółem system oceniał próbki trafniej niż słuchacze niebędący specjalistami. Poprawnie zakwalifikował on wszystkie 16 par mówców, które słuchacze ludzcy ocenili błędnie. Innymi słowy, ASRM lepiej niż człowiek ocenił 16 par. Były to wyłącznie pary różnych mówców, w tym 10 par kobiet i 6 par mężczyzn. Wyniki te są zbliżone do przeszłych obserwacji, według których głosy żeńskie są trudniejsze do interpretacji niż głosy męskie [16]. Wyniki uzyskane przez ludzi i system automatyczny przeważnie pokrywają się w niewielkim stopniu (operacjonalizowanym jako pewność w przypadku oceny przez człowieka i LLR w podejściu maszynowym).

Nie stwierdzono istotnych różnic pomiędzy parami bliźniaków i bliźniaczek w przypadku oceny zarówno przez człowieka, jak i ASRM. Wyniki nie odbiegają od tych uzyskanych we wcześniejszych badaniach, które sugerują, że głosy żeńskie rozpoznaje się trudniej niż męskie. Stwierdzono jednak, że ASRM dawał większy stopień pewności dla par różnych mówców, a ocena przez słuchaczy niebędących specjalistami – dla par tego samego mówcy. Wszystkie błędy ludzkie były związane z identycznością głosów, co w połączeniu z większymi wartościami pewności względem par tego samego mówcy pokazuje, że słuchacze byli pewniejsi wyników nieprawidłowych niż prawidłowych. Kwestię tę należy poddać dalszym badaniom.

Ograniczenia niniejszej pracy dotyczą próby populacyjnej; można temu zawsze zaradzić, zwiększając liczbę reprezentatywnych mówców. Innym ograniczeniem pracy był niewielki potencjał do wyjaśniania wyników otrzymanych podejściem maszynowym. Choć użyta w pracy próba populacyjna nie była dokładnie zrównoważona pod względem płci, stwierdzono zadowalającą skuteczność modelu przy akceptowalnych progach kalibracji [30]. Autorzy pracy zdają sobie też sprawę z ryzyka związanego ze stosowaniem skal pewności jako podstawy do porównania skuteczności oceny przez człowieka i przez system automatyczny.

5. Wnioski

Szybki rozwój technologii identyfikacji osób na podstawie głosu pozwolił na rozpowszechnienie stosowania kryminalistycznych ASRM przez organy ścigania i agencje śledcze. Podejście maszynowe do rozpoznawania mowy wymaga jednak dużej próby populacyjnej, reprezentującej warstwę demograficzną badanych mówców, a jego wyniki zależą od jakości porównywanych nagrań [31]. Ocena przez człowieka wymaga natomiast fachowych umiejętności i szkolenia. W niniejszej pracy wykorzystano podejście hybrydowe do sprawdzenia ludzkich zdolności słuchowych i skuteczności ASRM w celu rozróżnienia osób o skrajnie podobnych głosach.

Wykazano, że nawet nieprzeszkoleni rodzimi użytkownicy języka potrafią rozróżniać podobne próbki głosu pochodzące od bliźniąt jednojajowych. Gdyby doświadczenia powtórzyć z udziałem doświadczonych fonetyków i kryminalistyków, prawdopodobnie otrzymanoby jeszcze bardziej miarodajne wyniki. Ponadto uczestnicy badania wskazali cechy, które ich zdaniem były przydatne przy rozpoznawaniu mówców. O ile cechy, które słuchaczom wydają się dostarczać informacji, nadają się do interpretacji, o tyle sygnały, które dostarczają informacji kryminalistycznemu ASRM, są dla ludzi niezrozumiałe. Wyniki badania pokazują, że w warunkach idealnych ASRM pozwala rozróżniać głosy bliźniąt jednojajowych, a nawet robi to trafniej od ludzi nieprzeszkolonych w zakresie fonetyki, osiągając doskonałe wyniki. Można również porównać skuteczność innych popularnych ASRM, np. VOCALISE [32], ze skutecznością piątej generacji modelu Phonexia Voice Inspector.

Wyniki przeprowadzonych doświadczeń nie wpływają bezpośrednio na praktykę kryminalistyczną, głównie dlatego, że ich uczestnikami byli nieprzeszkoleni słuchacze, w odróżnieniu od specjalistów po dogłębnym szkoleniu w fonetyce sądowej. Dodatkowo materiał nie odzwierciedlał rzeczywistych warunków pracy kryminalistyków, które przeważnie polegają na tym, że do dyspozycji jest niewiele próbek mowy, i to zniekształconej, a przypadki z udziałem bliźniąt jednojajowych są niezwykle rzadkie. Z drugiej strony często zdarza się, że głos sprawcy jest mylony z głosem jego rodzeństwa, których struktura anatomiczna, jak można założyć, jest bardziej zbliżona do sprawcy niż struktura anatomiczna osoby niespokrewnionej. W takich przypadkach, sprawiających trudności nawet doświadczonym specjalistom, z pomocą mogą przyjść ASRM. Niniejsza praca uwidacznia znaczny postęp ASRM w zakresie rozróżniania głosów osób, których aparaty głosowe są niemal identyczne pod względem anatomicznym; por. Künzel, który w 2010 r. stwierdził wiele błędów podczas korzystania z systemu GMM-UBM (gaussowski model mieszany-uniwersalny model tła) [16]. Istnieją jednak pewne niezmienne źródła błędów w ASRM, np. dokładność

etykietowania VAD, arbitralne progi SNR czy jakość i liczność zestawu danych populacyjnych. Co więcej, nie ma podstaw do zakładania, że dane dwa głosy, które brzmią podobnie dla ludzkich słuchaczy, ale niekoniecznie powstały w podobnie ukształtowanych torach głosowych (zjawisko tzw. głosów bliźniaczych) [18], sprawią większe trudności systemowi automatycznemu niż dwa dowolne inne głosy.

Słuchacze bez wiedzy specjalistycznej wykazywali ogółem wyraźną tendencję do stwierdzania identyczności, tj. popełniali więcej błędów, kiedy mówcy byli różni, niż kiedy mówca był ten sam. Z drugiej strony ASRM popełnił więcej błędów (mierzonych stopniem pewności) w odniesieniu do próbek tego samego mówcy. Można więc powiedzieć, że system oceniał tu próbki bardziej zachowawczo niż człowiek. Co ciekawe, nie zaobserwowano istotnych różnic ze względu na płeć mówców. Niniejsza praca może również przyczynić się do analiz porównawczych skuteczności różnych kryminalistycznych ASRM. Analiza wyników testów ze znaną prawdą podstawową jest kluczowa dla sprawdzania działania systemów automatycznych.