NBP 2025, Vol. 30, Issue 2, pp. 110–126
https://doi.org/10.5937/nabepo30-56871
Original Scientific Paper

Speaker Recognition Based on Auditory Impression: The Role of Familiarity with the Speaker and Language

Mia Šešum[1], **Bojana Drljan**[2], **Maja Ivanović**[3], **Ivana Arsenić**[4]
University of Belgrade, Faculty of Special Education and Rehabilitation, Serbia

Submitted: 2025-02-18 • Accepted: 2025-05-12 • Published: 2025-05-23

Abstract: Speech is a fundamental means of interpersonal communication. Speaker identification based on voice and speech can be analysed through two perspectives, expert listening by trained phoneticians, for the purpose of forensic speaker identification, and by naive listeners. Factors related to the success of identification often include prior familiarity with the speaker and the language they speak. The aim of the study is to examine whether speaker recognition based on auditory impressions is influenced by prior familiarity with the speaker and the language being spoken. A total of 218 female students from the Faculty of Special Education and Rehabilitation participated in the experiment. An instrument was specially designed for the purposes of this research. The results indicated that familiarity with the spoken language did not influence speaker recognition. Similarly, familiarity with the speaker had no significant effect on recognition, except in the case of the English-speaking speaker, whose voice and speech were more accurately recognized by participants who had been previously familiar with her. Given that the research findings did not consistently support the hypothesis regarding the connection between prior familiarity with the speaker and the language they speak and the success of speaker identification, it is necessary for future research to focus on examining the connection between recognizing speakers and the acoustic characteristics of their voice and speech.

Keywords: forensics, perception, identification, speech, familiarity.

INTRODUCTION

Speech is a complex signal composed of interconnected tones that form meaningful units, perceived through the ear and brain's ability to process sound waves (Alkhatib & Kamal Eddin, 2020). It serves as a primary means for conveying information, where the speaker encodes a message into a variable waveform, and the listener decodes it upon reception

[1] Corresponding author: gia982@gmail.com • https://orcid.org/0000-0001-5192-878X • Phone: +381 64 24 73 97 7
[2] bojanad77@gmail.com • https://orcid.org/0000-0003-1501-3325
[3] majapivanovic@gmail.com • https://orcid.org/0000-0001-5194-1714
[4] ivana.arsenic@yahoo.com • https://orcid.org/0000-0003-2516-7811

eISSN 2620-0406
Citation: Šešum, M., Drljan, B., Ivanović, M., & Arsenić, I. (2025). Speaker recognition based on auditory impression: The role of familiarity with the speaker and language. *NBP. Nauka, bezbednost, policija*, *30*(2), pp. 110–126. https://doi.org/10.5937/nabepo30-56871

(Islam et al., 2022). Recognizing speakers based on their voice and speech is a crucial human ability, occurring at both naive and professional levels (Sharma & Sahu, 2018). At the naive level, recognizing familiar voices, such as those of family or friends, is a common daily experience, especially when visual cues are absent, like during phone conversations (Didla, 2020). However, recognizing unfamiliar voices relies on memory, which can become unreliable over time, reducing the accuracy of such recognition (Lindh, 2017). At the professional level, speaker recognition based on voice and speech plays a crucial role in forensics. Speaker recognition is a fundamental task in forensic phonetics, aimed at identifying an unknown speaker (Rana & Qureshi, 2024). Forensic phonetics relies on complex analyses of a speaker's voice and speech to uncover their identity (Jain et al., 2024). Voice is generally understood as the product of vocal tract activity, whereas speech denotes the specific realization of language (Šešum, 2021). Forensic phonetics is an interdisciplinary field that involves the application of knowledge from social, medical, and biological sciences (Carić & Širić, 2023). Speaker identification, as a core aspect of forensic phonetics, involves comparing the voice and speech of an unknown speaker from a disputed recording with that of a known suspect to establish identity (Zhou et al., 2022). Expert recognition through listening is a key component of the auditory-instrumental method, which overcomes challenges like poor recording quality and discrepancies in recording channels (Hansen & Hasan, 2015). The task of the forensic analyst is to listen to, analyse, and compare recordings of a known and a suspected speaker in order to determine whether they belong to the same individual (Jain et al., 2024).

Research in forensic phonetics has shown significant individual differences in speaker recognition abilities (Aglieri et al., 2017; Lavan, Burston, Ladwa, et al., 2019; Lavan, Merriman, Ladwa, et al., 2019; Mühl & Bestelmeyer, 2018, as cited in Jenkins et al., 2021). Identifying individuals with exceptional recognition skills can greatly benefit forensic investigations. Such individuals can assist in cases involving kidnappings, extortions, threats, and terrorist actions (Jenkins et al., 2021). Their expertise enhances the effectiveness of forensic investigations and contributes to societal safety by resolving serious criminal cases. Speaker identification based on voice and speech is becoming an increasingly intriguing topic in academic circles, as it is considered a key method of identification in the 21st century (Rana & Qureshi, 2024). Although voice and speech can serve as evidence in almost all types of criminal activity, they most commonly appear in cases involving organized crime, drug trafficking, extortion, threats, kidnappings, corruption, terrorism, rape, and murder (Šešum & Kovačević, 2015).

Most people find it easier to recognize faces than voices. Kreiman and Sidtis (2011) argue that recognizing familiar voices relies on a few distinctive vocal features, making it robust against moderate acoustic variations or non-machine masking. For unknown speakers, listeners must rely on any significant voice and speech features available (Nygaard, 2005). Nygaard (2005) highlights that linguistic and non-linguistic information are interconnected parts of the same speech signal, influencing speech perception. Petrini and Tagliapietra (2008) find that listeners are more sensitive to speech characteristics than voice characteristics, though separating the two can be challenging. Important characteristics of voice include features such as frequency, timbre, and intensity (although the latter is also influenced by speech manner), while characteristics of speech include rhythm, tempo, pauses, and lexical and syntactic specificities, among others (Šešum & Kovačević, 2015).

Using sentences instead of isolated words provides a richer phonetic environment, aiding recognition through intonation, stress patterns, and coarticulation (Goggin et al., 1991; Perrachione & Wong, 2007). However, Lavan et al. (2020) conclude that listener reliability decreases as the number of speech samples increases.

Voice and speech variability significantly impact the ability to recognize speakers (Zhou et al., 2022). This variability is substantial even for the same speaker, as voice and speech can change considerably across different contexts (Lavan et al., 2020). For example, when a person is excited or speaking on the phone, the volume and frequency range of their speech can differ significantly compared to when they are speaking calmly and face-to-face (Mukattash, 2016). The difference in the auditory impression of a person's voice when speaking over the phone compared to face-to-face communication is a result of the fact that telephone transmission limits the frequency range of the speech signal to approximately 3100 Hz (from 300 Hz to around 3400 Hz). As a result, voice qualities that lie in the higher frequency spectrum – extending beyond 10,000 Hz in natural speech – cannot be perceived by listeners. Additionally, physiological changes, such as those experienced by pregnant women due to the lifting of internal organs, can lead to notable voice alterations (Šešum, 2021). Similarly, a speaker's voice and speech can vary depending on their physical state – tiredness, hunger, or illness can affect speech production differently compared to when the speaker is well-rested, nourished, and healthy (Šešum, 2021). Furthermore, the use of drugs, certain medications, and alcohol can cause significant changes in a person's voice and speech, both during and after their influence (Babić et al., 2017). Variations in speaking style and recording conditions can also mislead listeners, making them perceive recordings of the same speaker as belonging to different individuals. This is more likely to occur when recordings are made under different conditions (e.g., microphone vs. telephone) compared to recordings made under consistent conditions (Morrison & Enzinger, 2019). Even small changes in intonation, voice quality, or speaking style can lead listeners to misidentify the same speaker as multiple individuals (Lavan et al., 2020). Therefore, listeners must not only distinguish between different speakers but also generalize across variations in a single speaker's voice and speech (Lavan et al., 2020).

The ability to recognize a speaker based on their voice and speech is often considered natural. However, identifying a speaker in an unknown language presents a challenge (Wester, 2012). Winters et al. (2008) concluded that listeners rely more on language-dependent information for familiar languages, while for unknown languages, they focus on language-independent voice and speech characteristics (Wester, 2012). Lavan et al. (2020) emphasize that familiarity with both the speaker and the language enhances recognition, linking this ability to broader linguistic skills. Scientific literature on gender differences in auditory perception mainly explores areas such as the accuracy of harmonic sound processing (Krizman et al., 2021), reaction times to auditory stimuli (Krizman et al., 2019), and response consistency (De Vos et al., 2020; Krizman et al., 2020), as well as the ability to recognize emotions from speech (Lausen & Schacht, 2018; Rezić & Bonett, 2021). However, research on gender differences in speaker recognition based on voice and speech is notably lacking. Furthermore, forensic literature does not suggest any anticipated gender-related differences in this area.

Scientific studies examining the impact of prior familiarity with the speaker on voice and speech recognition are limited but consistent in their findings. Familiarity with the speaker facilitates identification (Kreiman & Papcun, 1991; Van Lancker & Kreiman, 1987; Lavan et al., 2020; Kanber et al., 2022; Njie et al., 2023). These findings could be highly significant for forensic voice and speech analysis, particularly given the limited number of experts in the field and the potential media exposure of individuals whose voices and speech are subject to analysis. For this reason, the focus of this research is directed toward examining the aforementioned connection, as previous studies have been scarce, largely outdated, and have never been applied to the Serbian speech area. Similarly, the existing scientific literature consistently indicates that familiarity with the spoken language also aids in speaker identification (Goggin et al., 1991; Philippon et al., 2007; Perrachione et al., 2011; Wester, 2012; Zarate et al., 2015; Perrachione, 2017), though research on this topic remains sparse. Since the aforementioned results suggest that unfamiliarity with the speaker's language, alongside objective distractors in forensic voice and speech analysis (Šešum & Petrović, 2024), may also represent a significant factor in shaping the expert's subjective perceptual impression, the research is directed towards investigating this connection as well. Given that the subjectivity of experts is most frequently cited in the literature as a drawback of traditional forensic voice and speech analysis methods (Arjamand et al., 2024; Durán et al., 2024; Morrison et al., 2020), it is crucial to identify the factors that potentially contribute to it. Unlike previous studies, this research examines both variables – familiarity with the speaker and the language – using speech samples in four languages. Given that the speakers in the recordings used for this study were four university professors employed at the institution where the research was conducted, participants who had previously attended one or more of their courses during their studies were considered familiar with the speakers. Conversely, participants who had not had such experience were categorized as unfamiliar. Participants were deemed familiar with the language spoken if they were able to communicate in it, that is, if they had sufficient proficiency to understand the content of the recordings presented in that language. The simultaneous examination of these factors within the same sample of participants in the experiment, using both the native language and three foreign languages, is significant as it allows for an understanding of their mutual influences as well as an assessment of their individual importance for speaker recognition. Additionally, the value of this research is further enhanced by the fact that it was conducted on a large number of participants who are professionally oriented toward the study of language and speech. Given the importance of auditory perception in speaker identification, both in everyday life and criminal investigations, this study explores its implications. The aim is to examine whether speaker recognition based on auditory impressions is influenced by prior familiarity with the speaker and the language being spoken.

METHODS

Participants

The participants were 218 female students from the undergraduate and master's programs at the Faculty of Special Education and Rehabilitation in Belgrade, who are studying fields related to hearing impairments, as well as speech and language pathology and communication disorders, because, unlike students from other departments at the Faculty of Special Education and Rehabilitation, they can be considered trained listeners due to the experience gained during their studies. The research sample of participants was convenience-based because the participants were selected based on their availability and willingness to participate, rather than through random sampling. It included only female participants because there were no male participants in the relevant study programs. Given the fact that scientific literature has not documented any research findings indicating gender differences in auditory perception related to speaker recognition, this limitation of the study can be considered conditional. All participants had Serbian as their native language, and none reported having hearing problems. The participants were born between 2000 and 2002. All participants voluntarily agreed to participate in the research. The study was conducted anonymously.

Measurements and Procedures

The study involved the playback of two sets of audio recordings. The first set included a total of four monologue recordings, each approximately 30 seconds in duration, spoken in different languages (Serbian, English, German, and Farsi) by four different female speakers, all university professors employed at the faculty. Based on the regional context and the prevalence of foreign language use, the researchers assumed that the highest number of participants would be familiar with English, a significantly smaller number would be familiar with German, and that speakers of Farsi – a language not commonly used in Serbia – would not be represented in the participant sample. The speakers were selected based on their fluency in the respective language, age (all were in their fifth decade of life), absence of speech characteristics that could influence recognition, and their familiarity to the participants. The German-speaking speaker was known to all participants, the Serbian-speaking speaker was familiar to 162 participants, the English-speaking speaker to 119 participants, and the Farsi-speaking speaker to only 20 participants. The fundamental frequency of the Serbian speaker's voice was 144 Hz, the English speaker's 204 Hz, the German speaker's 230 Hz, and the Farsi speaker's 173 Hz. The average fundamental frequency was calculated based on the full duration of each recording.

The second set of recordings consisted of ten recordings of the same text read in Serbian by ten different female speakers, each reading the text once, with a duration of approximately 30 seconds per recording. This uniform duration of recordings was selected to prevent auditory fatigue and because such duration is considered sufficient for conducting relevant analyses in the context of forensic voice and speech examination (Šešum & Kovačević, 2015). The first half of the "Balanced Text" (Šešum, 2013), specially constructed for forensic purposes, was used for the readings. The balanced nature of the text refers to the natural distribution of syllable frequencies within semantic units of the Serbian language, as well as the inclusion of all phonemes. The text contains complex utterances suitable for speech analysis (Šešum, 2013). In addition to the four speakers featured in the first set of recordings, the second set included six additional female speakers unfamiliar to the participants, who belonged to the same age group as those from the first set. The average fundamental frequency of these speakers ranged from 138 Hz to 244 Hz, and none exhibited speech characteristics that would significantly affect speaker recognition. All speakers who participated in this study were native speakers of Serbian. All recordings were made using a condenser microphone (AKG C 535 EB) and a portable battery-powered recorder

(MARANTZ-DENON PMD 660). The sampling rate for all recordings was 44.100 Hz, as recordings of this quality allow for the coverage of the entire speech spectrum as well as the audible spectrum.

Since the speech of the same four speakers was present in both sets of recordings, participants were instructed to identify, after listening to each recording from the first set, which speaker from the second set matched the one they had just heard. Each of the four recordings from the first set was played twice consecutively, with 15-second pauses between them, followed by a single playback of all ten recordings from the second set. Since there were four initial recordings (in Serbian, English, German, and Farsi), the entire procedure was repeated four times. The study lasted up to 45 minutes for each group of participants, including breaks between the recordings to allow participants to rest (instruction time: 5 minutes; total recording duration: 24 minutes; pauses between recordings totalling: 16 minutes). A duration of 45 minutes optimally met the needs of the study as well as the maintenance of participants' attention, as it corresponds to the length of a school class period, to which they are accustomed from their education. The pauses between playing consecutive recordings in Serbian lasted 18 seconds, while the pauses between playing recordings in different languages lasted one minute. The length of the pauses was motivated by the need to prevent auditory fatigue, as well as the need for research efficiency to maintain optimal concentration among the participants.

The participants filled out a questionnaire specifically designed for this research. The first part of the questionnaire included data about the participants (gender, age, year of study, study group, native language, hearing status), while the second part required them to mark the ordinal number of the recording they believed was spoken by the requested speaker. The ordinal numbers of the recordings from the second set, which were played consecutively after the playback of each recording from the first set, were verbally announced during the testing prior to each playback. This part of the questionnaire was composed of four columns, each corresponding to the spoken production of a different language. The testing was conducted with larger groups of participants (up to 60), and it was repeated in four different sessions within the same month, in the same space, in order to gather results from 218 participants. Given that the groups consisted of the participants of the same gender, similar age, and professional orientation, who have known each other for a long time, no issues were expected or observed regarding differing group dynamics. After an oral introduction to the purpose and method of the research, the participants were given the questionnaire to complete. Prior to filling out the questionnaire, the participants were informed both in writing and orally that their participation in the research was entirely anonymous and voluntary. After the participants completed the general data section of the questionnaire, researchers used a public address system in the lecture hall where the study was conducted to play the recordings. The only researcher who communicated with the participants was the speaker – a co-author of the study – who was already familiar to all participants.

After completing the testing, the participants were not informed of the correct answers to avoid influencing subsequent groups. Based on the general data obtained from the participants, they were categorized, and the results were evaluated in relation to the accuracy of identification for each of the four designated speakers. The study was conducted in accordance with the ethical standards set forth by the Declaration of Helsinki.

Statistical analyses

The data analysis was conducted using Social Package for Social Sciences (IBM, SPSS statistics, version 26), employing descriptive statistical methods to describe the results of experiment. The chi-square test was used to compare categorical variables with significance at the standard threshold of p < .05. In cases with a small number of participants in a cell, Fisher's exact test was conducted.

RESULTS

Table 1 provides descriptive data on language familiarity, familiarity with the speaker, and speaker recognition.

Table 1. *Sociodemographic and Experimental Characteristics of the Participants*

Language familiarity	*n*	%
English	154	70.6
German	1	0.5
English and German	8	3.7
None	55	25.2
Familiarity with the speaker	**Yes n(%)**	**No n(%)**
Serbian speaker	162(74.3)	56(25.7)
English speaker	119(54.6)	99(45.4)
German speaker	218(100.0)	0(0.0)
Farsi speaker	20(9.2)	198(90.8)
Speaker recognition	**Yes n(%)**	**No n(%)**
Serbian speaker	211(96.8)	7(3.2)
English speaker	139(63.8)	79(36.2)
German speaker	53(24.3)	165(75.7)
Farsi speaker	199(91.3)	19(8.7)

Table 1 presents descriptive information about participants' familiarity with different languages and speakers, as well as their recognition of each speaker. Most participants reported familiarity with English language (70.6%), while a small number were familiar with both English and German (3.7%), and only one participant reported familiarity with German language alone (0.5%). A quarter of the sample (25.2%) reported no familiarity with any of the listed languages. Notably, none of the participants reported familiarity with the Farsi language. In terms of familiarity with the speakers (speaker previously known to the participant), most participants were familiar with the Serbian speaker (74.3%), and more than half were familiar with the English speaker (54.6%). Only 9.2% of participants reported familiarity with the Farsi speaker, while all participants (100%) reported familiarity with the German speaker. Interestingly, although all participants reported familiarity with the German speaker, this speaker was the least frequently recognized: only 24.3% correctly identified them. In contrast, the Serbian speaker was correctly recognized by 96.8% of participants, followed by the Farsi speaker (91.3%) and the English speaker (63.8%).

Table 2. *Differences in Speaker Recognition*

			Recognition of the speaker				
Recognition of the speaker			English speaker			FET	*p*
			Yes	No	Total		
	Serbian speaker	Yes	132	79	211		.039*
		No	7	0	7		
		Total	139	79	218		
			German speaker			FET	*p*
			Yes	No	Total		
	Serbian speaker	Yes	53	158	211		.143
		No	0	7	7		
		Total	53	165	218		
			Farsi speaker			FET	*p*
			Yes	No	Total		
	Serbian speaker	Yes	193	18	211		.493
		No	6	1	7		
		Total	199	19	218		
			German speaker			χ^2	*p*
			Yes	No	Total		
	English speaker	Yes	42	97	139	7.266	.005*
		No	11	68	79		
		Total	53	165	218		
			Farsi speaker			χ^2	*p*
			Yes	No	Total		
	English speaker	Yes	129	10	139	1.116	.208
		No	70	9	79		
		Total	199	19	218		
			Farsi speaker			FET	*p*
			Yes	No	Total		
	German speaker	Yes	51	2	53		.176
		No	148	17	165		
		Total	199	19	218		

Note: **p* < .05; ***p* < .01; FET – Fisher's exact test. Values indicate the number of participants who correctly ("yes") or incorrectly ("no") recognized each speaker. Crosstabulations show comparisons between pairs of speakers.

In order to investigate the differences in the frequency recognition of individual speakers, the relationship between familiarity and recognition of the speaker, as well as the relationship between language familiarity and speaker recognition, a chi-square test was used.

Crosstabulations show comparisons between speakers. Familiarity with the German speaker was not included in the analysis because all participants were familiar with her. In cases with a small number of participants in a cell, Fisher's exact test was conducted to compare differences in familiarity with the speaker.

Cross-comparisons of the frequency of recognition of all individual speakers are provided in Table 2. In cases with a small number of participants in a cell, Fisher's exact test was conducted to compare speaker recognition across different languages.

The analysis of speaker recognition showed that 211 participants correctly recognized the Serbian speaker and 139 recognized the English speaker. Recognition of the English speaker was significantly lower compared to the Serbian speaker (FET, p = .039), but significantly higher than recognition of the German speaker, who was recognized by only 53 participants ($\chi^2_{(1)}$ = 7.266, p = .005). No statistically significant differences were found in the comparisons involving the Farsi speaker: 199 participants recognized the Farsi speaker, compared to 211 for the Serbian speaker (FET, p = .493), 139 for the English speaker ($\chi^2_{(1)}$ = 1.116, p = .208), and 53 for the German speaker (FET, p = .176). Recognition of the German speaker was the lowest among all, with only 53 participants identifying her correctly (Table 2).

The results of the analysis of the relationship between familiarity and recognition of the speaker are presented in Table 3. As in the previous case, familiarity with the German speaker was not included in the analysis because all participants were familiar with her, so that variable remain constant in analyses. In cases with a small number of participants in a cell, Fisher's exact test was conducted to compare relationship between familiarity with and recognition of the speaker.

The relationship between familiarity with and recognition of the speaker was statistically significant only for the English speaker. Among the participants who reported familiarity with the English speaker (n = 119), 85 correctly recognized her – compared to only 34 who did not recognize her. Conversely, only 54 participants who were unfamiliar with the English speaker recognized her, while 45 did not ($\chi^2_{(1)}$ = 6.667, p = .007). No significant associations were found for the Serbian or Farsi speakers. Among 162 participants familiar with the Serbian speaker, 157 recognized her correctly, compared to 5 among those unfamiliar (p = .585, Fisher's exact test). For the Farsi speaker, 20 participants reported familiarity and all recognized her correctly, while out of 198 unfamiliar participants, 179 still recognized her (p = .121, Fisher's exact test) (Table 3).

Table 3. *Relationship Between Familiarity with and Recognition of the Speaker*

			Familiarity with the speaker				
			Serbian speaker			χ^2(FET)	p
Recognition of the speaker			Yes	No	Total		
	Serbian speaker	Yes	157	54	211	FET	.585
		No	5	2	7		
		Total	162	56	218		
			English speaker				
			Yes	No	Total		
	English speaker	Yes	85	54	139	6.667	$.007^{*}$
		No	34	45	79		
		Total	119	99	218		
			Farsi speaker				
			Yes	No	Total		
	Farsi speaker	Yes	20	179	199	FET	.121
		No	0	19	19		
		Total	20	198	218		

Note: *p < .05; **p < .01; FET – Fisher's exact test. Values indicate the number of participants who correctly ("yes") or incorrectly ("no") recognized each speaker, grouped by whether they reported familiarity with that speaker ("yes" or "no"). Crosstabulations show the relationship between familiarity and recognition for each speaker.

The results of the analysis of the relationship between familiarity with the language and speaker recognition are presented in Table 4. In the cross-analysis of the relationship between speaker recognition and language proficiency, only the data on the English speaker and the English language, as well as on the German speaker and the German language, were included. The data on the Serbian speaker and the Serbian language, as well as on the Farsi speaker and the Farsi language, were not analyzed because all participants were native speakers of Serbian, resulting in no variability in familiarity with the language, and none of the participants spoke Farsi, leading to uniformly low familiarity across the sample of participants. This is a consequence of the specific characteristics of the study sample, and the data related to familiarity with Serbian and Farsi languages remain constant in the analyses.

The analysis did not reveal any statistically significant differences in speaker recognition based on familiarity with the language. Among the participants who reported familiarity with the English language (n = 162), 98 correctly recognized the English speaker, while 64 did not. In the group unfamiliar with English (n = 56), 41 correctly recognized the English speaker, while 15 did not ($\chi^2_{(1)} = 1.843$, p = .071). Similarly, no significant relationship was found for the German speaker. Only 9 participants reported familiarity with the German language, of whom 1 recognized the German speaker. Among the 209 participants who did not report familiarity with German, 52 recognized the speaker (p = .691, Fisher's exact test). These findings suggest that familiarity with the language did not significantly affect the likelihood of correctly recognizing the corresponding speaker (Table 4).

Table 4. *Relationship Between Familiarity with the Language and Speaker Recognition*

			Familiarity with the language				
Recognition of the speake			English language			χ^2(FET)	*p*
			Yes	No	Total		
	English speaker	Yes	98	41	139	1.843	.081
		No	64	15	79		
		Total	162	56	218		
			German language				
			Yes	No	Total		
	German speaker	Yes	1	52	53	FET	.691
		No	8	157	165		
		Total	9	209	218		

Note: FET – Fisher's exact test. Values indicate the number of participants who correctly ("yes") or incorrectly ("no") recognized each speaker, grouped by whether they reported familiarity with the corresponding language. Crosstabulations show the relationship between language familiarity and speaker recognition.

DISCUSSION

Speaker recognition based on voice and speech is a human ability that holds significant importance in everyday communication as well as in the field of forensics. Although there are numerous factors contributing to successful speaker recognition, the level of their individual impact remains a subject of professional and scientific debate. This research was conducted to determine the relationship between factors frequently mentioned in the professional literature and the accuracy of speaker recognition. Based on the obtained results, it can be observed that the majority of participants are proficient in English as a foreign language, Farsi is unknown to everyone, and a small number of participants speak German, primarily those who also know English. The differences in the recognition of speakers are statistically significant among all speakers. The listeners most accurately recognized the voice of the speaker who spoke their native Serbian language, followed by the speaker who spoke Farsi, while they recognized the voice of the speaker who spoke German the least. However, the results of the statistical analyses indicate that significant differences were recorded only in the recognition of the English speaker, who was less recognized compared to the Serbian speaker, but more recognized than the German speaker. Additionally, regarding the connection between prior familiarity with the speaker and the recognition of their voice and speech, statistical significance was confirmed only for the English speaker. These findings are very interesting as they contradict the results of previous studies (Kreiman & Papcun, 1991; Van Lancker & Kreiman, 1987; Lavan et al., 2020), which consistently confirmed that familiarity with the speaker is a significant factor for identifying their voice and speech, and thus for distinguishing them from other speakers.

The obtained results indicate that the listeners best recognized the voice and speech of the speaker who spoke their native language, which aligns with findings from global studies. For instance, Goggin et al. (1991) investigated the ability to recognize speakers who speak in the listener's native language and in foreign languages. Listeners whose native language is English were tasked with identifying bilingual speakers who spoke either English or German. The researchers found that listeners more easily recognized speakers who spoke their native language compared to when the same speakers spoke a foreign language, such as German. These findings are also supported by the results of the study by Philippon et al. (2007). However, the statistical analysis within our research did not confirm the significance of speaker recognition based on whether the listeners were familiar with the foreign language of the speakers. This is further supported by the fact that, after the Serbian speaker, the highest number of participants recognized the voice and speech of the Farsi speaker, a language that none of the listeners know and which is structurally and prosodically a language that differs significantly from Serbian, English, and German. On the other hand, a significantly larger number of participants speak English (162) compared to German (9). It is possible that a more proportional representation of participants speaking these two languages would yield more reliable results.

To explain such findings, one could refer to the conclusion by Kreiman and Sidtis (2011), which suggests that, in addition to knowledge of the spoken language, other characteristics of voice and speech, such as accent and speaking style, may also help listeners in recognizing speakers. Winters et al. (2008) examined the impact of knowledge of the speaker's language on the perception of the characteristics of their speech by listeners, testing listeners' discrimination and identification of speech from bilingual speakers of German and English. The results showed that listeners are able to generalize their knowledge of a speaker's speech in the context of these two phonologically similar languages. The authors concluded that it cannot be assumed that the findings would be the same for languages that have less or even no phonological similarities. Lavan (2020) assumes that listeners are more likely to recognize a speaker more accurately when the speaker is speaking their native language, as opposed to a language they acquired later in life.

A study conducted by Wester (2012) aimed to determine how listeners assess the similarity of voice and speech when there is and when there is not a language barrier between them and the speaker. The results of the study suggested that it is significantly easier for listeners to recognize speech in their own language than in foreign languages. Although listeners were able to recognize the voice and speech of speakers who spoke in their native language and in a phonologically similar foreign language, as well as extend that ability to foreign languages that are not phonologically similar, Wester (2012) concluded that the certainty of identifying the speaker decreases when recognizing speech produced in foreign languages.

Given that the results obtained from our research lead to the conclusion that neither familiarity with the speaker nor the language spoken are reliable, decisive factors for speaker identification, it is likely that listeners rely more on other vocal characteristics of the speaker during recognition. Identifying these characteristics would contribute to a better understanding of the speaker recognition process. This assumption is supported by the findings of research by Baumann & Belin (2010), which shows that listeners depend on low-level acoustic characteristics, such as the fundamental frequency of the voice, which represents the number of vocal fold vibration cycles per second, or voice quality, in the process of identifying an unknown speaker. Given the fact that the voice and speech of the speakers were not characterized by specific features that would contribute to their recognition, the findings of our research could be partially explained by the influence of the fundamental frequency of the speaker's voice on recognition, as the reliability of speaker recognition decreased with an increase in the fundamental frequency of their voice. This is supported by the observed difference in familiarity with and recognition of English speaker, whereas this difference was absent in the case of Serbian and Farsi speakers. Specifically, the fundamental frequency of the voice of Serbian and Farsi speakers is lower, while the fundamental frequency of the voice of English and German speakers is higher. Given that the "depth" voice is perceived auditorily in relation to the its fundamental frequency, the findings of this study suggest that, female voices with deeper pitch tend to be more easily recognized than higher-pitched voices.

The fact that the research sample of participants was a convenience, consisting solely of female participants, can be considered a limitation of the study. Given that, in the environment where the research was conducted, language, voice, and speech are primarily professionally engaged with by women, it was not possible to achieve a comparable representation of both sexes. Although a sample of participants including both genders would certainly be more interesting and research-desirable, it is unlikely that it would have a significant impact on the study's results, given the fact that available scientific sources do not provide evidence of an advantage of one gender over the other in terms of auditory speaker recognition. Additionally, a limitation of the study is the small number of participants in some of the groups when comparing the recognition of different speakers, as well as the distinction between familiarity with and recognition of speakers. A larger sample of participants size would lead to more reliable conclusions regarding the observed differences or the absence of significant differences.

CONCLUSION

The results obtained in this study do not support the sustainability of the hypothesis regarding the connection between prior familiarity with the speaker and their language with the success of speaker recognition. These results are significant for the theory and practice of forensic voice and speech analysis, as they indicate that auditory speaker recognition does not require experts to share the same linguistic code as the individuals whose speech is being recognized. Additionally, the results show that prior familiarity with the speaker's voice, which is often the case when analysing the voices of public figures, does not significantly influence speaker recognition. Given the importance of forensic analysis within the judicial process, as well as the specificity of the profession and the limited number of experts in this field, it is crucial to identify all factors that could potentially affect the reliability of forensic analysis. The results obtained indicating that language and speaker familiarity are not significant factors in speaker recognition have both theoretical and practical implications for forensic phonetics, as these factors have traditionally been considered critical exclusion criteria when selecting experts for specific cases. Given that the reliability of automatic identification methods, which are not influenced by these factors, is still not acceptable for legal purposes, the recognition that language and speaker familiarity have little to no impact on speaker recognition could lead to a reassessment of the constraints traditionally imposed on forensic experts to enhance the objectivity of their findings. The fact that the examined factors in this study unexpectedly did not prove to be significant for auditory speaker recognition is important as it shifts research attention toward factors related to the speaker's speech production, such as individual voice and speech characteristics. To identify the voice and speech characteristics that are key to speaker recognition, future research should focus on examining the inherent acoustic properties of voice and speech, of which the fundamental frequency of the speaker's voice is undoubtedly the most important.

ACKNOWLEDGEMENTS

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. This research was supported by the Ministry of Science, Technological Development and Innovation [No 451-03-137/2025-03/200096], Republic of Serbia.

REFERENCES

Alkhatib, B., & Kamal Eddin, M. M. W. (2020). Voice identification using MFCC and vector quantization. *Baghdad Science Journal*, *17*(3), 1019–1028. https://doi.org/10.21123/bsj.2020.17.3(Suppl.).1019

Arjamand, M., Saleem, A., Basit, A., Iftikhar, S., Sharif, M., Cholistani, M. S., Farhan, M., Shumail, S., Khan, B. A., Ali, Z., Shahid, B., & Hasnain, M. (2024). The role of artificial intelligence in forensic science: Transforming investigations through technology. *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, *7*(5), 67–70. https://ijmrap.com/wp-content/uploads/2024/10/IJMRAP-V7N5P52Y24.pdf

Babić, I., Otuzbir, S., & Hodžić, I. (2017). The significance of the forensic phonetic in the voice identification as an effective protection and safety measure. *Nauka i tehnologija*, *5*(9), 157–165.

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychological Research*, *74*(1), 110–120. https://10.1007/s00426-008-0185-z

Carić, M., & Širić, L. (2023). Primjena forenzičke akustike i fonetike u kaznenom postupku s posebnim osvrtom na vještačenje glasovnih zapisa. *Zbornik radova Pravnog fakulteta u Splitu*, *60*(1), 189–217. https://doi.org/10.31141/zrpfs.2023.60.147.189

De Vos, A., Vanvooren, S., Ghesquière, P., & Wouters, J. (2020). Subcortical auditory neural synchronization is deficient in pre-reading children who develop dyslexia. *Developmental Science*, *23*(6), e12945. https://doi.org/10.1111/desc.12945

Didla, G. S. (2020). A review of voice disguise in a forensic phonetic context. *International Journal of English Literature and Social Sciences*, *5*(3), 721–725. https://doi.org/10.22161/ijels.53.25

Durán, J. M., van der Vloed, D., Ruifrok, A., & Ypma, R. J. F. (2024). From understanding to justifying: Computational reliabilism for AI-based forensic evidence evaluation. *Forensic Science International: Synergy*, *9*, 100554. https://doi.org/10.1016/j.fsisyn.2024.100554

Goggin, J., Thompson, C., Strube, G., & Simental, L. (1991). The role of language familiarity in voice identification. *Memory and Cognition*, *19*(5), 448–458.

Hansen, J. H. L., & Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine*, *32*(6), 74–99. https://doi.org /10.1109/MSP.2015.2462851

Islam, R., Abdel-Raheem, E., & Tarique, M. (2022). A novel pathological voice identification technique through simulated cochlear implant processing systems. *Applied Sciences*, *12*(5), 2398. https://doi.org/10.3390/app12052398

Jain, P., Chinmayee, P., Kaur, K., Chaudhary, S., Kaur, K., & Karunya, S. (2024). Advancements in forensic voice analysis: Legal frameworks and technology integration. *Asian Journal of Advances in Research*, *7*(1), 369–384. https://jasianresearch.com/index.php/AJOAIR/article/view/464

Jenkins, R. E., Tsermentseli, S., Monks, C. P, Robertson, D. J., Stevenage, S. V., Symons, A. E., & Davis, J. P. (2021). Are super-face-recognisers also super-voice recognisers? Evidence from cross-modal identification tasks. *Applied Cognitive Psychology*, *35*(3), 590–605. https://doi.org/10.1002/acp.3813

Kanber, E., Lavan, N., & McGettigan, C. (2022). Highly accurate and robust identity perception from personally familiar voices. *Journal of Experimental Psychology: General*, *151*(4), 897–911. https://doi.org/10.1037/xge0001112

Kreiman, J., & Papcun, G. (1991). Comparing discrimination and recognition of unfamiliar voices. *Speech Communication*, *10*(3), 265–275. https://doi.org/10.1016/0167-6393(91)90016-M

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons.

Krizman, J., Bonacina, S., & Kraus, N. (2019). Sex differences in subcortical auditory processing emerge across development. *Hearing Research*, *380*, 166–174. https://doi.org/10.1016/j.heares.2019.06.002

Krizman, J., Bonacina, S., & Kraus, N. (2020). Sex differences in subcortical auditory processing only partially explain higher prevalence of language disorders in males. *Hearing Research*, 398, 108075. https://doi.org/10.1016/j.heares.2020.108075

Krizman, J., Rotondo, E. K., Nicol, T., Kraus, N., & Bieszczad, K. M. (2021). Sex differences in auditory processing vary across estrous cycle. *Scientific Reports*, *11*, 22898. https://doi.org/10.1038/s41598-021-02272-5

Lausen, A., & Schacht, A. (2018). Gender differences in the recognition of vocal emotions. *Frontiers in Psychology*, *9*, 1–22. https://doi.org/10.3389/fpsyg.2018.00882

Lavan, N., Merriman, S. E., Ladwa, P., Burston, L. F., Knight, S., & McGettigan, C. (2020). 'Please sort these voice recordings into 2 identities': Effects of task instructions on performance in voice sorting studies. *British Journal of Psychology*, *111*(3), 556–569. https://doi.org/10.1111/bjop.12416

Lindh, J. (2017). *Forensic comparison of voices, speech and speakers: Tools and methods in forensic phonetics.* [PhD thesis, University of Gothenburg, Department of Philosophy, Linguistics and Theory of Science]. http://hdl.handle.net/2077/52188

Morrison, G., & Enzinger, E. (2019). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Introduction. *Speech Communication*, *85*, 119–126. https://doi.org/10.1016/j.specom.2019.06.007

Morrison, G. S., Enzinger, E., Ramos, D., González-Rodríguez, J., & Lozano-Díez, A. (2020). Statistical models in forensic voice comparison. In D. L. Banks, K. Kafadar, D. H. Kaye, & M. Tackett (Eds.), *Handbook of forensic statistics* (pp. 451–497). CRC Press.

Mukattash, B. (2016). The role of forensic phonetics in legal investigation: A case study of two speaker-identified/unidentified recorded samples. *Journal of Literature, Languages and Linguistics*, *29*, 31–37.

Njie, S., Lavan, N., & McGettigan, C. (2023). Talker and accent familiarity yield advantages for voice identity perception: A voice sorting study. *Memory & Cognition, 51*(2), 175–187. https://doi.org/10.3758/s13421-022-01296-0

Nygaard, L. (2005). *Perceptual integration of linguistic and nonlinguistic properties of speech*. In D. B. Pisoni, & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 390–413). Blackwell Publishing.

Perrachione, T., & Wong, P. (2007). Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. *Neuropsychologia*, *45*(8), 1899–1910. http://doi.org/ 10.1016/j.neuropsychologia.2006.11.015

Perrachione, T. K., Del Tufo, S. N., & Gabrieli, J. D. (2011). Human voice recognition depends on language ability. *Science*, *333*(6042), 595–595. https://doi.org/10.1126/science.1205993

Perrachione, T. K. (2017). Speaker recognition across languages. In S. Frühholz, & P. Belin (Eds.), *The Oxford handbook of voice perception* (pp. 1–17). Oxford University Press.

Petrini, K., & Tagliapietra, S. (2008). Cognitive maturation and the use of pitch and rate information in making similarity judgments of a single talker. *Journal of Speech, Language, and Hearing Research*, *51*(2), 485–501. https://doi.org/10.1044/1092-4388(2008/035)

Philippon, A. C., Cherryman, J., Bull, R., & Vrij, A. (2007). Earwitness identification performance: The effect of language, target, deliberate strategies and indirect measures. *Applied Cognitive Psychology*, *21*(4), 539–550. https://doi.org/10.1002/acp.1296

Rana, S., & Qureshi, M. A. (2024). A comprehensive review of forensic phonetics techniques. *The Asian Bulletin of Big Data Management*, *4*(2), 284–301. https://doi.org/10.62019/abbdm.v4i02.159

Rezić, A., & Bonett, A. (2021). Percepcija emocija putem vizualnog i auditivnog kanala. *Logopedija*, *11*(2), 50–60. https://doi.org/10.31299/log.11.2.3

Sharma, P., & Sahu, N. (2018). A review and analysis of voice identification system. *International Journal of Innovative Knowledge Concepts*, *6*(5), 189–191. https://doi.org/11.25835/IJIK-54

Šešum, M. (2013). Komparativna analiza formantnih struktura glasova sestara i glasova monozigotnih bliznakinja. *Beogradska defektološka škola*, *19*(3), 515–527.

Šešum, M. (2021). Forenzička fonetika-identifikacija govornika. In S. Knežević (Ed.), *Forenzičko računovodstvo, istražne radnje, ljudski faktor i primenjeni alati* (pp. 828–859). Fakultet organizacionih nauka Univerziteta u Beogradu.

Šešum, M., & Kovačević, J. (2015). Forenzička fonetika. In *Vodič za primenu Zakonika o krivičnom postupku Republike Srbije* (pp. 90–105). Projekat "Implementacija novog Zakonika o krivičnom postupku Republike Srbije".

Šešum, M., & Petrović, E. (2024). The complexity of speaker identification based on voice and speech and its application in forensics. In S. H. Fazeli (Ed.), *The ninth International Conference on languages, linguistics, translation and literature – full articles* (Vol. 1) (pp. 90–108). Ahwaz Publication of Research and Sciences.

Van Lancker, D., & Kreiman, J., (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia*, *25*(5), 829–834. https://doi.org/10.1016/0028-3932(87)90120-5

Wester, M. (2012). Talker discrimination across languages. *Speech Communication*, *54*(6), 781–790. https://doi.org/10.1016/j.specom.2012.01.006

Winters, S., Levi, S., & Pisoni, D. (2008). Identification and discrimination of bilingual talkers across languages. *The Journal of the Acoustical Society of America*, *123*(6), 4524–4538. https://doi.org/10.1121/1.2913046

Zarate, J. M., Tian, X., Woods, K. J., & Poeppel, D. (2015). Multiple levels of linguistic and paralinguistic features contribute to voice recognition. *Scientific Reports*, *5*, 11475. https://doi.org/10.1038/srep11475

Zhou, Y., Liu, Y., & Niu, H. (2022). Perceptual characteristics of voice identification in noisy environments. *Applied Sciences*, *12*(23), 12129. https:// doi.org/10.3390/app122312129