UDC 81'342.1:343.143
001.891.5:81'342.1
Kristina Tomić
University of Niš
Faculty of Philosophy

THE EFFECT OF LINGUISTIC CONTEXT ON SPEAKER RECOGNITION BY EARWITNESSES IN VOICE LINE-UPS

Abstract Construction of voice line-ups is a procedure performed by forensic phoneticians in earwitness testimonies that involves setting up a listening experiment for the purposes of speaker recognition by earwitnesses (Hollien 1990; 2002; 2012). Voice memory depends on a number of variables, including the lapse of time between the incident and the testimony, the length of utterances in a line-up, the number of 'foil' samples, the emotional state of the listeners, and many others (Hollien 2002; 2012). The goal of the current research is to determine whether, all things being equal, linguistic context, or understanding "what is being said", affects our ability to recognize voices. Thirty-two naïve listeners, native speakers of Serbian, were exposed to two different sets of voice line-ups. In the first set, voice was presented in linguistic context, i.e. the strings of utterances were complete sentences that express some meaningful information in Serbian, while, in the second set, the recording was modified to contain only portions of speech without complete words and sentences, thus voice was presented out of linguistic context. The recognition rate of the second experiment was significantly lower than that of the first (12.5% vs. 40.63%). The results indicate that listeners are more likely to remember voice, or at least more willing to perform the recognition task, if they are able to understand the meaning of what is being said.

Key words speaker recognition, naïve listeners, earwitnesses, voice line-ups, linguistic context

1. Introduction and Aims

Construction of voice line-ups or voice parades is a procedure performed by experts in earwitness testimonies or hearings that involves setting up a listening experiment for the purposes of speaker recognition (Hollien 1990; 2002; 2012). An earwitness is a person who has heard an incident/crime but has not seen the face of the perpetrator/offender. When the police believe they have apprehended the offender, they rely on earwitnesses for identification, which is performed by listening to the suspect's voice. The voice is supposed to be presented in a group of other voices—foils—and the witness is asked to identify which voice among the group belongs to the offender (Hollien 1990; 2002; 2012). As opposed to speaker identification by experts, who receive extensive training in listening throughout their career, earwitnesses are untrained individuals known as naïve listeners. Therefore, to obtain the best possible results, the voice line-up must be administered in a rather rigorous and structured manner.

So far, scientists have identified numerous linguistic and extralinguistic features that we rely on when recognizing someone's voice, including the pronunciation of individual segments, prosody, voice quality, and others (see Hollien 2002: 60–62). However, there has not been a scientific consensus on whether the content or "what is being said" influences voice recognition and to what extent (compare Read & Craik 1995; Yarmey 2001; Eriksson 2007). Furthermore, few studies have dealt with whether we are able to process and remember voices if we are exposed to speech signal that does not convey any meaningful information. The goal of the current research is to explore how the linguistic context or ability to understand what is being said affects speaker recognition by naïve listeners.

2. Theoretical Background

2.1. On voice line-ups

According to the Serbian Criminal Procedure Code, speaker recognition is supposed to be performed in accordance with face recognition (Sl. glasnik RS, br. 72/2011, 101/2011, 121/2012, 32/2013, 45/2013, 55/2014 i 35/2019, Article 90)), but there are very few regulations with regard to how the recognition task should be administered, which has caused a lot of controversy in Serbian legal circles. Speaker recognition by earwitnesses is often considered unreliable as it is "impossible to perform it without the suspect's consent, it is challenging to previously describe a person's voice, and it is difficult to find foils with similar voices" (Tintor 2015: 10). All of the above are reasons why, in Serbia, there is little to no court practice using voice recognition as evidence. However, voice recognition by earwitnesses is not a new notion in western countries and there is already substantial research on voice perception and recognition by naïve listeners. Therefore, provided that the procedure is conducted in an appropriate manner and with enough expertise, it may yield satisfactory results.

Comparing voice recognition to face recognition and performing it correspondingly is not considered good practice because humans process and remember voices differently compared to how they process and remember faces (Hollien 2002: 93; 2012: 3). There are numerous variables affecting voice perception and voice memory—one of the most prominent being whether we are dealing with familiar or unfamiliar voices, as these invoke different processes in our brain. Namely, when we hear a familiar voice, our brain performs a pattern recognition task, which is a discrimination process, while the perception of unfamiliar voices is a process which initiates feature analysis (Hollien 2002: 95).

Test structure and administration play a significant role in the result. For instance, it has been proven that longer foil samples improve the chances of successful recognition because listeners are exposed to a larger phonemic repertoire of the given speaker (Hollien 1990; 2012; Yarmey 1995). In addition, all of the samples must be of the same quality because poor audio quality reduces recognition accuracy (Nolan, McDougall, & Hudson 2008). The number and choice of foils is a crucial step in a voice line-up. It is usually recommended to have between five and eight foils that are neither too different nor too similar to each other (Hollien 1990; 2002; 2012;

Broeders 1996). If only one of the foils stands out, witnesses will be biased to choose it as the offender. Likewise, if all of the voices are too similar sounding, the witness will not be able to discern the offender and will thus choose randomly (Hollien 2002: 95).

Witness expectations have proven vital for successful voice recognition, which is why witnesses must be informed that the offender's voice may or may not be present in the line-up (Hollien 2002; 2012). In addition, witnesses must not be coerced to perform the recognition if they are not confident enough that they can recognize the criminal. Therefore, in both the actual test and the experiment, listeners should be allowed to opt out of the recognition (Broeders & van Amelsvoort 1999; Hollien 2002).

Accuracy of voice recognition also depends on numerous factors that experts are unable to control in an actual forensic case. The emotional state of the listener—as well as individual characteristics and talent—seems to play a significant role (Künzel 1994; de Jong 1998; Hollien 2002; 2012). Furthermore, in comparison to "face memory", "voice memory" is under a much greater influence of time lapse between the incident and the recognition task (Künzel 1994; Hollien 2002; 2012). Age has also proven to be an important factor, as listeners tend to be better at recognizing voices of speakers their own age range (Huntley & Pass 1995; Hollien 2002; 2012). In addition, elderly people do not perform as well in recognition tasks as younger and middle-aged people (Hollien 2002; 2012). Finally, there has been some debate whether male or female listeners perform better in voice recognition tasks; however, most of the research that has dealt with this topic has not confirmed any significant differences between the sexes (Yarmey & Matthys 1992; Atkinson 2015).

2.2. On linguistic context in speaker recognition

It has largely been agreed that there is no single definition of context and that it depends on what is being studied and how (Goodwin & Duranti 1992: 2). In this paper, context is observed as "a frame that surrounds the event being examined and provides resources for its appropriate interpretation" (Goffman 1974 as cited in Goodwin & Duranti 1992: 3). Thus, the notion of context involves two entities: a focal event, which is human voice in our case, and a field of action within which the event is embedded; that is, the text being spoken. If we rely on figure-ground theory from Gestalt Psychology that is often used in context analysis (see Talmy 1975; Wallace 1982; Hanks 1992; Kendond 1992), we may say that the acoustic signal of voice is the figure, while linguistic context is the ground. This paper's use of the term "linguistic context" is primarily used to refer to the semantic context on a discourse level, or "what is being said".

When listeners perform the task of speaker recognition, they actually perform a feature analysis of the speaker's voice. Some of these features have been identified in previous studies and these include: heard pitch, which may refer to fundamental frequency, intonation variability, phonation time ratio, and so on; articulation of individual consonants and vowels; voice quality, or the combination of laryngeal and pharyngeal features of our vocal tract that make our voices recognizable and unique; prosody or temporal parameters of speech; intensity or how loudly the speaker talks; and other speech characteristics such as dialect, unorthodox usage of stress/accent and various idiosyncratic features (Hollien 2002).

There are a number of studies testing whether and how content affects speaker recognition, but the results are often inconsistent. For instance, Read and Craik (1995) claim that the accuracy of speaker recognition increases in line with the similarity of the content. Similarly, Eriksson (2007) found that if the content of the voice line-up recordings differs from the training (offender) recording, certain age groups are likely to fail at the recognition task. On the other hand, Yarmey (2001) did not find any correlation between recognition accuracy and similarity of the speech content provided that the training (offender) recording contained longer speech material (Yarmey 2001).

There are not many recent studies exploring how speech presented out of linguistic context influences recognition rates. A study by Bricker and Pruzansky (1966) proved that stimuli consisting of random phonemes lacking in semantic meaning could result in higher recognition rates. They confirmed that listeners were able to recognize speakers with 87% accuracy when exposed to disyllabic nonsense words, which was only slightly lower compared to the experiment in which listeners were exposed to complete sentences.

The goal of this research is to determine whether speaker recognition is possible if the acoustic signal of voice is presented out of linguistic context. The results will help us understand whether speaker recognition by naïve listeners depends solely on the acoustic features of voice or whether linguistic information embedded in speech signal provides necessary ground for voice perception.

3. Methodology and Procedure

The current research is structured in accordance with the guidelines and recommendations for the voice line-up procedure in actual forensic cases (Broeders & van Amelsvoort 1999; de Jong-Lendle, Nolan, McDougall, & Hudson 2015) and it consists of two parts: recognition experiments with voice line-ups, and pre-test listening tasks for the selection of foils. Both tests were distributed online via a custom-made software solution developed using HTML, PHP, and JavaScript. Each entry made by the participants underwent an automatic server-side validation and was subsequently stored using the InnoDB/MySQL database framework.[1]

3.1. Pre-test – selection of foils

Thirty-five female native speakers of Serbian from the Prizren-Timok dialect area were recorded in spontaneous speech across two different tasks (Task 1 - expressing their opinion on the given topic and Task 2 - giving directions on a map)[2]. To ensure enough voice similarity between foils, we performed an acoustic analysis and created a subset of similar-sounding voices. To extract long-term acoustic features more easily, we modified the recordings in such a way to contain only the

[1] Acknowledgment: The software was created by Nikola Janković from NaissusWorks http://www.naissusworks.com/.

[2] The database is part of the corpus that is going to be used for the author's PhD thesis on forensic speaker comparison across languages.

vocalic portions of speech with a visible formant structure.[3] Using Praat (Boersma & Weenink 2018), we measured the following parameters: long-term fundamental frequency (LTF0), intensity, long-term formant frequencies of the third and fourth formant (LTF3 and LTF4)[4], and calculated the mean and standard deviation of the group. Any speakers whose values did not fall within one standard deviation of the group mean were dismissed as "too different", which left us with twelve speakers (Subset 1). Ten speakers out of the twelve were randomly chosen for the pre-test listening tasks, forming Subset 2.

The pre-test listener group was comprised of three male and three female naïve listeners without known hearing impairments. Their average age was 25.5 (SD 4.61) and they were all native speakers of Serbian coming from the same regional background as the speakers in the recordings.

In the first listening task, the listeners were presented with a 30-second-long voice sample from Speaking Task 1. They were asked to rate on a scale from 1 to 9 how likely it is for each of the speakers: (1) to be a singer, (2) to remember her voice, (3) to recognize her if he/she met her in person, and (4) to work as a radio presenter. High ratings of 8 or 9 on questions (2) and (3) would mean that the given speaker has a rather memorable voice and she would not be used in an actual voice line-up, while questions (1) and (4) acted as distractors.[5]

In the second listening task, the listeners were presented with six groups of recordings that each contained five 10-second-long voice samples from Speaking Task 2. They were instructed to listen to the recordings as many times as they needed and mark "the most different one" for each group. Each of the speakers was present in three of the six groups. If a certain speaker was constantly selected by several listeners, she would be considered "too different" to use in an actual voice line-up.

In the first run of the pre-test, two speakers were removed after being constantly singled out as "the most different" ones. Even though their voice quality was similar to that of the rest of the speakers, one of them had a distinct pronunciation of certain consonants and consonant clusters which made her "stand out" from the group, while the other was remarkable for talking rather slowly. Two of the remaining speakers from Subset 1 were added instead, forming Subset 3. The procedure was repeated after five days and the same test listeners gave more satisfactory results. Two speakers from Subset 3 were randomly chosen to act as "offenders" for the two different recognition tasks.

[3] For the exact procedure of audio modification, see Tomić and French (2019).

[4] Previous research has shown that the most important features for speaker identification are the fundamental frequency, the third and fourth formant, and the closing phase of the glottal wave (Lavner, Rosenhouse, & Gath 2001). Considering that extraction of the closing phase of the glottal wave would be time consuming to perform for 35 speakers, we instead opted for extraction of long-term vocal features. In addition, it is already known from previous studies that the third and fourth formant are indicative of individual voice quality and that their values vary significantly across speakers (Moos 2010; Gold, French & Harrison 2013; Gold 2014; Tomić & French 2019). We also decided to measure intensity, because a difference in the loudness of speech may affect voice perception and recognition (Hollien 2002).

[5] The experiment was structured after de Jong-Lendle et al. (2015). In an actual forensic case, the questions would be *How likely is it for the given speaker to be a criminal/to be a robber/to commit a violent assult, etc?*, however, since we are not dealing with actual case recordings and criminals, the questions were replaced with something more appropriate for the current research.

3.2. Recognition experiments

In the current research, we opted to follow the sequential approach to voice line-ups. The listeners were able to hear any of the foils as many times as they wanted, in any order, but only after having heard each of the foils at least once (Majewski & Basztura 1996). This approach is often characterized as more effective than the simultaneous approach, in which the recordings are presented on the same tape one after another and heard only once or several times but in a different order (Hollien 2002).

Listeners for the two recognition experiments were 32 native speakers of Serbian (16 male and 16 female), aged between 21 and 57 (mean 31.16, SD 9.75). The great majority of the listeners (27) were from the same geographic and dialect region as the speakers in the recordings, three were from the Kosovo-Resava dialect region, and the remaining two were from the Šumadija-Vojvodina dialect region. Regarding the listeners' education levels, 22 were university graduates and ten were high school graduates. None of them reported any hearing impairments. Three listeners disclosed having played a musical instrument and two stated they could sing well. The question about whether people describe you as musically talented yielded only one affirmative answer, thirteen 'sometimes' and eighteen 'no'. Three of the listeners marked their musical talent with a 4 or 5 on a 5-point Likert scale, and seven listeners described their current emotional state as highly stressed/upset.

In the first recognition experiment, the participants were presented with a 40-second-long recording of Person X talking about a city that left a huge impression on her. The recording was edited to exclude any toponyms or names. The software allowed the participants to listen to the recording only once, after which they were required to complete a short survey stating whether they recognized the speaker's voice and to briefly say what the speaker was talking about, in order to check the listener's comprehension of the content. In the recognition task, the listeners were presented with seven 30-second-long recordings. Six acted as foils and one was the offender sample (Subset 4). Two of the foils were talking about visiting a place, while the "offender" and the rest of the foils were expressing opinions on different topics such as housing, television, and games. The listeners were asked to recognize Person X from the listening task while given an option to choose any of the speakers, none of the speakers, or not to answer the question (since witnesses must not be coerced into giving an answer). They were asked to express their certainty on a 5-point Likert scale and give a brief explanation for their choice, if they wished.

The structure of the second recognition task was the same as that of the first one. Person Y was chosen from Subset 3, but was not part of Subset 4, that is, she did not act as a foil in Recognition Experiment 1. Since we aimed to explore the perception of unfamiliar voices, it was essential that the listeners had not heard Person Y's voice prior to Listening Experiment 2. The recording was modified to exclude some vowel-consonant clusters in such a way to obscure the content but to leave enough phonemes and speech for recognition to be possible. Recognition Task 2 contained seven 30-second-long recordings from Speaking Task 2 (Subset 5), where speakers gave directions on a map. Six recordings acted as foils, while one was the "offender". The foils in this part of the experiment did not include the offender from Recognition Task 1 and two of the foils were changed.

4. Results and Discussion

4.1. Pre-test results

In order to select the speakers with similar voices from the recorded corpus, we performed an acoustic analysis of long-term vocal features. The results are presented in the table below.

Long-term parameter	Mean	SD	CV	Subset 1 Mean	Subset 1 SD	Subset 1 CV
F0 (Hz)	214.3	20.81	9.71	214.19	8.62	4.02
Intensity (dB)	74.93 dB	3.07	4.09	74.76	1.98	2.65
F3 (Hz)	2,754.17	196.36	7.13	2,727.67	102.13	3.74
F4 (Hz)	3,845.63	122.91	3.2	3,808.5	69.96	1.84

Table 1. Acoustic analysis of vocal features of the recorded speakers; mean values; standard deviation (SD); and coefficient of variation (CV)

Subset 1 was formed after the speakers who did not fall within one standard deviation of the entire group were removed. As observed in Table 1, removing the outliers greatly reduced the between-speaker variability of the group. Significantly lower values of the coefficient of variation in Subset 1 indicate that acoustic characteristics of speakers' voices are similar.

Due to limited space, we will provide only the results of the second run of the pre-test listening tasks below. The first listening task of the pre-test involved grading the speakers on 9-point Likert scales across four questions. Since the first and fourth questions acted as distractors, the average score was calculated only for the second and third ones. The highest score in Table 2 is 6.92 for speaker 7, followed by the score of 6 for speaker 4 and the score of 5.75 for speakers 3 and 9.

Speaker	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Score	5.33	4.08	5.75	6	4.75	4.33	6.92	5.67	5.75	4.67

Table 2. Subset 3: average 9-point Likert scale scores showing how memorable speakers' voices are

The second listening task of the pre-test involved choosing the most different voice in a sequence of five voices. Each voice appeared in three out of six sequences, resulting in a total of 18 appearances, as there were six test listeners. We calculated how often each voice was chosen with regard to its appearance.

Speaker	Score (%)
S1	16.67
S2	5.56
S3	22.22
S4	44.44
S5	33.33

S6	11.11
S7	27.78
S8	27.78
S9	11.11
S10	0%

Table 3. Subset 3: the score of how often the given speaker was chosen to be "the most different one" in the sequence.

Looking at Table 3, we may note that speaker 4 has the highest score of 44.44%, which means that this speaker was selected as being the most different one 44.44% of the time (8 out of 18 appearances). Since the score is not above 50%, and this speaker did not score above 6 in the pre-test Listening Task 1, she was considered a suitable foil. Similarly, speaker 7, who has the highest score in the pre-test Listening Task 1, scored rather low in the second listening task (27.78%); therefore the results of the pre-test were considered satisfactory to proceed with the voice line-ups.

4.2. Recognition experiments

In Table 4 below, we can observe the results of the two recognition tasks with average certainty scores, which were marked on a 5-point Likert scale. In Experiment 1, the recognition rate was 40.63%, with 13 correct recognitions, seven by male and six by female listeners, while in Experiment 2, the recognition rate was as low as 12.5%, with only 4 correct recognitions by one female and three male listeners. The score of the z-test for the two population proportions indicates that there is a significant difference in the recognition rates of the two experiments ($z = 2.5472$, $p = 0.0178$). The same statistical test confirms that there is no difference in the performance of male and female listeners in either of the experiments, as well as that the level of education does not play a significant role. The results are in accordance with previous studies which dealt with voice recognition abilities of male and female listeners (see Yarmey & Matthys 1992; Atkinson 2015).

Only one of the listeners who successfully recognized the "offender" in Experiment 1 evaluated his musical talent with a high score (5 out of 5) and one of the listeners reported a high stress level (4 out of 5). In the second recognition task, all of the listeners who performed successful recognition reported low stress levels (1 or 2 out of 5). The average age of the listeners with correct recognition in Experiment 1 was 27 (SD 2.88), while, in Experiment 2, it was 26.75 (SD 3.7). None of them were older than 30. These results confirm that younger listeners are better at speaker recognition than older ones, as well as that it is easier to recognize voices of speakers closer to one's own age range (see Huntley & Pass 1995; Hollien 2002; 2012).

Exp.1 foils	No	Certainty	Exp.2 foils	No	Certainty
A	2	3	H	2	2.5
B	2	3.5	I	-	-
C	-	-	J	8	3.25
D (offend.)	13	3.69	K	2	2

E	5	3.4	L (offend.)	4	2.75
F	6	3.83	M	5	1.8
G	1	4	N	1	2
None	3	2.67	None	2	3
Skip	-	-	Skip	8	-

Table 5. Results of the two recognition experiments with average certainty marked on a 5-point Likert scale

All of the listeners were able to comprehend the meaning of what Person X was saying. The answers were considered satisfactory if they mentioned a city, a place, a location, etc.[6] In the second recognition experiment, four listeners reported that they were able to partly understand the recording, however, only one of them correctly described what Person Y was talking about. This listener, however, did not perform the correct identification in either Experiment 1 or 2.

Apart from the difference in recognition rates, the major difference in the two experiments was in the willingness of the listeners to perform the recognition task ($z = -3.0237$, $p = 00252$). Namely, while all listeners in Experiment 1 did choose an option (including the one that indicated Person X was not present among the foils), as many as eight listeners opted out of the recognition in Experiment 2. A common explanation was that they did not understand or hear the recording well. This confirms that even if there is enough acoustic signal for voice recognition, listeners do not feel they have enough information to perform it. Therefore, the absence of linguistic context significantly hinders speaker recognition.

Table 5 below summarizes the features the listeners relied on during the identification, according to their explanations. In Experiment 1, 25 listeners gave an explanation naming one or more features, while in Experiment 2, 14 listeners gave an explanation.

Feature	Exp. 1	Exp. 2
Voice quality	16	9
The way she talks / diction	6	2
Similar voice	5	3
Accentuation	3	
Tempo / pauses	2	
Pronunciation of words	2	
Intonation	2	
Vowels / consonants	1	2
City description	1	

Table 5. Features on which the listeners relied in recognition experiments

As many as 16 listeners in the first experiment and nine in the second named voice quality as one of the features that helped them in recognition (regardless of

[6] The question was: *What was Person X talking about?* Example answers: *about a city she visited, about a place she likes*, *about a tourist location.*

whether the recognition was successful or not). Another common explanation was that the voices were similar or that they recognized the way this person talks or her diction. Other features include accentuation, tempo and pauses, intonation, and pronunciation of individual words or segments. The only explanation that was not concerned with vocal features is by Listener 23, who stated that the person was describing a city, which is what Person X was talking about too. This isolated example may give insight into the rather low recognition rate in Experiment 1.

Namely, Listener 23 is one of the four listeners who correctly identified Person Y in Experiment 2, and the only one in this group who did not successfully identify the offender in Experiment 1. This clearly indicates that despite strong talent to recognize voices, this listener's decision was biased by the content of what the person was saying. Foil F, who talked about a city, was chosen six times as being Person X, with a rather high certainty score (3.83). Another popular choice was foil E, who talked about visiting a place (chosen five times). Therefore, the decision of Listener 23 is a strong indicator of how other listeners may have been biased in the same manner, even though they did not write this in the explanation box.

Finally, if we observe certainty scores, we can confirm that the listeners expressed a significantly lower level of certainty in the second experiment (paired t-test score $t = 3.52565$, $p = 0.00418$). A large number of listeners opted out of the recognition and there were fewer explanations for their choice. All of these are strong cues that it is very difficult for naïve listeners to process and remember speakers' voices if they are presented outside the linguistic context.

5. Conclusion

In the present research, our goal was to determine whether linguistic context affects recognition of voices by naïve listeners. The results confirmed that recognition accuracy rates are significantly higher in cases when the listeners are able to understand the meaning of what the speaker is saying than when voice is presented as nonsense. The recognition accuracy in the latter case is equal to chance. The fact that three out of four listeners who performed successful identification in the latter case also correctly identified the speaker in the former indicates, however, that the correct identification was not a product of random selection of answers, which confirms previous findings that certain people are more talented than others with regard to voice recognition (Hecker 1971; Huntley 1992; Künzel 1994; de Jong 1998). Listener talent is an intrinsic ability to recognize voices, and as such, it presents an uncontrollable factor in actual forensic cases which involve speaker recognition by untrained, naïve listeners (Hollien 2002: 103).

Furthermore, our research demonstrated that not being able to understand what the speaker is saying reduces the listeners' willingness to perform an identification. One of the possible interpretations may be that listeners *feel* that the linguistic information embedded in utterances is necessary for successful speaker recognition and that acoustic signal alone is not enough. That the listeners' instinct is correct is confirmed by the overall poor recognition rate in the latter experiment. However, from the obtained results it is difficult to conclude how much lack of meaning per se hinders recognition rates, since by stripping the utterances of meaning, we stripped

them of other non-acoustic information as well. These include accent, choice of vocabulary, and pronunciation of individual words, which all make an inseparable part of someone's speech and on which some listeners may rely when identifying an unfamiliar voice (Hollien 2002: 62).

The results also showed that recognition rates are affected by content in the sense that content can bias the listener into focusing on "what is being said" instead of the vocal features of the speaker. Therefore, in actual forensic cases, it is of utmost importance to record the foils and the suspect talking on the same topic. If this is impossible to arrange, it is important to warn the listeners that the content may vary throughout the recordings they will hear; however, this is not what they should base their judgements on.

While performing this research, we encountered a number of difficulties, some of which include the limitations of the administration of the experiment. Namely, while an online approach did increase the accessibility of the questionnaire and thus provided a larger number of participants, we were not able to control exactly how each participant completed the experiment. There might have been interruptions during the procedure, which could result in differences in time between when each listener heard the offender and foils, which would affect the results since voice memory does fade as this period of time increases (Künzel 1994; Hollien 2002; 2012). Moreover, there were some multiple attempts to complete the experiment, all of which were removed from the database and only the data provided at the first attempt was considered relevant and taken into account. Furthermore, even though the participants were asked to use headphones or earphones while completing the tasks and not perform the experiment in a loud room, there is no way of confirming that everyone followed the guidelines. In addition, due to the fact that in the first recognition experiment, two foils talked on the same topic as offender X instead of only one foil, it is impossible to determine whether these foils were often selected only because of their linguistic content or whether the similarity of the voices played a role as well. Finally, the listeners were not real witnesses; they did not experience an actual crime or assault and were thus not as motivated to remember the presented voices as they would have been in a real-life scenario. Previous research confirmed that being in a state of panic or arousal significantly increases voice memory, thus when victims hear the offender's voice during a voice-line up, their memories and trauma are triggered by the suspect's voice and they become upset the instant they hear it, which is a clear indication of voice recognition (Hollien 2002; 2012).

The significance of the current research is twofold. Not only is it useful for forensic researchers and experts administering voice line-ups, but also, the results we obtained here provide evidence in favor of the idea that speaker recognition depends on linguistic factors in addition to acoustic features. Observed through the theory of context, if the acoustic signal is taken to be figure, and the linguistic content is ground, then we can say that it is difficult to understand and interpret figure if it is not observed against the ground. In other words, the acoustic signal of voice appears to be inseparable from linguistic information when it comes to voice perception by naïve listeners. Since it is already known that phoneticians and trained experts perform better in speaker recognition tasks than naïve listeners (de Jong 1998), further research may focus on exploring how trained experts perform in cases when the speech is presented outside linguistic context. Such results may help us form more general views on speaker and voice recognition.

References

Atkinson, N. (2015). Variable factors affecting voice identification in forensic contexts. *PhD Thesis*. University of York.

Boersma, P., and Weenink, D. (2018). Praat: Doing Phonetics by Computer (Version 6.0.42) [Computer program}. Available:

http://www.fon.hum.uva.nl/praat/download_win.html [16.08.2018.]

Bricker, P., and Prozansky, S. (1966). Effects of Stimulus Content and Duration on Talker Identification. *The Journal of the Acoustical Society of America, 40*(6), 1441-1449. doi:10.1121/1.1910246

Broeders, A. (1996). Earwitness Identification: Common Ground, Disputed Territory and Unharted Areas. *International Journal of Speech, Language and the Law, 3*, 1-13. http://dx.doi.org/10.1121/1.1910246

Broeders, A., and van Amelsvoort, A. (1999). Lineup Construction for Forensic Earwitness Identification: A Practial Approach. *14th International Congress of Phonetic Sciences, San Francisco, CA*, (pp. 1373-1376).

de Jong, G. (1998). Earwitness Characteristics and Speaker Identification Accuracy. *PhD Thesis*. University of Florida, Gainesville, Fl.

de Jong-Lendle, G., Nolan, F., McDougall, K., and Hudson, T. (2015). Voice Lineups: A Practical Guide. *18th International Congress of Phonetic Sciences. Glasgow, Scotland*, (pp. 10-14).

Eriksson, E. J. (2007). That voice sounds familiar: Factors in speaker recognition. *PhD Thesis*. Umeå University.

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience.* Cambridge, MA, US: Harvard University Press.

Gold, E. A. (2014). Calculating likelihood ratios for forensic speaker comparisons using phonetic and linguistic parameters. *PhD Thesis*. The Unversity of York, Department of Language and Linguistic Science.

Gold, E., French, P., and Harrison, P. (2013). Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. *The Journal of the Acoustical Society of America, 133*(5), 3294. https://doi.org/10.1121/1.4800285

Goodwin, C., and Duranti, A. (1992). Rethinking context: an introduction. In A. Duranti, & C. Goodwin (Eds.), *Rethinking Context: Language as in Interactive Phenomenon* (pp. 1-42). Cambridge: Cambridge University Press.

Hanks, W. H. (1992). The indexical ground of deictic reference. In A. Duranti, & C. Goodwin (Eds.), *Rethinking Context: Language as in Interactive Phenomenon* (pp. 43-76). Cambridge: Cambridge University Press.

Hecker, M. H. (1971). Speaker recognition. An interpretive survey of the literature. *ASHA Monographs, 16*, pp. 1-103.

Hollien, H. (1990). *The Acoustics of Crime: The New Science of Forensic Phonetics.* New York: Springer.

Hollien, H. (2002). *Forensic Voice Identification.* London and New York: Academic Press.

Hollien, H. (2012). On Earwitness Lineups. *Investigative Sciences Journal, 4*(1), 1-17.

Huntley, R. A. (1992). Listener Skill in Voice Identification. *A paper presented at the American Aademy of Forensic Sciences*. New Orleans

Huntley, R., and Pass, K. (1995). Task Influences on Earwitness Reliability. In A. Braun, & J. Köster (Eds.), *Studies in Forensic Phonetics* (pp. 121-131). Trier: Wissenschaftlicher, Verlag, 64.

Kendon, A. (1992). The negotiation of context in face-to-face interaction. In A. Duranti, & C. Goodwin (Eds.), *Rethinking Context: Language as in Interactive Phenomenon.* Cambridge: Cambridge University Press.

Künzel, H. (1994). On the Problem of Speaker Identification by Victims and Witnesses. *International Journal of Speech, Language and the Law, 1*, 45-58. https://doi.org/10.1558/ijsll.v1i1.45

Lavner, Y., Rosenhouse, J., and Gath, I. (2001). The Prototype Model in Speaker Identification by Human Listeners. *International Journal of Speech Technology, 4*, 63-74. https://doi.org/10.1023/A:1009656816383

Majewski, W., and Basztura, C. (1996). Integrated approach to speaker recognition in forensic applications. *International Journal of Speech, Language and the Law, 3*(1), 50-64. https://doi.org/10.1558/ijsll.v3i1.50

Moos, A. (2010). Long-term formant distribution as a measure of speaker characteristics in read and spontaneous speech. *The Phonetician, 101*, 7-24. Available:

http://www.isphs.org/Phonetician/Phonetician_101.pdf [01.08.2018.]

Nolan, F., McDougall, K., and Hudson, T. (2008). Voice Similarity and the Effect of the Telephone: A Study of the Implications for Earwitness Evidence (VoiceSim). *Final report RES-000-22-2582m*. Swindon: ESRC.

Read, D., and Craik, F. I. (1995). Earwitness identification: Some influences on voice recognition. *Journal of Experimental Psychology: Applied, 1*(1), 6-18. http://dx.doi.org/10.1037/1076-898X.1.1.6

Talmy, L. (1975). Figure and Ground in Complex Sentences. *Proceedings of the First Annual Meeting of the Berkeley Linguistics*, (pp. 419-430).

Tintor, J. (2015). *Novi ZKP - Osnovni problemi odbrane.* Retrieved from Advokatska komora Srbije: Available:

https://blog.aks.org.rs/wp-content/uploads/2015/11/osnovni-problemi-odbrane.pdf [01.04.2019.]

Tomić, K., and French, P. (2019). Long-term Formant Frequencies in Cross-language Forensic Voice Comparison under Likelihood Ratio Framework. *A paper Ppesented at The 28th Annual Conference of the International Association for Forensic Phonetics and Acoustics in Istanbul, July 13th to 17th.*

Wallace, S. (1982). *Figure and Ground: The Interrelationships of Linguistic Categories.* Amsterdam: John Benjamins Publishing Company.

Yarmey, A. (2001). Earwitness descriptions and speaker identification. *International Journal of Speech, Language and the Law, 8*, 114-122. http://dx.doi.org/10.1037/1076-8971.1.4.792

Yarmey, A., and Matthys, E. (1992). Voice identification of an abductor. *Applied Cognitive Psychology, 6*, 367-377. https://doi.org/10.1002/acp.2350060502

Yarmey, D. A. (1995). Earwitness speaker identification. *Psychology, Public Policy, and Law, 1*(4), 792-816. http://dx.doi.org/10.1037/1076-8971.1.4.792

Zakon o krvičnom procesu. *Sl. glasnik RS(br. 72/2011, 101/2011, 121/2012, 32/2013, 45/2013, 55/2014 i 35/2019)*.

Kristina Tomić

UTICAJ JEZIČKOG KONTEKSTA NA PREPOZNAVANJE GOVORNIKA PRI SASLUŠANJU SVEDOKA LAIKA

Sažetak

Prepoznavanje govornika je procedura koja se sprovodi pri saslušanju svedoka laika koji su prisustvovali nekom zločinu ili incidentu i čuli glas počinioca ali ga nisu videli. U tom slučaju, stručnjaci, najčešće forenzičari fonetičari, sastavljaju eksperiment slušanja u kome je glas osumnjičenog predstavljen u nizu drugih sličnih glasova pri čemu se od njih očekuje da prepoznaju počinioca. Ustanovljeno je da pamćenje glasova zavisi od brojnih faktora kao što su: vreme koje je proteklo između incidenta i saslušanja, trajanje izričaka u eksperimentu, broj dodatnih glasova, emotivno stanje slušaoca i drugi. Cilj ovog istraživanja je da utvrdi da li jezički kontekst, tj. „ono što se govori", utiče na ljudsku sposobnost prepoznavanja glasova. Trideset dva izvorna govornika srpskog jezika predstavljali su slušaoce u dva eksperimenta prepoznavanja glasa u okviru ovog istraživanja. U prvom eksperimentu, glas je bio predstavljen u jezičkom kontekstu, takoreći, izričci su bili smisaone celine koje imaju neko značenje na srpskom jeziku. U drugom eksperimentu, snimak takozvanog počinioca bio je izmenjen tako da sadrži delove govora ali bez reči i rečenica u celosti. Možemo reći da je u drugom eksperimentu zvučni signal predstavljen van jezičkog konteksta. Rezultat prepoznavanja bio je znatno niži u drugom eksperimentu (12,5%) u odnosu na prvi (40,63%). Takođe, broj ispitanika koji je odbio da izvrši prepoznavanje je znatno veći u drugom eksperimentu (8 u odnosu na 0 u prvom). Ovi rezultati nam pokazuju da je veća verovatnoća da će slušaoci zapamtiti glas, ili se usuditi da ga prepoznaju, ako mogu da razumeju ono o čemu se govori. Ovakvi rezultati pokazuju nam da je percepija glasa neraskidivo vezana za jezičke informacije koje dobijamo slušajući nekoga kako govori i da akustički signal sam po sebi nije dovoljan za percipiranje glasa.

kristinatomic89@hotmail.com