TEKLA ETELKA GRÁCZI[1] – ATTILA FEJES[2] – VALÉRIA KREPSZ[1,3] – ANNA HUSZÁR[1]

[1]Nyelvtudományi Kutatóközpont
[2]Nemzetbiztonsági Szakszolgálat Szakértői Intézet
[3]Humboldt-Universität zu Berlin
graczi.tekla.etelka@nytud.hu, huszar.anna@nytud.hu, fejes.attila@nbsz.gov.hu,
krepsz.valeria@nytud.hu
doi:http://orcid.org/0000-0003-3351-9661, doi:http://orcid.org/0000-0002-2951-1918,
doi:http://orcid.org/0000-0003-4139-5718, doi:http://orcid.org/0000-0002-2099-6285

Tekla Etelka Gráczi – Attila Fejes – Valéria Krepsz – Anna Huszár: Speaker recognition over the course of 10 years and across speech styles
Alkalmazott Nyelvtudomány, Különszám: Alkalmazott nyelvészet és kriminalisztika, 2022, 94–109.
doi:http://dx.doi.org/10.18460/ANY.K.2022.007

Speaker recognition over the course of 10 years and across speech styles

A common task in forensic speaker recognition is to determine whether two speech samples recorded over a long period of time, perhaps years apart, are likely to be the same. The present study addresses one aspect of this problem. We have analyzed the possibility of speaker recognition in recordings made ten years apart.
Speech samples of various speech types were used from 6 female and 14 male speakers. Each speaker was recorded twice, 10 years apart, using a matching protocol. The speech samples were compared using a voice biometrics system (VOCALISE), a reportedly highly efficient software (for close timed recordings). The resulting score data were analyzed using linear mixed models by the means of R.
The results show that the recordings made 10 years apart have influenced the identification results. However, there is a large difference in scores between speakers. Comparing different types of speech samples, we found that the speech task played a more important role than the speech style. This difference is negligible in samples recorded ten years later, while it plays an important role within the same recording sessions.

Keywords: Voice Biometrics, error rates, forensic speaker recognition, speech style, longitudinal speech changes in voice biometrics

1. Introduction

A common task in forensic speaker recognition is to establish the likelihood of identity between two speech samples, one of an unknown person, recorded by the intelligence authorities and the other of a suspect, recorded later during the investigation. It could be years between the two speech samples. One of the longest periods reported is 27 years (French et al., 2006). This phenomenon calls for an analysis of the impact of longitudinal changes in speech characteristics on the effectiveness of speaker recognition.

The present paper seeks to answer the questions (i) to what extent the results of speaker recognition differ when comparing speech samples from speakers ten years apart with the results of comparing recordings from the same session, and (ii) whether speech style (in the present paper: reading, semi spontaneous speech,

spontaneous speech) plays a role in these results. Mapping speech changes and variability across different time intervals and analyzing their impact on speaker recognition results can help in growing forensic work, where the recognition task involves larger or even unknown time periods.

1.1. Biometric speaker recognition and its challenges

Biometrics is the identification of a person based on the unique, behavioural, or biological characteristics of a person.

Voice Biometrics uses speech for identification, which is suitable for this purpose due to the speaker-specific features of speech resulting from the uniqueness of the vocal tract (Anil et al., 2011). The first step is to extract the features. These features serve as data to the system to build a statistical model, separately for the two speech samples in question. The models of the two speech samples are compared, resulting in a score. In the speaker identification process, these score values are interpreted in the software's Likelihood Ratio (LR) framework as a final step. LR is the quotient of two probability values from the Bayesian analysis: The numerator is the probability that the two speech samples originate from the same person. The denominator is the probability that the two speech samples originate from two different speakers (Drygajlo–Haraksim, 2017). The resulting LR value shows how much more likely it is that the speaker of the two speech samples is the same rather than two different people. The LR approach is useful in forensics, especially when comparing individual characteristics, such as in speaker recognition (Drygajlo–Haraksim, 2017), as it is never possible to tell with absolute certainty if the two speakers are identical or not. Although speech includes largely speaker-specific characteristics, it varies and changes from occasion to occasion and over time.

1.2. Within speaker changes with aging

To say that speech characteristics change with aging is a cliché, but the way they change is not. Numerous studies have been published about changes in speech in samples recorded at long intervals apart. Far fewer studies have addressed changes in adult speakers over a short period of time.

Aging starts around the age of 30: Some speech organs start these changes in their mid-twenties, others later. Of course, the speed of development also varies between organs. These changes are also largely influenced by the genetics and individual variability of the subject. Not only the chronological age itself (primary aging) but also the subject's environment (e.g., air quality, secondary smoking), health history, and personal habits (e.g., abuse such as smoking, diet) (altogether: secondary aging) play a role in the between-subject variability of aging (Belsky et al., 2015; Busse, 2002). This alteration as well as sociolinguistic factors may induce speech-related changes, even in a short time interval (Schreier, 2021).

Cross-sectional studies have found that the fundamental frequency declined at a small rate along the age gradient between 20 and 50 years (Schötz, 2006; Stathopoulos et al., 2011) or between 30 and 60 years (Cox–Selent, 2015) for men, and the tendencies were similar or even more pronounced for women (Eichhorn et al., 2018; Stathopolus et al., 2011). Similar results were found in the speech of Queen Elisabeth II from a longitudinal aspect (with a more enhanced change between 25 and 45 years and above 70 years (Mwangi et al., 2009; Reuboldt et al., 2010).

Rhodes (2017) studied 8 speakers (6 men and 2 women) from a longitudinal aspect from a forensic phonetic point of view. The recordings were made between the ages of 21 and 49. He reported large interspeaker differences in all measurements, so no systematic change between the pairs of speech samples could be reported for shorter durations. The first formant showed more obvious changes between the 7-year comparisons and revealed a decrease from 21 to 49 years (for all but one vowel), however, this decrease showed large individual differences. F_2 and F_3 showed similar but less general tendencies for the long-term changes but no tendencies for the short-time comparisons. Russel and colleagues (1995) did find an f0-decrease between the ages of 18 and 25 of 6 female speakers, but no changes in the following 12 years.

Previous studies in Hungarian (on various subsets of the corpus also used in this study) have looked at the f0 changes over 10-11 years (Gráczi et al., 2022; Gráczi–Krepsz, 2020; Krepsz–Gráczi, 2018; Markó et al., 2021). The results indicate that the f0 tends to decrease, and that this tendency is clearer for women (especially in their 20s and 30s) than for men. As for male speakers, any change in fundamental frequency showed greater individual differences, with two speakers' f0 increasing by a large ratio.

1.3. Speech style and speech variability

It is well known that a number of prosodic features may differ from each other regarding the different speech styles (cf. e.g., Duchin–Mysak, 1987; Jacewicz et al., 2010). This can be explained by the fact that different speech situations require different speech planning strategies. On the one hand, in spontaneous narratives, speech planning and speech production take place simultaneously; the speaker plans the content of the message and is free to choose from words, as well as syntactical and grammatical structures. On the other hand, some processes of macro-design (the formulation of the message) and micro-design (the linguistic transformation; Levelt, 1989) do not play a role in reading since the linguistic material to be read is given. At the same time, conversation requires different speech strategies than monologue-type speech, as it creates a kind of "competition" between speakers, while the speech production of others gives the speaker time to plan their own speech. Accordingly, differences between speech

types are realized in their prosodic patterns, for example, (i) in temporal features, such as the articulation tempo itself (cf. Trouvain, 2003), (ii) in the realization of prosodic features (e.g., the appearance of boundary prosodic elements is much stronger in reading (e.g., longer phrase-final lengthening) than in spontaneous speech, cf. White et al., 2010) and the (iii) in fundamental frequency (cf. Daly–Zue, 1992; Skarnitzl–Vaňková, 2017).

Therefore, the speech task itself might affect speaker recognition due to differences within and across speakers. The reading task has the same segmental and prosodic purpose even if the speakers are different. These differences vary depending on the exact task when talking about spontaneous speech: In our study, speakers talk about their lives (job/education, family, hobbies). Even if we take the example of speakers talking about their experiences at the same faculty, these experiences can be very different, and therefore, speakers will word them differently, and have diverse and suprasegmental features due to different attitudes. On the other hand, the within-speaker difference may also vary across speakers due to the diversity of the intervening ten years in this specific study. During the semi-spontaneous speech, the speakers may summarize the same topic in different terms and phrases, while necessarily sharing keywords. Therefore, these three speech tasks can be viewed as a scale, with reading and spontaneous speech expressions pointing towards the end of this scale, while semi-spontaneous speech is somewhere in between.

1.4. Questions and Hypotheses

The question of the present study was how the time between speech samples recorded ten years apart and the variation of the speech task affected the results of speaker recognition. Our hypotheses were as follows: (i) Speaker recognition results are affected by the time between two recordings in the following way: (i/a) The comparison of speech samples from the same speaker, but ten years apart, will lead to results with a lower probability of speaker identity than the comparison of speech samples from the same speaker from the same recording session; however, (i/b) the results will show a higher probability between the recordings than when comparing different speakers. (ii) Speaker recognition results depend on the speech task used. (ii/a) If the same speaker is compared across multiple speech tasks, the results will show a lower probability of speaker identity between the two speech samples than if they are compared within the same speech task. (ii/b) The results will show the highest probability of speaker identity between the two speech samples when comparing spontaneous speech samples of the same speaker, and the lowest when comparing read speech samples of the same speakers.

2. Methods

2.1. Speakers

The speech material of 20 speakers was analyzed from a Hungarian Longitudinal Speech Corpus. This corpus is under development, so at the time of writing this study, we did not have access to the material of more speakers. The Hungarian Longitudinal Speech Corpus is based on the BEA (Neuberger et al., 2014) database, which contains a large number of speakers. The BEA recordings started in 2007 and were completed in 2017. The longitudinal extension (Gráczi et al., 2019) means that the speakers who are still available for recordings are invited to participate in a follow-up recording, 10 years after their BEA recording. The technical conditions are identical in the two recording sessions. Most of the speech tasks are the same in the two recording sessions. For information on the technical and speech task see: Gráczi et al., 2019. The recordings are mono audio recordings. A few BEA recordings were saved as stereo, but the two channels were identical, so one of them (always the first one for easier scripting) was kept.

The 20 speakers in this experiment included 6 female and 14 male subjects. Their age ranged from 19 to 45 years (28.7±6.7 years) at the time of the first recording. The second recording took place 10 years later. None of the speakers reported any speech or hearing impairment. Two female speakers reported smoking (20 and 25 years of smoking), and one male subject reported having given up smoking between the two recording sessions.

Female speakers are labelled F01, F02, …, F06, and male speakers are labelled M01, M02, …, M14.

2.2. Speech Material

Four speech materials were used in this experiment. The speech tasks were identical in the two recording sessions. The interview about the speaker's life (job, education, family, hobbies) was used as a quasi-monologue-like spontaneous speech. Two content summaries served as quasi-monologues and semi-spontaneous speech. In these tasks, the speaker listened to an informative text about plants and a historical anecdote, and were asked to tell the interviewer as accurately as possible what they had heard. The fourth task involved reading aloud a text about the unwanted side effects of pesticides.

The speech materials were prepared for the analysis manually. Noises, silent and filled pauses, and the interviewer's speech were removed from the analysis, i.e., only the parts of the speakers' speech appropriate for speaker recognition were retained in the speech samples.

2.3. Biometric speaker recognition task

Our results are based on the scores obtained from biometric speaker recognition using VOCALISE software, a DNN-based speaker recognition software in x-vector PLDA mode (Kelly et al., 2019). We opted for this software because its performance is well documented (Gerlach et al., 2019, 2021, etc.), and the results indicate high efficiency and reliable speaker recognition even under more complicated circumstances (e.g.: phone and studio recordings: Alexander et al., 2021; face covered recordings: Iszatt et al., 2021). For the present analysis, we used the "Xvector Default" settings.

Speaker recognition was performed on all possible speech sample pairs within the same gender. This means that the speech samples of female speakers were attested against their own and against the speech samples of other female speakers but not against those of the male speakers, and vice versa.

The VOCALISE system is a fourth-generation software (previous versions included GMM, GMM-UBM, i-vectors, x-vectors). It gives two numerical outputs: Score and Likelihood Ratio values. We use the score values in our analysis, and we consider the LR transformation as the second step in our work.

The Voice Biometrics consists of the following test: Feature extraction, Speaker modeling, Speaker comparison, and Score interpretation. In the methodology used, short-term frequency characteristics of speech were extracted by MFCC (Mel Frequency Cepstral Coefficients): The speech samples were split into 30 ms frames with a 50% overlap. Fast Fourier Transformation was applied to these sample parts, using a time-frequency representation of the speech window, and transformed into the Mel scale (Anil et al., 2016). The speaker modeling was performed using the x-vector method, which uses a pre-trained DNN (Deep Neural Network) to convert a set of audio features into a vector representation. The Deep Neural Networks are artificial neural networks that allow the modeling and representation of multiple relationships. The DNN is trained by large audio databases containing audio samples recorded under different conditions (e.g. technical conditions or different languages). The internal weights of the DNN are set to minimize errors at the output. During the Speaker comparison, a direct comparison is made between speaker models to obtain a score value.

In general, a higher score means a higher probability of speaker identity between the two speech samples compared. There is no threshold in this framework (as opposed to zero in the Log Likelihood Ratio (LR) interpretation). In this paper, only the score data and their interrelationships have been analyzed to show the connection between the results, since the score-to-LR mapping has to be done using a background database which includes a large number of speakers. However, there is no large speech database available where the speech samples of the same speaker are recorded years apart. Therefore, mapping the score results

into likelihood ratios would be biased. The conversion of score-to-LR is the next necessary step of this research to investigate the weak effect of this conversion, after understanding the basic effect of the time interval between the two recordings sessions on scores.

2.4. Statistical analyses

The scores received for speaker identification were analyzed using linear mixed models (lme4 package: Bates et al., 2015; R: R Core Team, 2021) separately for male and female speakers in each case. The models included the SCORES as dependent variables. SPEAKER IDENTITY (same speaker vs. different speakers), RECORDING SESSION IDENTITY (same session vs. 10 years apart), and SPEECH TASK (spontaneous speech vs. semi-spontaneous speech vs. reading) served as fixed factors. The interaction of the three fixed factors was allowed. The model selection was carried out starting from the most complex model to the less complex one. The models were compared using anova(). We chose the model with the lowest AIC number (Akaike, 1974) which did not differ significantly from the most complex model. In the selected model, the random effect terms were compared. Again, we moved from the most complex model, which included a random slope for the SPEECH TASK IDENTITY and RECORDING SESSION IDENTITY by speaker, to a model that included only a random intercept for the speaker. The model also included random intercepts according to the speech tasks of the two speech samples (SAMPLE TASK, TARGET TASK), the TARGET SPEAKER (whose speech sample was compared to in the given case), and factors not included as fixed factors (e.g., when the identity of the recording session was not included as a fixed factor, it was included as a random intercept in the model). The *p*-value was extracted with Satterthwaite approximation (lmerTest package: Kuznetsova et al., 2017). The marginal and conditional effect sizes of the best fitting model were calculated using the MuMIn (Bartoń, 2020). The pairwise comparison was carried out using the Tukey post hoc test (emmeans package: Lenth, 2021).

A second pair of linear mixed models were built to test the effect of SPEECH TASK IDENTITY and RECORDING SESSION IDENTITY – with their interaction allowed – in the same speaker conditions, i.e., where the two speech samples compared were from the same subject. The purpose of this second model was to better describe these factors, as the data analysis showed large differences between the tendencies in the full dataset used and those in this subset of the data, which are discussed in the Results section. The method of model selection was exactly the same as for the general linear mixed models described above. Random factors included random intercepts by SPEAKERS, TASKS, AND RECORDINGS of speech samples.

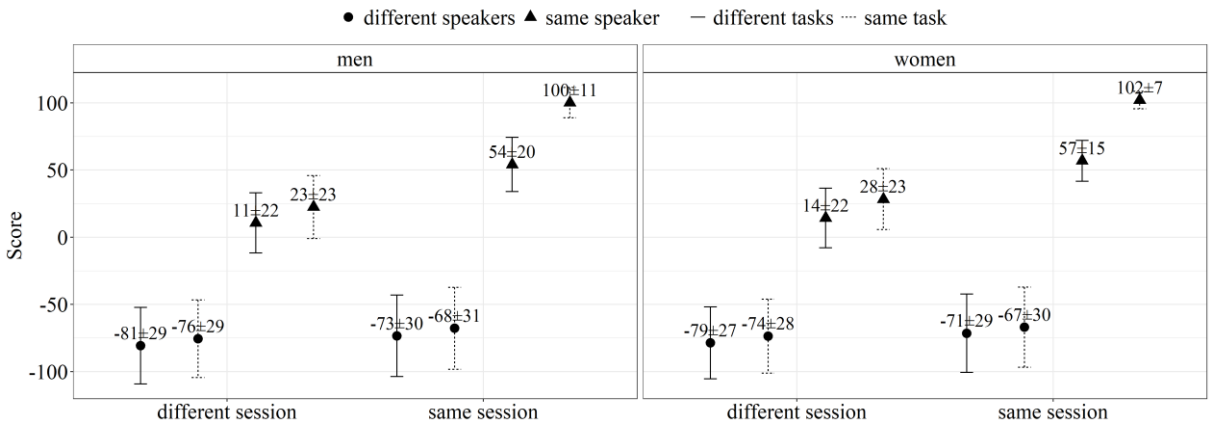
The figures were created using ggplot2 (Wickham, 2016).

3. Results

3.1. Results of the linear mixed models

The best-fitting model was the one that included all pairs of fixed effects: SPEAKER IDENTITY and SESSION IDENTITY, random slopes for the SPEAKER IDENTITY and SESSION IDENTITY by the speakers, and the random intercepts by the SAMPLE TASK and TARGET TASK, and TARGET SPEAKER, for both men and women (Figure 1). All fixed effects and all interactions were significant, therefore, we consider the threefold interaction of speaker identity, session identity and task identity to determine the results (women: $F(1, 2184.09) = 278.207$, $p < 0.001$, $r^2_m = 0.700$, $r^2_c = 0.815$; men: $F(1, 12482.0) = 589.14$, $p < 0.001$, $r^2_m = 0.515$, $r^2_c = 0.616$). The Tukey post hoc test showed considerably different results between the two gender groups. For women, scores on speech samples from the same speaker were significantly different for all possible combinations of TASK and SESSION IDENTITY, and scores on speech samples from different speakers were significantly different for all possible combinations of TASK and SESSION IDENTITY. However, scores obtained for speech samples from different speakers showed no significant differences in any possible combinations of TASK and SESSION IDENTITY. As for men, almost all possible SPEAKER, TASKS, AND SESSION IDENTITY combinations were significantly different, except for the different speaker + same session + different task compared to the different speaker + different session + same task. The difference between the two genders is due to the lower number of female speakers, which gives them more accurate speaker recognition results. Overall, we should consider that both the time between recordings and the speech task play a role in biometric speaker recognition.

Figure 1. Mean and sd values of speaker recognition scores grouped by speaker, task, and session identity

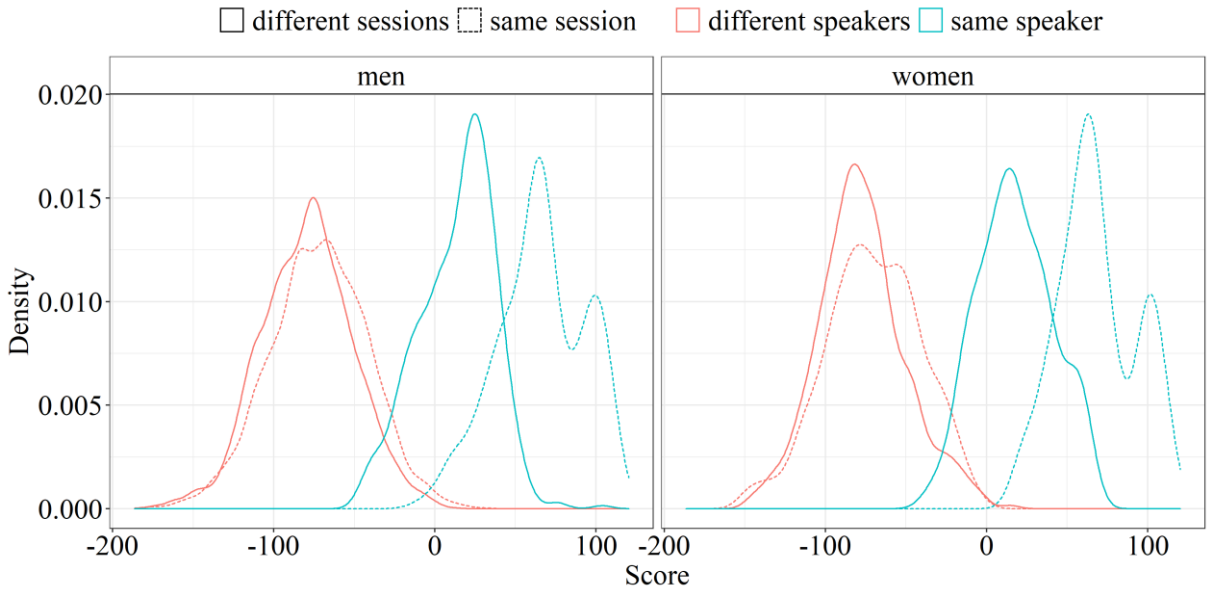


To understand the results more clearly, we discuss the differences in the results of comparisons between speakers. Speaker identification scores were lower for different speakers than for comparisons of identical speakers (Figure 2). In cases where the two speech samples were from two different speakers, the average score

was below −70 (of course without the effect of recording time), while it was above 60 in cases where the two speech samples were from the same speaker and the same recording session. However, if we consider the scores for the same speaker while comparing two speech samples ten years apart (= different recording sessions), the values are in between these two results. As figure 2 shows, the scores for the same speaker + same session speech sample comparisons are bimodal. This is because these values include both the same and different speech task recognition scores, between which have been found significant differences above.

The key result is that the scores are lower when comparing the same speaker + different recording sessions, which implies a lower score range in a lower region.

Figure 2. Speaker identification scores between speakers and recording sessions

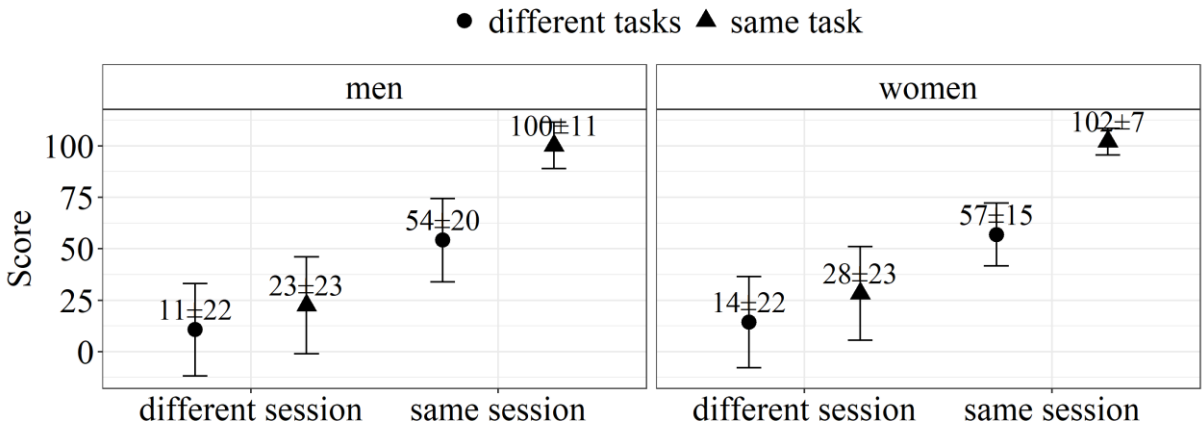


3.2. Speaker identification scores of within speakers across speech styles

Examining the effects of the RECORDING SESSION and the SPEECH TASK IDENTITY, results indicate that the effect of the time period is larger, but that TASK IDENTITY affects the scores of speech samples taken from the same session and 10 years apart (Figure 3). The difference due to the similarity or dissimilarity of the speech tasks compared is larger when the speech samples are 10 years apart. This may be due to the greater within speaker / within speech style variability over the 10 years.

The best-fitting model was the one that included both independent variables as fixed effects (including their interaction), and random intercepts by the speakers of the speech samples. According to the linear mixed model, the interaction of the two factors played a significant role (men: $F(1, 873) = 157.05$, $p < 0.001$, $r^2_m = 0.674$, $r^2_c = 0.764$; women: $F(1, 357.96) = 76.397$, $p < 0.001$, $r^2_m = 0.716$, $r^2_c = 0.815$).

Figure 3. Mean and sd scores of speaker recognition for the same speakers for session and task identity



The effects of the specific speech style show diverse results (Figure 4a&b). Regarding speaker recognition scores of speech samples 10 years apart, the only difference is that when both speech samples come from the same speaker, the score is slightly higher. (We do not wish to analyze in detail the possible effect of speaking styles on scores when comparing speech samples of different speakers, but we would like to point out that, similar to these results, we found slightly higher scores when comparing the two reading samples. This means that it is not exactly the identification efficiency itself that increases when comparing speech samples from the same speaker, when both are reading, or that the scores include the effect of the segmental and suprasegmental identity/similarity encoded in the MFCCs.) However, this is the least common scenario among real-life challenges. Besides, in terms of results for speech samples from the same session, it can be seen that the earlier findings in this chapter can be fleshed out in more detail. While the highest scores were found in cases where the speech tasks of the two speech samples were identical, as already discussed, we also observe that the scores for spontaneous and semi-spontaneous comparisons are higher than those for reading and spontaneous or semi-spontaneous comparisons. The differences are not large, but important to note for further consideration. This particular result may be due to the fact that spontaneous and semi-spontaneous speech styles differ considerably from reading in their suprasegmental features, with reading being closer to hyperarticulation, while less pre-planned speech styles are closer to hypoarticulation, leading to larger segmental variability and perhaps more speaker-specific data (if further influencing factors, like addressee, formality etc. are kept identical) (for a summary on speech styles see Krepsz, 2016).

Figure 4a. The mean and sd value of speaker recognition for speech samples from the same speaker from the same recording session, for the specific speech tasks of the speech samples

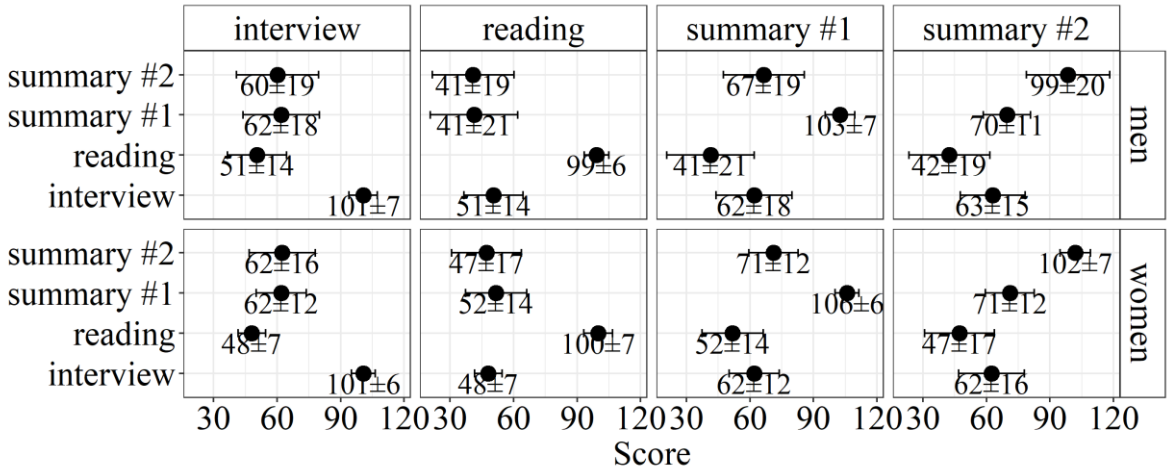
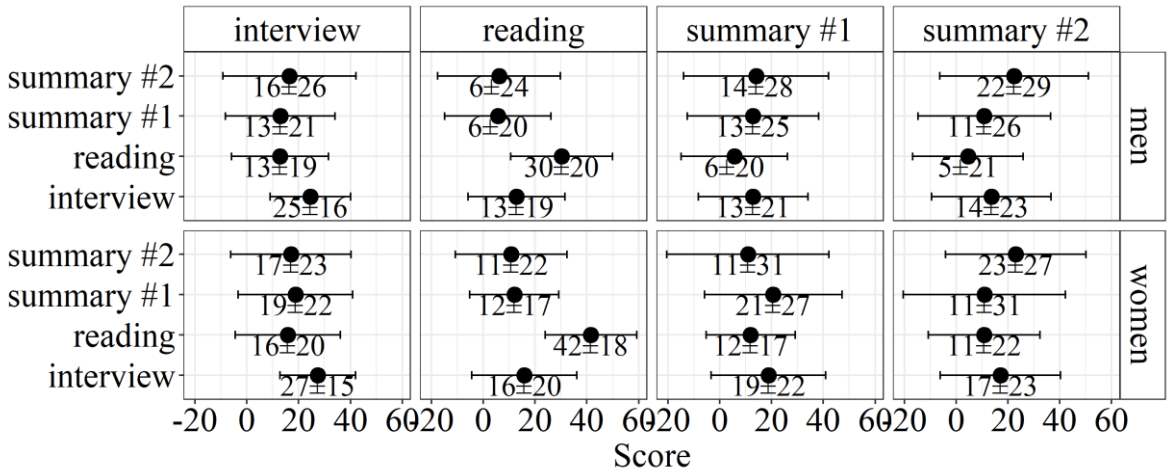


Figure 4b. The mean and sd value of speaker recognition for speech samples from two different recording sessions, for the specific speech tasks of the speech samples



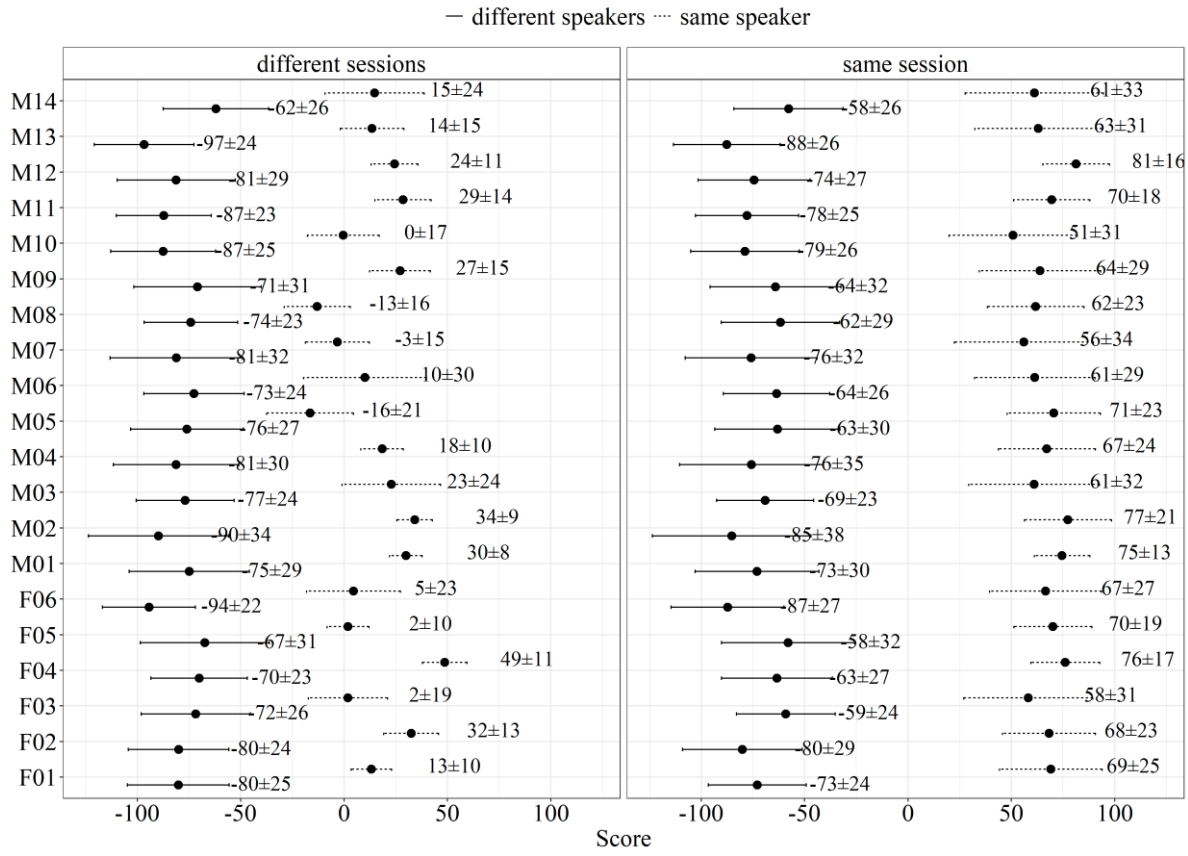
3.3. Individual differences

As described in the Introduction, changes in speech characteristics are heavily speaker specific. We can therefore assume that the results of the speaker recognition will also differ across subjects.

Figure 5 shows the mean and standard deviation of the scores obtained when comparing speech samples recorded during the same session and recorded 10 years apart, regardless of speech style. The mean and sd values of the speaker recognition scores indicate that successful identification and successful separation of speakers do not differ greatly between speakers when the speech samples are from the same recording session. There were speakers who received higher scores for their speech samples, such as M12, and F05, and others, who received low scores and had higher variability (larger sd), e.g., F03, and M10. However, individual differences increased for speech samples taken 10 years apart: The range of mean scores by speaker is between 51 and 81 for speech samples from

the same recordings and between −16 and 49 for speech samples from ten years apart, which is twice the mean score, indicating speaker recognition scores show a greater interspeaker variability between speakers when using speaker recognition based on speech samples from ten years apart than when using the same recordings. It should also be noted that the lowest mean score in cases where two speech samples from the same speaker were compared ten years apart is higher than the highest score when comparing speech samples from different speakers. However, the variability (indicated here by sd) shows that the range of scores for speech samples from the same speaker taken ten years apart overlap with the scores in the comparisons of speech samples from different speakers.

Figure 5. Mean and sd scores by speaker for speaker identity and recording session



4. Discussion and conclusions

Our study addressed two possible questions in biometric speaker recognition: (1) whether and how the longitudinal changes of the speech characteristics and, (2) whether and how the speech style of speech samples and the speech tasks analyzed affect the results of speaker recognition. The two questions were also addressed in interaction. The experiment was carried out on 20 speakers, recorded twice, ten years apart, and we analyzed four speech tasks from three speech styles: spontaneous, semi-spontaneous, and reading. VOCALISE software was chosen

for the biometric speaker recognition task, due to its well-documented, high performance.

Our first hypothesis was that speaker recognition results would be more successful on speech samples from the same speaker and from the same recording session, while those on speech samples recorded ten years apart would be lower. We therefore expected higher scores in the first case, and lower scores in the second. We also hypothesized that speaker recognition, however, would result in higher "similarity" scores even for speech samples taken ten years apart than for speech samples taken from different speakers. Our results confirmed this hypothesis.

Our second hypothesis was partially supported. First, we hypothesized that speaker recognition would be more efficient between speech samples of the same speech task. This was proven for the speech samples taken from the same recording session, but not over the course of ten years. The second part of the hypothesis did not hold. Scores were similar for any pair of speech samples taken from identical speech tasks. Comparing read speech over ten years gave slightly higher scores than comparing (semi-)spontaneous speech tasks, but the difference is small, and this is a hard-to-access scenario for forensic speaker recognition.

Both the results and the effect sizes of the linear mixed models showed differences between men and women. At the time of the experiment, only 6 female and 14 male speech samples were available, the variance is assumed to be mainly due to the low number of female speakers. It should be emphasized that the tendencies were similar for the two genders.

When we talk about speech tasks and speech styles, we need to enhance the results of the two semi-spontaneous speech tasks. The results were quite similar when comparing the same summaries, when comparing all identical speech tasks and when comparing any summary with another non-summary speech task. However, on the one hand, all mixed speech task comparisons gave similar results, while on the other hand, the scores for the comparison of the two summaries did not show higher results than for any other comparison of two different speech tasks. This may indicate that the speech task itself plays a more important role in speaker recognition than the speech style of the tasks. We also noted in the *Introduction* that the similarity of speech tasks within and between speakers may play a role in speaker recognition. When examining speaker recognition scores based on speech samples recorded ten years apart, scores were slightly higher when both speech samples were read than when both were from the same speech task, but not read. We noted that a similar effect of speech style was also detected when comparing speech samples from different speakers. This implies that segmental and suprasegmental similarity played a role in increasing the similarity of the MFCCs, but probably did not increase the efficiency of speaker recognition itself. However, this question requires more specific

investigation. Even if this scenario is not common in everyday forensic speaker recognition, the result is interesting from a theoretical point of view.

Findings on the speech corpus in question indicated that the fundamental frequency decreased in the speech of some speakers (mostly in women in their twenties and thirties) (Gráczi et al., 2022). However, the connection between the results of speaker recognition and f0 studies is not simply related, since Voice Biometrics and VOCALISE are not only based on f0 characteristics but use a more complex methodology based on various speech characteristics, similar to DNN methodology. Therefore, further acoustic features need to be studied.

Speaker identification results are prone to individual variability, even when comparing speech samples from close or the same recording sessions. In our study, the scores for speech samples from the same speaker were rather high, with a negligible interspeaker variability. However, scores for all speakers decreased when speech samples were from two different recording sessions, but with large interspeaker differences.

Our results indicate that biometric speaker recognition is less dependent on the speech style itself, however, the results are clearer when the speech tasks are left unchanged. We can also conclude that the success of biometric speaker recognition depends to a large extent on the speaker in question if there is a significant time difference between the recordings.

Acknowledgement

This research was supported by the Hungarian National Research, Development and Innovation Office of Hungary, project No. FK-128814.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19/6,* 716–723. DOI: 10.1109/TAC.1974.1100705

Alexander, A., Kelly, F. & Gold, E. (2021). A WYRED connection: x-vectors and forensic speech data. Abstract. In *International Association for Forensic Phonetics and Acoustics (IAFPA) Conference 2021*, Marburg. https://oxfordwaveresearch.com/wp-content/uploads/2021/08/Abstract-A-WYRED-connection-x-vectors-and-forensic-speech-data.pdf

Alexander, A., Forth, O., Atreya, A. A. & Finnian, K. & Kelly, F. (2016) *VOCALISE: A forensic automatic speaker recognition system supporting spectral, phonetic, and user-provided features*. In *Proceedings of Odyssey: The Speaker and Language Workshop*. https://oxfordwaveresearch.com/wp-content/uploads/2019/03/Alexander-et-al-VOCALISE-2016-Odyssey.pdf

Anil, K. J., Arun A. R. & Karthik N. (2011). *Introduction to Biometrics*. New York, Dordrecht, Heilderberg, London: Springer. DOI: 10.1007/978-0-387-77326-1 33.

Bartoń, K. (2020). *MuMIn: Multi-Model Inference.* R package version 1.43.17. URL: https://CRAN.R-project.org/package=MuMIn.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67,* 1–48.

Belsky, D. W., Caspi, A., Houts, R., Cohen, H. J., Corcoran, D. L., Danese, A., … & Moffitt, T. E. (2015). Quantification of biological aging in young adults. *Proceedings of the National Academy of Sciences of the United States of America 112*, 4104–4110. DOI: 10.1073/pnas.1506264112

Busse, E. W. (2002). General Theories of Aging. In: Copeland, J. R. M., Abou-Saleh, M. T., & Blazer, D. G. (eds.), *Principles and Practice of Geriatric Psychiatry* (19–22). West Sussex, UK: John Wiley & Sons Ltd.

Cox, V. O. & Selent, M. (2015). Acoustic and respiratory measures as a function of age in the male voice. *Journal of Phonetics and Audiology, 1,* 105. DOI: 10.4172/jpay.1000105

Daly, N. A. & Zue, V. W. (1992). Statistical and linguistic analyses of F0 in read and spontaneous speech. In *Proceedings of the International Conference on Spoken Language Processing 1,* 763–766.

Drygajlo A. & Haraksim R. (2017). Biometric Evidence in Forensic Automatic Speaker Recognition. In: Tistarelli, M. & Champod, C. (2017): *Handbook of Biometrics for Forensic Science* (221–239.) Cham: Springer International Publishing. DOI: 10.1007/978-3-319-50673-9.

Duchin, S. W. & Mysak, E. D. (1987). Disfluency and rate characteristics of young adult, middle-aged, and older males. *Journal of Communication Disorders, 20,* 245–257. DOI: 10.1016/0021-9924(87)90022-0

Eichhorn, J. T., Kent, R. D., Austin, D. & Vorperian, H. K. (2018). Effects of aging on vocal fundamental frequency and vowel formants in men and women. *Journal of Voice, 32,* 644.e1–644.e9. DOI: 10.1016/j.jvoice.2017.08.003

French, J. P. F., Harrison, P. & Windsor-Lewis, J. (2006) R v John Samuel Humble: The Yorkshire Ripper Hoaxer trial. *The International Journal of Speech, Language and the Law, 13 (2),* 256–273.

Gerlach, L., Kelly, F. & Alexander, A. (2019). More than just identity: speaker recognition and speaker profiling using the GBR-ENG database. In *Proceedings of International Association for Forensic Phonetics and Acoustics (IAFPA) Conference 2019*, Istanbul https://oxfordwaveresearch.com/wp-content/uploads/2020/02/IAFPA19_GBRENG_Gerlach_et_al_paper.pdf

Gerlach, L., McDougall, K., Kelly, F. & Alexander, A. (2021). How do automatic speaker recognition systems 'perceive' voice similarity? Further exploration of the relationship between human and machine voice similarity ratings. In International Association for Forensic Phonetics and Acoustics (IAFPA) conference 2021, Marburg, Germany, https://oxfordwaveresearch.com/wp-content/uploads/2021/08/Abstract-How-do-automatic-speaker-recognition-systems-perceive-voice-similarity-Further-exploration-of-the-relationship-between-human-and-machine-voice-similarity-ratings.pdf

Gráczi Tekla Etelka, Huszár Anna, Krepsz Valéria & Markó Alexandra (2022). Az alapfrekvencia-jellemzők longitudinális változása fiatal és középkorú felnőttek beszédében. In Mády Katalin & Markó Alexandra (szerk.), *Általános Nyelvészeti Tanulmányok, 34,* 111–142. Budapest: Akadémiai Kiadó.

Gráczi Tekla Etelka & Krepsz Valéria (2020). Évek múltán a zönge. Egyes zöngejellemzők változása 11 év alatt 6 férfi beszélő beszédében. In Fóris Ágota, Bölcskei Andrea, Bóna Judit, Gráczi Tekla Etelka & Markó Alexandra (szerk.): *Nyelv, kultúra, identitás. Alkalmazott nyelvészeti kutatások a 21. századi információs térben: III. Fonetika*. Budapest: Akadémiai Kiadó. https://mersz.hu/hivatkozas/m675nyki3f_23

Gráczi Tekla Etelka, Krepsz Valéria, Markó Alexandra, Huszár Anna & Száraz Bettina (2019). Az f0-jellemzők felolvasásban és spontán beszédben. *Alkalmazott Nyelvtudomány, 19.* Letöltés: http://alkalmazottnyelvtudomany.hu/wordpress/wp-content/uploads/Graczi_tan.pdf

Iszatt, T., Malkoc, E., Kelly, F. & Alexander, A. (2021). *Exploring the impact of face coverings on x-vector speaker recognition using VOCALISE.* In *Proccedings of International Association for Forensic Phonetics and Acoustics (IAFPA) Conference 2021, Marburg. https*://oxfordwaveresearch.com/wp-content/uploads/2021/08/Abstract-Exploring-the-impact-of-face-coverings-on-x-vector-speaker-recognition-using-VOCALISE.pdf

Jacewicz, E., Fox, R. A. & Wei, L. (2010). Between-speaker and within-speaker variation in speech tempo of American English. *The Journal of the Acoustical Society of America, 128,* 839. DOI: 10.1121/1.3459842

Kelly, F., Forth, O., Kent, S., Gerlach, L., & Alexander, A. (2019). Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors. In *Proceedings of Audio Engineering Society (AES) Forensics Conference 2019, Porto*. https://www.aes.org/tmpFiles/elib/20230127/20477.pdf

Krepsz Valéria (2016). Fonetikai hasonlóságok és különbözőségek a beszédtípusokban. In Bóna Judit (szerk.): *Fonetikai olvasókönyv* (175–188). Egyetemi e-jegyzet. Budapest: ELTE Fonetikai Tanszék.

Krepsz Valéria & Gráczi Tekla Etelka (2018). *Temporális és fonációs paraméterek beszélőn belüli variabilitása longitudinális szempontból – Esettanulmány*. Pszicholingvisztikai Nyári Egyetem 2018. Előadás, Balatonalmádi.

Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*, 1–26. DOI: 10.18637/jss.v082.i13

Lenth, Russell V. (2021). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.6.0. https://CRAN.Rproject.org/package=emmeans

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, Massachusetts, London: MIT Press.

Markó Alexandra, Huszár Anna, Krepsz Valéria & Gráczi Tekla Etelka (2021). Az alapfrekvencia jellemzőinek longitudinális összevetése felnőtt beszélők felolvasásában. *Beszédtudomány—Speech Science 2021*, 99–134. DOI: 10.15775/Besztud. 2021.

Mwangi, S., Spiegl, W., Hönig, F., Haderlein, T., Maier, A & Nöth, E. (2009). Effects of Vocal Aging on fundamental Frequency and Formants. In: *Proceedings of the International Conference on Acoustics NAG/DAGA* (1761–1764). Rotterdam.

Neuberger, T., Gyarmathy, D., Gráczi, T. E., Horváth, V., Gósy, M. & Beke, A. (2014). Development of a large spontaneous speech database of agglutinative Hungarian language. In Sojka, P., Horák, A., Kopeček, I. & Pala, K. (eds.), *International Conference on Text, Speech, and Dialogue* (424–431). New-York, Berlin, Heidelberg: Springer.

R Core Team (2021). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Reubold, U., Harrington, J. & Kleber, F. (2010). Vocal aging effects on F0 and the first formant: A longitudinal analysis in adult speakers. *Speech Communication, 52*, 638–651. DOI: 10.1016/j.specom.2010.02.012

Rhodes, R. (2017). Aging effects on voice features used in forensic speaker comparison. *International Journal of Speech Language and The Law, 24*, 177–199.

Russell A, Penny L, Pemberton C. (1995). Speaking fundamental frequency changes over time in women: a longitudinal study. *Journal of Speech, Language, and Hearing Research, 38*, 101–109.

Schreier, D. (2021). Variation and third age: A sociolinguistic perspective. *Linguistics Vanguard, 7(s2)*. DOI: 10.1515/lingvan-2020-0030

Schötz, S. (2006). *Perception, analysis and synthesis of speaker age* (Vol. 47). Lund: Lund University.

Skarnitzl, R. & Vaňková, J. (2017). Fundamental Frequency statistics for Male Speakers of Common Czech. *Acta Universitatis Carolinae Philologica, 3*, 7–17.

Stathopoulos, E. T., Huber, J. E. & Sussman, J. E. (2011). Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4–93 years of age. *Journal of Speech, Language, and Hearing Research, 54*, 1011–1021. DOI: 10.1044/1092-4388(2010/10-0036)

Trouvain, J. (2003). Tempo variation in speech production. Implications for speech synthesis. PhD dissertation. Saarbrücken: Saarland University. https://www.coli.uni-saarland.de/groups/BM/phonetics/contents/phonus-pdf/phonus8/Trouvain_Phonus8.pdf

White, L., Wiget, L., Rauch, O., & Mattys, S. L. (2010). Segmentation cues in spontaneous and read speech (100218:1-4.). In *Proceedings of the 5th Conference on Speech Prosody, 2010*, Chicago, USA.

Wickham, H. (2016). Data analysis. In Wicham H., Navarro, D. & Pederson L (eds.) *ggplot2 – Elegant Graphics for Data Analysis* (189–201). New York, USA: Springer.