2022 ACTA UNIVERSITATIS CAROLINAE PAG. 11–22
PHILOLOGICA 1 / PHONETICA PRAGENSIA

THE IMPACT OF MISMATCHED RECORDINGS ON AN AUTOMATIC-SPEAKER-RECOGNITION SYSTEM AND HUMAN LISTENERS

TOMÁŠ NECHANSKÝ, TOMÁŠ BOŘIL, ALŽBĚTA HOUZAR, RADEK SKARNITZL
Institute of Phonetics, Faculty of Arts, Charles University

ABSTRACT

The so-called 'mismatch' is a factor which experts in the forensic voice comparison field encounter regularly. Therefore, we decided to explore to what extent the automatic-speaker-recognition system's and the earwitness' ability to identify speakers is influenced when recordings are acquired in different languages and at different times. 100 voices in a database of 300 recordings (100 speakers recorded in three mutually mismatched sessions) were compared with an automatic-speaker-recognition software VOCALISE based on i-vectors and x-vectors, and by 39 respondents in simulated voice parades. Both the automatic and perceptual approach seem to have yielded similar results in that the less complex the mismatch type, the more successful the identification. The results point to the superiority of the x-vector approach, and also to varying identification abilities of listeners.

Keywords: forensic voice comparison, temporal mismatch, language mismatch, automatic speaker recognition, voice parade

1. Introduction

Forensic phoneticians, when identifying a speaker, encounter various cases differing in complexity. Even two realizations of the same word uttered right after each other will not be identical from the acoustic point of view, and this variability in speech must be acknowledged when comparing voices for forensic purposes. A ubiquitous characteristic of Forensic Voice Comparison (FVC) which increases the complexity of the process is *mismatch* between recordings, which can take several forms.

First, recordings examined in FVC typically differ in their technical aspects, particularly in the characteristics of the channel. For example, voice samples of the unknown speaker (typically the perpetrator) may originate from an intercepted mobile telephone call or from a wiretapped office, while those of the known speaker (the suspect) may be obtained in an interrogation room. The effect of channel variation has been investigated by numerous researchers, with special focus on telephone transmission (both landline and

https://doi.org/10.14712/24646830.2022.25

mobile); results of such studies are not surprising, with mismatched recordings yielding lower recognition scores (e.g., Alexander et al., 2005; Hughes et al., 2019; Bortlík, 2021). Technical mismatch also includes phenomena such as reverberation or various forms of background noise (see Guillemin, 2022 for a comprehensive summary).

Speakers themselves constitute sources of several kinds of mismatch. This kind of within-speaker variation stems from the incredible plasticity of our speech production mechanism. Some of the areas which have received considerable attention include various speech styles and the effect that they have on specific acoustic parameters (Jessen, 2009; McDougall & Duckworth, 2018; Ross et al., 2019), the impact of various affective or physiological states (Eriksson et al., 2007; Scherer, 2019), as well as phonetic accommodation to one's communication partner (see, e.g., Earnshaw, 2021; Šturm et al., 2021). In FVC, these are particularly important since the analyzed recordings tend to be mismatched in this respect. In general, research points to differences in the values of acoustic parameters and sometimes to decreased speaker recognition performance (e.g., Shriberg & Scheffer, 2009). However, some studies suggest that certain parameters remain relatively stable within speakers. For example, McDougall and Duckworth (2018) report considerable within-speaker consistency in various dysfluency features in telephone and interview styles. Other behavioural effects, discussed in more detail by Gold et al. (2022), include whispered speech, loud speech in the presence of Lombard effect, as well as disguised speech (Eriksson, 2010; Růžičková & Skarnitzl, 2017).

Another source of mismatch consists in the non-contemporary nature of FVC: the recordings which are compared in a forensic case must have been, by definition, obtained at different times. This has been examined by a number of researchers from the forensic-phonetic, as well as automatic speaker recognition (ASR) perspective. Their studies focused on the recognizability of speakers across different time spans, from several days (Ross et al., 2019; what is often referred to as 'session mismatch') and months (Kelly & Hansen, 2015) to years or even several decades (Hollien & Schwartz, 2000; Rhodes, 2017). It is not surprising that speech and voice patterns change throughout our lifetime, resulting in a drop of both human and machine recognition of speakers. For example, Rhodes' (2017) investigation of speakers over a span of 28 years showed a change in vowel formants of between 3 and 15%, with the most robust effect observed in *F*1. Likelihood ratios obtained from vowel formant data shifted towards incorrect decisions, and ASR performance dropped significantly at delays between recordings above 14 years.

The final source of within-speaker variability to be mentioned here is when different languages or accents are used by one speaker. Several studies have addressed foreign accent in FVC, focusing on the imitation of a foreign accent (Torstensson et al., 2004), on listeners' ability to identify authentic foreign accents (Neuhauser & Simpson, 2007; Sullivan & Schlichting, 2000), or on the degree to which non-native language background helps witness experts identify a speaker of another language (Schiller et al., 1997). Language-mismatched recordings have also been examined using ASR approaches. For example, Misra and Hansen (2014) found that when only English recordings were used for training the ASR system, language mismatch resulted in a drop in performance by a factor of 2.5; however, including non-English material at the training stage substantially improved performance. In a recent study, Bortlík (2021) examined the effect of foreign-accented speech on the performance of state-of-the-art ASR systems; he reported higher error rates

in language-mismatched comparisons – i.e., when a Czech speaker was speaking Czech in one condition and English in the other – than in matched comparisons.

The first aim of the present study is to investigate the combined effect on the performance of an ASR system of contemporary and non-contemporary recordings of speech produced in the same and in a foreign language, by the same speaker. The setting simulates two hypothetical situations which may be relevant for FVC. During the perpetration of a crime, the unknown speaker uses a foreign language (L2), while the suspect recording with the known speaker is in the speaker's mother tongue (L1). Since it is not unusual for the suspect recording to originate from a wiretap, language identity cannot be ensured, and cross-language comparisons will be required. The second aim is to explore the ability of listeners to identify the speaker in a simulated voice parade under the conditions described above.

2. Method

2.1 Material

The database for our research comprises 100 (78 female and 22 male, aged 20–25) speakers of Czech as L1 and English as L2 (with the CEFR level being B2 or C1). The speakers were studying English and American Studies at Charles University at the time of recording. The recordings were obtained in the sound-treated recording studio of the Institute of Phonetics in Prague, using the high-quality AKG C4500 B-BC condenser microphone, with 32-kHz sampling frequency and 16-bit resolution.

Three recording sessions are analyzed in this study. At the beginning of their studies, the speakers were asked to read: first, a phonetically rich text in Czech; and second, several pieces of BBC news in English. Four months later, the same students were recorded reading other BBC news texts in English again. On average, each participant produced ca. 1 minute of speech in Czech, and 3–4 minutes in English twice.

2.2 Automatic speaker recognition procedure

Since all the speakers are known but were recorded under three conditions, for each speaker we compared the following mismatches as if they were the unknown and suspect recordings:

- language mismatch (contemporary Czech and English recordings),
- temporal mismatch (non-contemporary English recordings),
- double mismatch (non-contemporary Czech and English recordings).

Speaker comparisons were performed in VOCALISE by Oxford Wave Research, using the i-vector (session VOCALISE i-vector 2017B) and x-vector (session VOCALISE x-vector 2019A-Beta-RC2) PLDA framework. In this framework, vectors of speakers (i-vector or x-vector) are compared using probabilistic linear discriminant analysis (PLDA); this post-processing method computes the likelihood of the vector pair originating from the same speaker versus coming from two different speakers. The i-vectors and x-vectors are different ways of speaker modelling in the speaker recognition pipeline. Whereas i-vectors make use of front-end factor analysis as the feature extractor, x-vectors rely on trained deep neural

networks (see Kelly et al., 2019 for more details); x-vectors are the most recent approach to speaker modelling. The resulting scores were calibrated using cross-validation in the BioMetrics software by Oxford Wave Research.

Apart from the three comparisons listed above, we conducted several partial comparisons to examine the effect of "tuning" (see Skarnitzl et al., 2019) using condition adaptation. Condition adaptation optimizes the ASR system to new conditions, specific to the given recordings, by adapting the LDA and PLDA models. By performing condition adaptation, the properties from dozens of i-/x-vectors in the adaptation set are used to adapt tens of thousands of i-/x-vectors in the training dataset of VOCALISE towards the new conditions; in other words, the statistics of the LDA and PLDA models were updated using a weighted combination of the original training data and the recordings provided by the authors. For this purpose, the three datasets were divided into two halves (50 speakers in set 1 and 50 in set 2; the division was random but identical across the three datasets). Subsequently, recordings of set-2 speakers were used for condition adaptation of set-1 comparisons, and vice versa.

We will report the equal error rate (EER) as the standard measure of ASR performance (EER is defined as the number when false-acceptance rate and false-rejection rate become equal; see Hansen & Hasan, 2015). Since some comparisons involve relatively small datasets, Convex Hull EER values are reported in all analyses. In addition, we will provide values of the log-likelihood-ratio cost (C_{llr}), a measure that evaluates the accuracy of an ASR system by capturing the gradient goodness of a set of likelihood ratios derived from test data, with values ideally not exceeding 1 (Morrison, 2011).

2.3 Listening test procedure

For our perception experiment – a simulated voice line-up in which an earwitness is supposed to identify the perpetrator's voice among recordings of suspects, we used recordings of 22 male speakers from the same database. Each line-up (or parade) featured six recordings: the perpetrator's voice and five suspects' voices available for matching with the perpetrator. Crucially, to approximate conditions of real-life voice parades, the perpetrator's voice could be either present among the five suspects (i.e., the so-called target voice), or absent. The perpetrator and suspect recordings (whether the target was present or absent) would differ in language (language mismatch), time of recording (temporal mismatch), both (double mismatch), or would not differ at all (no mismatch). The perception experiment comprised the following line-up conditions:

- 5 line-ups for recordings of no mismatch (contemporary Czech, or English),
- 4 line-ups for recordings of language mismatch (contemporary Czech and English),
- 2 line-ups for recordings of temporal mismatch (non-contemporary English),
- 4 line-ups for recordings of double mismatch (non-contemporary Czech and English).

In real-life voice parades, it is recommended that foils' voices (i.e., all suspect voices except the target) should be as similar to the target speaker's voice as possible (de Jong-Lendle et al., 2015). We used fundamental frequency (f_0) median as a measure of distance (i.e., similarity) between speakers. We selected the foils' voices to be closest to the perpetrator. This was not adhered to only when the target's and perpetrator's samples were mismatched; in this case, the most distant speakers were chosen.

In total, 90 samples (15 line-ups * 6 samples) of about 5 seconds in duration were used for the perception test. It was ensured that the samples within one line-up were not textually identical and that they were loudness-normalized. The perception test was designed in PsyToolkit (Stoet, 2010; 2017) and was coded as fifteen tasks requiring the participant to first listen to the perpetrator, then to the five suspects, and after that to either match one of the suspects with the perpetrator, or to check a box indicating that the perpetrator's voice was not one of the suspects. Participants were allowed to replay any of the recordings. In order to minimize the order effect, samples in each line-up, as well as the line-ups themselves were randomized.

Besides the experiment, we gathered basic demographic information from the respondents, who received monetary compensation for their participation. The perception experiment was completed by 33 female and 7 male respondents, aged 22–49, all coming from the Czech Republic. It was revealed later that one participant had not listened to the stimuli properly, and her responses were eliminated. Therefore, 39 listeners' answers were analyzed in the end. The total time spent ranged from 15 to 46 minutes; 80% of the participants finished the test (including the demographic survey) under 31 minutes.

3. Results

3.1 Automatic speaker recognition

Figure 1 compares equal error rates achieved by the i-vector and x-vector approaches: it is obvious (notice the difference in the scale of the two plots) that the i-vectors are significantly outperformed by the x-vectors.

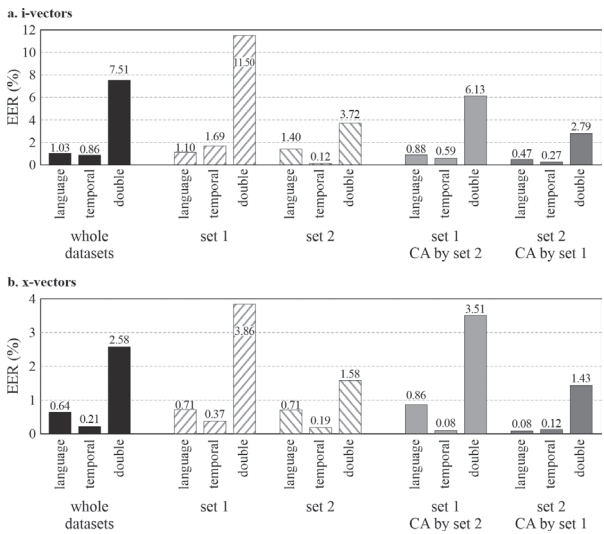

Figure 1. Equal error rates for i-vectors (a.) and x-vectors (b.) for three types of mismatch (language, temporal, double). In the left part (black), results for the entire datasets; in the middle (with stripes), for the half-size datasets; on the right (in grey), for the half-size datasets with the opposing half used for condition adaptation (CA).

What is particularly noteworthy is that single mismatch conditions (i.e., only temporal, or only language mismatch) result in very good performance, with EERs around 1% or lower, using both i-vectors and x-vectors. However, double mismatch conditions (both non-contemporary and language-mismatched) yield significantly higher error rates, 7.5% for i-vectors and 2.6% for x-vectors. The situation may also be illustrated using a combined equal error plot, with all three main comparisons (see Fig. 2); note that only results for x-vectors are shown in the figure. An equal error plot shows the false acceptance rate (FAR) and the false rejection rate (FRR) on the vertical axis against the threshold score on the horizontal one; the intersection of the two curves corresponds to the EER. The better the curves are separated, the better the recognition.

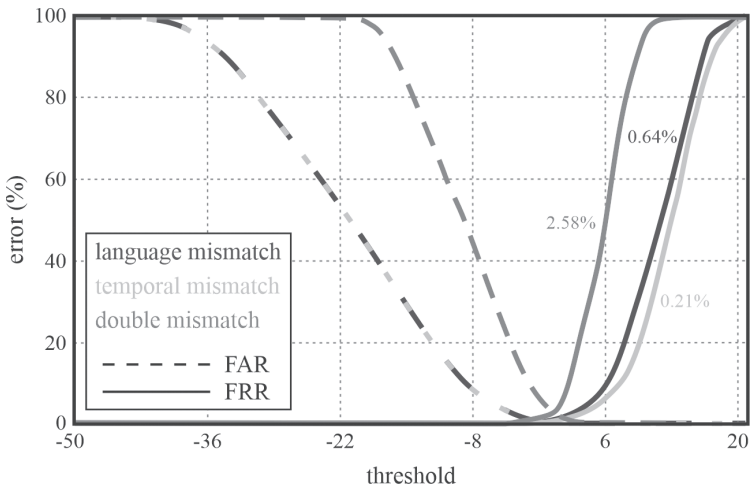

Figure 2. Equal error plot comparing the three main comparisons using x-vectors; EER values are shown in the corresponding colours. FAR = false acceptance rate, FRR = false rejection rate

The same tendencies can be observed also for the comparisons of partial datasets (shown with stripes in Fig. 1) and for partial datasets "tuned" by the corresponding opposing half using condition adaptation (in shades of grey). At the same time, EER clearly depends on the particular selection of speakers under comparison: results for set 1 and set 2 are far from identical. System accuracy can be regarded as high for all comparisons performed, with C_{llr} being 0.4 for the double-mismatched condition in i-vector set 1, and considerably lower in all other (i-vector and x-vector) comparisons ($C_{llr} < 0.1$ for all single-mismatched comparisons, and $0.07 \leq C_{llr} \leq 0.4$ for the double-mismatched conditions).

What remains to be discussed is the hypothesized benefit of condition adaptation on ASR performance; in other words, we are interested in finding out whether using the opposing half of the dataset for PLDA adaptation yields lower error rates. Overall, this benefit was slightly more salient in the case of i-vectors, where we can see a considerable improvement in most of the scores (cf. the striped and grey bars in Fig. 1a); in the case of x-vectors, condition adaptation yielded a lower EER in five out of the six comparisons (Fig. 1b). Crucially, the benefit turned out to be greatest with the double-mismatched conditions.

3.2 Listening test

First, we wanted to know how listeners scored individually and what the overall successful identification rate was. The results for individual participants are presented in Figure 3 for the target-present scenario and in Figure 4 for the target-absent scenario. The figures provide overviews of hits (correctly identified targets), foils (incorrectly identified suspects), correct rejections (correctly rejected all suspects), and incorrect rejections (incorrectly rejected all suspects).

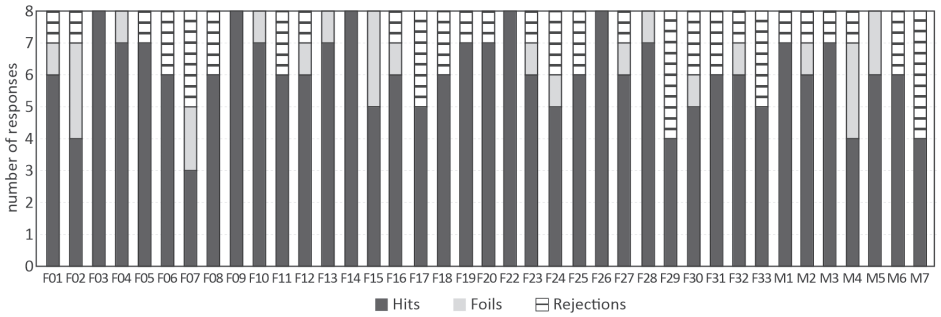

Figure 3. Individual responses (hits, foils, and incorrect rejections) for target-present parades (see text).

Since there were 8 target-present line-ups (see Fig. 3), the maximum number of correct answers (labelled as hits in the figure) was 8, which was achieved by five listeners. On the other hand, this condition allowed for two types of mistakes – incorrectly identifying another suspect as the target (labelled as foils) and incorrectly rejecting all suspects (assuming the target was absent; marked as rejections). The highest number of incorrect answers combined was 5, which was produced by only one listener (i.e., a successful identification rate of only 37.5%).

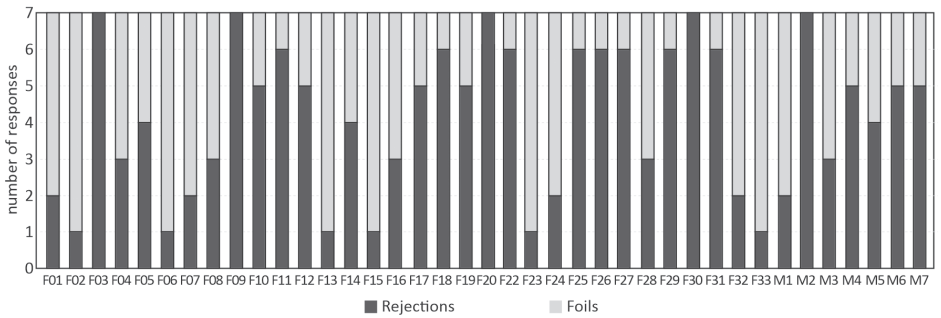

Figure 4. Individual responses (correct rejections and foils) for target-absent parades (see text).

As for target-absent parades (Fig. 4), it was possible to either answer correctly (rejecting all suspects, as the target was not included; labelled as rejections in the figure), or to incorrectly choose a suspect (foils). We had 7 of such parades, and a 100% successful identification rate was again achieved by five listeners (of which two also scored 100% in

the target-present setup). Overall, for both scenarios, out of 15 parades 61.5% of listeners managed to solve 10 or more, 38.5% solved 12 or more correctly, and only 2 listeners (0.5%) succeeded in all.

Second, we were interested in whether any of the collected demographic data corresponded with identification rates; however, none of respondents' sex, age, education, nor level of English proved significant. Note that education was treated as a binary variable, with participants either with or without linguistic background.

Third, we explored whether respondents were able to perform better (i.e., scored a higher number of correct answers) in a specific type of mismatch. To find out, all parades were divided into target-present and target-absent groups and according to mismatch type (none, temporal, language, double). The results are shown in Table 1.

Table 1. Percentages of all correct answers for all line-up types. * marks comparisons with the no-mismatch target-present condition (°) which turned out significant ($p < 0.05$).

TARGET-PRESENT SCENARIO	
mismatch type	**correct answers**
none °	98.3%
temporal	79.5%
language *	60.3%
double *	56.4%

TARGET-ABSENT SCENARIO	
mismatch type	**correct answers**
none	66.7%
temporal	64.1%
language *	61.5%
double *	46.2%

To assess the statistical significance of the reported relationships, we used R (R Core Team, 2022) and the *lme4* (Bates et al., 2015) and *afex* (Singmann et al., 2022) packages to perform a mixed effects logistic regression analysis (bobyqa optimizer) of correct and incorrect responses (correct responses include hits and correct rejections). As fixed effects, we entered Target and Mismatch with an interaction term into the model. As random effects, we used intercepts for Suspect and Participant, as well as by-Participant random slopes for the effect of Target. P-values were obtained by performing pairwise post-hoc tests with Tukey method of p-value adjustment for comparing a family of 8 estimates using the *emmeans* package (Lenth, 2022). We found significant differences ($p < 0.05$) between target-present line-ups without any mismatch (marked ° in Table 1) and the four parade groups which are marked with an asterisk in the table.

Finally, we wanted to see whether it is true that the more times the listener played recordings in a parade, the more likely it was for them to answer correctly. As Figure 5 reveals, this is not the case. For target-absent line-ups, no correlation was found (Pearson correlation coefficient $r = 0.18$; $p = 0.707$). On the other hand, for target-present parades, we discovered a strong negative correlation between the number of playbacks and correct answers ($r = -0.92$; $p = 0.00138$). In other words, repeated playback of the voices in the parade did not result in higher accuracy of the listeners; on the other hand, it strongly correlated with the listeners' decision-making uncertainty in line-ups featuring a lower successful identification rate.

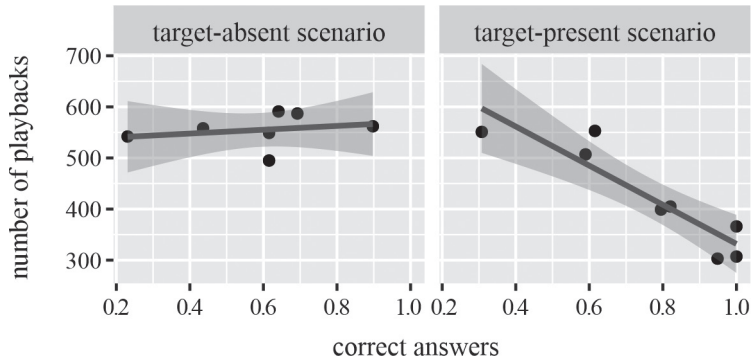

Figure 5. Relationship between the number of replayed recordings and correct answers.

4. Discussion

In this paper, we aimed to find whether mismatched recordings have any impact on speaker identification performance of an ASR system and of human listeners. Our data clearly prove, in line with previous research, that once voices come from mismatched sources, performance does worsen.

As for ASR, we expected that the global validity of the system would decline with increasing "dissimilarity" of the compared datasets. Since the span between the English recordings was only four months, we believed that the speakers' already advanced level of English had not improved considerably, and thus their production remained similar; nevertheless, it was certainly possible for the temporally mismatched recording to have been affected by other changes, such as illness. On the other hand, we supposed that speakers had used different phonetic settings and produced acoustically different phones and prosody in Czech and English. Then, it seemed logical that the combination of these two mismatches would reflect in the results.

For the original datasets, divided datasets, and adapted datasets comparisons (by both i- and x-vectors), our assumptions were confirmed although the difference in EER between language and temporal mismatch was relatively small. The exceptions are set 1 under i-vectors and set 2 calibrated with set 1 under x-vectors when VOCALISE performed slightly better in language than temporal mismatch. As discussed in Section 3.1, system tuning by means of condition adaptation did not always turn out to be beneficial.

Regarding our perception experiment, we added a no-mismatch condition (representing the most similar type on our scale) and wondered whether listeners as well would confirm our "dissimilarity" hypothesis stated above. Even though the percentages of correct answers seemed promising (see Tab. 1), statistically we were able to confirm only a fraction of significant pairs. It would be interesting to replicate the experiment with a higher number of respondents to establish that the more complex the mismatch is, the less successful in speaker identification people are.

It is worth mentioning that there were considerable identification differences amongst individual participants. Whereas we witnessed two "super-recognizers" who solved all 15 parades, there were seven respondents who responded incorrectly in more than 50%

of line-ups. It is worth pointing out that the two successful participants are university students of phonetics; and one of them managed to complete the experiment in 15 minutes (compared to the mean of 26.3) without even having to listen to all suspect recordings in six of the fifteen line-ups.

There were more listeners who did not need to listen to all suspect voices: over 56% of respondents correctly chose a target in at least one parade in this manner (and 28% in two or more parades). Clearly, the experiment featured speakers whose voice characteristics were so salient that it was not necessary for the respondent to continue listening to others. In fact, two speakers were identified in 100% and one in 95% of cases. Conversely, there were two speakers that were much more difficult to recognize – only in 23% and 31% of cases. Also, we registered a suspect speaker who was incorrectly identified as the perpetrator in 37 cases in the target-absent scenario. Instrumental analyses of these speakers' voices could reveal further details as to why he is easily mistaken for other speakers; however, this is beyond the scope of this study.

To conclude, we have shown that ASR systems perform noticeably worse when analyzing voices recorded in different languages and at different times. Nevertheless, in our perception experiment the listeners' ability to identify the perpetrator also dropped considerably as compared to recognizing matched voices. The comparison of known and unknown recordings originating from mismatched sources is far from trivial and it is something of which forensic experts, when drawing conclusions, should be aware.

Acknowledgements

This study was supported by the project the Grant Schemes at CU, reg. no. CZ.02.2.69/0.0/0.0/19_073/0016935.

REFERENCES

Alexander, A., Dessimoz, D., Botti, F., & Drygajlo, A. (2005). Aural and automatic forensic speaker recognition in mismatched conditions. *International Journal of Speech, Language and the Law*, 12(2), 214–234.

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

Bortlík, J. F. (2021). *Czech accent in English: Linguistics and biometric speech technologies*. Palacký University Olomouc. (unpublished PhD dissertation)

de Jong-Lendle, G., Nolan, F., McDougall, K., & Hudson, T. (2015). Voice lineups: A practical guide. In: *Proceedings of ICPhS 2015*, paper 0598.

Earnshaw, K. (2021). Examining the implications of speech accommodation for forensic speaker comparison casework: A case study of the West Yorkshire Face vowel. *Journal of Phonetics, 87*, 101062.

Eriksson, A. (2010). The disguised voice: Imitating accents or speech styles and impersonating individuals. In: Llamas, C., & Watt, D. (Eds.), *Language and identities* (pp. 86–96). Edinburgh University Press.

Eriksson, E. J., Rodman, R. D., Hubal, R. C. (2007). Emotions in speech: Juristic implications. In: Müller, C. (Ed.), *Speaker classification I* (pp. 152–173). Springer-Verlag.

Gold, E., Ross, R., & Earnshaw, K. (2022, accepted). Within-speaker variation: Speaker-based causes. In: Nolan, F., McDougall K., & Hudson, T. (Eds.), *Oxford handbook of forensic phonetics*. Oxford University Press.

Guillemin, B. (2022, accepted). Within-speaker variation: External causes. In: Nolan, F., McDougall K., & Hudson, T. (Eds.), *Oxford handbook of forensic phonetics*. Oxford University Press.

Hansen, J. H. L., & Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine, 32*(November), 74–99.
Hollien, H., & Schwartz, R. (2000). Aural-perceptual speaker identification: Problems with noncontemporary samples. *Forensic Linguistics*, 7(2), 199–211.
Hughes, V., Harrison, P., Foulkes, P., French, P., & Gully, A. J. (2019). Effects of formant analysis settings and channel mismatch on semi-automatic forensic voice comparison. In: *Proceedings of ICPhS 2019*, 3080–3084.
Jessen, M. (2009). Forensic phonetics and the influence of speaking style on global measures of fundamental frequency. In: Grewendorf, G., & Rathert, M. (Eds.), *Formal linguistics and law* (pp. 115–139). Mouton de Gruyter.
Kelly, F., Forth, O., Kent, S., Gerlach, L., & Alexander, A. (2019). Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors. Presented at the *Audio Engineering Society (AES) Forensics Conference 2019*, Porto, Portugal, 2019. Retrieved from https://www.aes.org/e-lib/browse.cfm?elib=20477
Kelly, F., & Hansen, J. H. L. (2015). Evaluation and calibration of short-term aging effects in speaker verification. In: *Proceedings of Interspeech 2015*, 224–228.
Lenth, R. (2022). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.7.4-1, <https://cran.r-project.org/package=emmeans>.
McDougall, K., & Duckworth, M. (2018). Individual patterns of disfluency across speaking styles: A forensic phonetic investigation of Standard Southern British English. *International Journal of Speech, Language and the Law*, 25(2), 205–230.
Misra, A., & Hansen, J. H. L. (2014). Spoken language mismatch in speaker verification: An investigation with NIST-SRE and CRSS BI-LING corpora. In: *Proceedings of the IEEE Spoken Language Technology Workshop*, 372–377.
Morrison, G. S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice, 51*(3), 91–98.
Neuhauser, S., & Simpson, A. P. (2007). Imitated or authentic? Listeners' judgements of foreign accents. In: *Proceedings of ICPhS 2007*, 1805–1808.
R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
Rhodes, R. (2017). Aging effects on voice features used in forensic speaker comparison. *International Journal of Speech, Language and the Law*, 24(2), 177–199.
Rogers, H. (1998). Foreign accent in voice discrimination: A case study. *Forensic Linguistics, 5*(2), 203–208.
Ross, S., Earnshaw, K., & Gold, E. (2019). A cautionary tale for phonetic analysis: the variability of speech between and within recording sessions. In: *Proceedings of ICPhS 2019*, 3090–3094.
Růžičková, A., & Skarnitzl, R. (2017). Voice disguise strategies in Czech male speakers. *Acta Universitatis Carolinae – Philologica 3, Phonetica Pragensia XIV*, 19–34.
Scherer, K. R. (2019). Acoustic patterning of emotion vocalization. In: Frühholz, S., & Belin, P. (Eds.), *Oxford handbook of voice perception* (pp. 61–91). Oxford University Press.
Schiller, N. O., Köster, O., & Duckworth, M. (1997). The effect of removing linguistic information upon identifying speakers of a foreign language. *International Journal of Speech, Language and the Law, 4*(1), 1–17.
Shriberg, E., & Scheffer, N. (2009). Does session variability compensation in speaker recognition model intrinsic variation under mismatched conditions? In: *Proceedings of Interspeech 2009*, 1551–1554.
Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. (2022). afex: Analysis of Factorial Experiments. R package version 1.1-1, <https://cran.r-project.org/package=afex>.
Skarnitzl, R., Asiaee, M., & Nourbakhsh, M. (2019). Tuning the performance of automatic speaker recognition in different conditions: Effects of language and simulated voice disguise. *International Journal of Speech, Language and the Law, 26*(2), 209–229.
Stoet, G. (2010). PsyToolkit – A software package for programming psychological experiments using Linux. *Behavior Research Methods*, 42(4), 1096–1104.
Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, 44(1), 24–31.
Šturm, P., Skarnitzl, R., & Nechanský, T. (2021). Prosodic accommodation in face-to-face and telephone dialogues. In: *Proceedings of Interspeech 2021*, 1444–1448.

Sullivan, K., & Schlichting, F. (2000). Speaker discrimination in a foreign language: First language environment, second language learners. *Forensic Linguistics, 7*(1), 95–111.
Torstensson, N., Eriksson, E. J., & Sullivan, K. P. H. (2004). Mimicked accents – Do speakers have similar cognitive prototypes? In: *Proceedings of SST2004: the 10th Australian international conference on speech science and technology*, 271–276.

Tomáš Nechanský
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
E-mail: tomas.nechansky@seznam.cz

Tomáš Bořil
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
E-mail: tomas.boril@ff.cuni.cz

Alžběta Houzar
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
E-mail: alzbeta.houzar@ff.cuni.cz

Radek Skarnitzl
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
E-mail: radek.skarnitzl@ff.cuni.cz