Earwitness evidence accuracy revisited: Estimating age, weight, height, education, and geographical origin

Adrian Leemann, Péter Jeszenszky, Carina Steiner, Hannah Hedegard
University of Bern, Switzerland

In forensic casework, there are instances where a victim has heard rather than seen the perpetrator – such as when victims are blindfolded and robbed, and police have independently identified a suspect through other investigative means. Voice line-ups can be constructed (cf. Nolan & Grabe 1996), testing the witness' ability to identify the suspect from a roster of voices. There are contrastive approaches regarding the selection of the foils, specifically whether it should be based on the suspect's voice or the witness' description of it. Proponents of the latter method stipulate that the foils 'meet the speech profile of the suspect […] and be matched for features such as biological and social gender, perceived age, accent and dialect' (Broeders and Van Amelsvoort 2001:241). More recent research advocates the opposing method (cf. Nolan 2003; de Jong-Lendle et al. 2015), that witness statements should not play a role in the construction of the line-up. Indeed, studies have shown that naïve listeners' estimations of age, weight, height, geographical origin, and education are subject to substantial variation in accuracy (cf. Gonzalez 2003, Tomkinson & Watt 2018). What remains to be seen, however, are a) the accuracy of these estimations in relation to each other, because these studies have all been independent, and b) other factors such as sexual orientation or physical size, that may be used by witnesses to describe the voices they heard. Without necessarily supporting either method of foil selection, this study seeks to yield more comprehensive information regarding the ability of lay people to identify social characteristics in voices, an existing knowledge gap that contributes to the controversiality and questioned validity of this forensic application.

More broadly, there are several other factors that affect both approaches, about which we still know relatively little about, and therefore by extension the degree to which they render earwitness evidence error prone and inaccurate (e.g. decay of long-term auditory memory; exposure length and (un)familiarity of speakers etc., cf. Saslove & Yarmey 1980, Stevenage et al. 2011, Öhman et al. 2013). This study also seeks to shed more light on the third issue listed, by investigating any potential correlation between accuracy of dialect identification and extent of dialect awareness.

In this contribution – a proof-of-concept study – we examined how 36 naïve listeners from across German-speaking Switzerland rated a short snippet of speech (one sentence of c. 20 syllables read aloud) from 16 speakers of the SDATS database (Leemann et al. 2020) in terms of age, height, weight, geographical origin, and educational background. The 16 speakers (8F, 8M) were selected so as to be maximally diverse in terms of age, geographical origin, educational background, height, and weight. The 36 listeners (26F, 10M; mean age 31, SD=17) came from the cantons of Aargau, Bern, Luzern, Graubünden, St. Gallen, Schaffhausen, Solothurn, Schwyz, Wallis, and Zurich. A large proportion of them had a higher education degree (50%), marginally fewer a school-leaving certificate (Matura) (42%), and the remainder a vocational school degree (8%). Each listener was able to play the sentence multiple times via a web-interface (see Fig. 1). Sentences were presented in random order. For age, height, and weight we measured the mean absolute difference (rather than the 'average difference', which would cancel out deviations from true values). For geographical origin, the listeners had to guess the canton of origin. We measured the distance in km between their guess and the true canton of origin, based on the geographic centroids of cantons.

Results revealed the following patterns: for age (Fig. 2), there is an average difference between estimation and true age of 12.8 years. For height and weight (Figs. 3 & 4), the average differences amount to 4.3cm and 7.7kg. For geographical origin (Figs. 5 & 6), the average difference between estimated and true origin is 49.2km. Finally, for educational background (Fig. 7), results showed that for ten out of 16 speakers, education estimations are above or right at chance level (chance level at 25%). For six speakers, guessing the speaker's educational level was below chance.

In terms of a (at this stage impressionistic) ranking of accuracy, then, it appears that estimations of age and geographical origin deserve strong caution, while height and weight guesses are surprisingly more accurate. Note, though, that height and weight guesses are strongly related to the gender of the speaker, which is almost unanimously guessed correctly (and therefore not shown in graphs). The fact that for a majority of speakers, listeners were able to assign educational background above chance is remarkable in the context of Swiss German: typically, it is assumed that variation between speakers is horizontal, i.e. geographically determined – and not vertical (i.e. socially determined – with 'educational background' as a proxy). Perhaps between-speaker variation in reading competence in dialect was a tell-tale sign of the speakers' educational background. In terms of geographical origin, it is conceivable that the inaccuracy of estimations may have to do with the short exposure to the audio samples in the current design. Estimations of geographical origin would likely have been higher, had there been a longer exposure phase.

In future, this study will be expanded to a substantially larger speaker and listener set, using longer exposure times. This will consolidate (or refute) some of the findings of this first pilot study. In terms of ranking of parameters, we will also need to find ways of how the estimations reported can be compared to each other in a more nuanced fashion. This research could form the basis of further study into how auditory memory decay affects the ability to identify region, age, educational background etc., as some parameters may be more resistant than others. In terms of implications for voice line-ups, this study's results indicate that several parameters are less reliable than others when taking witness' testimony into account, namely geographical origin and age.

References

Broeders, A., & van Amelsvoort, A. (2001). A practical approach to forensic earwitness identification: constructing a voice line-up. Problems of Forensic Sciences, 47, 237-245.

de Jong-Lendle, G., Nolan, F., McDougall, K., & Hudson, T. (2015). Voice lineups: a practical guide. In *Proceedings of the International Congress of Phonetic Sciences,* p. 10-14.

Gonzalez, J. (2003). Estimation of speakers' weight and height from speech: A re-analysis of data from multiple studies by Lass and colleagues. Perceptual and motor skills, 96(1), 297-304.

Leemann, A., Jeszenszky, P., Steiner, C., Studerus, M., Messerli, J. (2020). SDATS Corpus – Swiss German Dialects Across Time and Space. Retrieved from osf.io/s9z4q

Nolan, F. (2003) A recent voice parade. Forensic Linguistics 10(2), 277-291.

Öhman, L., Eriksson, A., & Granhag, P. A. (2013). Angry voices from the past and present: Effects on adults' and children's earwitness memory. Journal of Investigative Psychology and Offender Profiling, 10(1), 57-70.

Saslove, H., & Yarmey, A. D. (1980). Long-term auditory memory: Speaker identification. Journal of Applied Psychology, 65(1), 111.

Stevenage, S. V., Howland, A., & Tippelt, A. (2011). Interference in eyewitness and earwitness recognition. Applied Cognitive Psychology, 25(1), 112-118.

Tomkinson, J. and Watt, D. (2018). Assessing the abilities of phonetically untrained listeners to determine pitch and speaker accent in unfamiliar voices. Language and Law/Linguagem e Direito, 5(1), 19–37.

Fig. 1: Testing interface.

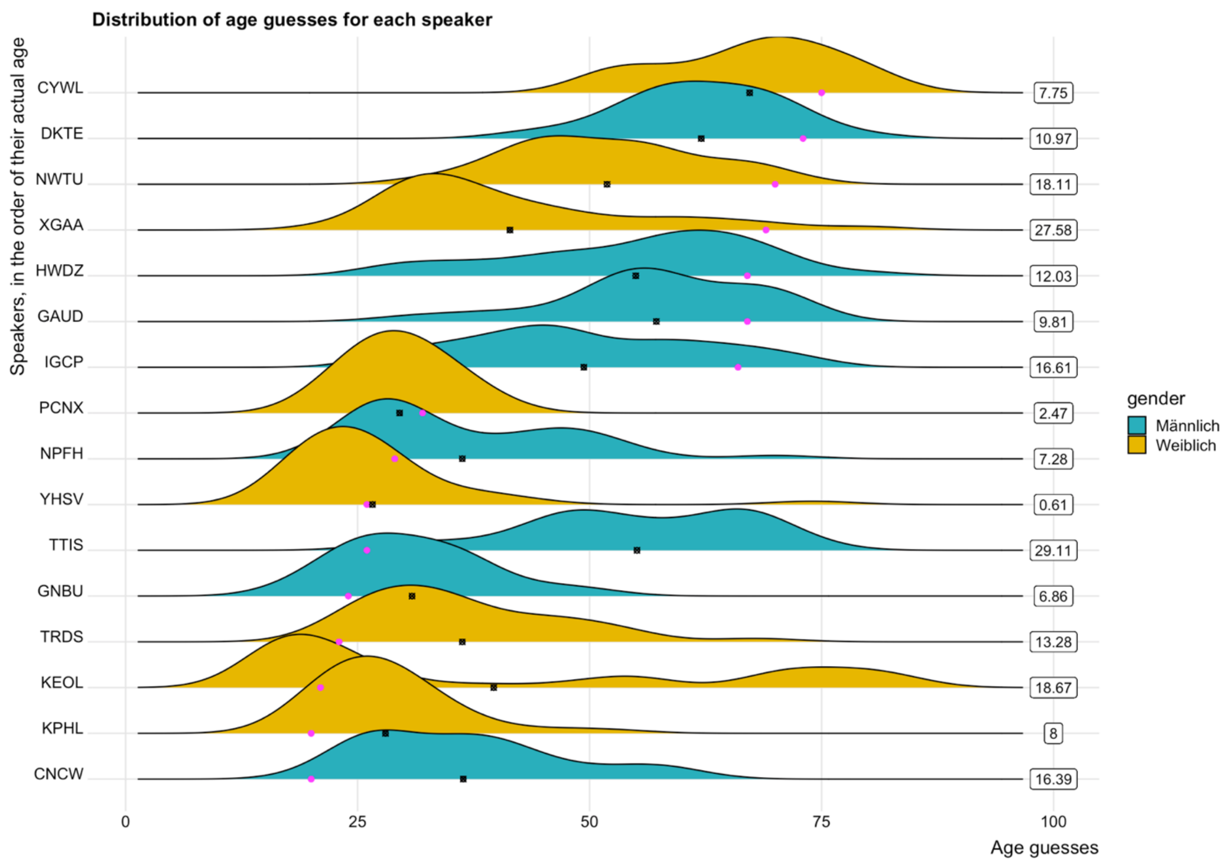

Fig. 2: Density plot showing age guesses for each speaker. The pink dot shows true age; the black dot shows the average estimation. The number on the right shows the deviation from pink to black dot in years.

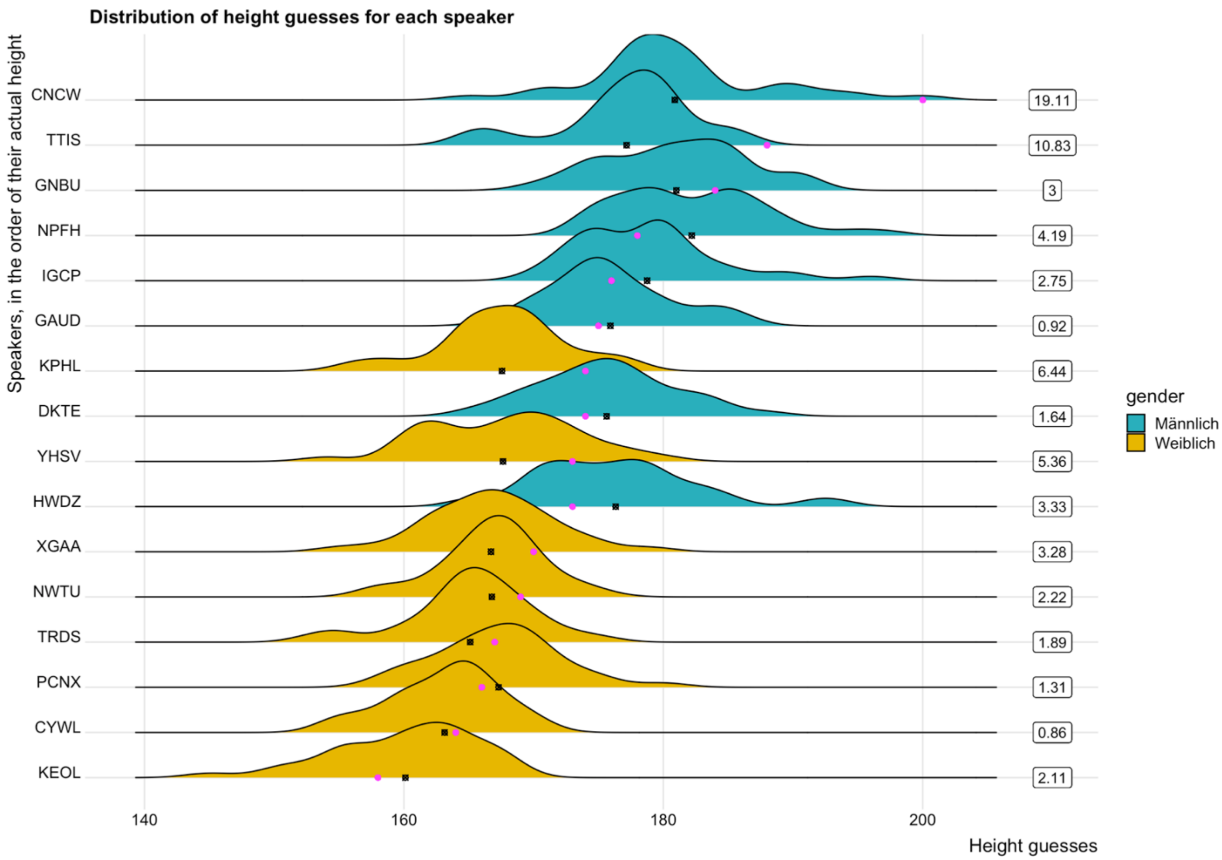

Fig. 3: Density plot showing height guesses for each speaker. The pink dot shows true height; the black dot shows the average estimation. The number on the right shows the deviation from pink to black dot in cm.

Fig. 4: Density plot showing weight guesses for each speaker. The pink dot shows true weight; the black dot shows the average estimation. The number on the right shows the deviation from pink to black dot in cm.

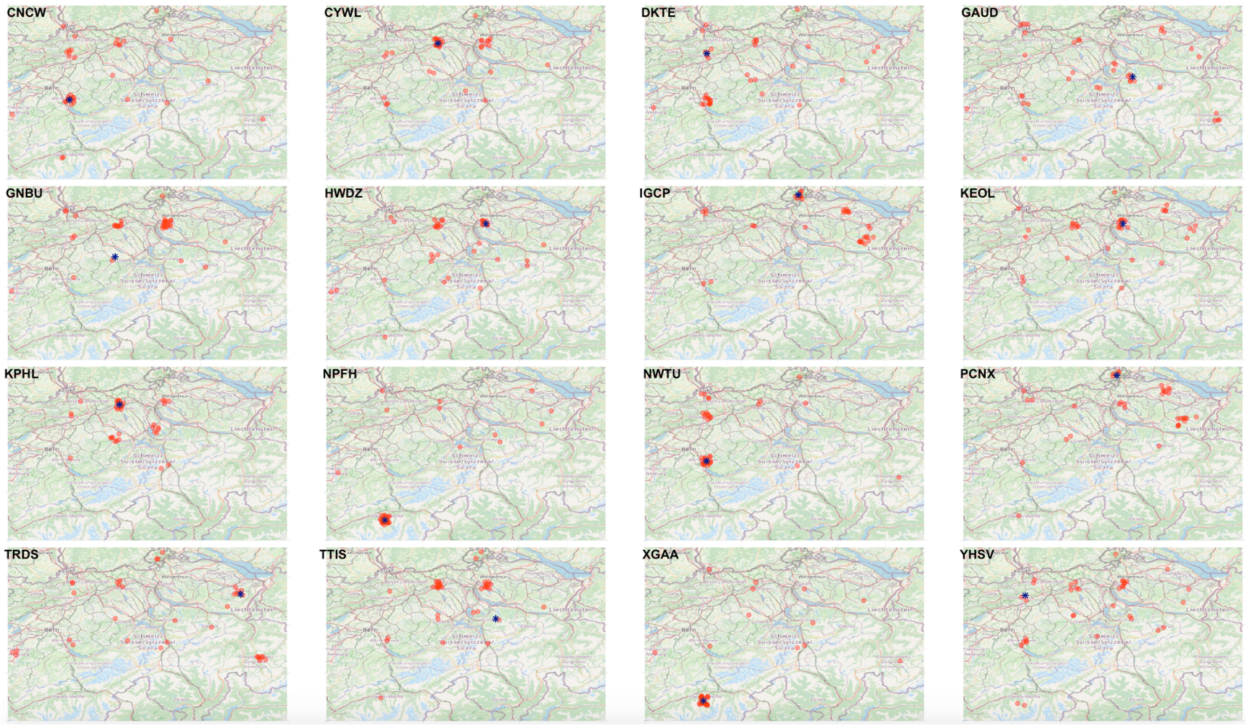

Fig. 5: Map showing true geographic origins (asterisks) and guessed geographical origins (red dots).

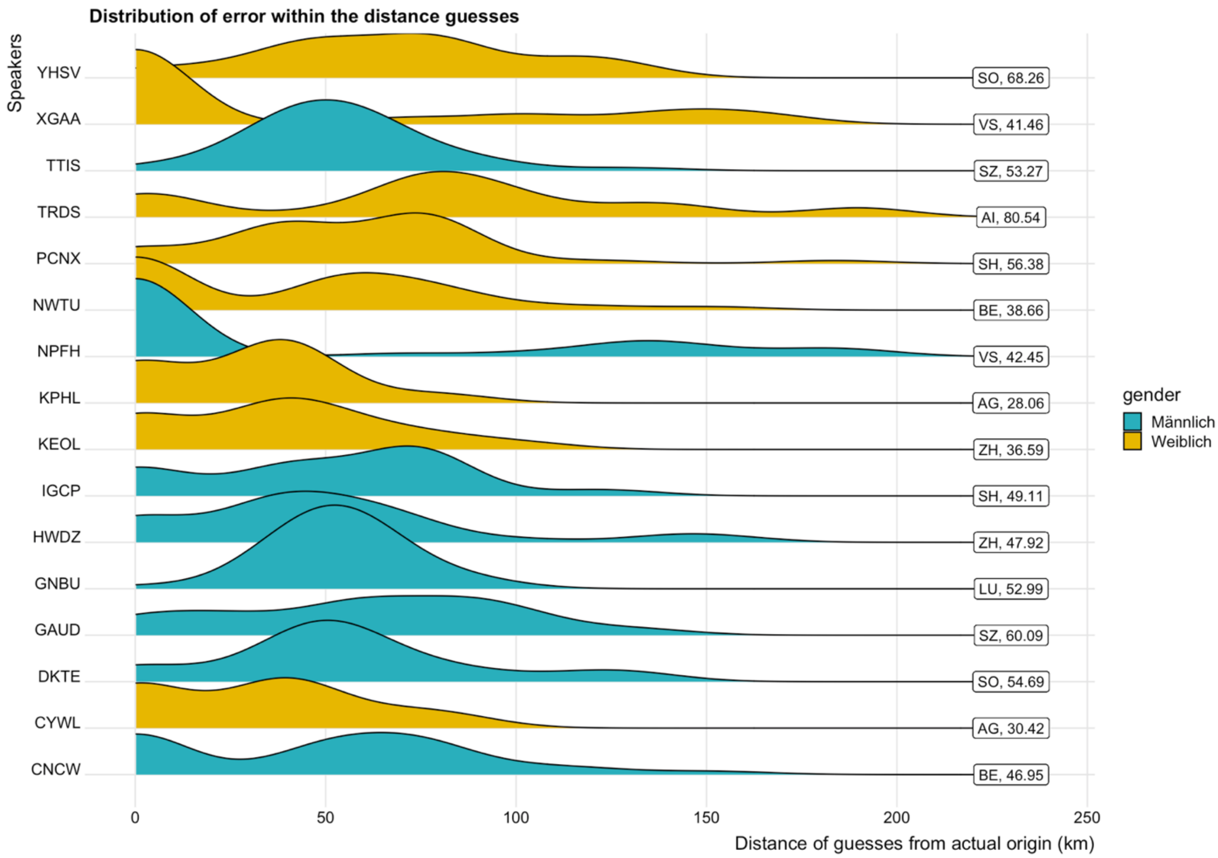

Fig. 6: Density plot showing the distribution of errors of distance guesses. The label on the right shows the true canton of origin and the deviation from true to estimated origin in km (cantonal centroid as benchmarks).

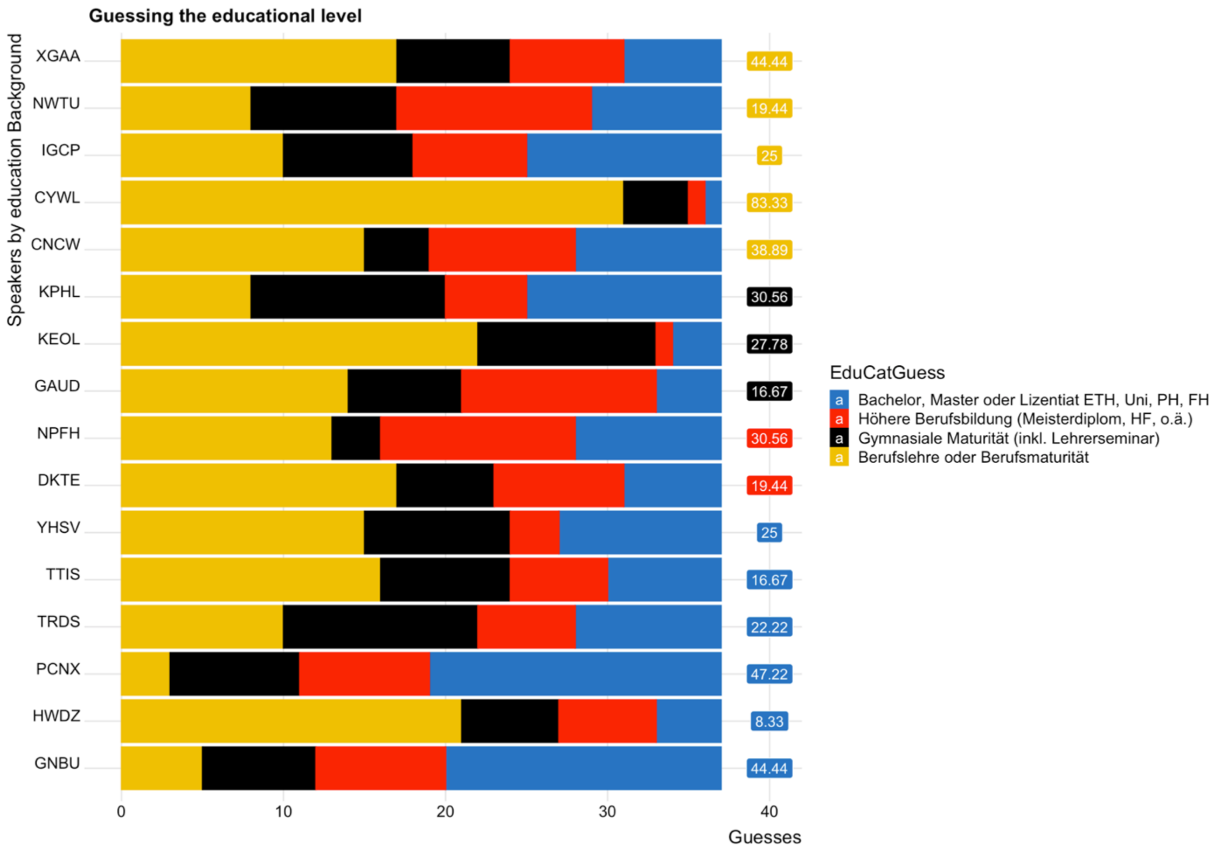

Fig. 7: Stacked bar plot showing guesses of educational background. The number on the right shows the % of correct responses (the color coding of the box represents the true educational background).