The quest for speaker individuality – a challenge for forensic phonetics

Angelika Braun, University of Trier, Germany

brauna@uni-trier.de

This contribution takes a principled approach towards speaker individuality with the forensic application of phonetics in mind. It is argued from the perspective of a forensic practitioner with extensive casework experience. The question of individuality can be split up into two sub-questions: (a) is "speaking" individual, and (b) can the individuality be detected under forensic circumstances? In examining (b), a further question is how the human listener and the computer deal with detecting speaker individuality. These issues will be addressed consecutively.

Intuitively, there is little doubt that speaking is highly speaker specific and that it is among the elements which lend themselves to be used as a biometric. Indeed, listeners are generally able to recognize familiar speakers even when they whisper or use falsetto voice (Braun / Kraft 2013). The pivot point of this view is the notion of "voiceprint" as an obvious analogy to fingerprint, which was set out by Lawrence Kersta in 1962 (Kersta 1962) and is still widespread among lay listeners today.[1] This fingerprint analogy fails to recognize that – other than fingerprints – speaking is highly variable even within one speaker. Illnesses affecting the speech organs in general and the larynx in particular may render a speaker hard to recognize even by his close friends and relatives. The same is true for extreme emotional states and, e.g., excessive consumption of substances like alcohol or psychodelic drugs.

Speaking forms part of human *behavior* rather than being a mere consequence of anatomical configurations like e.g., the size of the resonance cavities within the vocal tract. Francis Nolan has coined the term *plasticity of the vocal tract* (Nolan 1983, pp. 27-28) in order to describe the within-speaker variation caused by, e.g., pulling the larynx upwards in a stressful situation and thereby reducing the size of the pharyngeal cavity or by protruding one's lips and thereby enlarging the oral cavity.

We can draw the interim conclusion that speaker individuality is still an essentially unresolved issue. There are anatomical limitations within which a speaker can produce "sound"[2]. It is within that range that all speaking behavior takes place. The trouble is, however, that neither the anatomical basis nor the behavioral component is constant. The anatomical/physiological basis changes as soon as the speaker e.g. catches a cold and gets congested; behavioral patterns may change with the situation or a second speaker involved. All of this does not yet include a deliberate change of voice and speech features, i.e. disguise. Consequently, there is a plethora of combinations between speaker anatomy/physiology and behavioral factors including the possibility that between-speaker variability may be smaller than within-speaker variability under certain circumstances (Bolt et al. 1976).

This brings us to the second question, i.e. the robustness of speaker specific features to certain factors which may be present in the forensic setting. In addition to behavioral factors, these include technical issues like telephone transmission or signal reduction by way of coding, such as MP3. Thus the features to rely on in the forensic setting not only have to be speaker specific but also resistant to technical issues like transmission and coding.

---

[1] The present author was confronted with the fingerprint analogy when recently being interviewed for a popular science TV program, and it proved quite difficult to convince a well-known TV host that the notion of *voiceprint* is invalid.

[2] For reasons of brevity, I take this to include respiratory, phonatory, articulatory and linguistic features alike.

It has been suggested in the past that automatic SR systems may not be particularly sensitive to intraspeaker variability caused by e.g. emotional states because the Mel frequency cepstral coefficients (MFCCs) which many of them rely on are said to represent the vocal tract anatomy. If, however, vocal tract configuration cannot assumed to be a constant, then systems relying on it should be expected not to perform very well if e.g. the speaking situation changes. This actually seems to be the case: Automatic systems are not only extremely susceptible to channel mismatch (Becker 2011), but also to behavioral mismatch and cannot deal with disguise at all (González Hautamäki et al. 2017; 2019). This may lead to totally erratic results and inexplicable errors.

Another issue which is affected by these considerations is the phrasing of conclusions. For about the past 20 years, expressing conclusions in terms of likelihood ratios has been a frequent demand because the traditional probability scales were said to be logically and statistically flawed (e.g. Champod and Meuwly 2000). LRs have a reputation of being more "objective" in that they render one numeric measure based on extensive statistical background data. While this is certainly correct from a strictly statistical point of view, there are practical considerations which raise some questions. One of them is whether LRs are really so unambiguous or whether they would not have to be tailored to the emotional and physiological states of the speaker. This implies that there could be more than one LR per recording, depending on which background data are used. – On a different strand, LRs do not allow for attaching weight to certain findings. For instance, the forensic practitioner will easily identify findings which will effectively rule out identity. These, however, are not reflected as such in the LRs, which may lead to absurd results of, e.g., two speakers from different parts of the country being taken to be one and the same. The trained human listener is in a much better position to compensate for behavioral and technical mismatches as well as consider parameter salience, and s/he will work these into her/his conclusion on a given probability scale. We may after all have to accept the thought that the "objectivity" of LRs can be challenged on some counts and that the much criticized probability scales, which admittedly involve subjective judgement on the part of the expert, do have some forensic merit.

References:

Becker, T. *(*2011): *Automatischer forensischer Stimmenvergleich*. PhD. Diss. Trier.

Bolt, R.H./ Cooper, F.S. / Green, D.M./ Hamlet, S.L./ Hogan, D.L./ McKnight, J.G./ Pickett, J.M. / Tosi, O./ Underwood, B.D. (1979): *On the Theory and Practice of Voice Identification*. Washington, D.C.: National Academy of Sciences.

Braun, A. / Kraft, L. (2013): "Die Erkennbarkeit vertrauter Stimmen bei Verstellung", In: Mehnert, D. / Kordon, U. / Wolff, M. (Hg.): *Systemtheorie. Signalverarbeitung. Sprachtechnologie. Rüdiger Hoffmann zum 65. Geburtstag.* Dresden: TUDpress, pp. 226-233.

Champod, C. / Meuwly, D. (2000): "The inference of identity in forensic speaker recognition", *Speech Communication* 31, 193-203.

González Hautamäki, R., Sahidullah, M., Hautamäki, V., and Kinnunen, T. (2017). "Acoustical and perceptual study of voice disguise by age modification in speaker verification," *Speech Communication* 95, 1–15.

González Hautamäki, R. Hautamäki, V. and Tomi Kinnunen (2019): "On the limits of automatic speaker verification: Explaining degraded recognizer scores through acoustic changes resulting from voice disguise", *The Journal of the Acoustical Society of America* 146, 693-704.

Jessen, Michael (2008): "Forensic Phonetics", *Language and Linguistics Compass* 2, 1-41.

Kersta, Lawrence G. (1962): "Voiceprint Identification", *Nature* 196: 1253-7.

Nolan, Francis (1983): *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press.