EXAMPLES OF CASEWORK IN FORENSIC SPEAKER COMPARISON

Isolde Wagner

Bundeskriminalamt
isolde.wagner@bka.bund.de

ABSTRACT

Forensic speaker comparison, as carried out at the German federal criminal police office 'Bundeskriminalamt' (BKA), is an accredited method that combines auditory phonetic-linguistic analysis and acoustic procedures of digital audio processing. Auditory phonetic-linguistic analysis is used to describe audible speech features. Acoustic measurements and calculations are used to quantify auditory perceptions and to detect inaudible features. In the context of acoustic procedures validated forensic automatic and semi-automatic speaker recognition are applied for further objectivity.

Examples of casework on telephone recordings of drug dealers and other criminals illustrate the application of the method and the challenges for the analysis of forensic audio material, which is typically characterized by poor acoustic quality, short duration and mismatches of various factors.

Keywords: Forensic Speaker Comparison, Forensic Automatic Speaker Recognition.

1. INTRODUCTION

The methodology of forensic speaker comparison in Germany was developed in the 1980s for the purpose of identifying unknown voices of telephone interceptions in criminal cases [8]. It was continuously improved until today. Whereas in the first years detailed auditory and predominantly linguistically based analyses and descriptions played a substantial role, in the last three decades, along with technological and digital developments more and more acoustic measurements and calculations were added to the methodology providing additional objectivity, such as time-domain and frequency-domain measurements. Finally, automatic and semi-automatic procedures were introduced into the process of analyses. Since 2008, at the BKA forensic speaker comparison became an accredited standard operation procedure of inspection according to DIN EN ISO/IEC 17020.

In spite of technological progress the specific nature of the forensic material is often characterized by very short duration, non-cooperative speakers and reduced acoustic quality. Therefore, the forensic material still remains a huge challenge in forensic speaker comparison.

2. METHODOLOGY

Forensic speaker comparison has the aim to help answer the question of the same-speaker-hypothesis or different-speaker-hypothesis between unknown and suspected speakers by comparing their acoustic traces on audio recordings. Speech features are analysed in many dimensions on which speakers can be distinguished. As it is expressed in the literature (e.g. [4], [6], [7] and [9]) speaker-discriminatory features should be as independent from one another as possible. The relevant discriminatory information is provided from the relation between intra- and inter-individual speaker variation. All findings are compared and evaluated on the basis of the (dis-) similarity and the typicality of speaker-specific characteristics. After this comparison and evaluation process a conclusion statement is given on a verbal probability scale of identity or non-identity of the speakers.

2.1. Principles

As shown in figure 1 below, the methodology combines auditory phonetic-linguistic perception and descriptions of speech features on the one hand and acoustic measurements and calculations of the speech signal on the other hand. The analyses cover three traditional categories: (1) speech and language, (2) voice and (3) manner of speaking as well as an additional feature in the context of voice that comprises automatic and semiautomatic speaker recognition.

The category 'speech and language' covers phonetic analyses of speech sounds along with linguistic analyses of lexical and grammatical features. As a result, descriptions of native language, dialect, foreign accent and socially or individually distinctive features can be provided.

The category 'voice' comprises vibration characteristics of the vocal folds and characteristics of the vocal tract. Pitch and speech melody is quantified by the acoustical correlates mean fundamental frequency and its variability, like standard deviation or variation coefficient. The latter is not – like standard deviation – correlated with the absolute level of fundamental frequency and thus an independent feature (cf. [5], [6]; for the formula). Voice quality can mainly be described on an auditory basis and is evaluated based on the experience and training of the expert. Due to the limitations of the forensic material with its reduced acoustic quality, micro vibrations of the vocal folds, respectively jitter and shimmer, can hardly be measured. Characteristics of the vocal tract, however, can be captured with formant frequency measurements and calculations of cepstral coefficients.

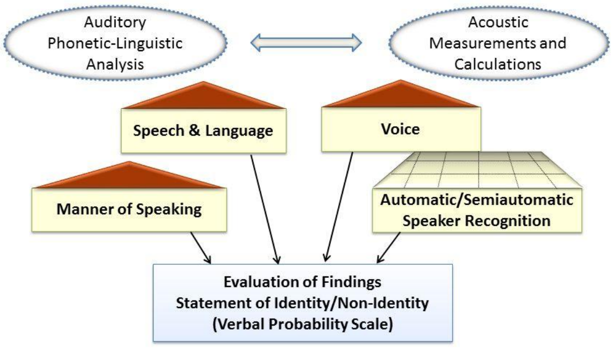
In the context of cepstral coefficients and as an additional methodology in the category of 'voice' forensic automatic and semi-automatic speaker recognition procedures (FASR and FSASR) are applied if the material satisfies the criteria. As described in the Methodological Guidelines for Best Practice in FASR and FSASR the procedures are embedded in the Bayesian interpretation framework ([2]). FASR operates in its central processing stages automatically. These consist of at least feature extraction, feature modelling, similarity scoring and Likelihood Ratio (LR) computation. FSASR operates in its central processing stages partially automatically and partially with human intervention. Human intervention in FSASR focusses on the level of feature extraction: acoustic-phonetic information is measured manually or is supervised based on tracking algorithms. The remaining processing steps proceed automatically and, like in FASR, result in a LR.

At the BKA a variety of different commercial FASR systems are used. The systems are tested regularly on forensic data under different conditions such as duration, type of recording and language. Some results on a German telephone-based forensic corpus called GFS 2.0 (German Forensic Speech corpus) are published in Solewicz et al. ([11]).

The category 'manner of speaking' mainly contains descriptive supra-segmental features, like e.g. articulation precision, speech fluency, some speech disorders, intonation and respiration or speech tempo. The latter can be measured in terms of articulation rate.

In the whole speaker comparison process the knowledge and competence of the expert plays a substantial role. They have to make decisions at every step of the speaker comparison analysis using available background information, as well as their experience. Discriminatory power of the features, as well as case-specific similarity and typicality are evaluated in all the different features analysed. The evaluation of all findings results in a statement on a verbal probability scale of identity or non-identity of speakers. The results of the investigation are documented in an expert report.

Figure 1: Standard operation procedure of forensic speaker comparison at the BKA.


Auditory Phonetic-Linguistic Analysis
Acoustic Measurements and Calculations
Speech & Language
Voice
Manner of Speaking
Automatic/Semiautomatic Speaker Recognition
Evaluation of Findings
Statement of Identity/Non-Identity
(Verbal Probability Scale)

2.2. Validation

The method of forensic speaker comparison is developed and improved since more than three decades. Since ten years its standard operation procedure is regularly validated in inter-laboratory proficiency tests and collaborative exercises ([1]; for an example of a collaborative exercise). Because of the subjective component of the method and for quality assurance and control purposes it is also required that results and opinions are to be examined by a second expert.

2.3. Case Assessment

Before the method of forensic speaker comparison is applied, all the relevant audio material has to be tested as to whether it satisfies the criteria of the analyses and procedures. The criteria concern quantity and quality factors as well as match or mismatch conditions of the speech material in various aspects, such as digital audio format or acoustic environment of the recordings, spoken language, age or situational factors affecting speech behaviour, like e.g. stress, disease, voice disguise or intoxication. The result of this testing determines if and to what extent the audio material is appropriate for the method. It also shows the influence that can be expected on the evaluation and the results of the forensic speaker comparison.

3. EXAMPLES OF CASEWORK

Three examples of casework in different but representative forensic scenarios illustrate the application of the methodology.

3.1. Analysing short material

In a case of drug dealing the recordings of five telephone conversations served as evidence for the criminal act, because the interlocutors talked about details of the drug dealing. One of the interlocutors was unknown for the police. Thus, the recordings were submitted to the forensic science institute of the BKA for an investigation of forensic speaker comparison.

The acoustic quality of the audio material had no remarkable distortions, but the durations of the recordings of the questioned interlocutors were very short to short (5 to 17 seconds of net speech). For the comparison ten recordings of a suspected speaker were provided from the police. These recordings were even shorter (1 to 9 seconds of net speech). Only little information on speech behaviour was expected from these recordings. But despite the very short durations of the material the findings led to an evaluation on a high identification level for all speakers. Numerous strong consistencies could be found between all the questioned interlocutors and the suspected speaker. The questioned interlocutors as well as the suspected speaker spoke German with the same peculiarities of an Albanian accent, showed similar grammatical errors and used lots of identical empty and confirmation phrases in a specific combination. All of those speakers had a peculiar harsh and creaky voice quality, salient respiration and showed a high articulation rate of 6 to 8, locally up to 10 syllables per second.

3.2. The problem of brothers

In another case a member of an extended family clan was accused of criminal activities, where telephone conversations in German and Arabic were involved. Since a private expert opinion addressed only the German recordings, questions were raised during the trial and additional investigations of the Arabic recordings were considered to be necessary. Furthermore, the defence attorneys pointed out that one of the suspects' brothers could also be the speaker in question. Hence, a second expert opinion was necessary. Speaker comparisons for all German and Arabic recordings were requested (1) between the questioned speakers among each other, (2) between the questioned speakers and the suspected speaker and (3) also between the questioned speakers and the brother of the suspected speaker. Finally, it was asked if a forensic speaker comparison can generally come to a reliable result when brothers are involved.

Thirty telephone recordings with different durations (4 to 150 seconds of net speech) and different acoustic qualities (partly distortions, reverberations, background noise) were provided: 12 recordings of the questioned speakers, 5 recordings of the suspected speaker and 13 recordings of the brother of the suspected speaker. There were many mismatches in the conversation situations. Some of the situations involved a calm or sleepy mood some others lead to very loud and excited or upset speech. The variations of the characteristics of some speech features, especially fundamental frequency and voice quality were correspondingly high between the recordings, even in the intra-speaker conditions of the two suspected relatives. But during the analyses it became obvious that the questioned speakers as well as the suspected speaker used a kind of youth slang and a continuous idiosyncratic code switching between the German and the Arabic language. They switched even within small units of utterances from one word to the next. The brother did neither show youth slang nor code switching in any of the 13 recordings, but had strong disfluencies and a lax and mumbled articulation. The utterances in Arabic language were investigated by an Arabic expert, who found two different Arabic dialects between the questioned speakers and the suspected speaker on the one hand and the brother on the other hand.

After the evaluation of all findings the investigation came to the conclusion that there is a very high probability of identity between the questioned speakers among each other as well as between the questioned speakers and the suspected speaker. However, the brother of the suspected speaker could not at all be thought of as one of the questioned speakers. Considering these findings the general question of reliable results in the context of siblings which was raised in court was answered as follows. A forensic speaker comparison can absolutely come to a reliable result with brothers involved. Voices of brothers can in principle be distinct, but discriminability could be rather difficult with similar voices, not only between relatives and especially in poor acoustic quality recordings.

3.3. Integrating an automatic procedure

In a case of illegal financial transactions, where more than one minute of net speech in German telephone recordings was available for each speaker, the procedure of forensic automatic speaker recognition was applied in addition to the auditory phonetic-linguistic and acoustic analyses.

In the first step the two recordings were analysed in the traditional categories. A number of consistencies were found between the questioned and the suspected speaker. Both speakers had some markers of an eastern German dialect, they had a special realisation of a farewell phrase (/tschü-üs/), the voices were monotonous with only little variation and slightly nasal and breathy. The articulation rates were at the upper limit of the population distribution. In the second step an i-vector based calibrated commercial automatic speaker recognition system, which was validated on forensic material, was applied. In this procedure after feature extraction and modelling the similarity between the questioned and the suspected speaker is calculated in relation to the distribution of same-speaker comparisons as well as in relation to the distribution of different-speaker comparisons. The result is given as a LR value or as $\log_{10}$LR. The value represents the strength of evidence. If $\log_{10}$ LR is greater than 0, there is more evidence for identity, if $\log_{10}$ LR is smaller than 0, there is more evidence against identity. In this case the $\log_{10}$ LR was 5. This result supports identity and was evaluated with all other results. The final statement was given on a verbal probability scale.

4. DISCUSSION OF CASEWORK EXAMPLES

The casework examples raise a set of more general issues that are addressed in the following subsections. Since the examples are a mere snapshot of forensic practice, the general issues addressed here are necessarily selective and non-exhaustive as well.

4.1. Duration requirements

Duration of available speech is an important factor in forensic speaker comparison. There can be limits of net speech below which no analysis is possible. However, such limits have to be established separately for each different speaker-discriminatory feature; there should not be any one-size-fits-all duration threshold. For example, the systematic analysis of filled pauses generally requires longer speech passages than the auditory assessment of voice quality. The example in 3.1 shows that even net durations below ten seconds can provide important speaker information. What the example also shows is that foreign accent can be a rich source of idiosyncratic features, as discussed in Jessen ([6]).

4.2. Dealing with non-twin siblings

Although twins tend to show relatively high levels of similarity in their voice and speech patterns (monozygotic more so than dizygotic), non-twin siblings, which are forensically more frequent than twins, often show lower levels of similarity than twins ([10]). Feiser ([3]) shows that non-twin male siblings can quite often be distinguished based on methods commonly used in forensic phonetics and acoustics, such as the analysis of formants, fundamental frequency and speech tempo. Practitioners in our group are often asked in court to make general statements about the problematic or non-problematic nature of siblings' speech. Research as well as practice shows that there can be no uniform answer. In the case example shown, siblings differed strikingly, but there can also be more challenging situations.

4.3. Combining evidence

The casework example in 3.3 mentioned a variety of speaker-discriminatory features, including dialectal patterns, greeting expressions, pitch variability, auditory voice quality, and the results of an automatic (i-vector-based) system. The question arises as to how these features are combined in order to reach a final conclusion to the speaker comparison task. As a special case of this question, it is often asked what consequence for the conclusion it would have if results from an automatic system did not agree with results obtained from traditional auditory-acoustic features. The way of combining evidence practiced at the BKA is roughly as follows: For each of the features found within the four categories shown in Fig. 1 an assessment or quantification is provided of the similarity and typicality of the feature values found within a case, as established in the likelihood ratio framework ([9]). In that sense, FASR is also based on a feature, viz. cepstral parameters. Attention is also paid to the overall speaker-discriminatory performance of the feature at hand – being well aware that the performance of FASR can be very high, especially when comparing between telephone recorded samples. It is possible that in a particular case a feature of the auditory-acoustic categories 'speech & language', 'voice' and 'manner of speaking' – for example a very rare greeting expression – outweighs automatic speaker recognition in the final conclusion, especially when its results are close to the inconclusive area. In other words, we do not see FASR as a separate domain but look at each of the features, their similarity and typicality in a case and their overall performance and then combine the evidence in a qualitative way in order to arrive at a conclusion. Although currently still verbalized as posterior probability (probability of same-speaker and different-speaker hypotheses) the underlying methodology is oriented towards the likelihood ratio framework.

Acknowledgment:
Thanks to my colleague Michael Jessen for his support in the interpretation of the casework examples.

4. REFERENCES

[1] Cambier-Langeveld, T. 2007. Current methods in forensic speaker identification: Results of a collaborative exercise. *The International Journal of Speech, Language and the Law* 14, 223–243.

[2] Drygajlo, A., Jessen, M., Gfroerer, S. Wagner, I., Vermeulen, J. & Niemi, T. 2015. *Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition.* Frankfurt: Verlag für Polizeiwissenschaft.

[3] Feiser, H.S. 2015: *Untersuchung auditiver und akustischer Merkmale zur Evaluation der Stimmähnlichkeit von Brüderpaaren unter forensischen Aspekten.* Frankfurt: Verlag für Polizeiwissenschaft.

[4] Gfroerer, S. 2014. Sprechererkennung und Tonträgerauswertung. In: Widmaier, G. (ed), *Münchner Anwaltshandbuch Strafverteidigung*. Munich: Beck, 2682-2707.

[5] Jessen, M., Köster, O. & Gfroerer, S. 2005. Influence of vocal effort on average and variability of fundamental frequency. *The International Journal of Speech, Language and the Law* 12, 174–213.

[6] Jessen, M. 2012. *Phonetische und linguistische Prinzipien des forensischen Stimmenvergleichs.* Munich: Lincom Europa.

[7] Jessen, M. 2018. Forensic voice comparison. In: Visconti, J. (ed), *Handbook of Communication in the Legal Sphere*. Berlin: Mouton de Gruyter, 219-255.

[8] Künzel, H.J. 1987: *Sprechererkennung. Grundzüge forensischer Sprachverarbeitung.* Heidelberg: Kriminalistik Verlag.

[9] Rose, P. 2002: *Forensic speaker identification.* London: Taylor and Francis.

[10] San Segundo, E. 2014: Forensic speaker comparison of Spanish twins and non-twin siblings: A phonetic-acoustic analysis of formant trajectories in vocalic sequences, glottal source parameters and cepstral characteristics. Ph.D. Dissertation, Consejo Superior de Investigationes Científicas.

[11] Solewicz, Y.A., Jessen, M. & van der Vloed, D. 2017. Null-Hypothesis LLR: A proposal for forensic automatic speaker recognition. Proceedings of INTERSPEECH 2017 (Stockholm), 2849-2853.