Forensic Speaker Recognition

Introduction

The global success of mobile and multimedia communication provides an opportunity for extensive use of audio recordings to search and establish links between individuals and criminal activities. However, establishing these links through forensic speaker recognition often remains the last resort for forensic investigation and evaluation in court, in the absence of other pieces of evidence. The reason lies in the following difficulties: firstly, speech is the result of the combination of physical and behavioral characteristics of the speaker, the behavioral component being prevailing the physical one. This prevents the display of a fixed and known number of *highly* discriminatory features. Secondly, the quality of audio recordings captured in forensic conditions is most of the time far from ideal. Thirdly, the absence of any known underlying model prevents a symbolic representation of the speaker-dependent information, limiting the study to recognition approaches. To tackle these pitfalls, forensic speaker recognition has been established as a multidisciplinary area of study, combining mainly phonetics, linguistics, speech signal processing, and forensic statistics.

Speech as a Trace

Speech is a behavior developed by human beings for communication. The speech production originates in two main stages: the language generation and the voice production. This complex process depends partly on physiological traits, such as the length and shape of the vocal tract and the dynamic configuration of the organs involved in articulation. It also depends on environmental and socio-linguistic factors, for example, the level of education, the dialectal particularities and the linguistic context [1]. Through speech, language can convey up to six communication levels simultaneously (referential, poetic, emotive, conative, phatic, and metalingual). Speech is an inherently transient phenomenon that can only be imperfectly captured as a signal. The actual forensic trace comprises speakers' utterances recorded as an analogue or digital signal on a storage medium. In some cases, information about the voice of interest only exists in the memory of a victim or witness and available in the form of an earwitness testimony.

Where the recording consists of a wiretapped telephone call, the trace is generally recorded in uncontrolled conditions, including undesired variations in signal quality because of background noise, transmission channels, and recording devices. Anonymous calls are generally short, from seconds to minutes. Where it is a monologue, it may be a prerecorded message, possibly modified by a filtering or editing procedure and even be the result of speech synthesis. From a lexical point of view, the themes are targeted, e.g., abuse, extortion, obscenity, and/or threats. Results of wiretapping procedures can reach hundreds of hours of recording. Their lexical content is varied, depending on, e.g., the relation between the interlocutors, the type of conversation, and the emotions involved. Some utterances may also refer to internal codes for groups or organizations.

At the source level, the inference of identity of speakers remains a challenge. This is mainly due to the absence of a fixed and known number of highly discriminatory features in speech and to uncontrolled recording conditions. However, when the question of the inference at source level is solved, inferences at activity and even offense level are, then, often straightforward.

Forensic Speech Analysis

The speaker-dependent features are spread out over the information conveyed in the different levels of speech communication. For speaker recognition purposes, they are classified in higher-level features if they relate to the language generation and lower-level features if they relate to the voice production. Higher-level features refer more to the linguistic and extra linguistic information of speech as a behavior and its understanding by human beings, whereas lower-level features refer more to the speaker's vocal physiology and its processing by automatic approaches [2].

The main requirements for the methods applied by human beings and computers to forensic speaker recognition are the independence of the utterances to the lexical context, named text-independent methods, the ability to handle minimal length recordings, and a superior robustness regarding noise, transmission channel, and variations in recording conditions. Efficient analysis strategies combine the use of both high and low level features: many of the high level features need more speech material than the low level features for a reliable analysis, but they are generally more robust regarding channel variation. Therefore, the methodology for forensic analysis of human speech should ideally combine expert-based and automatic approaches.

Expert-Based Approaches

Expert knowledge has developed around three main approaches: auditory-perceptual, visual, and phonetic-acoustic [3]. The auditory-perceptual approach was developed during the first part of the twentieth century. It consists of a detailed auditory analysis of the parameters of the voice such as the pitch, the timbre, and the voice quality; the parameters of the speech such as the articulation, the diction, the prosody (speaking rate, pauses, intonation), any speech defects; and the parameters of the language spoken. It includes linguistic observations of lexical, phonological, morphological, syntactic, and idiomatic features, the study of the extent and the variation of the dialect, the accent, and the idiolect, and it also considers paralinguistic features such as breathing patterns. The auditory-perceptual approach does not easily lend itself to validation; in this respect, it is clearly insufficient if used alone, and it remains a first step in the analysis [4].

The visual approach was developed in the 1960s in the United States under the name of "voiceprint" technique [5]. It comprised a visual comparison and interpretation of representations of the spectrum of frequencies of the speech signal, the broadband speech spectrograms. This task was performed by "experts", most of whom had no scientific education in a speech science or related area. From its conception, the theoretical and logical grounds of the technique were contested. Firstly, the term voiceprint is a misleading analogy to the fingerprint and its associated qualities of distinctiveness, inalterability, and permanence; none of them apply to the voice and to the speech spectrogram. Secondly, confusion exists between the degree of reproducibility of the production of spectrograms (analysis) and the high degree of uncertainty associated with the individualization process (interpretation), which can be described as closer to art than science. Since the publication of the National Academy of Science (NAS) report in 1979, some US jurisdictions reject the admissibility of the technique although some still allow it. The Daubert decision (1993), advocating a reliability standard for the admissibility of scientific evidence, clearly militates against the use of this kind of nonvalidated approach, but the US jurisdictions have not adopted a uniform practice in this matter yet [6, 7].

The acoustic-phonetic approach was developed during the 1980s [8]. The methodology consists of a set of instrumental measurements of particular acoustic speech parameters present in the segmental and suprasegmental levels of the speech. The segmental analysis focuses on the duration and the spectral distribution of energy of the segments, the vowel formant frequencies (resonances of the vocal tract), their trajectory and, more recently, their dynamics. The suprasegmental analysis concentrates on the measurements of the long-term features, for example, the long-term average spectrum of the speech signal, the long-term distribution of formants, and different parameters of the fundamental frequency, such as the long-term F_0, the mean F_0, and the deviation of F_0. Once measured, the parameters are interpreted using statistical information on their distributions [9]. Procedures to set up earwitness line-ups have been designed in order to minimize the biases when auditory testimonies are exploited for recognition purpose [10, 11].

Automatic Approaches

Attempts to use computers for forensic speech analysis started in the 1970s, first using semiautomatic methods and from the 1990s using text-independent automatic methods [4]. These automatic methods rely on speech processing, to extract speaker-dependent feature vectors from the speech signal, on pattern recognition to compute comparison scores between feature vectors and on Bayesian statistics to estimate the evidential value of these comparison scores [12].

Current automatic speaker recognition systems mainly rely on low level speaker-dependent features, extracted from the short-time spectral level. The spectrum of the speech signal, analyzed in short-term windows, is directly related to the shape of the vocal tract, which presents speaker dependencies. The spectral envelope is described efficiently using linear predictive coding (LPC) in an all-pole model with 10 to 16 coefficients. However, as these coefficients are correlated, the cepstrum transform has been proposed in order to obtain pseudo-orthogonal linear prediction cepstrum coefficients (LPCC). These coefficients may also be obtained from a perceptually based mel-filter spectral analysis (mel frequency cepstral coefficients – MFCC). Other low-level features, which capture the dynamic speech patterns, such as delta and even delta–delta features have also been proposed, but LPCC and MFCC are the most widely used low level features for automatic speaker recognition. State-of-the-art systems also intend to take advantage from high level speaker-dependent features, mainly contained in the phonotactics, the prosody, and the idiolect. In phonotactics, speaker-dependent information is embedded in the particular use and realizations of the phones and syllables, which presents a highly language-dependent variability. Prosody is the combination of instantaneous energy, intonation, speech rate, and unit duration which all exhibit speaker dependencies. The idiolect contains the information related to the speaker-specific use of a language [13].

Numerous pattern recognition methods have been developed to model and compare the speaker-dependent spectral features, i.e., vector quantization (VQ), ergodic hidden Markov models (E-HMM), artificial neural networks (ANN), and support-vector machines (SVM), but most of the current systems are based on universal background models and Gaussian mixture models (UBM-GMM). The UBM-GMM approach is at the basis of text-independent automatic speaker recognition. It is a generative model where a mixture of multivariate Gaussians model the probability distribution of speech features. The latest developments focus on subspace modeling of speaker and session/channel variability. On the basis of joint factor analysis techniques that simultaneously model the speaker, channel, and residual variability, the current approaches represent speech utterances in terms of i-vectors, low dimensional, and fixed-length representations that preserve the speaker identity information. The hyper space of total variability is formed by merging the speaker, channel-session, and residual variability. Probabilistic linear discriminant analysis then models within and between speaker variations in the i-vectors. Various methods exist to transform the recognition scores to calibrated log-likelihood ratios, such that these systems can be used for forensic evaluation. This is carried out using empirical validation in a Bayesian framework, rather than using an ad hoc normalization scheme, world-models or cohort-based normalizations [14].

Until the last decade of the twentieth century [15], the forensic interpretation of the results given by automatic approaches remained difficult, as solutions concentrated on decision theory and frameworks applied in commercial applications of automatic speaker recognition: speaker verification (1 : 1) or speaker identification (1:N or 1:N+1). Nowadays, the evaluation of forensic evidence in speaker recognition cases is based on the latest developments of forensic statistics. The automatic systems not only deliver likelihood ratios, but the accuracy and the calibration of the probabilities inferred by these likelihood ratio values are tested [13, 16].

Forensic Individualization through Speech

Forensic individualization through speech is a question of inference of identity of source. In essence, the answer to the question that is provided by science will remain inductive and therefore relative to the data analyzed and the hypotheses tested. Nevertheless, the use of a framework based on logic and forensic statistics to interpret the results of forensic speech analysis allows for the most scientifically valid answer to be reached.

With this framework, a practitioner may report logical, robust, and balanced statistics to a court of justice, presenting the strength of the evidence according to the prosecution and the defense hypothesis, in the form of a likelihood ratio. Moreover, the practitioner may combine likelihood ratios assigned using personal probabilities for the auditory-perceptual and acoustic-phonetic approaches with the likelihood ratios assigned more objectively from automatic approaches.

The adoption of this interpretation framework by the practitioners is an ongoing process, but the forensic speaker recognition community has pioneered this matter for a long time. The goal has been, and is, to search actively for solutions that will resolve the logical flaws of an interpretation given in terms of posterior probabilities of common origin for the trace and the source [17].

Validation and Practice

Most of the forensic laboratories still favor one of the two approaches, even if several attempts to combine the results of the expert-based and automatic approaches are ongoing, as they appear to be in many respects complementary. For example, much larger amounts of speech data can be handled automatically than manually, which affects both validation and practice. Currently, the validation of expert-based methods is limited to the assessment of the competence of practitioners using limited datasets, which may be considered as an insufficient procedure [18]. Automatic methods may be validated against defined performance characteristics and metrics using large-scale datasets. This brings a far clearer picture on the possibilities and limits of the automatic than the expert-based method. Automatic methods are also less dependent on the language spoken, making their validation possible for different languages and their versatility superior for practical use [19].

In practice, automatic approaches can be used in about one-third of cases, particularly for the ones with a large amount of speech material. Expert-based methods are claimed to be more flexible for cases with qualitative and quantitative limitations and more robust for speech specimens with strongly mismatched behavioral and technical conditions, or those containing linguistic or dialectal particularities. Finally, as most practitioners are phoneticians or linguists, the expert-based methods are more easily explained in a court of justice, where they often perceive automatic methods as “black boxes” [20].

Research and Development

Improvement in forensic speaker recognition requires a study of how to combine the results obtained with the expert-based and automatic approaches, as it appears that they are complementary. This in turn can only be achieved through the coordination of research and development.

Methods may be developed and implemented to monitor and validate more adequately the inevitable subjective components in the expert-based approaches. The development of double-blind approaches for selecting and grouping speech specimens in the preanalysis phase of cases may reduce the effect of confirmation bias [21]. It may also increase the calibration of the practitioner regarding the voice, the speech, and the language particularities of the speakers involved in the case. Studies to estimate and combine the statistical probabilities of the features used in the expert-based approaches could help the practitioners to improve and calibrate the subjective probabilities they assign to these features on the basis of their training and experience.

The relevance of the features selected may be studied using Bayesian networks and their correlations analyzed with multivariate likelihood ratio approaches. Collaborative exercises and proficiency testing are recognized tools to monitor the expert-based practice, but they encounter two major difficulties in the field of forensic speaker recognition: they are very time consuming and the language dependencies constitute a barrier to their organization at an international level.

The robustness of the automatic approaches to mismatch conditions may also be improved with the development of noise, channel, and recording distortion compensation strategies specifically oriented to the forensic application. Phonetics and linguistics may further contribute to the development of more robust and automatic feature extraction and processing of higher level features from speech specimens of forensic quality. Knowledge of the language dependency of both approaches, expert based, and automatic should also progress, as cases may involve a broad variety of languages.

References

[1] Eriksson, A. (2005). Tutorial on Forensic Speech Science. Part I: Forensic Phonetics. in *Interspeech 2005 – Eurospeech 2005: Proceedings of the 9th European conference on speech communication and technology*, Lisbon, Portugal.

[2] Shriberg, E. (2007). Higher-level features in speaker recognition, in *Speaker Classification I*, Springer, pp. 241–259.

[3] Gold, E. & French, P. (2011). International practices in forensic speaker comparison, *International Journal of Speech, Language and the Law* **18**, 293–307.

[4] Meuwly, D. (2001). Reconnaissance de Locuteurs en Sciences Forensiques: l’apport d’une Approche Automatique. Université de Lausanne, Faculté de droit, Institut de police scientifique et de criminologie.

[5] Bolt, R., Cooper, F., David, E., Denes, P., Pickett, J. & Stevens, K. (1970). Speaker Identification by Speech Spectrograms: A Scientists’ View of its Reliability for Legal Purposes, *J. Acoustic. Soc. Am.* **47**(2), 597–612.

[6] Gruber, J.S. & Poza, F.T. (1995). *Voicegram Identification Evidence*, Lawyers Cooperative Publishing, Rochester, NY, Vol. 54.

[7] Moenssens, A.A., Henderson, C.E. & Portwood, S.G. (2007). Spectrographic voice recognition, in *Scientific Evidence in Civil and Criminal Cases*, Foundation Press, Eagen, MN, 789–815

[8] Nolan, F. (1997). Speaker recognition and forensic phonetics, in *A Handbook of Phonetic Science*, W. Hardcastle & J. Laver, eds, Blackwell Handbooks in Linguistics, Oxford, pp. 744–767.

[9] Morrison, G.S. (2010). in *Forensic voice comparison. Expert evidence (Ch. 99)*, I. Freckelton & H. Selby, eds, Thomson Reuters, Sydney, Australia, pp. 1–105.

[10] Broeders, A.P.A. (1996). Earwitness identification: common grounds, disputed territory and uncharted areas, *Forensic Linguistics* **3**(1), 3–13.

[11] Nolan, F. & Grabe, E. (1996). Preparing a voice lineup, *Forensic Linguistics* **3**(1), 74–94.

[12] Champod, C. & Meuwly, D. (2000). The inference of identity in Forensic speaker recognition, *Speech Communication* **31**(2–3), 193–203.

[13] Ramos Castro, D. (2007). *Forensic Evaluation of the Evidence Using Automatic Speaker Recognition Systems* Universidad autónoma de Madrid.

[14] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P. & Ouellet, P. (2011). Front-End Factor Analysis for Speaker Verification, *IEEE Transactions on Audio, Speech, and Language Processing* **19**(4), 788–798.

[15] Künzel, H.J. (1994). Current approaches to forensic speaker recognition, Proc. *ESCA Workshop on automatic speaker recognition, identification and verification*, Martigny, Switzerland, 135–141.

[16] Brümmer, N. & du Preez, J. (2006). Application-independent evaluation of speaker detection, *Computer Speech and Language* **20**(2–3), 230–275.

[17] Meuwly, D. (2006). Forensic individualisation from biometric data, *Science and Justice* **46**(4), 205–213.

[18] Morrison, G.S. (2013). Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison, *Science & Justice : journal of the Forensic Science Society*. http://www.science and justice journal. com/article/S1355-0306(13)00061-0/abstract.

[19] Van Leeuwen, D., Martin, A., Przybocki, M. & Bouten, J. (2006). NIST and NFI-TNO evaluations of automatic speaker recognition, *Comp. Speech and Lang.* **20**, 128–158.

[20] Jessen, M. (2008). Forensic linguistics, *Language and Linguistics Compass* **2**(4), 671–711.

[21] Broeders, A.P.A. (2006). Of earprints, fingerprints, scent dogs, cot deaths and cognitive contamination – a brief look at the present state of play in the forensic arena, *Forensic Science International* **159**, 148–157.

DIDIER MEUWLY