THE ROLE OF AUTOMATIC SPEAKER RECOGNITION TECHNIQUES IN FORENSIC INVESTIGATIONS

A.P.A. Broeders

National Forensic Science Laboratory, Rijswijk, Netherlands
Department of Language and Speech, University of Nijmegen, Netherlands

ABSTRACT

There are several major differences between the sphere of application of automatic speaker recognition techniques and the conditions attending speaker identification in the forensic context. Some of the factors involved are discussed below. The prevailing view that these differences preclude the introduction of even the more powerful automatic verification techniques in forensic work is questioned, and an indication is given of ways in which this question may profitably be addressed.

1 INTRODUCTION

At first sight, the proposition which serves as the central theme for this session - the definition of the speaker can be expected to come from the laboratory in the next few decades - looks simple enough. It seems to call for either wholehearted support or utter rejection, and it was no doubt phrased with the express aim of provoking such primitive responses. However, the simplicity of the proposition is somewhat deceptive. In its present form, I find myself unable to react to it in unequivocal terms. I have therefore taken the liberty of reformulating it in terms of what, from my perspective anyway, seems to be the real question underlying it: "Can automatic speaker recognition techniques be expected to play a role in the forensic context in the foreseeable future?"

In fact, this question must itself be rephrased in several ways, with each subquestion addressing a different aspect of the central issue. Some of these questions are discussed below, and some indications are given of the way in which they might be resolved.

2 IS THERE A VOICEPRINT?

It is widely accepted today that the term voiceprint is a misnomer for what is basically simply a spectrographic representation of a particular utterance by a particular speaker. Indeed many would argue that the term is better avoided altogether. However, there is a sense in which the term can usefully be employed in a manner which rather more closely resembles the parallel use of the term fingerprint, i.e. to refer to a unique representation of a particular individual. For the sake of the present discussion we could conceive of a voiceprint as a representation, in whatever shape or form, of such acoustic information as will uniquely characterize each individual speaker. This would enable us to address a more specific question, viz. whether a voiceprint in the sense just defined is in fact a real possibility.

Obviously, such a unique representation can only fully serve its purpose if we can rely on the signal under examination to contain the acoustic information that is required for a unique identification. However, we know that on the physical plane speech is marked by constant variation. The representation we are looking for would therefore have to reside in a continuously varying signal. But we know that as ordinary language users, even when dealing with speakers with whom we are very familiar, we are liable to make identification mistakes, especially - but not exclusively - in

situations where we are expecting a particular speaker but are in fact exposed not to the expected speaker but to a close soundalike. We may think we hear a friend answering the phone, only to find that we are talking to his son. This suggests that, for human listeners at any rate, there is a very real sense in which we cannot be sure that no two speakers speak exactly alike (Nolan [1]), and that we must at least consider the possibility that there is not always enough speaker-specific information in the signal to enable us to verify the identity of a familiar speaker, let alone that of an unfamiliar speaker.

Over and above the inherent variability of speech as a physical phenomenon, there is of course the variation inherent in speech on the linguistic plane. Anyone who has been in a position to listen to even a moderate amount of unmonitored speech will have been struck by the wide variety of speech styles used by many speakers in different communicative contexts. As language users we may be able to identify speakers on the basis of utterances produced in a quiet conversational style with reasonable success but we have great difficulty doing this if the utterances are produced with different degrees of intensity. Similarly, we do not feel confident about extrapolating the quality of a speaker's whisper, headvoice or loud voice from speech produced by the same speaker with a modal voice quality (Broeders and Rietveld [2]).

Given the variability of speech in different communicative contexts, it is doubtful whether any representation can be made which will capture the unique acoustic information required for the identification of the speaker from signals as diverse as those found in real-world conditions. Or, phrased differently, it is doubtful whether the speaker-specific information contained in the signal will be sufficiently uniform and consistent across various speech styles to serve as a basis for automatic speaker recognition in situations that are more challenging than the typical closed-set automatic verification context. So it appears that no matter what type of speaker profile we conceive of, it is likely to lack the one property that, together with uniqueness, makes the fingerprint such a powerful means of identification, i.e. invariance.

It is worth noting though that in spite of the lack of reliability that has been shown to be associated with the traditional voiceprint technique in forensic speaker identification (Bolt et al. [3]), one still occasionally comes across unwarranted claims like that recently found in a brochure advertising the 'Kreutler Computerised Speech Lab'. Next to the photograph of a computer screen display which, on closer examination, turns out to bear a more than remarkable resemblance to the Kay CSL-system, the law enforcement and security services type clientele that the brochure seeks to address are offered the following information: *'Forensic analysis is a widely spread technique to identify persons by their voice prints. These voice prints are specific for each person and can not be altered.'* ([4], p. 6)

3 FORENSIC VS COMMERCIAL APPLICATIONS

Various authors, including both Künzel [5] and French [6], have drawn attention to the severe limitations imposed by real-world conditions on forensic speaker identification and discuss the implications this has for the application of automatic speaker recognition procedures, as used in commercial applications. There are five major factors that need to be taken into account here. They are:

Text dependence

In automatic speaker verification

systems the utterances that are used for the verification test can in principle be pre-selected for best performance. In the forensic context the nature (and size) of the contested material is normally entirely beyond the investigator's control, and the nature of the reference material, i.e. the material that is known to have been produced by the known speaker (usually the suspect), is often determined by what happens to be available in a particular case.

Speaker cooperation

Even if reference material can be collected expressly for the purpose of an identification test, the investigator will, even at the best of times, have to be mindful of the observer's paradox (Labov [7]). As it will not normally be legally possible to collect a speech sample without the suspect being aware of it, let alone without the suspect's consent, there is the very real danger that the reference material that is collected does not constitute a representative sample of the suspect's speech.

Obviously, speakers may deliberately set out to systematically alter their speech style, may choose to be less than forthcoming and may more generally try to avoid producing a representative speech sample. However, even cooperative speakers may, as a result of the stressful nature of the situation they find themselves in, produce speech that varies considerably from their usual repertoire. In the automatic speaker verification context however, the situation is unlikely to be experienced as stressful and speakers can normally be relied upon to be cooperative since they stand to gain from a positive result.

Of course, the questioned material itself may show signs of varying degrees of deliberate disguise or, more generally, be of a nature which virtually precludes its being subjected to any type of systematic investigation.

Recording and transmission conditions

Telephone recordings account for a very large proportion of all forensic material. In addition to the major frequency bandwidth reduction of the telephone system, the effect of the handset and various less predictable signal modifications introduced by the telephone system, there are the effects of a wide range of recording equipment to be reckoned with. Between them they may give rise to a variety of signal degradations and distortions which may vary quite considerably from one call to the next. In the verification context, of course, none of these complications will normally arise, since great pains will be taken to control the quality of recording equipment and transmission channels.

Class size

A major problem for the application of automatic speaker recognition techniques in the forensic context lies in the size of the speaker set in real-world forensic conditions. Automatic procedures are typically geared to applications with a known or closed set of speakers. On the other hand, in the forensic context, the unknown speaker cannot be assumed to be one of a small set of speakers but must normally be taken to be one of a class whose membership, if not of indefinite size, may very often be quite large and is typically unknown.

Cost of errors

There is an even more fundamental difference between the usual sphere of application of automatic techniques and the forensic context. As is well-known, in closed-set verification systems, there is a trade-off between the false acceptance of unknown speakers or impostors and the false rejection of known speakers or customers. In a commercial application, the cost incurred by the false acceptance of an imposter in the form of unauthorized access to information, services or facilities can be balanced against the frustration and loss of time generated by the false rejection of a bona fide customer. But in the legal setting, such a cost-benefit analysis in essentially financial terms would be unthinkable. Indeed, it has often been argued that in the forensic context any method that, in addition to correct identifications, will produce even a single incorrect identification is unacceptable, since it conflicts with one of the fundamental principles that any judicial system may be required to subscribe to, which says that it is better to have a guilty person acquitted than an innocent suspect convicted.

4 COMMON PROBLEMS

Although it is fair to say that the factors discussed above present formidable obstacles to the introduction of automatic speaker recognition techniques in the forensic context, this should not be taken to imply that there is no point in investigating conditions in which benefits may be derived from their application. It may well be the case that an automatic speaker identification technique, in the sense of a set of decision procedures that is carried out entirely independently of human interpretation, is an unrealistic scenario but that is not to say that there is no room for these methods at all.

In fact, there may be good reasons for a somewhat more optimistic view than is taken by many commentators. Part of the explanation for the lack of progress may lie in the gap separating what, perhaps somewhat disrespectfully, may be termed the engineering approach as opposed to the linguistically-oriented approach to speaker recognition. Leaving aside the decreasing number of adherents of the voiceprint technique, practising forensic phoneticians, especially those associated with the International Association for Forensic Phonetics (IAFP), are keenly aware of the need to bridge this gap. There are indications that speech technologists too are aware of the need to take more account of both the linguistic and the judicial aspects of forensic speaker identification (Bimbot et al. [8], p. 82). This development may well be aided by a growing awareness that the factors limiting the applicability of automatic procedures do not in fact always constitute absolute impediments.

The text

A good example is text dependence. The use of a limited number of fixed passwords obviously tends to render the older automatic verification systems vulnerable to fraud. After all, with the increasingly widespread availability of low-cost, high-quality digital speech processing technology, it is not too difficult to record the voice of a bona fide customer and subsequently replay it to gain unauthorized access to a particular system or service. So the need arises for text-independent or text-prompted formats. A possible solution is a combination of speech and speaker recognition techniques which allows the system to freely prompt random utterances and to check not only whether the voice is that of the customer but also whether the required text is produced (Furui [9]). On the other hand, there are many forensic situations where the requirement of text dependence, i.e. the availability of identical utterances in both questioned and reference materials, can easily be met.

The speaker

The same applies to speaker cooperation. Again, there are situations when reference material is available whose status is not contested by either party and which also satisfies the major demands that it is representative of the speaker's linguistic repertoire and is produced in a communicative context which is similar to that in which the

questioned material was produced, so that an adequate basis for comparison exists. Conversely, in commercial speaker verification environments, there are obviously also limits to the degree of cooperation with which the speaker can realistically be relied upon to interact with the machine. Ironically enough, the use of such a pre-eminently human faculty as language by machines will often cause frequent users to lose patience with other, less than human characteristics of the machine and to develop a reluctance to adapt their performance to the machine's requirements. Possible effects on speech include a loss of articulatory precision and lower overall intensity.

The telephone line

Telephone transmission conditions do not in actual forensic casework necessarily always vary more than they would in commercial verification applications. In fact, it is quite common for recorded telephone conversations that are the subject of a forensic inquiry to have been made from the same location, through the same extension and on the same day. Recording conditions are also frequently at least potentially controllable to the point where they may be sufficiently uniform to meet the same technical requirements that must be met in commercial applications. Traditional analogue telephone logging and tapping devices are increasingly being replaced with advanced digital facilities, with calls being stored in a digital format.

The speaker set

Class size is probably ultimately the more intractable problem. This is sometimes obscured by the confusion that is created by the use of the terms identification versus verification. In fact, forensic phoneticians are typically involved not in speaker identification but in speaker verification albeit - and here lies the real difference - with an open set of speakers rather than a closed set. But the question that poses itself in the forensic context is essentially a verification, not an identification problem: is the questioned material produced by the same speaker as the reference material? In more concrete terms: were all the questioned calls made by the same person, and if so, do they originate from the person who is believed to have made them?

The complication introduced by the circumstance that in the forensic context the unknown speaker is not normally claimed to be one of a closed set of speakers but must be assumed to be one of an open class creates problems that are essentially of a statistical nature. What an objective forensic procedure would be required to do is not just to quantify the degree of similarity between questioned and reference samples and make a decision based on a comparison with a pre-determined threshold, as occurs in closed-set verification applications, but to give a statistically sound indication of the probability of this degree of similarity occurring by chance. Or, to phrase the question in Bayesian terms, it should allow one to calculate the likelihood ratio of the probabilities that the findings would arise under the two conditions that the defendant was, and was not the unknown speaker (Evett [10]).

The consequences

Finally, there is the cost of error aspect. Obviously, erroneous conclusions can do a great deal of harm, especially if findings are presented without an indication of the reliability of the methodology used with reference to the specifics of a particular case. On the other hand, if our final criterion is that a method be demonstrated to produce no false positives, it may well be unnecessarily strict. What is important is that reliable statistics can be given, or that, if a probability scale is used,

the relative position on this scale of the particular degree of probability arrived at in a particular case is indicated, and a clear statement is given of the limitations of the methodology employed (Nolan [11]). If this requirement can be met, speaker identification evidence does not compare unfavourably with other types of expertise that are regularly sought by courts of law. By the nature of their work, judges are constantly involved in weighing probabilities and uncertainties. Deference to experts of whatever designation is a threat to any judicial system (Nijboer et al. [12]), although the danger may well be greater in adversarial systems where 'rival' experts find themselves in the business of explaining their findings to a jury, whose critical faculties may well be taxed beyond capacity by the level of abstraction required to follow the argument.

Also, there is an as yet largely uncharted demand for forensic speaker recognition expertise for investigative rather than evidential purposes. In large-scale police investigations a degree of uncertainty may be less problematic and an informed use of automatic procedures may improve the quality of decisions and lead to considerable savings in time and staff expenditure.

5 COMBINED RESEARCH

A particularly promising approach is that described by Boves et al. [13]. Within the design of the Dutch POLYPHONE speaker database a number of operational conditions are systematically varied so that their effects can be investigated. The recording platform used to collect the speech of the 5,000 speakers in the POLYPHONE database proper, was also used to collect an additional 2 groups of 50 speakers each, specially selected to examine the effect of variables like kinship and linguistic background. The speakers are 100 adult males, all native residents of two distinct parts of the Netherlands, the cities of The Hague in the West and Nijmegen in the East, who between them form some 50 pairs made up of two or more brothers, or a father and a son. The composition of this speaker set was partly inspired by the sort of questions that are particularly relevant in the forensic real-world context, where the pertinent statistic is not how likely a speaker is to be confused with a random 'imposter' but with a speaker with a similar linguistic background. Forensic phoneticians are rarely asked to compare samples involving clearly different accents but suspects or their barristers may well claim that the speaker in the questioned recording is the suspect's brother, and the circumstances of the case are often such that this possibility cannot be ruled out.

The design makes it possible to investigate a variety of questions that are particularly relevant to the forensic field. The project includes experiments to compare identification performance among the two sets of closely matched speakers with that among the larger group of male POLYPHONE speakers, and to investigate within-dialect as opposed to between-dialect confusions, as well as experiments to study the effect of close kinship on error rates. As all speakers in both sets of 50 each made 8 phone calls using two different handsets, intra-speaker and inter-phone variation can also be studied.

Preparations are also under way to test the performance of the arithmetic-harmonic sphericity measure developed by Bimbot and Mathan [14,15] on the material produced by the two sets of 50 speakers.

6 THE DEBATE CONTINUES

In some countries, speaker identification in the forensic context is a very controversial issue. To some extent,

this may be due to the exaggerated claims made by those who were responsible for the introduction of the so-called voiceprint technique. At the same time though, the short-lived popularity of the voiceprint may serve as a vivid reminder of the need for phoneticians to take an active interest in forensic questions, if only to expose phonetically unsound testimony offered by non-phoneticians of various denominations.

Of course, individual phoneticians must decide for themselves whether they wish to do forensic research or take on actual casework. But, as argued elsewhere (Broeders [16]), it would be wrong for phoneticians or linguists as a body to refuse to be involved in forensic work for the sole reason that they feel their discipline cannot provide incontrovertible evidence. That is nevertheless exactly what the motion adopted by the Groupe Communication Parlée de la Société Française d'Acoustique [17] would seem to advocate, inasmuch as it effectively calls for the withdrawal of all phonetic expertise from the field of forensic speaker identification. However, ironically enough, the overriding importance of the need for speech scientists and phoneticians to collaborate with those with first-hand knowledge of real-world forensic conditions could hardly have been demonstrated more forcibly than by the text of the motion. It reflects a sad lack of understanding of the type of question that poses itself in the forensic context, of the way in which these questions are handled by practising forensic phoneticians in countries like Britain, Germany and The Netherlands, and of the role and the responsibility of the expert witness in a judicial investigation.

That this position is unlikely to stimulate the necessary collaboration between forensic practitioners and other phoneticians and speech scientists is all the more unfortunate as phonetics as a science only stands to gain from the type of questions that emerge from the real-world conditions that apply in the forensic context. Fortunately, though, there are also indications that more and more phoneticians and speech scientists are taking an active interest in the problems posed by forensic speaker identification, with symposia like the present providing an ideal opportunity to exchange views, clear up some of the more persistent misunderstandings and define common research aims.

7 CONCLUSION

Recent developments have led to a situation where closed set speaker verification and open class forensic speaker identification have come to share a greater number of problems than has so far been the case. It follows that there is every reason to look into the possibility of combined research. The projects described in section 5 provide good examples of this approach. It is based on the premise that, in forensic applications too, performance of automatic recognition techniques will be dependent on the amount of control that can be exerted on operational conditions (Doddington [18]). It implies that in carefully controlled forensic conditions automatic procedures may in due course also come to play a role, if only for investigative rather than evidential purposes.

However, even here the process will never be fully automatic. It will always take an experienced phonetician or a linguistically informed speech scientist to decide what parts of the speech samples under examination are linguistically sufficiently similar to be used as suitable test material. Ultimately, then, it is the variation along the linguistic dimension that may well prove to be least amenable to efforts to bring automatic speaker verification techniques to bear on forensic material. In other words, it is unrealistic to anticipate a fully automatic procedure that will be able to extract a sufficiently comprehensive speaker profile from a questioned speech sample, given the variety of speech styles encountered in forensic conditions.

REFERENCES

[1] Nolan, F. (1991), 'Forensic Phonetics', *Journal of Linguistics* 27, 483-493.

[2] Broeders, A.P.A. & A.C.M. Rietveld (1989) 'Segmental Marking as a Cue in Auditory Voice Identification of Telephone Speech', in: J.P. Tubach & J.J. Mariani (eds.), *Eurospeech 89*, CEP Consultants, Edinburgh, 71-74.

[3] Bolt, R.H. et al. (1979) *On the Theory and Practice of Voice Identification*, Washington DC: National Academy of Sciences.

[4] 'Professional Telecommunication Systems', brochure published by Kreutler, Brussels.

[5] Künzel, H.J. (1994) 'Current Approaches to Forensic Speaker Recognition', *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, 135-141.

[6] French, P. (1994) 'An Overview of Forensic Phonetics with Particular Reference to Speaker Identification', *Forensic Linguistics* 1(2), 169-181.

[7] Labov, W. (1972) *Sociolinguistic Patterns*, Oxford: Blackwell.

[8] Bimbot, F., Chollet G., Paoloni A. (1994) 'Assessment Methodology for Speaker Identification and Verification Systems' *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, 75-82.

[9] Furui, S. (1994) 'An Overview of Speaker Recognition Technology', *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, 1-9.

[10] Evett, I.W. (1991) 'Interpretation: A Personal Odyssey', in: Aitken, C.G.G. and Stoney, D.A. (eds.) *The Use of Statistics in Forensic Science*, New York: Ellis Horwood.

[11] Nolan, F. (1992) 'Code of Practice', *Journal of the International Phonetic Association* 22(1 & 2), 80-81.

[12] Nijboer, J.F., Callen, C.R., Kwak, N. (eds.) (1993) *Forensic Expertise and the Law of Evidence*, Amsterdam: North-Holland.

[13] Boves, L., Boogaart, T., Bos, L. (1994) 'Design and Recording of large Databases for Use in Speaker Verification and Identification', *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, 43-46.

[14] Bimbot, F., Mathan, L. (1993) 'Text-free Speaker Recognition using an arithmetic-harmonic Sphericity Measure' *Proceedings of Eurospeech*, Berlin, 169-172.

[15] Bimbot, F., Mathan, L. (1994) 'Second-Order Statistical Measures for Text-independent Speaker Identification' *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, 51-54.

[16] Broeders, A.P.A. (1991) 'Great Debate on....' *Nesca - The ESCA Newsletter* 5, 50-51.

[17] Bureau du Groupe Communication Parlée de la Société Française d'Acoustique (1990), Motion adopted on September 7, *Nesca - The ESCA Newsletter* 4, 39.

[18] Doddington, G.R. (1985) 'Speaker Recognition - Identifying People by their Voices', *Proceedings of the IEEE*, 73(11), 1651-1664.