THE FUTURE OF SPEAKER IDENTIFICATION: A MODEL

Harry Hollien
Institute for Advanced Study of The Communication Processes
University of Florida, Gainesville, FL USA

ABSTRACT

One of the major areas within Forensic Phonetics, and to some extent Phonetics in general, is that of speaker recognition--and especially speaker identification. To date, most of the problems attendant to this issue have escaped resolution. The chaos here may be due to the fact that several types of professionals (Phoneticians, Engineers, Psychologists and the Police) all are working in the area but in a fairly uncoordinated manner. The strengths and abilities each bring to it are incomplete and often are further degraded by their weaknesses. The result is that no robust method of speaker identification currently exists. The following presentation will provide a review of the basic problems in the field, its boundaries, past approaches to the problem, the strengths and limitations of the relevant specialists and a model which could lead to its resolution in "the next decades." Forensic Phoneticians are central here; however, the model specifies that objective means must be employed if a valid and effective speaker identification system is to become a reality.

1. INTRODUCTION

One of the fairly new--and certainly exciting--areas within the Phonetic Sciences is that of Forensic Phonetics. Specialists with-in this area are making significant contributions both to relevant research and in response to problems faced by members of the legal and law enforcement communities. This interface ranges from speech enhancement and/or decoding to tape authentication, from detection of stress in voice to the vocal cues which signal intoxication. However, of all the problems encountered, that of speaker identification-is probably the most challenging and (perhaps) the most important. For one thing, it involves issues that are fundamental to the Phonetic Sciences; indeed, it appears appropriate to state that research here should claim a measurable portion of our time and energy. Second, it holds substantial social significance.

By now we should have defined and structured the issue. If we could not "solve" it, we should have, at least, approached it in a coordinated manner so that the relevant relationships could be systematically researched. Unfortunately, we have not done so. As a discipline, our members have tended more to react to positions taken by, or requests from, members of other disciplines rather than to have organized the necessary models and carried out appropriate research. Whether we like it or not, this area of our field is in near chaos.

In the preceding paper, Nolan has provided most of the basic definitions relative to the speaker identification task and has outlined certain of the specific problems and difficulties we face. In the effort to follow, an attempt will be made to supplement his perspective and provide a model which could lead to a solution "in the next decades". To do so, several issues must be addressed; they include reviews of 1) the bases for speaker identification; is it possible to do it in the first place? 2) the boundaries of, and approaches to the problem; which of the available approaches may ultimately lead to a successful resolution? 3) the other classes of professionals who are relevant to the area; what are their responsibilities and what contributions can they make? 4) guidelines for future speaker identification efforts; i.e., the proposed model.

2. BASES FOR SPEAKER IDENTIFICATION

Two related questions may be asked about the identification of speakers by voice. They are: 1) does each human speak in a manner so idiosyncratic that, overall, he/she is different from all others and 2) is inter-speaker variability always greater than intra-speaker variability? The answer to both of these questions is a resounding "probably not!" Worse yet, it is to the discredit of our discipline that we have not already researched these fundamental issues to any great extent. Indeed, nearly all the authors of the over 700 presentations on speaker recognition listed by Hollien and Alderman [1] have addressed only narrow issues or relationships--and many of them involve "application." Application? At first glance it would appear counterproductive (if not ludicrous) to attempt the "solving" of a problem before its nature is understood. None-the-less, this situation functionally constitutes the present State-of-the-Science re: speaker identification.

What is needed, of course, is a major research thrust in the basic areas. For one thing, researchers should attempt to determine if talkers actually do exhibit unique enough characteristics to permit universal speaker identification to be developed. At the very least, an effort of this type would establish the limits and boundaries of the problem and, possibly, lead to techniques and/or procedures which would permit a valid, if restricted, response.

In reality, there is little-to-no possibility that such a massive effort would be supported by any agency or group. This is surprising as there is no question but the need for valid speaker identification and verification methods is a critical one. It exists in nearly every sector of society. What is lacking is the foresight by any of the relevant agencies to see beyond an end-product. Sadly enough, the need is for basic research; about all that will be supported is "product development."

Given the unlikelihood that basic speaker identification issues can be addressed in any meaningful way, only a single recourse appears available. That is, all that may be possible is to generate a working model by the synthesis and interpolation of current information as supplemented by research conducted on a piece-meal basis. Actually, some of the necessary relationships have been established--at least enough of them for researchers to attempt advances in this area. Useful data already can be found in a number of published articles and reports; four books (Baldwin and French, 2; Hollien, 3; Küenzel, 4 and Nolan, 5) provide summaries of most of the important relationships; further, they suggest some useful models. It now appears evident that the two questions cited above must be answered in the negative only if a binary answer is required. It also appears evident that a given talker may be differentiated from other individuals within specific sets of speakers if a critical number of his or her features are measured and appro-priate metrics (in multidimensional space) are established and compared. On a simpler level, it appears that establishing speaker profiles may very well be of merit. What appears both needed and realistic is completion of a number of investigations in which attempts are made to identify and validate those individual parameters--plus constellations of parameters--which are robust to the task. Subsequently, the resistance of these individual parameters and profiles to forensic type degradation can be studied. However, it should be noted that this element within the model does not reject human decisions for some mathematically derived metric or group of metrics. The fundamental focus here still would be on human performance and it's assessment by humans.

3. THE BOUNDARIES OF SPEAKER IDENTIFICATION

As was pointed out in the prior paper by Nolan (see also Hollien, 3 for a definition), speaker identification is only one element within the general rubric of speaker recognition. Here the task is to determine if a given (and known) speaker is the same person as the one who produced the target utter-ances (i.e., the "unknown" talker). This task is a very difficult one primarily due to the nature of speech and the situation within which it exists. The speech will be noncontemporary; all sorts of channel and speaker distortions may (and usually do) exist. For example, 1) there may be many competing speakers, 2) the unknown talker may have provided only a limited speech sample, 3) the process is an "open" one (i.e., the unknown may not be in the suspect pool), 4) speakers usually are uncooperative, 5) poor recordings may exist, and so on. In this milieu, the scientist (or practitioner) will have little control over the available signals.

On the other hand, speaker verifica-tion is a process where an attempt is made to authenticate the identity of a given speaker by comparing his utterances to those in a closed set of voices of which he is a member (that is, unless he is an imposter). Because of the high control practitioners enjoy in this situation (a closed set, speaker cooperation, continual updates, exten-sive samples, sophisticated equipment, etc.) the challenge of speaker verification is a much less rigorous one than is that of identification. As would be expected, far more research has been carried out in the verification area (than on identification) as it is easier to manage and can lead to substantial monetary returns. What few people appear to realize is that, due to the severe challenge created by the identification task, there is little chance that even successful verification approaches can be applied to "solve" identification. It also is unfortunate that very few people understand that any method which is successful for speaker identification will simultaneously solve the verification problem.

Boundaries to speaker identification also have been established in other domains. In one case, it is the decision-making process that is controlling. There are three "entities" which can be involved in this process: 1) laymen, 2) professionals (usually Phoneticians) and non-humans (i.e., computers, other machines). These divisions, while even more complex than they seem, have essentially been defined in Courts-of-Law. They will be discussed in turn.

3a Laymen.

The Courts have pretty much established, defined and limited acceptable behavior for the first of these cohorts, i.e., laymen. Ordinarily, the process involved takes one of two forms. In the first instance, an individual who can demonstrate a close familiarity with the unknown talker is allowed to testify that he or she can recognize and identify him or her as the (otherwise) "unknown" speaker. In support, there is very good research evidence as a basis for this postulate; that is; people who really know a talker usually can identify that person from speech samples at very high degrees of accuracy. On the other hand, there is no research available which will allow predictions to be made about how often a given individual will be correct in a specific situation. Moreover, the question must be asked as to whether or not this procedure is part of the speaker identification milieu? Of course it is. While not central at all to the fundamental requisites of the area (or the model to follow), it is the responsibility of Phoneticians to study such behaviors and to define them and their limits.

The second subgroup within the untrained cohort involves people who do not know the talker but have heard him. We know from research that untrained individuals, while usually not particularly good at this task, exhibit great variation in their natural ability and that environmental circumstances may have a substantial affect in upgrading or degrading their performance. Additionally, the process here often culminates in what are referred to as earwitness lineups or "voice parades." These lineups are a reality and cannot be ignored. The procedures used in their conduct vary wildly both with respect to their nature and quality; currently, a lively controversy exists as to who should control the earwitness process in the first place--Phoneticians, the police or relevant Psychologists. Again the problems associated with earwitness lineups are rather peripheral to core speaker identification. Nevertheless, the issues here are the responsibility of the Forensic Phonetician. Relevant procedures are, and will continue to be, employed by law enforcement agencies and the courts. While they cannot be central to our model, they must be taken researched
and understood.

3b Professionals.

The second group includes trained professionals--usually Forensic Phoneticians--who are responsible for the judgements about a speaker's identity. Since the professional but rarely knows the unknown talker, their procedures must involve systematic comparisons of some type. They certainly require that a stored sample of the unknown talker, plus one for the suspect, are available. As is well known, Phoneticians often employ panels of trained and untrained auditors to perceptually judge whether a particular unknown voice was produced by the same person as was the "known" voice. This procedure usually involves direct comparisons of samples which are embedded in a field provided by foils or controls. The Phonetician uses the resulting scores to aid him or her in making decisions; machine processing also may be carried out for the same purposes. But what he or she most commonly does is listen to samples of the unknown voice plus that of the suspect (possibly within a field of foils) over and over again. This process ordinarily involves assessment of these talkers' specific features (dialect, fundamental frequency, voice quality, articulation, etc.,) one at a time. It has been shown that techniques wherein the segmentals and suprasegmentals of speech are systematically evaluated work pretty well and they do so under a variety of conditions. Nonetheless, not much is known about the efficiency or, even, the validity of these approaches. There does not even appear to be a methodological consistency among the Forensic Phoneticians who work in this area. Which of these professionals is better at it than others, what are their "hit" rates, how do the various techniques stack up against each other, how does effectiveness vary as a function of different situations? The questions are many but the answers few. Moreover, this area is absolutely central to the speaker identification process.

3c Machines.

The third approach is that of machine processing of the speech signal for speaker identification purposes. Again the procedures employed take two directions. The first involves traditional signal processing techniques such as axis crossings, HMM, LPC, Cepstral approaches and/or related methods. In the second, researchers attempt to duplicate human auditory processing of the signal; they seek out those features that auditors employ in making identification decisions, attempt to develop appropriate algorithms and, subsequently, program computers to mimic the process. These several approaches have a longer history than is generally appreciated. Early attempts at development reach back to the World War II era and are contemporary with the "voiceprint" technique (i.e., subjective pattern matching of time-frequency-amplitude sound spectrograms). Some of the early attempts were sited at government or industrial laboratories; others were commercial efforts at speaker verification. Unfortunately the thrust was primarily on system development rather than on data gathering relative to basic identification. Hence, a number of excellent beginnings were abandoned when field trials proved disappointing. Even the few sustained, long-term programs have progressed but slowly. It must be said that even the near-magic of modern technology is not inadequate to the task when the establishment of applied techniques is required before the basic relationships are understood.

Therein lies the functional challenge to the forensic application of speaker identification--or to speaker identification procedures developed for any reason. It will be difficult to establish any kind of effective system until at least reasonable information is available about the natural boundaries of this area and the inter- and intra-speaker variability confusions resolved. Once relevant relationships here have been established, the ways by which application can be carried out also will become available.

4. PARALLEL BUT UNCOORDINATED EFFORTS

A second rather serious problem also exists in the speaker identification area. It results from the well intentioned, but sometime misguided, efforts of the three major groups of professionals working on speaker identification problems. They are the Phoneticians, Audio-Engineers and relevant Psychologists. The insularity and narrowness within each of these groups is creating a serious impediment to orderly progress in the area.

For example, the expertise of Psychologists and Phoneticians overlap in the earwitness identification area. Of course, the Psychologists appear to be almost exclusively concerned with voice parades, whereas Forensic Phoneticians have tended to downgrade this procedure as a risky one at best. It is only very recently that each of these groups has become aware of the relevant philosophies and activities of the other. Procedures here certainly would benefit from a melding of the behavioral skills/knowledge of the Psychologists (and their research/experience with eye-witness lineups) and the Phonetician's fundamental understanding of hearing and aural-perceptual speaker identification. Further, an even more active role, by Psychologists, directed at other speaker identification issues should result in better understanding of all of the behaviors involved.

A problem with even more serious consequences is that which exists between Phoneticians and relevant Engineers. Many Phoneticians are quite unwilling to extend their identification efforts beyond the traditional aural-perceptual techniques and employ modern technology. On the other hand, Engineers often view the identification process as a simple signal analysis exercise and do not seem to understand how the effects of social pressures, the enormous variability in human behavior and the vagaries of the forensic milieu itself can disrupt machine processing of any type. Accordingly, with but few exceptions, the Phonetician's computer-based efforts have been rather feeble and Engineers' attempts to fit their procedures into the real world have been equally disappointing. On the one hand, many Phoneticians refuse to accept the possibility that the only solution to the speaker identification challenge will involve the use of modern technology. Yet, the reality here is quite apparent. On the other hand, the Engineer typically cites what he or she perceives as inadequate quantitative skills on the part of the Phonetician as well as the contradictions-confusions to be found in their literature. Engineers suggest that the answer is in the signal and a good solution can be easily achieved if they only were allowed to address the problem. Perhaps so. However, if this is true, why is it that progress is relatively nonexistent when the much more malleable issue of speaker verification is considered? More important, even after decades of great effort, closure still has not been realized with respect to the challenge of speech recognition by machine. Perhaps it is because Engineers have not been willing to address problems related to speech and speakers as well as the myriad of other

distortions (environmental, channel, speaker) found in the communicative act.

Unlike the difficulties outlined in the previous section, a reasonable solution re: the differences among professionals, may be possible. That is, after nearly a half century of frustration, these groups may be realizing that they need to establish a functional interface with each other. Further, since Phoneticians are central to the problem, it would appear that they bear the primary responsibility in fostering such cooperation. Not an easy task, of course, but one that is mandatory if an effective solution is to be realized.

5. A MODEL

As stated, there currently appears to be only one reasonable solution to the challenge of identifying speakers by voice. It is to develop a machine-based system which can be used to decode and analyze the identity information contained within the speech signal in much the same manner as does the human being. The responsibility for each decision would be the same (i.e., the professional); the primary difference being that software would be substituted for neuroprocessing. One such approach has been to identify, and single out (for processing) those features which people use in this manner (Hollien, 3; Stevens, 6). For example, fundamental frequency level and variability, vocal intensity patterns, prosody plus voice and speech quality are among those elements which have been specified. Segmentals also can be included but they are a little more difficult to process on an automatic or semiautomatic basis. Nonetheless, patterns of vowel formant usage are important here as are articulatory gestures and especially dialect. The advantages of using an approach such as this one is that the data from aural-perceptual speaker identification research can be used to structure the effort; after all humans actually attend to such features and generally are reasonably successful in using them as identity cues. Perhaps even more important, auditors appear capable of carrying out this task even in the face of severely degraded listening conditions. Thus, it should be clear that, if machines can be taught to focus on these same relationships, and process them properly, a reasonable solution to the cited problem should be achievable. Certainly, a given set of procedures can be established, applied and tested; as a result, system strengths and weaknesses can be understood. It is only by this approach, or a similar one, that a valid and effective speaker identification procedure can be developed. Most important, its use would eliminate most, if not all, of the very subjective methods currently being employed. Indeed, it is difficult to understand, much less assess (on any reasonable basis anyway) the effectiveness of Forensic Phoneticians no matter how well trained, talented and motivat-ed they are. Worse yet, some of their techniques may be considered proprie-tary and, hence, cannot be assessed at all.

Please note, however, that it is not being suggested that only machine (computer) assessment of natural speech features is a viable approach to speaker identification. There probably are other elements within the speech signal that can serve as effective identify cues also. The fact that traditional signal analysis approaches have proven grossly inadequate should not preclude efforts to identify still other cues that maybe more robust. Further, it must be remembered that many assumptions must be made even if signal analyses of the natural speech feature type are employed. That is, it is not presently known just how robust each of the "natural" attributes are when environmental distortions (noise, passband, speaker distortions) are present; nor is it known how they can be combined to effect good decisions. While logic and available data will allow a few predictions to be made, it is not possible to specify just how robust each parameter will be under all (or even some) of the conditions which will occur. Nor is it known just how they should be normalized and weigh-ted within the speaker profile. Of course, any signal analysis approach will suffer from these same restrictions. Hence, experiments will have to be carried out to establish these relation-ships before vectors are applied.

Is it possible to proceed even in the face of questions about speaker variability and the differential effects of speakers, recording equipment and the environment? This query probably can be answered in the affirmative if a model is established and safeguards are included in its structure. A suggested model is as follows.

1. It must be assumed first that only digital analysis of the signal will yield a method that ultimately can be established as: a) stable, b) robust, c) efficient and d) universal.

2. The ultimate decisions made must be the responsibility of Forensic Phoneticians and/or other professionals--not the machines themselves.

3. The limitations (cited in the text) must be addressed or, at least, taken into account. That is, compensation for the possibility that intra-speaker variability may exceed inter-speaker variability must be made both with respect to relatively small and very large populations of talkers. The system also must be resistive to channel and speaker distortions.

4. It must be recognized that any attempt to establish a functioning method must be programmatic in nature. That is, it is doubtful that one or even a few experiments will yield information sufficient to develop a working system; a substantial program of research will have to be carried out.

5. The parameters, features and/or vectors (within the signal) which provide the identity cues must be identified and tested. As has been implied, enough information must be gathered about each of them that their behavior can be predicted. The situations in which they are effective and not effective must be established.

6. The ability of a proposed "system" to respond to a variety of situations and challenges should be researched and system robustness inductively specified. Test selection and administration is critical to this process. There is little chance that a system designed even for limited use can be developed unless users have information about the specific types of situations to which it can be successfully applied.

7. The system must be multidimen-sional in nature. Indeed, there probably is no single (or even small group of) feature(s) that will permit a particular speaker to be identified even under the most restricted of circumstances. Further, identification of the number and class of situations in which the method will be effective will require additional analysis--and ultimately the merging of a number of features. The number and class of situations in which the method will be effective will require additional analysis--and ultimately the merging of a number of features. A profile approach should be a effective in this regard.

As may be seen, the model cited specifies that an objective (rather than subjective) approach must be taken if the speaker identification problem is to be resolved. So too must a concerted effort be mounted to permit rational decisions to be made as to what may and may not be a accomplished. This discourse should not be interpreted as one of fault-finding as many researchers have contributed materially to the corpus of information now available. Nor is fair to fault individuals for carrying out finite (rather than programmatic) projects; often the culprit was the simple lack of funding. Perhaps, the only blame to be assessed here is one which can be directed at those practitioners who make sweeping claims about their methods; some show promise but all presently are of limited scope.

The solution also demands a change in the work patterns and philosophies of the specialists involved. Anyone--including Forensic Phoneticians--who believes that good resolution will emerge solely from efforts within his or her specialty, is not being realistic. It will take the combined efforts of members from all three of the cited professions to affect a solution.

REFERENCES

[1] Hollien, H. and Alderman, G. A. (1995) Speaker Identification and Recognition: Current References, *Miszellen Beiträge zur Phonetik und Linguistik*, 64, in press.

[2] Baldwin, J. and French, P. (1990), *Forensic Phonetics*, London, Pinter.

[3] Hollien, H. (1990) *Acoustics of Crime*, New York, Plenum Press.

[4] Küenzel, H. (1987) *Sprechererkennung: Grundzüge Forensischer Sprachverarbeitung*, Jeidelber, Kriminalistik-Verlag.

[5] Nolan, J.F. (1983) *The Phonetic Basis of Speaker Recognition*, Cambridge, UK, University Press.

[6] Stevens, K. N. (1971) Sources of Inter- and Intra- Speaker Variability in the Acoustic Properties of Speech Sounds, *Proceed., Seventh Inter. Cong. Phonetic Sci.*, Montreal, 206-232.