Available online at www.sciencedirect.com
SCIENCE DIRECT®
Forensic Science International 146S (2004) S107–S111
Forensic Science International
www.elsevier.com/locate/forsciint

Towards a protocol in speaker recognition analysis

L. Romito, V. Galatà*

Laboratory of Phonetics, University of Calabria, via P. Bucci-cubo 20/A, I-87036 Arcavacata di Rende (CS), Italy

Available online 18 October 2004

Abstract

On all sides there is an increasing need for a protocol in speaker recognition analysis. This need arises not only in the scientific word, but corresponds also to a requirement of guarantee and protection of all the individuals involved in criminal lawsuits. This study aims neither to resolve the problem of the protocol and of the fundamental steps in a speaker recognition analysis, nor to impose a precise methodology to be followed, but is an attempt to stir the stagnant waters of the debate on expert report phonic analysis.

The most important steps in the realisation of a phonic expertise are outlined as a basis and a starting point for a future protocol to which should follow the establishment of a phonic experts list.

Keywords: Speaker recognition; Forensic phonetics; Protocol

On all sides there is an increasing need for a protocol in speaker recognition analysis. This need arises not only in the scientific world, but corresponds also to a requirement of guarantee and protection of all the individuals involved in criminal lawsuits. It very often occurs that one arrives at the highest levels of judicial process with expert reports coming to different conclusions because of the use of different methodologies, although it is well known that an experiment can be considered scientific if and only if it answers to a known methodological protocol and only if the same results are achieved by the application of the same method [1].

This study aims neither to resolve the problem of the protocol and of the fundamental steps in a speaker recognition analysis, nor to impose a precise methodology to be followed, but is an attempt to stir the stagnant waters of the debate on expert report phonic analysis and investigation.

We believe that the discussion should begin by establishing if the expert report constitutes the research, within a signal, of a particularity that can be recognised and compared with a known or an unknown voice, or, on the other hand, if it can be the application of a known, programmed procedure. In the first case, each examination would assume the role of research rather than a technical consultation, while in the second case, we could try programming a methodological protocol.

It is beyond doubt that the research has to continue in order to obtain more reliable results, to lower the error threshold or, further, to identify new parameters that provide more certain answers to the questions asked by the court. At the same time it is beyond doubt that every research activity requires an experimentation period in the laboratory and certainly not in the courtroom. Only after this period of experimentation carried out by different researchers in different laboratories, can a scientific publication give value to a new procedure and render it applicable.

The resolution of the protocol problem becomes more pressing when we note that within the same method, the

* Corresponding author.
E-mail addresses: luciano.romito@unical.it (L. Romito), vgalata@libero.it (V. Galatà).

0379-0738/$ – see front matter
doi:10.1016/j.forsciint.2004.09.033

differences can be substantial and can depend either upon the operator or the consultant [1]. The perceptual method, for example, can be applied using non-expert listeners, and therefore a large number of people, who are asked to listen to an ad hoc constructed [2] voice line-up, expressing only a judgement of similarity or dissimilarity between two voices. Or it can be applied using expert listeners (therefore a small number of people) who listen to a whole passage identifying or not identifying the speaker on the basis of personal intuition and competence. The listening is in this case based on acoustic parameters such as pitch, voice quality, intensity, articulation, prosody and speech disorders.[1]

After this short introduction, and basing what follows on a scientific bibliography, we are going to outline the most important steps in the realisation of phonic expertise.

A comparative phonic examination should be differentiated in at least four important and propedeutic steps. The first step concerns signal acquisition, while the second and the third regard the identification and measurement of the parameters and the last concerns statistical comparison.

1. The first step

This is the most important step of the whole procedure because improving a bad recording is not easy and even after complicated signal analysis, using sophisticated equipment for filtering operations, the results are not brilliant. What therefore should be analysed is the recording method following already existing standards, which are not applied or taken into consideration.[2] Each recording that is destined for a courtroom should be accompanied by a report on paper indicating the essential characteristics of the recording modality, that is to say: the position of the microphone, the recording speed, the equipment used, the equipment brand, the recording material (cassette, tape, DAT etc.) and everything that allows the expert to recreate, if necessary, the same recording environment and quality.[3]

Before beginning any scientifically valid operation of transcription or comparison, it is necessary to check that the recording satisfies some essential electro-acoustical characteristics. These characteristics are in first place the signal/noise ratio (SNR) that should generally be equal to 10 dB.[4] This means that the background noise should not be so high as to hinder the analysis or the transcription of a recording.[5]

As is well known, a voice or a discussion can be recorded live (environmental interception) or by telephone (telephonic interception).[6] During a telephone call, normally, the speaker maintains the same distance from the microphone that is positioned in the phone handset, so that the interception produces recordings with a good intensity level and with an acceptable SNR. Unfortunately, we cannot say the same for an environmental interception, where the microphone or the 'bug' can be positioned directly on a person who functions as informer or in an environment to be checked. As regards the environment, the location of the 'bug' becomes very important. If it has to transmit signals for a number of hours that is greater than 10 or 15, it has to be supplied with direct current (no battery, in fact, is capable of sustaining a recording for a high number of hours). This implies, therefore, that the place has to be near an electric socket or equipment fed by electricity. In this case, the principal drawback comes from the non constant distance that the speaker assumes from the microphone causing recordings with extremely variable intensity levels that oscillate from distortion to zero, in addition to the environment noise that is impossible to prevent or to control. A protocol should therefore give precise indications on the limit of 10 dB.[7]

The second important characteristic is the bandwidth, that is to say, the interval of frequency used for recording and that has to be at least 3000 Hz.[8] The last characteristic, perhaps the most important, but also the most difficult to deal with, is length. This parameter should not be considered in operations involving transcription but only in those concerning linguistic–dialectological analysis in speaker recognition. It is obvious that in order to obtain a sufficient quantity of data, so that both comparison and linguistic analysis have sense, the material should have a certain length. On the other hand, establishing a limit as in the cases of both SNR and bandwidth is not easy because this parameter is strictly bound to the speaker.[9] The limit should,

[1] See Aural-perceptual approach in [3].

[2] Consider, for example of [4] recommendations and standards which lead the way in this field and supply a valid methodological reference.

[3] Recreating the same environment could be useful in those cases of voice sampling for perceptual comparison, so that the unique variable between two recordings leaves only the voice to be compared (intercepted unknown voice and recorded known voice) and not the channel or the background noise, etc.

[4] This because otherwise the presence of too high background noise could affect negatively the measurement of some formant contours.

[5] In [5] there is an accurate analysis of SNR in transcriptions and in the perception of 'noisy' signals. On the basis of studies during the 1950s, the SNR is not considered as it is in the field of acoustics, but rather as an index of intelligibility between a signal considered information signal and a signal considered as a disturbing signal. From this point of view even a dialog in a silent environment can have a low index of intelligibility if the voice to be transcribed (information signal) is covered by a disturbing other voice (noise).

[6] See [2] pp. 210–214.

[7] Very often recordings in a car with the radio switched on can give SNR values much higher than 10 dB even if the material is unusable. This is not a problem if we consider the radio signal as a disturbing signal and therefore as noise.

[8] With modern recording technologies the difference lies in the limits of the channel rather than in the recording itself.

[9] In 5 s a man with a high speech rate can produce sufficient material for any analysis. On the other hand, in the same time, he can produce so little material as to nullify the investigation.

therefore, not be fixed in the domain of time but should be indicated as the number of necessary parameters extracted to realise a phonic comparison which is scientifically acceptable, that is to say with an intrinsic false acceptance rate below a certain threshold.

2. The second step

In this second step, after the valuation of the 'analysable' signal, we can choose the vowels to measure. The parameters normally used in studies for identification purposes are the vowel values of fundamental frequency and formant frequencies.[10] Any other procedure up until this moment has to be considered experimental and should not find space in courtrooms.[11]

Even if the choice of formant frequencies appears to be easy, there are serious problems. Let us begin with the main question, or rather, the individuation of vowels.

The individuation of a vowel implies:

1. determining of the exact phonological inventory of the linguistic system to analyse;
2. definition of the status of each vowel in the system: which ones among all the vowels present can and should be measured; establishment of a dependence between the stressed vowels system and the unstressed vowels subsystem; the consideration, first of all theoretically, and the verification in practical terms, by measurement of processes like reduction, centralization and, in general, the effects of co-articulation;
3. verification, for each occurrence, of the influence of diaphasic, diastratic and diatopic variables on the speech to analyse.

The absence of a clear vision regarding the phonetic-phonological status of a vowel, in its reference system, can cause grave misunderstandings in terms of causing the consideration of units as equal when they are not, or the contrary.

Even in cases of apparently easy interpretation of the inventory of the linguistic system, like spoken Italian, the solution is not always unique. For example the procedure for expert report investigation consists of the detailed analysis of an hypo-system, considered, a priori, penta-vocalic such as /i, e, a, o, u/ considering as equal units such as /e, E/ and /o, ⊣/. As has been argued many times, there exist many types of spoken Italian, characterised by a phonological system with seven vowels /i, e, E, a, ⊣, o, u/. Hypothesizing a penta-vocalic system in some areas of Italia, both in the north and in the south, brings one to consider similar realisations of different vowels[12] such as [e] and [E], characterised by different formant values (/e/ $F1 = 360$, $F2 = 2040$; /E/ $F1 = 560$, $F2 = 1840$).[13]

If this creates serious problems in compiling the inventory of the Italian linguistic system, more attention has to be paid to those systems, like, for example, dialects, where the phonological inventory is extremely different from the Italian one. Some examples could be southern Italian dialects where we do not only have oppositions like /e, E/ but also /i, I/.[14] Without considering these important distinctions, the measurements of two different units are considered as belonging to the same unit, increasing in that way, the error and nullifying the analysis itself.

For this reason our starting point is the individuation of the effective speaker's phonological system: if it is an epta-vocalic system, the analysis has then to consider seven vocalic units and the same if the number is higher or lower.

An added problem is the production of mixed utterances, that is to say the production of phrases mixed both in Italian and dialect,[15] by the same speaker.

In the presence of the production of mixed utterances, the problem which arises is theoretical, because stems from the legitimacy of considering on the same level systems belonging to different codes, and practical, because the genetic and/or structural relationship between codes and the relationships of typological affinity have to be evaluated. Only after an accurate analysis of the phonological inventory of the two codes but, above all, of the phonetic distribution of the vocalic segments' areas,[16] can the linguistic decide to use the measures of vowels belonging to dialectal passages and those belonging to Italian ones as a unique linguistic system.

Once the vocalic hypo-system of the speaker's community has been established, the next passage consists in the segmentation of the units of this hypo-system. The units that are usually considered are the vowels in their continuum phonicum. At this point, another problem arises regarding the choice of vowel. Choosing only stressed vowels or unstressed ones, or to consider them as belonging to a unique system is another problem that requires phonetic deepening to be resolved.

[10] In lot of expert reports even short- or long-term spectra are included, but there is not yet a scientific acceptance of such parameters in the field of speaker recognition.

[11] We refer here to VOT measurements, occlusive length, nasal formants etc.

[12] Such as [pesca] *fruit* and [pEsca] from the verb *to fish*, [botte] *recipient for wine* and [b⊣tte] *physical blows*.

[13] The values here reported come from [6] p. 161 (Table III).

[14] See [7,8] § 1.1.5.3, [1].

[15] Mixed utterances can nowadays be Albanian-Italian, Slavonic-Italian, Wolof/Serer or other and Italian because of high rates of immigration.

[16] See [7], where the problem of the vocalic systems comparison pertaining to different languages or dialects that present the same number of vowels is investigated both phonologically and phonetically. The experiment are conducted on epta-vocalic systems such as Tuscan, Neapolitan, neo Venetian and northern Calabrian dialects (Lausberg area) and penta-vocalic such as the Italian spoken at Catanzaro, Cosenza etc. The tables and the graphs reported at pp. 61–68 are highly explicative.

Studies make a difference between stressed and unstressed vowels and particularly differentiate the space occupied inside the vocal quadrilateral from the system of stressed vowels with regard to that occupied by the subsystem of unstressed vowels. The unstressed vowels, in fact, are more affected by co-articulation processes like reduction, centralization and neutralization until we reach the complete cancellation. The target of a vowel is reached only during the production of a segment that has a precise status in the phonic chain, that is to say a stressed vowel. Only these segments characterised by a dynamic accent will reach an articulatory target. Unstressed vowels, on the contrary, are often considered transitive elements passing between consonant and consonant. They suffer therefore from limitations from the articulatory point of view, and from those of energy and length.[17] Such influences clearly have an effect on the acoustic production resulting in different formant values [10].

As regards the production of unstressed vowels in final position, the problem is different. There is not only reduction with neutralization of the opposition of medial vowels /e, E/ and /o, ⊣/ but the system can in some cases reduce the number of elements as happens for example in many dialects.

In conclusion with regard to this important step of the identification of the segments to analyse, we can state that before beginning any analysis on recorded material, it is necessary to analyse the linguistic variety that is in question in order to identify the phonological inventory (we cannot in any case assume a priori the existence of a unique phonological inventory of five or seven vowels) and the prosodic rules that underlie the code and, only after that, to continue with the analysis and measurement of the identified vowels.

3. The third step

The third step deals with analysis in external sense, that is to say identification of the steady-state portion[18] inside the segmented vowel and the consequent collection of data.

The measurement process of the segmented vowels formant values, matters at least for two fundamental aspects: the operator and the used equipment. In the case of non high quality equipment, once specified its characteristics, it is possible to calculate the error margin that is reflected in the result and that will be constant in all measurements. In fact, an approximation of a certain weight in the formant measurements will be present in all the measures in the same way. In the case of the operator, on the other hand, being inexpert or without linguistic competence, the validity of the results would be in crisis. There is not, in fact, way to evaluate, a posteriori, the error present in the measurements. Therefore, despite the fact that the method is considered objective in terms of criteria, the vowel identification process and of its steady-state portion to be measured and the data collection remain still exclusively bound to the operator's experience, capacity and competence.

In this context, therefore, it is necessary that the expert be assisted by operators with linguistic competencies, that he explicates clearly what analysis he has adopted in his investigation and the technical characteristics of the equipment used to allow everyone else to reproduce the experiment with the same results.

4. The fourth step

This step regards statistical analysis and comparison. Such an analysis should give as a result not just percentages of similarity, but complex answers, including the system's false error probability and the methodological protocol, strictly bound to the specificity of the voice analysed, to the inadequacy of the measured parameters etc. In conclusion it should give the judge all the necessary information permitting him to come to a judgement on the question of two utterances belonging to the same speaker or not, avoiding the substitution of the judge by an expert, as is often requested.

5. Conclusions

Finally, a methodological protocol should provide for:

1. analysis of recorded material to verify if it is scientifically acceptable and when, on the other hand, it has no sense to do any analysis because of the possibility of high error deriving from it. This means establishing the limits of the characteristics a signal should have in order to be analysed. These characteristics could be: the SNR,[19] the bandwidth, the length of the signal to analyse or better the number of necessary parameters to be collected;
2. identifying exactly the methodology: formant measures, length, identification of VOT, etc. and the measurement technique;
3. identifying exactly the segments to analyse: stressed vowels, unstressed ones, in a penta- or epta-vocalic system etc.;

[17] From the data reported in [9,10] the percentage of a stressed vowel lengthening compared with an unstressed vowel in Cosenza dialect swings from 57.4 to 126.2% in a word list. This value increases if we consider spontaneous speech.

[18] It is argued that during spontaneous speech, the vocal tract remains 'frozen' assuming a precise setting, for at least 20–25 ms during the production of a stressed vowel. This segment, having stability in the formant contours, is steady-state.

[19] As identified in [5], that is to say as an index of intelligibility's value.

4. managing mixed utterances: making different comparisons dialect–dialect, Italian–Italian etc.;
5. establishing exactly the statistical analysis to apply.

Once a methodological protocol has been established for speaker recognition, the next passage has to be the identification of a professional figure, creating a phonic experts list.

The constitution of a phonic experts list and the shared definition of a methodological protocol should be a guarantee for all the individuals involved and would besides create order in the great market of experts and more or less scientific methods. The phonic experts list should contain all the expert professionals' names active in the study of recorded voice, not only with regard to speaker recognition, but about everything that has to do with forensic phonetics (Speaker profiling, intelligibility enhancement of audio recordings, transcription and analysis of disputed utterances, authenticity or integrity examination of audio recordings). This is necessary first of all because for a long time the professionalism of an expert transcriber, expert in general linguistics, psycholinguistics, phonetics and phonology and dialectology, is confused with the figure of court audience transcriber. The audience transcriptions are very often entrusted to societies or cooperatives of copyists, photocopy operators, printers and to people whose professionalism is more strictly bound to administrative matters than to scientific analysis of the recorded voice. Unfortunately, the audience transcriber carries out transcription examinations or is called as a consultant nominated by the court. The reasons for all this are to be studied in the rates of payment applied (equal both for audience transcriptions and forensic transcription's examinations), as well as the handing in time, regular presence in the courtrooms. All this happens both in the case of transcription examinations and examinations concerned with speaker recognition despite the law art. 220, 1° comma of the code of penal procedure recites: "*La perizia è ammessa quando occorre svolgere indagini o acquisire dati o valutazioni che richiedono specifiche competenze tecniche, scientifiche o artistiche*." and the law art. 221, 1° comma c.p.p. recites: "*Il giudice nomina il perito scegliendolo tra gli iscritti negli appositi albi o tra persone fornite di particolare competenza nella specifica disciplina*."

Not applying what has been described so far belittles the professional figure and the phonic expert's expertise who, in the popular imagination, needs no more than a recorder, a pair of headphones and a lot of patience.

References

[1] L. Romito, M. Maddalon, J. Trumper, La parametrizzazione nei test di riconoscimento, in Atti delle VIe Giornate di Studio del gruppo di Fonetica Sperimentale (G.F.S.), sul tema: caratterizzazione del parlatore, Collana degli atti dell'associazione italiana di acustica, Roma, 1996, pp. 87–93.

[2] L. Romito, Manuale di Fonetica Articolatoria, Acustica e Forense, Centro editoriale e librario università degli studi della calabria, Cosenza, 2000.

[3] L. Romito, C. Paucci, I metodi di riconoscimento del parlatore in Europa—Tesi di Laurea A.A. 2002–2003.

[4] Audio Engineering Society, AES standard for forensic purposes. Criteria for the authentication of analog audio tape recordings, Audio Engineering Society Inc., 2000.

[5] L. Romito, Rumore, percezione ed indici di intelligibilità (in stampa).

[6] A. Giannini, M. Pettorino, La Fonetica Sperimentale, Edizioni Scientifiche Italiane, Napoli, 1992.

[7] J. Trumper, L. Romito, M. Maddalon, Vowel systems and areas compared: definitional problems, in Atti del Convegno "L'interfaccia tra Fonologia e Fonetica" a cura di Emanuela Magno Caldognetto e Paola Benincà (Padova 15 Dicembre 1989), Unipress, 1991, pp.43–72.

[8] L. Romito, M. Maddalon, J. Trumper, Atteggiamento della Magistratura nei confronti delle perizie foniche, in Atti delle VIe Giornate di Studio del gruppo di Fonetica Sperimentale (G.F.S.), sul tema: caratterizzazione del parlatore, Collana degli atti dell'associazione italiana di acustica, Roma, 1996, pp. 34–45.

[9] L. Romito, Cenni sui correlati elettroacustici dell'accento in alcune varietà di italiano, in Atti delle IVe Giornate di Studio del Gruppo di Fonetica Sperimentale (G.F.S.), Torino, 1993.

[10] L. Romito, E M. Lorenzi, Considerazioni generali sul comportamento di alcune varietà dialettali meridionali e settentrionali rispetto all'accento intensivo, Quaderni del Dipartimento di Linguistica 15, Università della Calabria, Serie Linguistica 6, Benvenuto (CS), 1997, pp. 11–34.