

The use of statistical methods in forensic speaker identification in Russian Federation

Timur Svirava

Saint-Petersburg State University, Russian Federation

timsvir@mail.ru

(Term paper in Statistical Methods, spring 2009)

ABSTRACT

This paper is aimed at making an outline of statistical methods used in forensic speaker identification in the forensic institutions of Ministry of Justice of Russian Federation.

INTRODUCTION

Forensic expertise of sound records is 1) an investigation of magnetic or other types of sound records with the purpose of establishing facts that could serve as evidence and 2) drawing up a report that could be used in a legal procedure. One of the most common tasks that are solved within this type of expertise is that of forensic speaker identification. The essence of this task consists in establishing of such a stable complex of features that would be sufficient to postulate an individual identity between voice and speech recorded on a source phonogram and those recorded on a sample phonogram¹. In the majority of cases forensic speaker identification, at least in Russia, is supposed to answer the question “*Is it him or not?*” and in this way it is somehow opposed to such science fields as mathematics or cybernetics where this question is solved within the task of verification. So, the notion of forensic identification is different from the notion of identification in these fields (Kaganov, 2005).

While the birth of speaker identification in the United States of America is relatively well traced back and goes back from the paper of Lawrence Kersta (Kersta, 1962, Lindh, 2004)² this is not the case for the former USSR where

¹ According to the theory of forensic identification adopted in Soviet Union and then in Russia the results of the process of identification could be following: 1) establishment of presence of identity, 2) establishment of absence of identity, 3) conclusion about impossibility to solve the task of identification (Koldin, 2003). These results are presented categorically and no probability is used.

² Kersta was convinced that the spectrograms could easily be used to identify speakers and called his method as “voiceprint” in direct analogy with the term “fingerprint”. His

these investigations had been classified as secret for a long time. One of the first proofs of their existence can be found in a novel by Aleksandr Solzhenitsyn released in 1968 (Solzhenitsyn, 2006) and in memoirs of Lev Kopelev (Kopelev, 1991) where an example of application is given. According to them, it took place in 1949 during a criminal investigation of a phone call made to the embassy of the United States of America by a Soviet diplomat who tried to sell the data about Soviet secret agents abroad. There were five persons suspected in this crime but the investigation bodies were unaware of the identity of the person who had made the call. So, the scientists from the Acoustic laboratory were given a tape with the record of the call along with the samples of voice and speech of the five suspects. The description of the methods and devices used for the forensic speaker identification is not detailed enough but the terms used by Solzhenitsyn (“vocoder”, “voiceprints”) might imply that the scientists were aware of the studies being made at that time outside the Soviet Union (Eriksson, 2005).

However, it is only in the 1960-s when forensic speaker identification appeared as a separate type of legal expertise. The use of speech recording in a criminal procedure (more precisely, during an investigative action) was introduced in 1966 when a new article 141.1 “The use of sound recordings during the interrogation” appeared in the Procedural criminal code. Nevertheless, a possibility of records obtained during investigative actions to serve as evidence was at large for several decades until 1990 when the Code of criminal proceedings was expanded with an article 35 “Interception of telephone and other conversations”. Nowadays, the article 189.4 of the Procedural criminal code of Russian Federation permits recording during the interrogation of a person (accused, suspected, witness or victim). Thus, sound records have gradually acquired a status of a document being able to serve as evidence.

At the present time, forensic speaker identification studies are made in the state institutions belonging to 1) Ministry of Internal Affairs (MAI), 2) Ministry of Justice and 3) Federal Security Service (FSS). The techniques used in the latter are secret and not public. The techniques used in MAI were adopted in 1995 and are described in (Fesenko, 1995) and we will focus on the techniques approved in the forensic laboratories of Ministry of Justice which are published in (Gurzhey, 1995).

The major difference between the approaches used in Russia and in other major countries is in the way how the strength of evidence is expressed. If in

paper touches the surface of this visual pattern matching procedure, but concentrates on presenting the results from an experiment that he conducted. In order to demonstrate the simplicity in the procedure he used high school girls who performed the identification of subjects very successfully (Lindh, 2004).

these countries the result of forensic speaker identification expertise is defined via the likelihood ratio³ that we will discuss below, the situation in Russia where courts cannot accept any probabilistic evidence leads to the fact that the strength of the evidence should be expressed in a categorical way⁴. The law allows giving a probabilistic conclusion if a categorical one cannot be given (i.e. “*probably it is his voice*”) but the degree of probability is not stated and the results of such an expertise are *de jure* equal to a conclusion about impossibility to solve the task of forensic identification.

PROBABILISTIC APPROACH

The probabilistic approach used in a number of countries⁵ for example Australia (Rose, 2002). The strength of the evidence in forensic speaker identification here is given by the ratio of two probabilities: the probability of the evidence given the hypothesis that the two samples are from the same speaker, and the probability of the evidence assuming that the samples are from different speakers. This ratio is called the Likelihood Ratio (LR), and can be expressed in the formula

$$LR(E) = \frac{p(E | H_p)}{p(E | H_d)}$$

where H_p is the prosecution hypothesis (the two speech samples come from the same speaker), H_d is the defense hypothesis (the samples come from different speakers) and E for evidence. To put it differently, numerator represents the degree of similarity between the speech samples – the greater the difference between the two speech samples the lower the probability that they were produced by the same speaker. As for the denominator, it represents the typicality of the properties of the speech samples – if the acoustic properties of the speech samples are very common in the population then there is a high probability of obtaining such acoustic properties by picking samples from randomly chosen speakers. Pairs of speech samples which are more similar or more atypical will produce higher likelihood ratios than pairs of speech samples which are more different or more typical.

One of the important features of the approach based on likelihood ratio is that it allows a combination of evidence taken from different sources: either

³ On the basis of the high degree of similarity (this is the evidence), the expert says that it is very likely (a probability judgement) that speech samples come from the same speaker (the prosecution hypothesis) (Rose, 2005).

⁴ In other terms, an expert has to prove or disprove the prosecution hypothesis.

⁵ In the United Kingdom UK Position Statement also gives the recommendation that the expert refrain from giving the probability of hypothesis, given evidence $p(H|E)$. Moreover, it emphasizes that its proposal is not to be considered identification, but comparison, and this is the term they use (Rose, Morrison, 2008).

from different forensic investigations or from different aspects of the same investigation. For example, if the voices on the two samples are compared with respect to many features and a likelihood ratio can be estimated for each of these features, an overall forensic-phonetic Likelihood Ratio (OvLR) can be derived from the ratios for the individual features in the following way

$$\text{OvLR} = f(\text{LR}(E_1), \text{LR}(E_2), \dots, \text{LR}(E_N)).$$

The exact way of the combination function depends on inter(in)dependence of the items of evidence on which single likelihood ratios are based. If they are mutually independent the combined likelihood ratio is the product of their associated likelihood ratios. If the different items of evidence are not independent, combination can become very much more complicated.

Whichever way likelihood ratios from different items of evidence are to be combined, it is clear that, even though single likelihood ratios from individual features are not big but exceed 1, a large number of such ratios can build up into a high-valued combined likelihood ratio. For example, by multiplying four single likelihood ratios equal to 3 we will get the combined likelihood ratio equal to $3^4=81$. On the other hand, a single likelihood ratio value much less than 1 could reverse the overall positive evidence to a negative one. So, if we add one more ratio equal to 0.01 we will have the combined likelihood ratio 0.83 i.e. lower than one.

Likelihood ratio	Proposed verbal equivalent
>10 000	Very strong evidence to support . . .
1000 to 10 000	Strong evidence to support . . .
100 to 1000	Moderately strong evidence to support . . .
10 to 100	Moderate evidence to support . . .
1 to 10	Limited evidence to support . . .
1 to 0.1	Limited evidence against . . .
0.1 to 0.01	Moderate evidence against . . .
0.01 to 0.001	Moderately strong evidence against . . .
0.001 to 0.0001	Strong evidence against . . .
<0.0001	Very strong evidence against . . .

Table 1. Verbal equivalents for likelihood ratios (Champod and Evett 2000, 240).

The value of the likelihood ratio thus quantifies the strength of the evidence. Since the numerical form of likelihood ratio is not always easily interpretable to the court, translations into verbal scales have been proposed. One proposal, used at the Forensic Science Service, is shown in Table 1.

An extremely important role in the evaluation of forensic identification evidence is then played by the concept of prior odds i.e. (in case of forensic speaker identification) the odds in favor of the hypothesis of common origin for

two or more speech samples before the forensic-phonetic evidence. It is simply the ratio of the probability of the assertion being true, $p(A_p)$ (i.e. that the samples come from the same speaker) to that of it being false, $p(\sim A_p)$.

After an expertise has been made and an overall likelihood ratio has been obtained, basing on the prior odds and the voice evidence, we can calculate the value of the posterior odds for believing the assertion (A_p), given the initial conditions and the forensic-phonetic evidence (E) using the formula which is in fact the odds version of Bayes' theorem:

$$\frac{p(A_p | E)}{p(\sim A_p | E)} = \frac{p(A_p)}{p(\sim A_p)} \times \frac{p(E | A_p)}{p(E | \sim A_p)}$$

THE SITUATION IN FORENSIC LABORATORIES IN RUSSIA

In order to describe and understand the role of statistical methods for forensic speaker identification studies effectuated within the forensic laboratories belonging to the Ministry of Justice we should start with a brief description of such a study. An identification study is a multidimensional complex study and consists of three main parts which are auditive, linguistic and instrumental (Gurzhey, 1995; Kaganov, 2005). They are concluded with the synthesis part where, on the basis of the results obtained in the three preceding parts, an expert comes to the final conclusion.

In the course of the auditive part of the study, the following features are studied perceptually and with the use of instrumental control devices: general auditive impression properties (e.g. voice perception, manner of articulation, voice timbre etc.), properties of speech process realization (such as loudness, speech rate, pitch stress properties etc.) and individual properties such as sex, age, emotional state etc.

Linguistic part of the study is aimed at finding similarities and differences between the speech samples. This is made on the basis of their distinctive features of realization of speech units of different levels. Properties of speech units are usually grouped in phonetic, syntactic and lexical groups. These properties, once determined by audition are then checked with instrumental analysis of the speech signal (sonograms, oscillogram, cepstral analysis and so on).

The results of the first two part of the study are usually presented in the final paper as a table or a text where the similarities or differences between two speech records are presented⁶. No statistics are usually used within these parts

⁶ We could add that in some, but very vague, sense this enumeration could be said to resemble the nominator in the formula:

$$LR(E) = \frac{p(E | H_p)}{p(E | H_d)}$$

and the result of the study is presented for example in the following way: “comparative linguistic analysis has elicited a coincidence (of a number) of identification-significant features extracted from the speech on the phonograms”. As we can see, neither the strength of this coincidence nor any indications on the co-occurrence of such features in the speech of different persons are given.

The only part of the study where some statistical analysis of the data has to be made explicitly is its third part, the instrumental analysis. Although according to the special law describing the activity of state forensic institutions (Federal law, 2003) an expert is free to choose the methods that will be used in his expertise, three techniques are the most widely used in this part, these are pitch analysis (Koval, 2002), formants analysis (Kaganov, 2005) and formants matching technique (Koval, 2007).

Pitch analysis

This technique is aimed at an evaluation of individual acoustic features that characterize the functioning of the source of incitation of speaker’s vocal path. Here, statistical characteristics of pitch and pitch contour are analyzed. It is accepted (Kaganov, 2005) that these features are the most time-independent and their values are characteristic of speaker individuality.

The analysis is made by the special software used for speech analysis in the forensic laboratories of the Ministry of Justice. The most widespread software is Speech Interactive System (SIS) designed by Speech Technologies Centre (Saint Petersburg), Justiphone by Aim Technologies (Moscow) and OTEExpert by OT Contact (Moscow). We will consider the first of the three as it is the most common and has been used for more than ten years already.

SIS in its current versions (6.x) offers almost fully automated pitch analysis of two samples of speech and provides a decision whether they belong to the same speaker or not. The main task of an expert is to find two adequate samples from the data he has, as they should be emotionally and intonationally comparable, the noises that could impede pitch extraction should be absent etc. There are also some restrictions on the length of speech data in the technique. So, the manuals (Methodical recommendations..., 2000, 2002) insist on having at least 30 seconds of voiced speech data (excluding pauses and voiceless

However, an expert usually tries to give support to the prosecution hypothesis by showing the number of similarities in the two samples; the fact whether similar acoustic or phonetic features are widespread among the native speakers or not is usually not referred to. Moreover, it often happens that an expert tries to explain the differences he found in comparable acoustic or phonetic features (for example, differences in speech rate) by the fact that records had been made under different conditions (for instance a suspected person had been full of emotions speaking over the phone but was highly depressed when he was giving his speech samples as he had already been under arrest etc).

consonants) in each sample in order to obtain enough statistics. In the current versions 16 parameters of pitch are used, a list of them is given in Table 27. False acceptance and false rejection probabilities are computed for each parameter, they are based on built-in statistics used in software training. Nevertheless, the criteria are not explicit in the manual so the confidence level rests unclear. FA and FR values for the weighted relative deviation of pitch features are computed then. We should also note that each feature has its own built-in weight pitch median having the heaviest one. The overall decision, based on FA and FR values, is made afterwards. The decision of the software is always categorical but not probabilistic. If FR exceeds FA the answer is positive (i.e. “the voices belong to the same speaker”), if not it is negative.

Parameter	Signal I	Signal II	FA%	FR%
Mean value, Hz	121.7	120.4	4.5	86.1
Max. value, Hz	217.9	185.7	45.6	27.8
Min value, Hz	79.6	74.5	25.5	44.0
Max. value - 3%, Hz	173.0	159.0	29.2	37.7
Min. value + 1%, Hz	88.7	94.7	25.7	31.3
Median, Hz	119.5	118.3	4.2	87.1
Percentage of sections with rising pitch, %	44.2	47.6	49.6	42.0
Dispersion of pitch logarithm	0.0050	0.0028	61.3	5.6
Skewness of pitch logarithm	0.56	0.57	1.7	97.5
Kurtosis of pitch logarithm	2.99	3.77	54.4	32.7
Mean speed of pitch change, % per second	-18	-13	19.6	73.8
Dispersion of pitch log. derivative	2.44	2.62	16.6	71.7
Skewness of pitch log. derivative	-0.17	-0.31	26.5	65.9
Kurtosis of pitch log. derivative	15.78	16.40	12.2	78.5
Mean value of pitch rise, % per second	695	635	16.0	72.0
Mean value of pitch fall, % per second	646	671	7.4	86.4
Weighted relative deviation of pitch features			6.4	46.7
FA – portability of false acceptance, i.e. probability of an error if we decide that there is the same person speaking on the two phonograms; the lower FA value is the higher is the probability that it is the same person speaking. FR – probability of false rejection, i.e. probability of an error if we decide that there is not the same person speaking on the two phonograms; the lower FA value is the higher is the probability that it is not the same person speaking. Overall decision based on an analysis of pitch statistics: the voices on the phonograms belong to the same speaker.				

Table 2. An example of comparison of pitch statistics made by SIS software.

On the other hand, such software as Justiphone and OTExpert do not take the decision themselves but make an expert responsible for it. For this purposes weighted average relative deviation is used (Kaganov, 2005):

⁷ The data are taken from a real forensic speaker identification study.

$$\Delta = \frac{\sum_j \left| \frac{x_j - y_j}{y_j} \right|}{j} \times 100\%,$$

where x_j and y_j are the values of parameter j in the two samples of speech respectively. The resulting Δ is compared to 20% this value being set as maximum allowable intra-speaker variation (Kaganov, 2008).

This type of analysis is the most widespread due to its relative simplicity and the fact that it is less dependent on the signal quality and the signal/noise ratio than the following techniques.

Formants analysis

If the former analysis is aimed at the comparison of voice incitation source features during the formant analysis tries to reveal the features of the “filter” part if we refer to Fant’s source-filter model (Fant, 1970), namely, resonant frequencies or formants. So, the properties of formants of similar speech sounds are compared. Usually, vowels like /a/, /u/, /i/ in comparable phonetic positions (left and right context, stress position, comparable loudness and emotional state of the speakers and so on). During such a comparison a table is being filled where the values of the F0-F4 (or F2/F1, F3/F2 and so on ratios) measured on the central parts of the vowels are indicated. Normally, up to 20 pairs of realizations are needed (Methodical recommendations..., 2002). If the mean values for the two samples approach one another to a considerable degree it could be possible to say that the speakers on the samples have comparable speech habits in realization of these sounds (Koval, Zubova, 2007).

Formants matching technique

Finally, formants matching technique developed by Koval et al. (Koval, 2007; Koval, Labutin, Raev, 1995) from the theoretical point of view corresponds to indirect comparison of geometrical features of anatomy of vocal path of speakers. The main idea of this technique is based on the assumption that a speaker, while pronouncing speech sounds, is able to change configuration of his vocal path only within the limits imposed by the strict anatomic constraints. Every configuration can be controlled by a speaker only in its general geometrical measurements that ensure the realization of target acoustic resonant properties in the low-frequency domain of the spectrum, i.e. only for the first two or three formants. Resonant properties of each vocal path configuration for the fourth, fifth and so on formants are not usually controlled by a speaker but are conditioned by existing anatomic restrictions on possible configuration changes of the vocal path of the speaker. To put it differently, if the first three formants’ values are fixed then higher formants in the speech of the same speaker could take up only more or less stable individual positions.

High-frequency formant structure of speech of a given person (if the low-frequency formant structure is fixed) is thus often stable, invariable in time and is not subject to premeditated or unintentional distortion by a speaker i.e. high-frequency formant structure could serve as his stable biometric description and a comparison (identification) of these formant structures – and, via them, of geometrical properties of the speech path – makes possible speaker identification with a high degree of reliability. So, an expert's task is to find a number of coincident vocal path configurations and, if he can consider such a coincidence being not accidental, take a positive identification decision. Three conditions have to be fulfilled:

- 1) the number of coincident vocal path configurations is sufficient;
- 2) this coincidence is sufficiently precise;
- 3) if these coincidences represent relatively independent configurations of vocal path (i.e. articulations of sounds whose vocal path configuration are very different).

This technique is claimed (Methodical recommendations..., 2000, 2002) to be reliable enough. So, the values of formant frequencies can be measured with a more or less high degree of accuracy depending on the signal/noise ratio in the speech signal, pitch value⁸ and so on. The third formant's exact value might have 5 to 10 different positions as its position is not so strictly fixed by the phonological system of a language as the values of the first and the second formants. If we assume that we can distinguish between 5 positions of the third formant the probability of coincidence of the first three formants in the two samples of speech is claimed to be 1/25. Similarly, if we find 8 pairs of segments where vocal path configurations are independent but the first four formants coincide we might claim to have the target threshold probability of an error being as high as 10⁸, it approximately corresponds to an accidentally coinciding pair of speakers in a set of 14000 speakers (Koval, 2007). If we can distinguish between more than five positions of the fourth formant or can measure the values of the fifth formant we will need fewer pairs to ascertain the same probability.

We will not discuss here the weak theoretical points of this technique as it is semi-heuristic, like the majority of techniques used for forensic speaker identification. As for its application we could say that, in theory, it allows language-independent forensic speaker identification as it deals with similar articulation events but not with the pronunciation of the same allophones in the same context⁹. On the other hand, the practice shows that the parameters

⁸ A formant frequency value is not supposed to be measured with a precision exceeding one half of the pitch value. So if FO value is 100Hz in some part of a signal, maximum accuracy of such a measurement will be 50Hz.

⁹ This fact could be important enough for the forensic practice as the methods of identification approved by the Ministry of Justice can deal only with the speech in Russian.

of the speech signal (especially the records made via mobile phone channel with the bandwidth of 3500Hz) do not make possible distinguishing of the fourth and the higher formants especially for high-pitched female voices.

CONCLUSION

Having made an outline of the use of statistical methods in forensic speaker recognition in the laboratories of Ministry of Justice we are now able to make some conclusions.

- the most crucial difference between the essence of the forensic speaker identification expertise in Russia and some other countries consists in the fact that an expert evaluates the probability of hypothesis, given evidence;
- due to the lack of likelihood ratio equivalent and the fact that only a categorical conclusion may serve as an evidence in a court, the results of an expertise are intended to be presented in non-probabilistic way as ascertained identity of speakers on the samples;
- the use of statistical methods to evaluate similarity/differences in speech data does not seem to be adequate enough. It concerns auditive and linguistic analysis to a greater extent;
- as the preliminary results after each part of the study are presented in a more or less probabilistic way (“*resemblance or equality of the features*”) it does not seem that there are enough theoretical, statistical and other bases to ascertain the identity of the speaker in the synthesis part of the study.

REFERENCES

1. Champod, C. and Evett, I. (2000): Commentary on Broeders (1999), in *Forensic linguistics* 7/2, p. 238–43.
2. Eriksson, A. (2005): Tutorial on forensic speech science. In *Interspeech*, Lisbon, Portugal.
3. Fant, G. (1970): *Acoustic theory of speech production*. The Hague, Netherlands: Mouton, 2nd ed.
4. Federal Law on State forensic activity in the Russian Federation №73-FZ, May 31st of 2003.
5. Fesenko, A.V., editor (1995): The technique of speaker identification by phonograms in Russian using the automated system “Dialect”, Moscow: Voyskovaya chast’ 34435 (in Russian).
6. Gurzhey, T.I., editor (1995): Speaker identification by magnet records of his speech (Method manual for experts, investigators and judges), Moscow: Russian Federal Center of Forensic expertise of the Ministry of Justice on the Russian Federation (in Russian).
7. Kaganov, A.Sh. (2005): Forensic expertise of sound records, Moscow: Jurlitinform (in Russian).

8. Kaganov, A.Sh. (2008): The instrumental study of characteristics of speech path excitation for speaker identification purposes, in *Forensic expertise № 3 (15)*, Saratov (in Russian).
9. Kersta, L. G. (1962): Voiceprint identification. In *Nature* 196: 1253-1257.
10. Kolidin, V. Ya. (2003): Forensic identification, Moscow: Leks-Est (in Russian)
11. Kopelev L. (1991): Slake my sads. Memoires. Moscow: Slovo (in Russian).
12. Koval S.L., editor (2002): The book of methodical recommendations for the realization of forensic expertise of speech records, Saint Petersburg: Speech Technologies centre (in Russian).
13. Koval, S.L. (2007): The usage of formants matching technique in the instrumental part of forensic speaker identification expertise, in *Theory and practice of forensic expertise № 3 (7)*, Moscow: Nauka, p. 160-174 (in Russian).
14. Koval, S.L., Labutin, P.V., Raev, A.N. (1995): Automatic speaker recognition using formants-based nearest-neighbor distance measure, in *Proceeding EUROSPEECH'95*, Madrid, vol. 2, p. 341-344.
15. Koval, S.L., Zubova, P.I. (2007): Speaker identification by his voice and speech on the basis of complex analysis of phonograms, in *Theory and practice of forensic expertise № 3 (7)*, Moscow: Nauka, p. 68-76 (in Russian).
16. Lindh, J. (2004): Handling the "Voiceprint" Issue, in *Proceedings, FONETIK 2004, Dept. of Linguistics, Stockholm University*, p.72-75.
17. Methodical recommendations on the practical usage of SIS software for the analysis of speech signals (2000), Saint-Petersburg: Speech Technologies Centre (in Russian).
18. Methodical recommendations on the practical usage of SIS software for the analysis of speech signals (2002), Saint-Petersburg: Speech Technologies Centre (in Russian).
19. Rose, P. (2002): Forensic speaker identification, London: Taylor & Francis.
20. Rose, P. (2005) Forensic speaker recognition at the beginning of the twenty-first century – An overview and a demonstration. *Australian Journal of Forensic Sciences*, 37, p. 49–72.
21. Rose, P., & Morrison, G. S. (2008) A response to the UK position statement on forensic speaker comparison. Manuscript submitted 30 October 2008 for publication in *the International Journal of Speech Language & the Law*.
22. Solzhenitsyn A. I. (2006): *The first circle*, Moscow: AST: AST Moskva (in Russian).