

E.I. Galyasina

**OBJECTIVE AND SUBJECTIVE COMPONENTS OF THE INTEGRATED
SPEAKER RECOGNITION SYSTEMS**

*Bureau of an independent expert appraisal "Versia"
Russia, 117630, Moscow, Starokaluzhskoye shosse, 58
Ph.: (095) 394-71-45; a fax: (095) 330-80-44;
E-mail: galyasina@mtu-net.ru*

The integration of an objective and subjective (expert) components in the solution of some applied problems of aural-perceptual and automatic speech and speaker recognition is discussed. The perception of a subjective acoustical image perceived by an acoustical analyser of the person, and audio signal by the way of parametric description and mathematical representation, their correlation on different analysis stages and decision making in forensically applied combined expert-computer systems of speech and speaker recognition and are considered.

In modern speech technologies and applications alongside with subjective recognition of acoustical images the different automated methods and combined (expert - computer) systems are widely applied. It is obvious that completely automatic systems of speech recognition, speaker identification and verifications display poor efficiency and reliability that is not adequate to criteria and the inquiries of forensic applications in Russia. Thus the degree of participation of the person - operator (or expert) in the automated voice technologies can be different. The computer can execute the role of an accessory instrument; labour saving of the person, and all analysis stages and decision-making implement are all at direct participation and control of the expert. And the person can fulfil a pure operator function when the computer carries out the whole procedure of investigation and conducts the rules for the decision.

So, in forensic researches in Russia the share of a subjective and objective component is not identical and can considerably vary depending on a particular problem, facing the expert. For example, for the practical solution of a problem of speech recognition on a phonogram engaged in criminal cases, especially followed by noise, the subjective method prevails. The speech recognition (transformation a signal to text) implements the expert (or commission of experts) to use a perception method of repeated listening of a recorded signal. A method of reading "of blind" sonograms, that is subjective visual perception, sometimes will auxiliary be used. Thus the reliability of speech recognition is determined by auditor abilities, skills of the expert, and also in many respects depend on quality of sound recording. The objectivity of the solution of such problem is reached by application of monitoring listening by several independent commissions of experts. The comfort of perception and efficiency of speech and speaker recognition recorded on a tape is increased at the expense of application of instrument methods of noise reduction or filtration. However the final solution on the text contents of an audio signal is received extremely on the basis of subjective perception of the expert.

Other problem – determination an authenticity of sound recording, vice-versa, in the greater degree will involve objective - hardware methods and in a minimum degree is determined by the subject of an expertise. Here relevant role is played by robust acoustic parameters, precisely measured by objective methods and tracing their abnormalities, that sometimes and can be detected by expert perception, but are finally valued by results of precise measurements through appropriate measuring instruments and computer devices.

It is necessary to mark, that the solution on originality of phonogram and identity of the utilised means of sound recording, is received by the expert, on the basis of estimations of the indications of measuring devices and computer processing of their results. In that case reliability of the solutions and conclusions of the expert depends on accuracy and resolution of used instrumentation, reliability of instrument methods of acoustic measurements.

The problem of forensic speaker recognition (including identification, verification and discrimination of the speaker), equally will involve both subjective and objective components. The expert participates in research, search of features and decision making - at all stages of segmentation and allocation of speech units from a continuous signal, their qualitative and quantitative assessment on the basis of integration of results of a perception and precise acoustic measurements. Under the

control of the expert all objective researches are conducted. In cases, when the results of parametric processing of a signal do not suit the expert or have no indispensable reliability, a new training of the automated system is made, or the expert selects diverse strategy of acoustic measurements.

The forensic recognition of the speaker by speech samples can be made purely subjective and by automatic methods. The subjective methods can be audio-perceptual or visual, and the objective methods are completely grounded on an automatic procedure of matching of speech samples, when the computer offers the solution on an identity or distinction of the speakers. The subjective perceptual method is grounded on psychophysical matching and together with linguistic analysis is a mandatory and integral part of any forensic speaker research. Thus, special training of the expert, availability of acoustical experience, excellence of acoustical memory, ability to trap finest nuances of sounding speech, difficult for instrument measurement, high ability of adapting of hearing to a high level of noise and distortions, highest resolution of an acoustical analyser of the person are taken into account.

Extreme case of application of subjective speaker recognition method is the so-called acoustical identification, when the witnesses hearing a voice of the suspect, attempt to identify him by memory on the basis of a homogeneous line up of voice recordings. In forensic practice such method can be applied, when there is no tape recording, and a unique source of the proof becomes an acoustical image kept in memory of the witness. When a tape-recording is available the method of an acoustical perception, will be used by the experts for allocation of phonetic and other linguistic characteristics of speech. The reliability of such speech recognition in many respects depends on personal abilities of the expert, his acoustical memory, degree of acquaintance with a voice and other linguistic and extra linguistic factors. The method of a visual perception is grounded on matching of sonogram representation of voice samples. The reliability of such method is stipulated by qualification of the expert, his experience "of reading" of voiceprints (spectrograms), abilities to correlate the graphic information to an appropriate acoustical image. The automatic speaker identification is due to progress in the field of computer and voice technologies and is grounded on an objective method of measurement of voice parameters and matching of the parametric descriptions of researched speech samples. Thus the number of measured and computed parameters can be considerable [1].

For usage of results of speech recognition in expertise there is a lot of known limitations. Essential is that the results of speaker identification should be reliable enough to use in court and pursuant to the standards of a procedural law of Russia. In forensics the identification should be only 100-percent. At any different result (let even 99,999 %) it can be spoken only about a probabilistic conclusion on expertise, and it is treated for the benefit of a suspected person. Besides the limitations are superimposed by different conditions and circumstances of sound recording (availability of distortions, not desire of the suspect to cooperate with automatic recognition systems, capability of masking and imitations, simulation of a voice characteristics). The problem also that a voice of the speaker being a subject to identification can be similar to voices of homogeneous group of people of close social environment and one age group, and also undergo changes in period between records of the measurement standard sample and voice of a suspected one. Because of such factors in practice it is difficult to receive reliable results of speaker identification using only subjective or strictly objective automatic methods.

Therefore in practice of expert testimony in Russian courts the increasing popularity is conquered by the combined systems operating an integrated procedure, complex of subjective and objective methods. The combined expert-computer systems allow the expert to use both objective instrument methods of voice analysis and subjective - at a level of an acoustical perception and visual analysis of mathematical representation of a speech sample and, as the domestic practice of production of expertises displays, are optimal. It is necessary to underline, that the speaker recognition decision is carried out not on the basis of mechanical summing of results of matching of the parametric description obtained through special computing procedures and an acoustical perception but on the integration of knowledge received by the expert. So it is impossible to make an authentic conclusion about a person identity on the basis of mechanical summing of two, three etc. probable conclusions about voice likeness.

The perceptual analysis of sounding speech is a process of identification of sound images with the measurement standards of perceptual base. By means of the measurement standards it is possible

to perceive not only such soundings, which coincide the measurement standards, known for the expert, but also such, which one are only similar to them, resemble them. The measurement standards of perceptual base have the zoned nature, which one integrates points appropriate to identically perceived implementations. For the expert the measurement standards of sounding of phonetic units are formed by an inductive and deductive way as a result of numerous acoustical experiences and directional training. The process of voice perception in combined system of speaker recognition has multilevel stratification including phonetic, morphological, syntactical and semantic levels [1].

Apparently, the first stage of process of speech perception is frequency analyses of an audio signal by an acoustical analyser of the person and conversion it in the form accessible to further processing of an indicated signal by a nervous system. The availability of the phoneme description of words in perceptual base allows the expert to analyse a lexical structure of the sentence. Its partitioning on phonetic words and linguistic description of syntactical structure is made for understanding sense of the expression. The eventual result of a perception is an understanding of sense of signal information. The process of the perceptual analysis, recognition and interpretation of the message is not limited by the analysis of a physical (acoustic) signal and summit levels of the language. Great value is coming from preliminary knowledge, expert experience, motivations, purposes, intentions and characteristics of the speaker and the listener. The acoustic properties of sounds at perception of sounding speech do not play for listening the main roles. This information, certainly, will be used but is largely supplemented by a language data, including semantic. Thus it is impossible to select in hierarchy of levels of a language data operating at perception of speech during the perceptual analysis by the expert, any one level as main or carrying on. The perception represents composite process, and the indispensable language data will be used simultaneously and in the complex, and all levels are interdependent and supplement one another, derivate a stratified whole [2].

The function of the expert is segmentation of a speech sample on discrete language units. During partitioning micro segments, which are divided on intra sound units (parametric events, inter sound transitions etc.), sound segments and combinations of sounds (syllables) are segmented. Macro segmentation on phonetic words, phrases, super phrase unities (paragraphs) and text is also conducted [1]. In practice the court experts will use miscellaneous stock (set) of speech elements, voice units and acoustic segments. There is no a unified generally accepted algorithm of segmentation on the basis of precise delimitation of each unit on changes in the acoustic schedule (on the basis of measurement of energy, fundamental frequency, spectral changes etc.). There is a great demand of an unambiguous correlation between acoustic phenomena of speech and elementary sound units, lying in their fundamentals stably working alongside with signal distortions, noise and interferences. The segmentation is carried out paralleling with phonetic identification of acoustic segments or speech elements on the basis of an acoustical perception of the expert and visual analysis of oscillograph and spectral voice representations. Thus, the problem of precise presence of boundaries of acoustic phenomena in a speech sample and their unequivocal segmentation on the basis of unified stock micro and macronutrients of speech remains actual and perspective.

Other function of the expert is sampling suitable for research and comparable units of speech, selection of the strategy and tactics of the instrument analysis, processing of speech elements by instrument investigation, obtaining of the primary parametric descriptions, secondary statistical processing, estimation of results and making the solution. On a share of the instrument analysis is a precise converting a speech signal being recorded on a tape to a digital form. The hardware component of a combined system also executes function of visualization and mathematical representation of an audio signal in spectral and time-domain. The toolkit for segmentation of a sound wave on micro and macro speech segments, appropriate to language units, and also for measurements of acoustic parameters and calculation of values of demanded identification features is granted to the expert. In a number of the automated systems the instrument component actuates also proposal of the candidate solutions on an identity or distinction of the compared speakers on the basis of matching vectors of parameters with a definite threshold, a capability of additional training of a automated system, choosing combinatory of measured spectral - temporary parameters and computed values of acoustic tags [1].

The main problem of the used today integrated expert-computer systems of speech and speaker recognition is a reliability prediction and reliabilities of received results. To evaluate reliability of an integrated expert - computer system means, to evaluate reliability of each component and their combinations. Reliability of a subjective component is knowledge, ability, skills, experience and qualification, and also subjective emotional and psycho physiological condition of the expert (registration of the factor of uneasiness of the operator).

But a cardinal problem is installation of equivalence of a perception of a subjective acoustical image perceived by an acoustical analyser of the person, and audio signal by the way of its parametric description and mathematical representation, their correlation on different analysis stages and decision marking in combined expert-computer systems of speech and speaker recognition.

In practice the reliability is reached by highest qualification of the subject of expertise. An expert possessing miscellaneous method of speech testing should have the demanded competence for the integrated estimation of the results of audio and instrument analysis. But the difficulties of a synthesizing stage arise, when the expertise is carried out not by one encyclopaedically educated expert but two or three experts of miscellaneous specialities, for example, radio engineer, philologist, physics. Each expert is competent in fulfilling the research by one method and is not skilled to evaluate the results, obtained by the colleague. Thus each expert can use as objects of research different units of speech. The unequivocal rules of decision marking on the basis of the complex of parametric and perceptual estimations of miscellaneous speech units is not fulfilled yet and the final solution frequently is received by a so-called method of fluidised voting. Thus it is considered, that the qualitative value of speech feature detected by an acoustical perception of an audio signal, is such a characteristic that in acoustical space of expert perception is recognized as nearest to a parsed feature of acoustic stimuli. The expert estimation and perceptual, psycho-acoustic measurement perceived by the acoustical analysers of the person of parameters of an audio signal is reduced to psychophysical measurements of acoustical incentives by a method of pair matching of mental and acoustical images and is recognized as physical representation of acoustic event.

Here it is necessary briefly consider some terms and definitions used in practice of expert testimonies in court: «acoustic parameter », « acoustic correlate», «perceptual feature», «a phonetic feature» in their inter coupling. It is represented, that it is possible to consider the use of terms «spectral parameter» and «temporary parameter» in value of the objective indications on an output of measuring device used at measurement of the characteristics of speech segments in spectral and a time-domain. For the denotation of results of the acoustic, perceptual and linguistic analyses it is possible to use combinations of terms accordingly «acoustic feature », «perceptual feature», «linguistic feature». In such context the term «acoustic correlate » and «a phonetic feature» can be used for the denotation of the acoustic character perceived by listening, and appropriate identifiable phonetic sequence of speech segments. However it is necessary to recognize, that the problem of unification of a used nomenclature of terms and expressions remains actual.

For a reliability augmentation and reliability of a subjective component of combined forensic expert-computer systems of speech and speaker recognition especial education and training of the experts of miscellaneous specialities for obtaining general qualification in comprehensive analysis of oral speech is necessary. The reliability of an objective component of the automated systems is determined by reliability of results of acoustic measurements and efficiency of used instrumental base, application of metering devices, with an authentically established error. Thus measuring instruments should be introduced in the state register of means of measurements and pass metrology check (law of Russian Federation (1993):"On maintenance of unity of measurements "). The methods of measurement of basic voice parameters (duration of voice segments, frequency of a fundamental frequency, formant frequencies etc.) are to be standardized and certified. The Russian standard 8.563-96 states that the technique of fulfilment of measurements is determined as set of operations and rules, obtaining data with a known error. Thus different errors are taken into account. Instrument errors can be caused by means of measurement devices. Mathematical models influence methodical errors of measurements and algorithms used in a measuring procedure.

Usage of combined integrated expert-computer systems for forensic speaker and speech recognition will be effective and reliable under condition of integration of a subjective and instrument

components. Development and implantation of the standardized technique of acoustic measurements of the main voice parameters alongside with special expert education and hard training are perspective for the development of forensic speech technology in Russia.

REFERENCES

1. Potapova R.K. Speech: the communication, computer science, and cybernetics. Ì. «Radio and communication», 1997ã. (In Russian).
2. Zlatoustova L.V., Galyasina È.I. Speaker recognition by acoustic-perceptual characteristics of sounding speech. In “The theory and practice of speech researches”. Proceedings of the Conference. Ì, September 14-18, 1999, page 60-80