DEVELOPMENT OF METHOD AND DEVICE FOR IMPROVED REAL-TIME SPEECH RECOGNITION RELIABILITY

K.P. MAISTRENKO

V.M. Glushkov Institute of Cybernetics
Academy of Sciences of the Ukrainian SSR
Kiev 207, USSR 252207

ABSTRACT

The method and the device for invariant voice images recognition are suggested aimed at improving the real-time voice images recognition reliability when an arbitrary number of operators are involved.

The experience of numerous investigations dealing with the problem of designing the voice signal recognition systems testifies to the great difficulties involved in providing high reliability for these systems even in the case of one dictating operator. This is explained by great variability of the principal voice signals parameters which considerably grows when a wider circle of operators is involved. The evaluation of variational boundaries of the voice signal parameters variation shows that the interval of its intensity variation makes approximately 60 db, the frequency spectrum ranges from 20 Hz to 20-22 kHz, the continuity of pronouncing the voice images by different operators can be twice as quick or slow varying from operator to operator.

Of importance is also the fact of the great variability of the Russian language phonemes, which is related to the position of a phoneme in a word; the qualitative composition and the form of neighboring phonemes surrounding this one; the fact whether the phoneme is stressed or not; the rhythmic-dynamic structure of word combinations and word-forms, etc.

However, the human aural system possesses the unique capabilities with respect to fruitful voice signals recognition irrespective of the voice, rate and sound intensity of pronunciation. These properties, inherent in the aural system, have recently attracted the great attention of investigators and designers of speech recognition systems.

Working with the voice signals whose boundary parameters and possible variation of intensity, the frequency spectrum, the duration and rate of sounds pronunciation were considered above, one gets convinced that the following procedures are expedient:

efficient signal reception in the environment and formation of its acoustic analog (or establishment of interface between a recognition system and a signal source);

- normalization of voice signal with regard to intensity (amplitude) taking care that the optimal level is constant
- normalization of voice signal with regard to frequency characteristics at the expense of maximum restriction of possible variation of timbre, prosodic and emotional coloration (and/or variation of the principal tone frequency);
- normalization of voice signal with regard to duration of vocal speech units pronunciation adopted to the rate of information inflow.

The solution of these problems affords formation of the invariant description of the voice signal which would be the least influenced by the negative effects of speech variability and obvious redundancy interfering with the recognition system reliability.

Proceeding from this concept, we developed the method of voice images recognition using invariant voice signals processing [1,2] . The essence of this method is better understood on considering the troublespots of the known methods of voice signal recognition.

Thus, there are voice signal recognition methods where isolation of the voice signals' attributes is realized through the use of coding and of articulation attributes [3] .

However, standardization with regard to frequency of the principal tone results in the accuracy reduction when isolating the voice signals attributes.

The well-known method and the device for its implementation imply conversion of the voice images into an electric signal, amplification, phonemes separation, dynamic spectral analysis, quantization, separation of phoneme's attributes, their normalization and comparison with the reference signal [4] .

The reference signal is formed as a sum of power functions possessing the fixed transfer characteristics.

The weak point of this method consists in low accuracy and rate of voice images recognition because the accurate power functions can be obtained only with the help of the ideal multiplier.

The objective of the suggested method resides in improving the accuracy and rate of voice images recognition. To this end and according to the given method, the amplified electric signal is standardized with regard to continuity, after quantization it is normalized in frequency and amplitude and according to the obtained signals the short pulses are formed which are integrated and normalized with regard to continuity and compared with the invariant reference signals of the voice images obtained in the process of teaching the recognition system. The reference functions are formed by choosing the scale factors with respect to the signal of mismatch between the function being compared and the reference one.

The preliminary standardization of the voice image signals with respect to continuity and quantization ensures the consequent and synchronized operation of the entire analysis route. Formation of invariants with respect to frequency by way of simultaneous speech analysis and synthesis eliminates the effects produced by scattering of speech sounds tonality. The normal well-articulated voice of the synthesized sounds will always be heard at the frequency invariantor's output no matter how high the operator's voice tonality. Standardization of voice image signals with respect to amplitude, when uniformly weak or strong voice signals are amplified, expands the dynamic input signal range, preserves the highest formants in the spectrum which are usually lost when the speech is clipped.

Preliminary voice signal processing and normalization in continuity by way of invariants formation with regard to continuity increases the rate of voice image recognition providing high recognition accuracy. What is more, the comparison of voice images converted in the above-mentioned way with the reference signals obtained by functional conversion $X \to f_i(X)$

where X and $f_i(X)$ -

are independent functions, and by preliminary record of scale factors (during teaching) whose sampling is performed at high speed in the process of recognition, considerably increases the accuracy and rate of voice image recognition.

The investigations carried out with the use of recognition system models and experimental breadboards made it possible to suggest the variants of hardware implementation of invariantors of voice image amplitude, frequency and continuity which sufficiently reduce the redundancy of voice signals, enhance the invariance of recognition systems intended for the real-time operation with an arbitrary circle of operators [5,6] .

The suggested way of invariant voice signal processing can also be applied to processing the sound and acoustic signals for the purpose of their analysis, synthesis and recognition [7,8] .

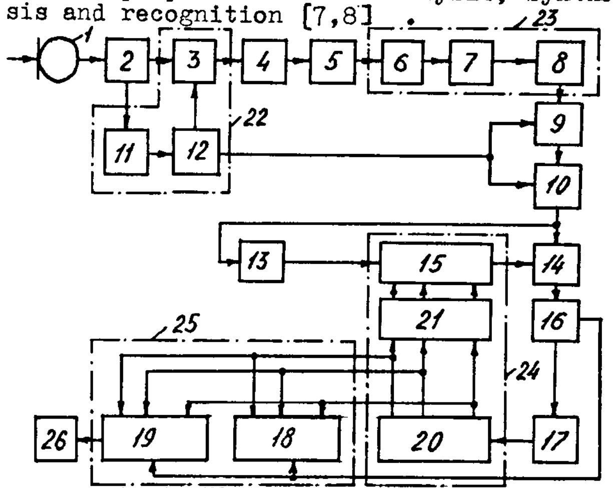

Fig.1. Flowchart of the device for invariant voice images recognition.

Fig.1 shows the device flowchart of the invariant voice image recognition implementing the suggested method. The device contains a microphone 1, an amplifier 2, an electron key 3, a frequency invariantor 4, an amplitude invariantor 5, an amplifier-limiter 6, Schmitt-trigger 7, a pulse shaper 8, an integrator 9, a sound continuity invariantor 10, a sound continuity generator 11, a sound continuity quantization unit 12, a saw-tooth voltage generator 13, a comparator 14, a reference functions generator 15, a zero-organ 16, a pulse generator 17, a printer 18, an electronic digital computer 19, a decoder 20, a code-analog converter 21, a synthesizer 26.

The elements 3, 11, 12 make up a sound continuity standardizer 22. The elements 6, 7, 8 make up a short pulse shaper 23. The elements 15, 20, 21 make up a reference signal unit 24, the elements 18 and 19 make up a register 25. The elements connection is realized as shown in Fig.1.

The device functions in the following way. The operator's voice is transformed by a microphone 1 into electric signals which are intensified by an amplifier 2 and arrive at an electron key 3 and a sound continuity generator 11 of sound continuity standardizer 22. Since the informative part of the elementary sound resides in its initial stage, then these devices ensure the normal passage of the initial sound energy over the interval of approximately 0,1 sec, and then the channel is switched off till the new sound energy pulse appears. The quantized pulses arrive at a frequency invariantor 4. This device carries out the dynamic spectral analysis of the voice signal and converts the spectrum of the operator's voice signal in such a way that the voice of the synthesized sounds becomes independent of the tone's pitch of the operator's voice. The normal voice will always be heard at the frequency invariantor output, independently of the speaking operator. Artificial voice of frequency invariantor 4 arrives at an amplitude invariantor 5 which converts the voice signal in such a way that the signal at its output becomes no longer dependent on the amplitude but the main characteristics of sound information are completely retained. This is attained through functional transformations in the amplitude invariantor 5 which ensure: sampling of all weak signals, self-sustained or in combination with strong signals within the whole dynamic spectrum, their amplification to the normalized level with regard to amplitude, and comparison with each other followed by summation. As a result, different sound signals turn out to be equal in amplitude, and the output signal reminds of a clipped signal though it is of a higher quality.The amplitude invariantor's 5 output signal is clipped with the aid of an amplifier-limiter 6 and Schmitt-trigger 7, the dependence on amplitude of an output signal of invariantor unit 5 is completely eliminated, then the signal is differentiated and formed as a microsecond pulses package. Having been shaped in the pulse shaper 8, the pulses are integrated in the integrator 9, then the integrated pulses arrive at the continuity invariantor 10 intended for storing integration function and its compression in time for reproducing the integration function with higher frequency. The continuity invariantor 10 makes it possible to record integration function,0,1 sec in continuity at 100 descrete points, to an accuracy of 1% and to reproduce this function periodically repeating it at output, with frequency 200 kHz which ensures high frequency of comparison between the integration function within a comparator 14 and the reference functions generated by a reference functions generator 15 of a reference signal unit 24 and permits of recognitions within short time intervals. Single-argument function converter, preliminary trained to integrated functions of the elementary speech images can be used as the reference functions generator 15. Sampling of the reference functions available in the generator is realized in conformity with the output signal of the sawtooth voltage generator 13. The reference functions are fed in succession and at high speed from the reference functions generator 15 to the comparator 14. In case of disagreement between the reference function and voice image recognition function arriving from the sound continuity invariantor 10 at the comparator 14 output, there appears a signal , passing to zero-organ's 16 input, the zero-organ starts the pulse generator 17 which, in its turn, starts the decoder 20. The spectra of scale factors are correlated with respect to the decoder's 20 codes through the use of the code-analog converter 21. If the selected spectrum of scale factors ensures the similarity of the compared functions in the comparator 14 when the reference functions generator 15 is questioned, then the zero-organ 16 generates the switch off signal for the pulse generator 17 and the signal for fixing the code as an alphabetic record corresponding to the recognized sound in the printer unit 18. It is stored in computer memory and fed into the synthesizer 26. Then the device is cleared and gets ready to recognize the next sound image.

The reference functions shaping in the reference functions generator 15 is realized in the following way. The test voice images are distinctly pronounced before the microphone. Just in this mode the zero-organ 16 controls the system of scale factor adjustment, the latter consists of the pulse generator 17, the decoder 20, the code-analog converter 21. The determined spectra of scale factors are preliminary recorded before dictation of test voice images, and then introduced into the code-analog converter 21. If the scale factors vary smoothly during the reference function generator 15 tuning, so after the data is fed into the code-analog converter, the scale factors instantly assume those values at which the trained curve will be reproduced.The total number of scale factor spectrum variations equals the number of code combinations. For the described device the decoder is designed for ten bit binary code when the number of decoder's combinations amounts to 1024.

The described device is easily tuned, trained and implemented, its high accuracy of voice signals recognition ensures its utilization in the systems of man-machine interaction when robots of "ear-intelligence" type are designed and in other engineering domains.

REFERENCES

[1] K.P. Majstrenko, "The Method and Device for Invariant Voice Images Recognition, Transactions of All-Union School-Seminar "Psychological Bionics", Charkov, 1986, p.20.

[2] B.V. Bolotov, K.P. Majstrenko, G.G. Chub,"The Method of Voice Images Recognition", Certificate of Copyright of the USSR N 621003, BI N 31 of 25.08. 1978.

[3] B.N. Sorokin, Certificate of Copyright of the USSR N 432581, BI N 22 of 15.06. 1974.

[4] V.Ju. Trachtman, "The Device for Voice Signals Analysis", Certificate of Copyright of the USSR N 298943, BI N 11 of 31.03.1971.

[5] B.V. Bolotov, K.P. Majstrenko, "The Device for Frequency Voice Images Normalization", Certificate of Copyright of the USSR N 643959, BI N 3 of 23.08.1979.

[6] B.V. Bolotov, K.P. Majstrenko, G.G. Chub, "The Device for Voice Information Recognition", Certificate of Copyright of the USSR N 758238, BI N 31 of 23.08.1980.

[7] K.P. Majstrenko, A.A. Tyshko, "The Device for Sound Signals Processing", Certificate of Copyright of the USSR N 771709, BI N 38 of 15.10.1980.

[8] K.P. Majstrenko, A.A. Tyshko, "The Device for Acoustic Information Processing", Certificate of Copyright of the USSR N 822248, BI N 14 of 15.01.1981.