Programming and Computer Software, Vol. 29, No. 4, 2003, pp. 210–218. Translated from Programmirovanie, Vol. 29, No. 4, 2003.
Original Russian Text Copyright © 2003 by Zhirkov, Kortchagine, Lukin, Krylov, Bayakovskii.

Graphic Representation Method and Neural Network Recognition of Time-Frequency Vectors of Speech Information

A. O. Zhirkov, D. N. Kortchagine, A. S. Lukin, A. S. Krylov, and Yu. M. Bayakovskii

Department of Computational Mathematics and Cybernetics, Moscow State University, Vorob'evy gory, Moscow, 119992 Russia

e-mail: kryl@cs.msu.su

Received February 19, 2003

Abstract—Currently, various time-frequency representations are often used for sound analysis. These representations, on the one hand, are convenient for visible sensation of sound by a human and, on the other hand, can be used for automatically analyzing sound pictures. In this paper, various methods for representation of sound as two-dimensional time-frequency vectors of a fixed dimension and their use for speech and speaker recognition problems are discussed. Probabilistic, distance-based, and neural-network methods for the recognition of these vectors by examples of separate words are considered. Numerical experiments showed that the best among them is the method based on a three-layer neural network, the short-time Fourier transform, and the two-dimensional wavelet transformation. For the speaker recognition problem, a distance-based recognition method employing the adaptive Hermite transform turned out the best among all.

1. INTRODUCTION

In the majority of speech recognition and analysis systems, sound is considered to be a stream of frequency feature vectors. A feature vector usually consists of a cepstrum and a vector of its derivatives [1]. The cepstrum is constructed from frequency normalization, taking the logarithm, and a *discrete cosine transform* (DCT) of the amplitude component of the *discrete Fourier transform* (DFT). The discrete Fourier transform is successively applied to the input audio signal in the Hamming window with the stride of 20–30 ms. The recognition system reduces to the training of a syntactical model, which is based on hidden Markov models (HMMs) [2]. The hidden Markov models are used jointly with *artificial neural networks* (ANNs) [3, 4], which improves the probabilistic characteristics of the former.

The authors of this paper decided not to follow the accepted scheme. One alteration of the scheme is turning the column vector to the matrix vector through the addition of the time component. The other alteration consists in replacing the so-called short-time Fourier transform (STFT) by a transform with an adaptive window length and using the Hermite functions instead of the trigonometric Fourier series.

Note that the idea to include the time component into the feature vector is not new. To improve the reliability of the recognition, a part of the load on the analysis of the time component is transferred from the HMM to the feature vectors; however, in real systems, only one additional parameter—the derivative vector—is used. The main problem associated with the extension of the time component is that the steadiness of the dimension is required and that the feature vector is to be maximally decorrelated. Several methods for solving this problem has recently been suggested. One of them consists in obtaining the cepstrum from several (rather than one) frequency vectors located close to each other by means of the two-dimensional discrete cosine transform [5], which made it possible to considerably improve the recognition robustness. In the work [6], a number of possible two-dimensional representations for the time-frequency vectors are considered, including those based on the two-dimensional wavelet localization. In our work, the two-dimensional wavelet transformations are used for separating the most stable, low-frequency, components of the time-frequency information. This work also studies various methods of classification of the feature vectors obtained and shows that the most stable among them is the neural network classification.

To illustrate the adequacy of the representation of speech information by time-frequency vectors, we present examples of the 3D visualization of a smoothened STFT surface. Note that the idea of graphic representation and the use of images in the sound-related studies are currently widely used [7].

Figures 1a and 1b show the spectrograms corresponding to the word "test" said by one speaker with different rates. As can be seen, the distributions of the time-frequency energy depicted in these figures are the same up to the scaling factor. Figures 1c and 1d show the spectrograms corresponding to the word "nine" said by a man and a woman, respectively. In this case, the energy distributions for the two signals are not identical; however, one can see that both pictures still have some structures in common. Thus, we arrive at the

0361-7688/03/2904-0210$25.00 © 2003 MAIK "Nauka / Interperiodica"

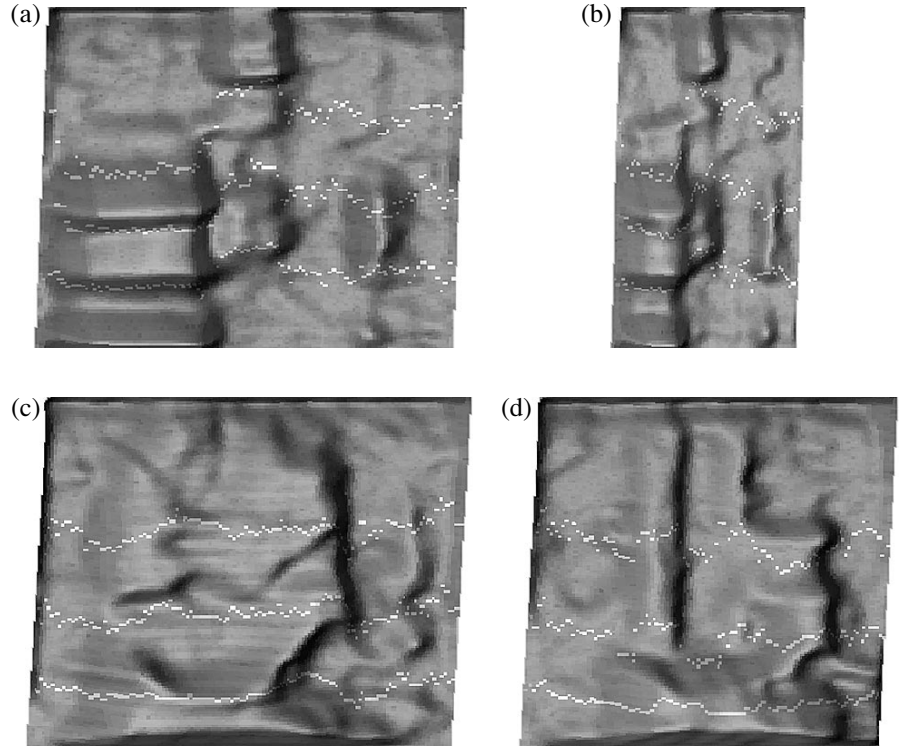

Fig. 1. Smoothened logarithmic spectrograms composed of the STFT amplitudes (the horizontal direction corresponds to time and vertical direction corresponds to frequencies from 150 Hz to 5 kHz; the white lines are centroids).

question of whether it is possible to classify/recognize words based on only time-frequency energy distribution, without considering details of the sound information? If this is possible, then how to better organize time-frequency vectors and what recognition method to use? The major part of this paper addresses just these questions. The discussion of the method follows the order of the audio signal transformations, starting from the input audio signal through the result of the recognition:

- short-time spectral analysis,
- algorithm for determining boundaries of the feature vectors,
- method for obtaining the feature vector, and
- neural network pattern recognition.

2. TIME-FREQUENCY TRANSFORMATIONS

Any system for automated analysis of audio information receives a sampled audio signal with quantized amplitude. Figure 2a shows an oscillogram of a signal corresponding to a speech fragment consisting of two words. Applying the short-time spectral analysis based on the fast Fourier transform (FFT) to this signal, we obtain a sequence of complex DFT vectors. The phase component of the signal contains, basically, information about space orientation of the signal source. Unlike the speaker recognition problem, the phase information in the speaker-independent speech recognition problem is of no importance and, therefore, is not considered. The amplitude component of this vector as a function of time, i.e., the spectrogram, can be viewed as an ordinary image. It is well known that the most informative frequencies of human's voice lie in the range from 100 Hz to 5 kHz. Therefore, only harmonics with the frequencies belonging to this interval are left in the spectrogram. Next, the logarithms of the harmonic amplitudes are taken, and the so-called mel-scale filtering is applied,

$$melscale(f) = 2595\log_2(1 + f/700),$$

where f is the frequency in Hertz.

The spectrogram obtained in this way (Fig. 2b) contains noise components, which are due to the noise in the input signal and certain properties of the DFT. The distribution of the noise is assumed to obey the normal (Gaussian) law with a non-zero expectation. By using the fact that both frequency and time dimensions of the spectrograms obtained are excessive, it is possible to get rid of this noise by reducing the image and applying the low-pass filtering. The noise in the input signal can be divided into two components. One of them is a stationary noise, which may be assumed to be additive. In this case, the noise suppression reduces to subtracting the noise spectrum from the columns of the spectrogram. The spectrogram obtained in this way is shown in Fig. 2c.

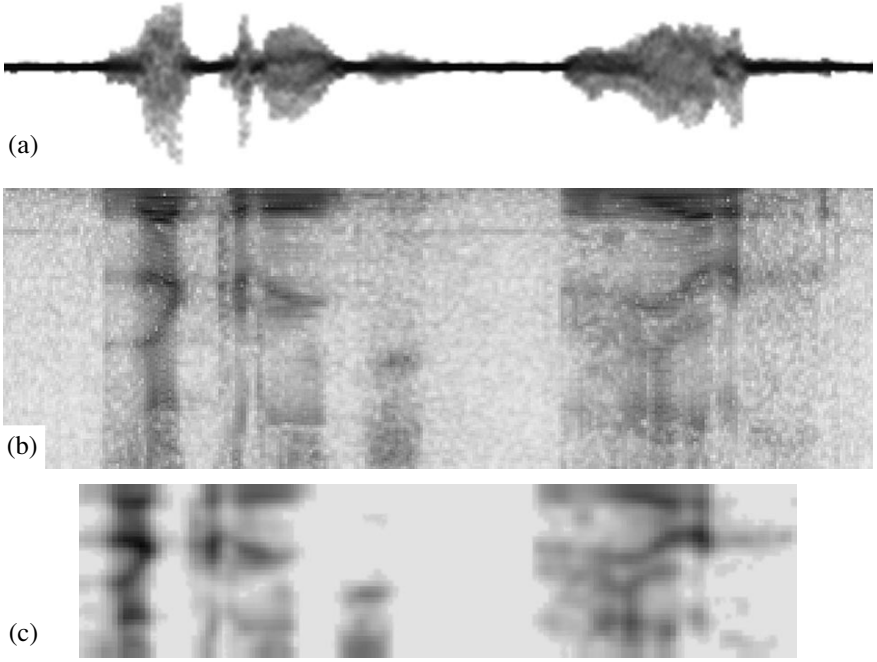

Fig. 2. Successive signal transformations: (a) oscillogram, (b) spectrogram (lower frequencies lie at the top), (c) spectrogram after the dimension and noise reduction.

3. FINDING OF THE FEATURE VECTORS

The spectrogram of each word (see, e.g., Fig. 3a) is scaled to a square form. Then, the two-dimensional wavelet transform is applied to it as many times as required for obtaining a low-frequency $T \times F$ matrix, where T and F are temporal and frequency resolutions, respectively. It has been shown experimentally that the optimal resolution (from the standpoint of the subsequent neural recognition) is 8×8. To improve the robustness of the feature vector recognition, the vectors are normalized to zero expectation and unit variance (Fig. 3b).

4. PATTERN RECOGNITION

Consider the simplest case where a speech fragment is a sequence of a limited number of words. In this case, the speech recognition reduces to classifying two-dimensional feature vectors. Each class corresponds to a certain word said by different speakers. Examples of two-dimensional feature vectors for different words are shown in Fig. 4, where each column contains words said by one speaker, and the rows correspond to the word classes.

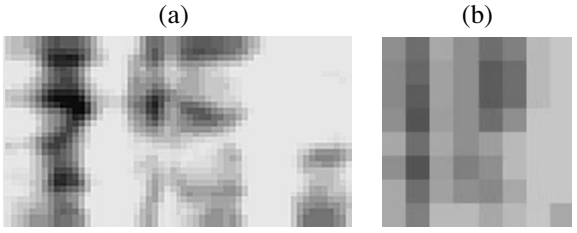

Fig. 3. Final scaling and normalization: (a) word spectrogram, (b) word feature vector.

There exist many classification methods with learning. Let us discuss basic—statistical, distance-based, and neural-network—approaches to classifying the vectors.

4.1. Statistical Classification Method

Let us introduce the following notation: $W^j_{i,x,y}$ is a feature vector from the learning sample, where j is the class number, i is the example number, and (x, y) are the coordinates of the feature vector; $X_{x,y}$ is a vector from the learning sample; and $P(C_j|X_{i,j}, \theta^j_{x,y})$ is the probability that X belongs to the class j with the coordinates (x, y) in the feature vector for the normal probability distribution.

Then, the classification function *class*(X) is defined as follows:

$$M^j_{x,y} = \frac{1}{n_j}\sum_{i=1}^{n_j} W^j_{i,x,y},$$

$$\sigma^j_{x,y} = \frac{1}{n-1}\sum_{i=1}^{n_j} (M^j_{x,y} - W^j_{i,x,y})^2,$$

$$\theta_{norm} = \{M, \sigma\},$$

$$class(X) = \operatorname{argmax}_j \left[\prod_{x,y} P(C_j | X_{x,y}, \theta^j_{x,y}) \right],$$

i.e., the maximum a priori probability that the input vector belongs to a certain class is found under the assumptions that the distribution is normal and the vector components are independent.

4.2. The Least-Distance Classification Method

The distance-based classification method, in the case of the component-wise comparison of the vectors, can be written as

$$class(X) = \operatorname{argmin}_i \left[\min_j \sum_{x,y} \rho(X_{x,y} W^i_{j,x,y}) \right],$$

where ρ is the distance. In this case, the nearest vector from the learning sample set ("code book") is found, and the vector X is assumed to belong to the same class as the nearest code book vector.

We compared several distance functions, such as the sum of the absolute values of differences between the components, the mean square deviation of the vectors, and the maximum of the absolute values of differences between the components. The best recognition results were obtained when the commonly used mean squares metrics was employed.

4.3. Neural Network Classification

Neural network classification is implemented by means of a neural network. A feature vector comes to the first layer of the network, and the results of fuzzy classification are obtained on the output layer. The network architecture between the input and output layers may be different and is selected with regard to the input data space. For the majority of problems, however, one hidden intermediate layer with the varying number of neurons is quite sufficient. Figure 5 shows a network state occurring upon receiving a time-frequency vector on its input (the left neuron column). The different brightness of neural connections means different synapse weights. The values that are less than a certain threshold are not shown. On the output, we obtain a vector with the number of neurons equal to the number of classes considered, i.e., to the number of words. As can be seen from the figure, the values of all neurons, but one, are close to zero (are white), which suggests that the input vector belongs to the class with the number equal to the number of the darkest neuron.

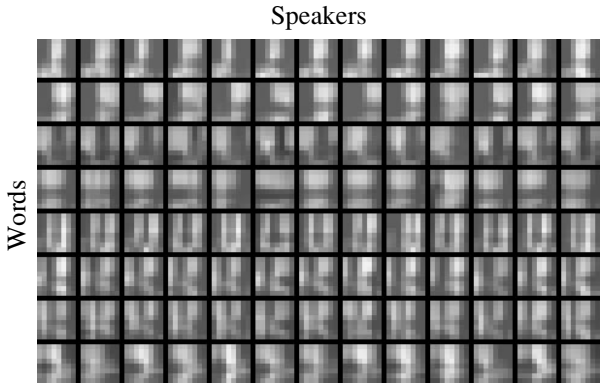

Fig. 4. Graphic representation of time-frequency vectors for different speakers and different words.

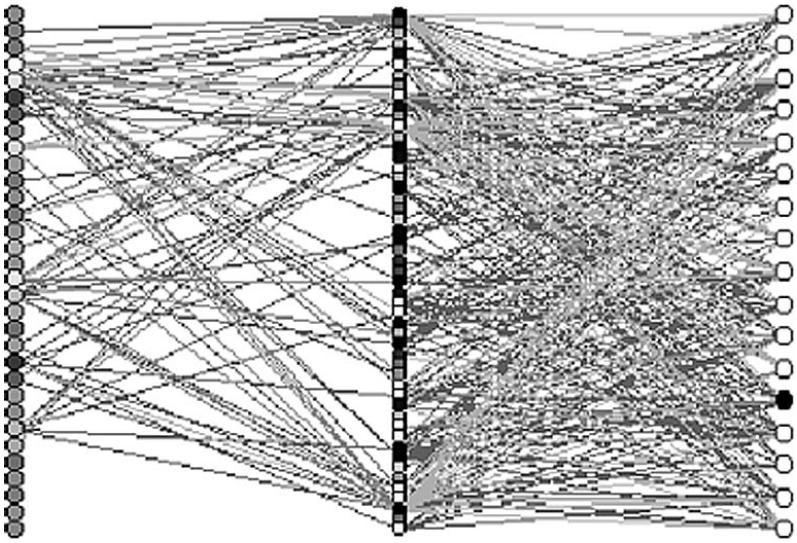

Fig. 5. Neural network used for the speaker-independent recognition.

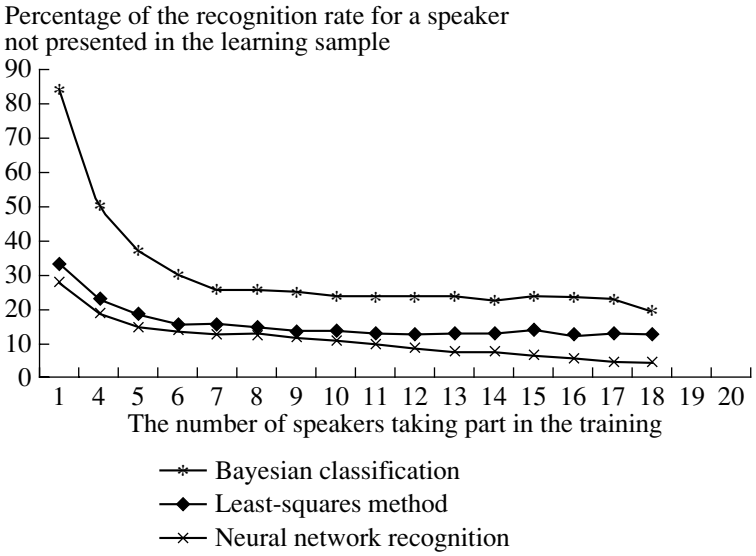

Fig. 6. Dependence of the error recognition rate on the number of speakers used in the learning sample for different classification methods.

Mathematically, this classification can be written as

$$class(X) = \operatorname{argmin}_i \left[\min_j \prod_{x,y} \rho(EVec(ANN(X)), ANN(X)_{j,x,y}) \right],$$

$$EVec(x) = \left((y_1, y_2, \ldots, y_n) | y_i = \begin{cases} 1, & (i = \operatorname{argmin}_j(x_j)) \\ 0 \end{cases} \right),$$

where ρ is the Euclidean distance and $ANN(x)$ is the vector of the last layer of the neural network.

Thus, the vector containing results of the fuzzy classification is compared with the patterns corresponding to different "ideal" classifications of the form {0, 0, …, 1, …, 0}, where "1" corresponds to the class number.

Table 1. Error recognition rate for various classification methods

Method		The number of examples in the learning sample		
		2	4	16
Metric approach	Sum of squares	33.1%	18.1%	12.3%
	Maximum magnitude	34.8%	19.5%	13.8%
	Sum of magnitudes	33.6%	19.2%	13.7%
Statistical classification	Gauss distribution	80%	35.2%	32.0%
	Cauchy distribution	49%	33.7%	32.2%
Three-layer neural network		28%	13.3%	5.9%

4.4. Comparison of the Classification Methods

We used a database that contained 20 different words said by 20 different speakers. Note that the set of speakers contains both males and females of different age. The word database was divided into two classes: a *learning set* and a *test set*. The objective of the classification method is to make as few errors as possible on the test set after learning from the learning set. Figure 6 shows the dependence of the percentage of the recognition errors made on the test set on the number of speakers in the learning set for different classification methods. Clearly, the more the number of the learning samples, the better the recognition accuracy should be, and the experimental results substantiate this point for all classification methods (not counting two outlier points). The slopes and asymptotes of the curves corresponding to different methods are, of course, different.

The Bayesian classification has the greatest decrease rate and the highest asymptote (see also Table 1). This phenomenon can be explained by the fact that this classification model is based on two following a priori assumptions: (1) the normal distribution of vector components belonging to one class and (2) statistical independence of the vector components.

The distance-based approach, where the assumptions on the space of classes are less restrictive, demonstrates much better recognition performance. The basic assumption in this method is that the contributions of all components of the time-frequency vector in the total sum are linear.

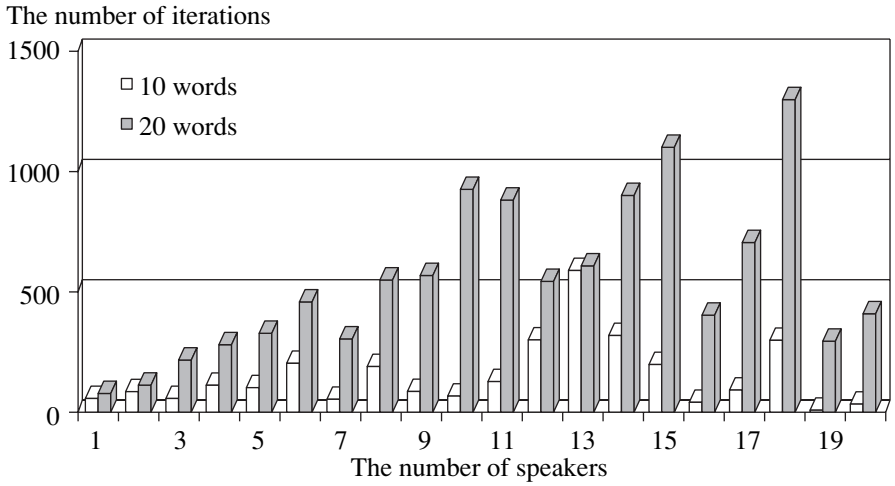

Fig. 7. Dependence of the number of iterations required for the network training on the number of speakers and the number of words.

The neural network classification is the only method where no restrictions on the set of feature vectors are imposed. It is known that the three-layer neural network with a continuous sigmoid activation function can approximate an arbitrary continuous function with any desired degree of accuracy. This implies that such a network is capable of classifying any continuous finite sets. A disadvantage of neural networks is that, in the general case, the learning time exponentially depends on the number of the learning examples. However, such a dependence takes place only when the vectors from the base set are weakly correlated. The dependence of the number of iterations on the number of examples for a practical test set is shown in Fig. 7.

Specific features of the classification methods considered are presented in Table 2.

The basic performance criterion for the speech recognition systems is the recognition error rate and the reliability of the recognition. Recognition error rates of the neural network classification for two values of noise intensity and different noise types are presented in Table 3. In this example, the test database consisted of 20 words said by 20 different speakers. The learning set contained words said by 17 speakers; the other words were used for testing.

5. ADAPTIVE HERMITE TRANSFORM

Along with the short-time Fourier transform, we consider an alternative representation of audio signals, which takes into account the fact that the speech structure is quasi-periodic. This representation is based on the adaptive Hermite transform [10, 11], which is efficiently used for processing and analyzing images [10–12].

The Hermite functions

$$\psi_n(x) = \frac{(-1)^n e^{x^2/2}}{\sqrt{2^n n! \sqrt{\pi}}} \frac{d^n(e^{-x^2})}{dx^n}$$

form a complete orthonormal set of functions [8]. These functions can also be defined by the recurrent relations

$$\psi_0 = \frac{1}{\sqrt[4]{\pi}} e^{-x^2/2},$$

Table 2. Properties of the classification methods

	Distance-based methods	Probabilistic approach	Neural classifier
Dependence of the learning rate on the number of the examples	Does not depend	Linear rate	Exponential rate (in the general case)
Dependence of the recognition rate on the number of the learning examples	Depends	Does not depend	Does not depend
Parametrization of the space of classes (n is the dimension of the feature vector and m is adaptively selected dimension)	Set of n-dimensional vectors	n-Dimensional probabilistic ellipsoid	m-Dimensional polyhedron

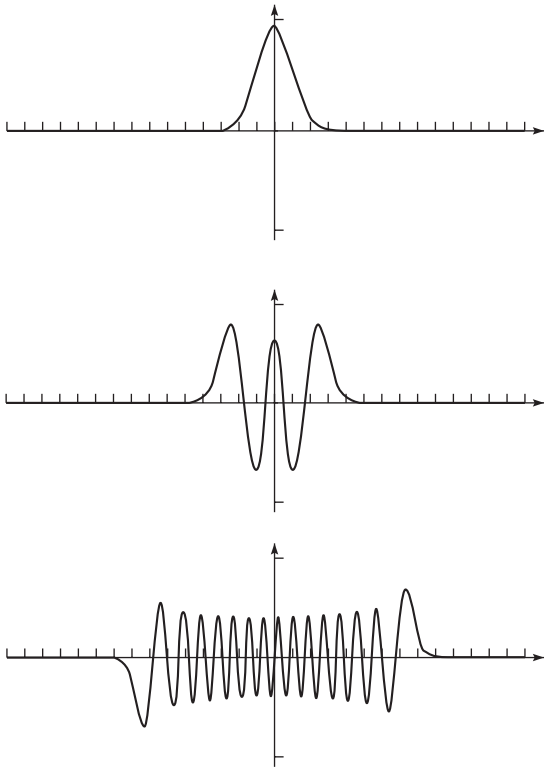

Fig. 8.

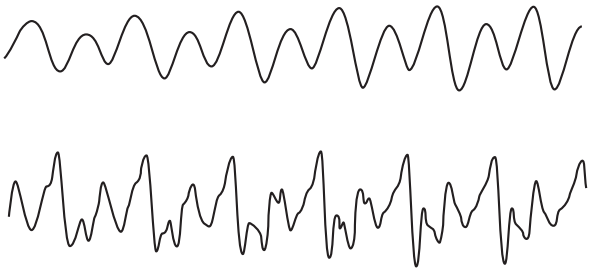

Fig. 9.

$$\psi_n = \frac{\sqrt{2x}}{\sqrt[4]{\pi}} e^{-x^2/2},$$

Table 3. Dependence of the error recognition rate on noise type and intensity

Noise source	Ratio of the signal to noise	
	22 dB	28 dB
Music	15%	2.5%
Speech	33%	2%
Stationary nois	10%	3%

$$\psi_n = x\sqrt{\frac{2}{n}}\psi_{n-1} - \sqrt{\frac{n-1}{n}}\psi_{n-2}, \quad \forall n \geq 2.$$

In addition, the Hermite functions are eigenfunctions of the Fourier transform [13],

$$F(\psi_n) = i^n \psi_n,$$

where F denotes the Fourier transform operator.

Plots of three Hermite functions are depicted in Fig. 8.

To apply the Hermite transform, it is required to determine the integration interval [12]. As can be seen from the form of the speech signal shown in Fig. 9, many speech fragments have a quasi-periodic structure (it should be emphasized that the lengths of neighboring quasi-periods, as well as the forms of the signals on the neighboring quasi-periods, may slightly differ).

For the input interval, we successively take these quasi-periods [14]. The endpoints of these intervals are selected such that the extremum of the signal on the interval is attained, approximately, in the middle of the interval, and the values at the endpoints are close to zero. Further, the obtained approximation interval $[-A_0, A_0]$ is extended to give the interval $[-A_1, A_1]$ determined from the equation

$$\int_{-A_1}^{A_1} \psi_n^2(x)dx = 0.99,$$

where n is the number of the Hermite functions used for the approximation.

Next, the input signal is expanded into the Fourier series in terms of the Hermite functions,

$$value(x) = \sum_{i=0}^{n-1} c_i \psi_i(x),$$

$$c_i = \int_{-A_1}^{A_1} f(x)\psi_i(x)dx.$$

Since the Hermite functions are the eigenfunctions of the Fourier transform, we simultaneously obtain the Fourier transform of the given signal.

In addition to the linear coding, it is possible to use a hierarchical coding, which, on the one hand, makes results more stable and, on the other hand, points to the analogy between the Hermite coefficients and formants. The essence of this coding is as follows. First, a quasi-period is approximated by one function; then, the difference is found and extended to the interval required for the approximation by two functions; the function obtained is approximated by two functions; and so on. It should be emphasized that, although such a representation is redundant, it makes it possible to perform a complete analysis both in the frequency and

time domains, which improves the possibilities of a finer analysis of individual features of the speakers.

6. COMPARISON OF DIFFERENT REPRESENTATIONS OF TIME-FREQUENCY FEATURE VECTORS IN THE SPEECH AND SPEAKER RECOGNITION PROBLEMS

We carried out several tests with the use of different recognition and classification methods. The aim of the experiments was to compare the short-time Fourier transform and the adaptive Hermite transform. One test database with 20 words and 20 speakers was used. Two following classification problems were considered: the speaker-independent word recognition and context-independent speaker recognition.

The recognition methods are based on the classification method for two-dimensional time-frequency feature vectors considered in the previous sections. The experiments showed that the best speaker recognition rate (26%) is obtained when the Hermite transform based on the adaptive expansion into quasi-periods is used. Note that, in this case, the linear coding with 32 coefficients and the subsequent sign summation was used. Note that the accuracy of the indexing based on the hierarchical Hermite coding with additional filters and manual tuning of the threshold value was as high as 95%. The text-independent speaker recognition is a considerably more difficult task than the recognition based on the use of known words or phonemes. Similarly, the speaker-independent recognition method is considerably more complicated compared to the speaker-dependent recognition method. In the first case, the recognition rate may be as high as 99%, or even greater if the acoustic environment is appropriate. In the case of the speaker-independent word recognition problem, the best result (96%) has been obtained when fixed-length windows with the Fourier basis and logarithmic summation were used.

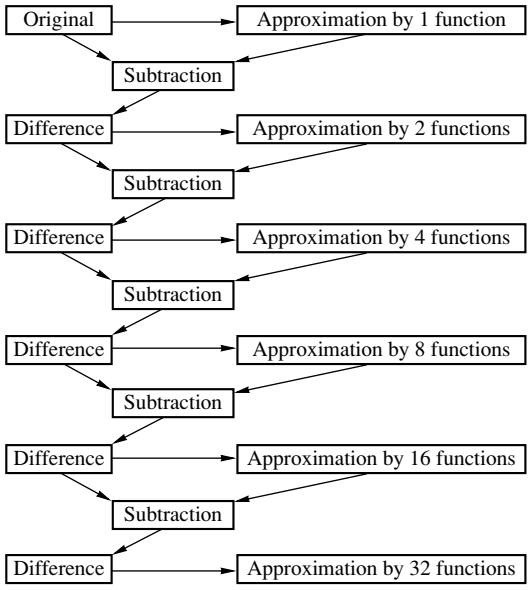

Fig. 10.

Results of the experiments are presented in Table 4 and illustrated in Fig. 11.

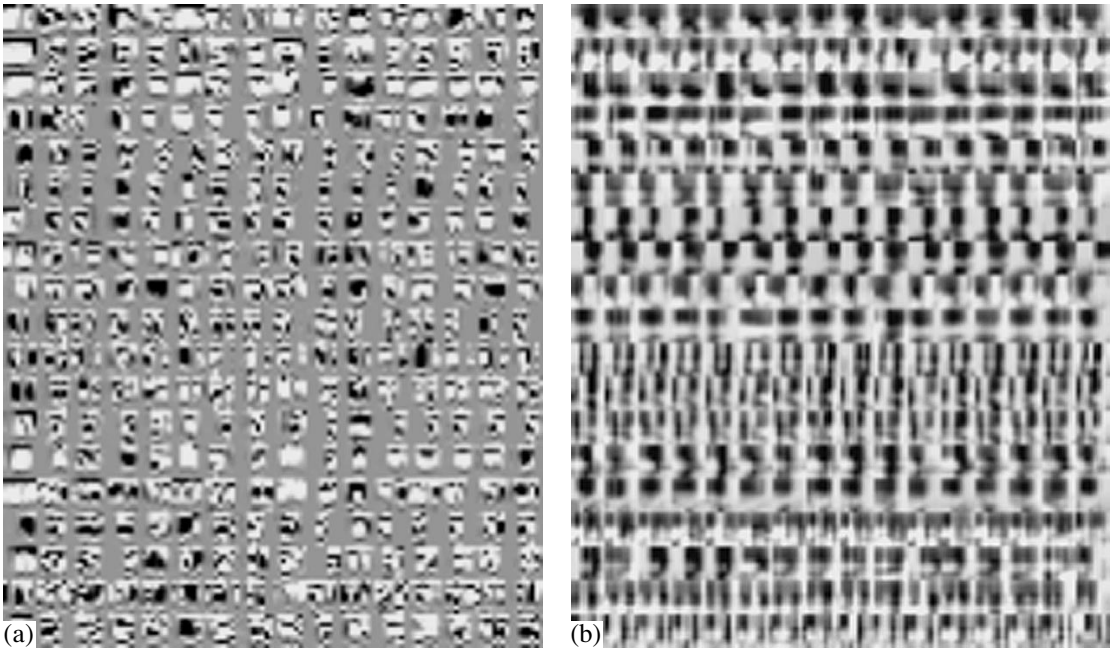

Fig. 11. The best representations of the feature vectors. (a) The library of the feature vectors based on the Hermite transform and quasi-periods, which is used of the speaker recognition. (b) The library of the feature vectors based on the short-time Fourier transform with fixed-length windows, which is used of the word recognition.

Table 4. Correct recognition rate for different representations of the feature vectors

	Feature vector representation				
	hermite basis		fourier basis		
	quasi-periods		quasi-periods		short-time Fourier analysis with absolute logarithmic summation
	sign summation	summation of magnitudes	sign summation	summation of magnitudes	
Word recognition	38%	14%	16%	29%	96%
Speaker recognition	26%	25%	22%	24%	12%

It should be noted that the method based on adaptive quasi-periods contains more information about the speaker than the transforms with fixed-length windows. The adaptive selection of the quasi-period length and location takes into account the localization of the Hermite functions both in the frequency and spatial domains, which makes it possible to take into account more information about individual features of the particular speaker.

7. CONCLUSIONS

The speech and speaker recognition methods considered in this paper compare well with the existing methods and, in some respects, even outperform them. The main advantage of the word recognition method based on the two-dimensional feature vectors and neural network classification is that the most important noise-resistant time-frequency information is concentrated in a compact time-frequency vector. Unlike this approach, the method based on the adaptive Hermite transform works with a finer time-frequency localization. This representation of audio information is less stable against noise; however, it contains more specific voice features compared to the approach based on the Fourier transform. Owing to this feature, this technology works well when solving the speaker recognition problems.

The authors of this paper believe that the methods employing the time-frequency vectors as a base element and the artificial neural networks, which form probabilistic characteristics on the basis of these vectors, are highly promising for solving speech recognition problems. In some identification systems, the speech recognition and speaker recognition can be combined, for example, with the use of neural or Bayesian networks. These combined systems, in turn, can interact with other sensors and audio–visual information sources, such as videocameras and range finders. The data exchange between various information sources can considerably improve the recognition rate, as well as reliability of the recognition of audio–visual information.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research, project no. 01-01-00981, and by the Intel Technologies.

REFERENCES

1. Bourland, H. and Morgan, N., Hybrid HMM/ANN Systems for Speech Recognition: Overview and New Research Directions, http://www.tzi.org/ik98/prog/kursunterlagen/-t2/bourland.html.
2. Tran, D., Wagner, M., and Zheng, T., A Fuzzy Approach to Statistical Models in Speech and Speaker Recognition, *Proc. of 1999 IEEE Int. Fuzzy Systems Conf.*, Korea, pp. 1275–1280.
3. Lippman, R. and Gold, B., Neural Classifiers Useful for Speech Recognition, *Proc. IEEE First Int. Conf. on Neural Networks*, 1987, vol. 4, pp. 417–422.
4. Gold, B. and Morgan, N., *Speech and Audio Signal Processing*, Wiley, 1999.
5. Demars, C., Two-Dimensional Representations of Speech Signal. Time-frequency Representation and Parametrizations, 1999, http://www.limsi.fr/Individu/-chrd/tablematniE2001.html.html.
6. Chan, C.P., Lee, T., and Ching, P.C., Two-Dimensional Multi-Resolution Analysis of Speech Signals and Its Application to Speech Recognition, *Speech and Signal Processing* (Proc. of 1999 IEEE Int. Conf. on Acoustics), 1999, vol. 1, pp. 405–408.
7. Dvoryankin, S., Relationship between Digits and Graphics, Sound and Image, *Otkrytye sistemy*, 2000, no. 3, pp. 25–32.
8. Szego, G., Orthogonal Polynomials, *Am. Math. Soc. Colloquium Publications*, 1959, vol. 23.
9. Jeckson, D., Fourier Series and Orthogonal Polynomials, in *Carus Mathematical Monographs*, 1941, no. 6.
10. Martens, J.-B., The Hermite Transform—Theory, *IEEE Trans. Acoustics, Speech, Signal Processing*, 1990, vol. 38, pp. 1595–1606.
11. Martens, J.-B., The Hermite Transform—Applications, *IEEE Trans. Acoustics, Speech, Signal Processing*, 1990, vol. 38, pp. 1607–1618.
12. Krylov, A. and Kortchagine, D.N., Projection Filtering in Image Processing, *Proc. of the Conf. Graphicon' 2000*, Moscow, 2000, pp. 42–45.
13. Krylov, A. and Liakishev, A.V., Numerical Projection Method for Inverse Fourier Transform and Its Application, *Numerical Functional Analysis Optimization*, 2000, vol. 21, pp. 205–216.
14. Krylov, A.S., Kortchagine, D.N., and Lukin, A.S., Streaming Waveform Data Processing by Hermite Expansion for Text-Independent Speaker Indexing from Continuous Speech, *Proc. of the Conf. Graphicon' 2002*, Nizhni Novgorod, 2002, pp. 91–98.