ISSN 1063-7710, Acoustical Physics, 2014, Vol. 60, No. 4, pp. 466–473. © Pleiades Publishing, Ltd., 2014.

ACOUSTIC ECOLOGY. NOISE AND VIBRATION

Generalized Fractional S-Transform and Its Application to Discriminate Environmental Background Acoustic Noise Signals[1]

D. Jhanwar[a], K. K. Sharma[b], and S. G. Modani[b]

[a]*Govt Engineering College Ajmer, Badliya Chauraha, Ajmer, 305001 India*

[b]*Malaviya National Institute of Technology, J.L.N. Marg, Jaipur, India*

e-mail: deepakjhanwar2001@gmail.com; kksharma_mrec@yahoo.com; shrigmodani@gmail.com

Received June 6, 2013

Abstract—We propose a modification of S-transform (ST) by changing the kernel of Fourier transform (FT) with that of fractional Fourier transform (FRFT) and call it generalized fractional ST (GFST). The FRFT is a generalization of FT and it has been shown more useful than the FT for signals with changing frequencies such as chirp signals. The proposed GFST is applied to analyze and classify different environmental background sound mixed with speech signal in the form of additive noise. The simulation results demonstrate that Euclidean distance between the feature vectors computed from generalized fractional ST corresponding to different background noise is increased as compared to ST for the same set of feature vectors and signals.

Keywords: S-transform, fractional Fourier transform, additive noise, Euclidean distance, environmental background acoustic noise

DOI: 10.1134/S1063771014040058

1. INTRODUCTION

Environmental sound recognition is a basic audio signal processing task having important applications in navigation, assistive robotics and other mobile device-based services. The audio based scene denotes a location with different acoustic characteristics like train, airport, traffic area or quiet hallway. There have been recent trends in finding solutions to provide hearing information for mobile robots to enhance their context awareness with audio information. In the context aware applications, e.g., a mobile device like cellphone can automatically change the notification mode based on the knowledge of user's surroundings. It can switch to the silent mode in a meditation room, theater or lecture room or to provide information regarding location of user. Discrimination of the background noise signals mixed with speech signal is required for providing information regarding the physical location of person and several methodologies for the same are discussed in [1–7]. These methods utilize the short-time Fourier transform (STFT), wavelet transform (WT) and other time–frequency representations of signals such as Wigner distribution (WD), the ambiguity function and the spectrogram, etc. However, the WD of a non-stationary multi-component signal has many undesirable cross-terms. The STFT is also known to have poor frequency resolution in low frequency range and poor time resolution in high frequency range. Both of these shortcomings are due to fixed width of the STFT window. The WT does not provide phase information of local spectrum as their entire waveform translates in time without change in shape.

The S-transform (ST), which is also a time–frequency representation of the signal, localizes the real and imaginary spectrum [8, 9]. It combines the features of WT and STFT in the time–frequency representations of the signal. Several important modifications have already been made in it by changing the properties and width of Gaussian window used in it [10–12]. To provide the user a specified time and frequency resolution in the time–frequency plane, the width of Gaussian window (which is proportional to inverse frequency (f) in original ST) is replaced by (f/γ) where γ is a real number [10]. It means that one standard deviation of Gaussian window contains γ wavelengths of the complex sinusoid at any frequency. To obtain better time resolution, the method of bi-Gaussian window is introduced in [11]. The problem of resolving waveforms with changing frequency is addressed by introducing a complex Gaussian window with a custom built complex phase function [12]. A method is discussed to extract information from seismic noise [13]. A method of estimating the isotropic sea noise level with a horizontal array in the presence of uncorrelated interference and interference with a

[1] The article is published in the original.

complex spatial structure is proposed and experimentally tested [14].

In this paper, we propose a modification of ST by changing the kernel of Fourier transform (FT) with that of fractional Fourier transform (FRFT) and call it generalized fractional ST (GFST). The FRFT is a generalization of FT and it has been shown more useful than the FT for signals with changing frequencies such as chirp signals [15–19]. The proposed GFST is applied to analyze and classify different environmental background sound mixed with speech signal in the form of additive noise. The simulation results demonstrate that Euclidean distance between the feature vectors computed from GFST corresponding to different background noise is increased as compared to ST for the same set of feature vectors and signals.

2. GFST

The ST which gives us the time frequency representation of a signal combines the features of STFT and WT [8, 9]. The ST of signal $s(t)$ is expressed as [9]

$$ST_s(\tau, f, \sigma) = \int_{-\infty}^{\infty} s(t)g(t-\tau)\exp(-j2\pi ft)dt, \quad (1)$$

where $g(t)$ is a Gaussian window function satisfying the condition

$$\int_{-\infty}^{\infty} g(t)dt = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{t^2}{2\sigma^2}\right)dt = 1 \quad (2)$$

and σ is the width of the Gaussian window. The width of the window is made function of frequency $\sigma(f)$ and if it is taken as inverse of the frequency,

$$\sigma_f = \sigma(f) = \frac{1}{|f|},$$

then (1) simplifies to

$$ST_s(\tau, f, \sigma_f) = \frac{|f|}{\sqrt{2\pi}}\int_{-\infty}^{\infty} s(t)\exp\left(-\frac{(t-\tau)^2 f^2}{2}\right) \times \exp(-j2\pi ft)dt. \quad (3)$$

We propose a modification in (3) by replacing the kernel of FT by the kernel of FRFT. Thus, the proposed GFST is defined as

$$ST_s^{\alpha}(\tau, u, \sigma_u) = \int_{-\infty}^{\infty} s(t)g(t-\tau)K_{\alpha}(t, u)dt, \quad (4)$$

u is the FRFT domain parameter at an angle α from the time axis,

$$\alpha = \frac{a\pi}{2}, \quad (5)$$

where a is a real number, $0 \le |a| \le 2$. The width of the Gaussian window σ_u is made function of u and it is taken as

$$\sigma_u = \sigma(u) = \frac{1}{|u|}.$$

The GFST which is formed by substituting the kernel of FRFT is merely a rotated version of ST:

$$ST_s^{\alpha}(\tau, f) = ST_s(\tau\cos\alpha - f\sin\alpha, \tau\sin\alpha + f\cos\alpha).$$

The kernel of the FRFT is expressed by [15]

$$K_{\alpha}(t, u) = \sqrt{\frac{1-j\cot\alpha}{2\pi}} \times \exp\left(j\frac{(u^2+t^2)\cot\alpha}{2}\right)\exp(jut\,\mathrm{cosec}\,\alpha), \quad (6)$$

if α is not a multiple of π.

2.1. Interpretation of GFST Equation

The rotation type relationship between both pairs of axes is

$$\begin{pmatrix} t \\ f \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}.$$

The GFST is the Gaussian windowed signal which is decomposed in terms of "chirp" basis functions. These basis functions only differ by a time shift and by a phase factor that depends on u for different values of u [17]:

$$K_{\alpha}(t, u) = \exp\left(-j\frac{u^2}{2}\tan\alpha\right)K_{\alpha}(t - u\sec\alpha, 0).$$

Another convenient way of looking at the integral in (4) is to define

$$m(t, u) = s(t)K_{\alpha}(t, u). \quad (7)$$

Substituting (7) in (4), we get

$$ST_s^{\alpha}(t, u, \sigma_u) = \int_{-\infty}^{\infty} m(t, u)g(t-\tau)dt = m(t, u) \otimes g(t, \sigma_u),$$

where $\otimes$ denote the symbol for convolution. If FRFT of $s(t)$,

$$S_{\alpha}(u_i) = \int_{-\infty}^{\infty} s(t)K_{\alpha}(t, u_i)dt,$$

and FRFT of Gaussian window $g(t, \sigma_u)$,

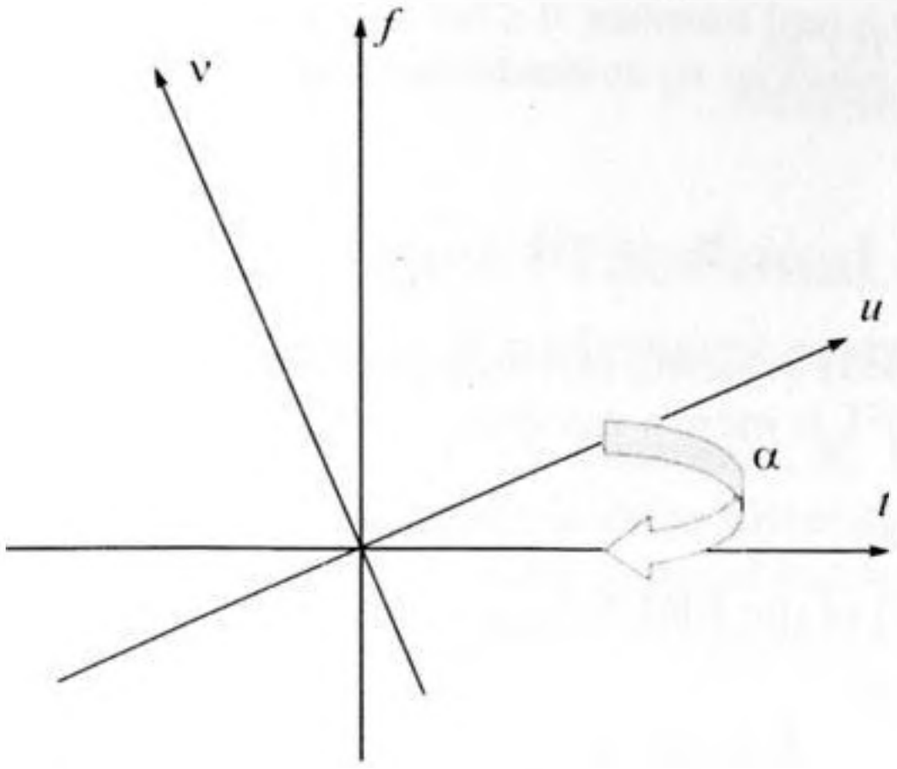

Fig. 1. Representation of u and v axes at an angle α.

$$G(u_i, t, \sigma_u) = \int_{-\infty}^{\infty} g(t, \sigma_u) K_\alpha(t, u_i) dt$$

$$= \frac{\exp\left(j\frac{\alpha}{2}\right)}{\sqrt{\cos\alpha + j\frac{1}{2\pi\sigma_u^2}\sin\alpha}}$$

$$\times \exp\left(-\frac{u^2}{2\sigma_u^2}\left(\frac{1+\tan^2\alpha - j\left(\frac{1}{2\pi\sigma_u^2} - 2\pi\sigma_u^2\right)\tan\alpha}{1+\frac{1}{4\pi^2\sigma_u^4}\tan^2\alpha}\right)\right),$$

then $ST_s^\alpha(\tau, u, \sigma_u)$ may be expressed as

$$ST_s^\alpha(\tau, u, \sigma_u) = \int_{-\infty}^{\infty} \{[S(u_i) \otimes \delta(u_i - u)] G(u_i, t, \sigma_u)\} \times K_{-\alpha}(\tau, u_i) du_i,$$

where $[S(u_i) \otimes \delta(u_i - u)]$ is the forward translation of $S(u_i)$ on u.

It can be shown using the properties of kernel of the FRFT [17]:

$$ST_s^\alpha(\tau, u) = ST_s(\tau, f) \text{ for } \alpha = \frac{\pi}{2}.$$

The GFST transform combines the Gaussian window function which is slowly varying envelope and localizes in time, and the FRFT kernel which selects the time–frequency being localized on new axis, u. It is the time–frequency localizing Gaussian that is shifted or translated on new axis, u at an angle α. The GFST produces a localized time–frequency representation.

2.2. Recovery of Original Signal

To recover the original signal from GFST, we take the average of (4) to obtain

$$\int_{-\infty}^{\infty} ST_s^\alpha(\tau - t, u, \sigma_u) d\tau = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s(t) g(\tau - t) K_\alpha(t, u) d\tau dt$$

$$= \int_{-\infty}^{\infty} s(t) K_\alpha(t, u) dt \int_{-\infty}^{\infty} g(\tau - t) d\tau.$$

Using (2) we obtain

$$\int_{-\infty}^{\infty} ST_s^\alpha(\tau - t, u, \sigma_u) d\tau = S_\alpha(u),$$

where $S_\alpha(u)$ is FRFT of $s(t)$.

The signal $s(t)$ can be recovered from $S_\alpha(u)$ using

$$s(t) = \int_{-\infty}^{\infty} S_\alpha(u) K_{-\alpha}(t, u) du.$$

The generalized fractional ST is expected to perform better for signals with changing frequency with regard to their time–frequency representation. The merits of ST over FT are enhanced for GFST against FRFT. The features are quite discriminating for different values of angle α in GFST whereas in ST the flexibility in choosing the value of angle is absent. Considering the characteristic of noisy speech signal as highly non-stationary, the GFST have different axes (as per Fig. 1) in the time–frequency plane and are segmented by the Gaussian window of varying width as per equation (6). The features extracted from the segmented axis at specific angle in the time–frequency plane are found more discriminating (max at $a = 0.5$) as compared to simple FRFT based features. The work on FRFT based discriminating features is presented by the author in [20].

3. k-NN CLASSIFIER

The k-NN classifier [21] simply places the different points (feature vectors) of the training set in the feature space, and the decision of classification is done according to "voting" of the nearest neighbors to a point under test. The voting is done by picking the k points nearest to the test point, and the selected class is the class that is most often picked. The distance between can be measured with different metrics, the most often used are the well-known Euclidean distance and the Mahalanobis distance. In our methodologies, the distance is measured using the Mahalanobis metric [22].

The square Mahalanobis distance between the feature vectors FV_1 and FV_2 is given by

$$d^2 = (FV_1 - FV_2)^T \Sigma^{-1} (FV_1 - FV_2),$$

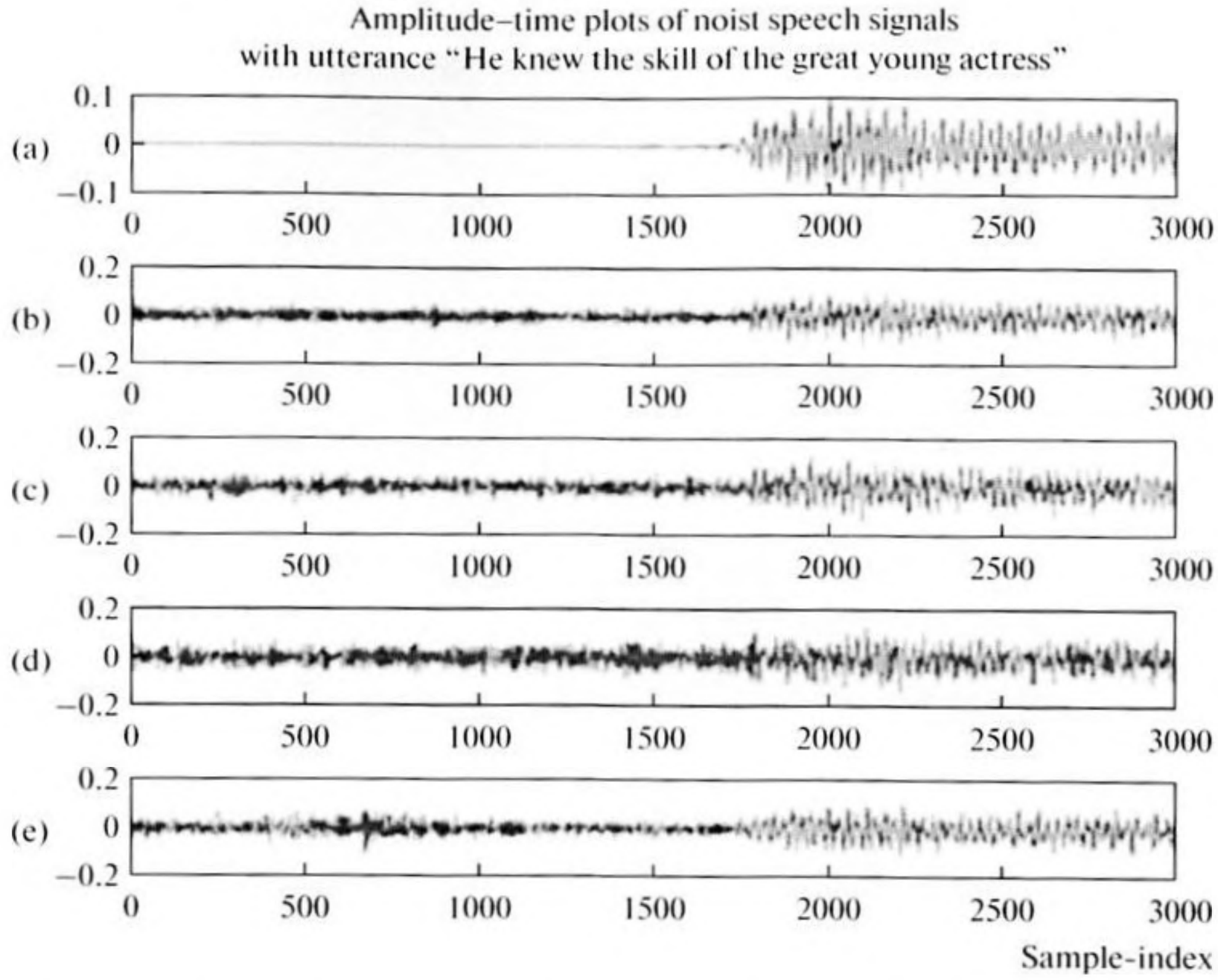

Fig. 2. Plots of the clean and speech signals with background noise: (a) clean speech, (b) train noise, (c) airport noise, (d) car noise, (e) restaurant noise.

where Σ is the covariance matrix of the training data. Here, the algorithm used relies on the assumption that the covariance matrix is the same for all classes, which is not true for a majority of the practical cases. There are several advantages in using the Mahalanobis metric instead of the Euclidean metric which are as follows:

—the Mahalanobis metric automatically scales the coordinate axes of the feature space;

—it decorrelates the different features to the whole set of training samples as one entity;

—the Mahalanobis metric is able to accommodate curved as well as linear decision boundaries.

4. RESULTS AND ANALYSIS

The proposed generalized fractional ST is applied to discriminate the different environmental background noise sources mixed with speech signals. The signals used in the simulations are taken from a NOIZEUS Speech Database. Here only four types of background noise sources, i.e., train, airport, car and restaurant mixed with human speech signals are taken into consideration. Figure 2 shows plots of some of the clean and noisy speech signals at sampling rate 8000 samples per second.

Figures 3 and 4 show the pseudo-color plots (checker-board plots) of ST corresponding to background train noise in the presence of speech utterances. It is in the form of continuous regions, hence little cumbersome to find feature vectors for discrimination from other noise sources.

In Fig. 4, the rotation of axis by 0.3π (at FRFT order = 0.6) makes the energy scattered as compared to Fig. 3, hence time and frequency localized points contain more information regarding the characteristics of signal as compared to simple ST. The following features are extracted from the discrete version of GFST and ST of corresponding noisy speech signals. These features are extracted in terms of frequency (row) value corresponding to time-frame (column) values (from maximum and up to 90% of max value) for each column:

—maximum frequency value per time-frame;

—minimum frequency value per time-frame;

—mean frequency value per time-frame;

—standard deviation of frequency values per time-frame.

These feature values are taken as average of first 100 frames of the signal. The frame period is chosen as 50 ms. The set of four features as mentioned for train and airport background noisy speech signals with SNR 5 dB corresponding to ST and GFST (for $a = 0.5$) matrices are shown in Figs. 5 and 6 respectively with respect to time-frame for the first 100 time-frames. The utterances used corresponding to the reported

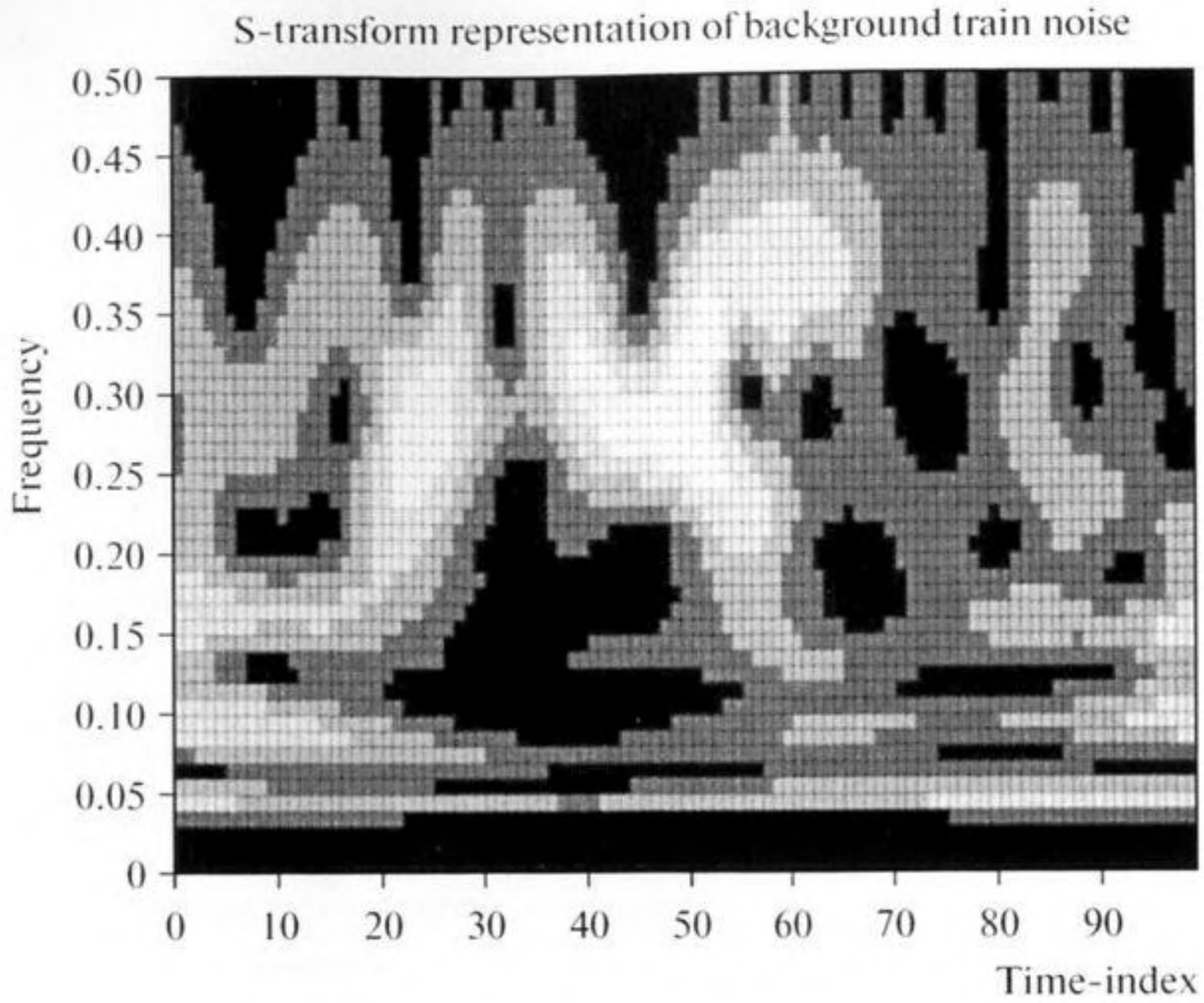

Fig. 3. S-transform representation of train noise.

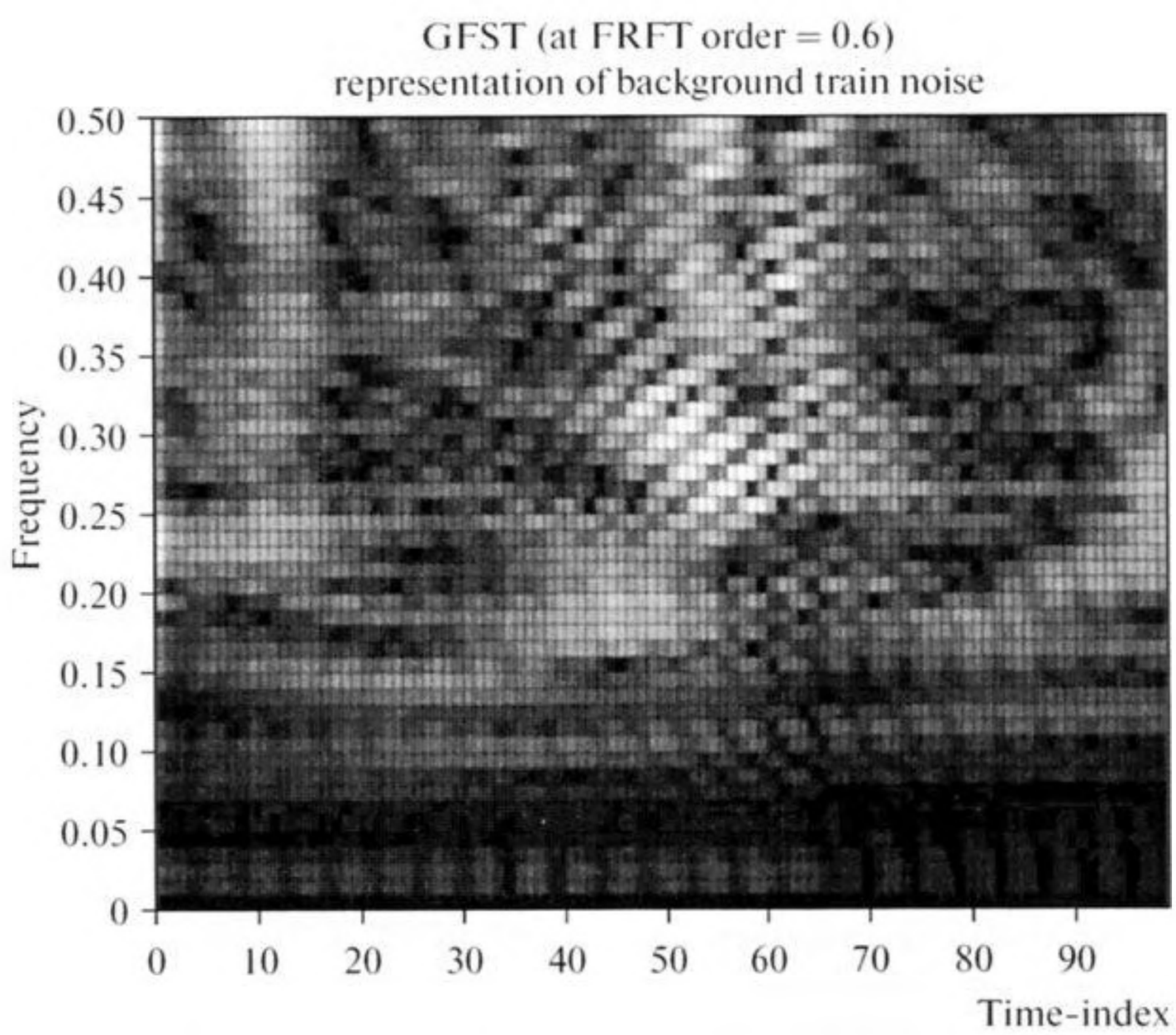

Fig. 4. GFST (at FRFT order = 0.6) representation of train noise.

plots are "He knew the skill of the great young actress." The same features for train and restaurant background noisy speech signals with SNR 5 dB corresponding to ST and GFST (for $a = 0.5$) matrices for same utterances are shown in Figs. 7 and 8 respectively with respect to time-frame for the first 100 time-frames.

The Euclidean distance between feature vectors of different pairs of background noise sources is plotted on Fig. 9 with respect to different values of fractional

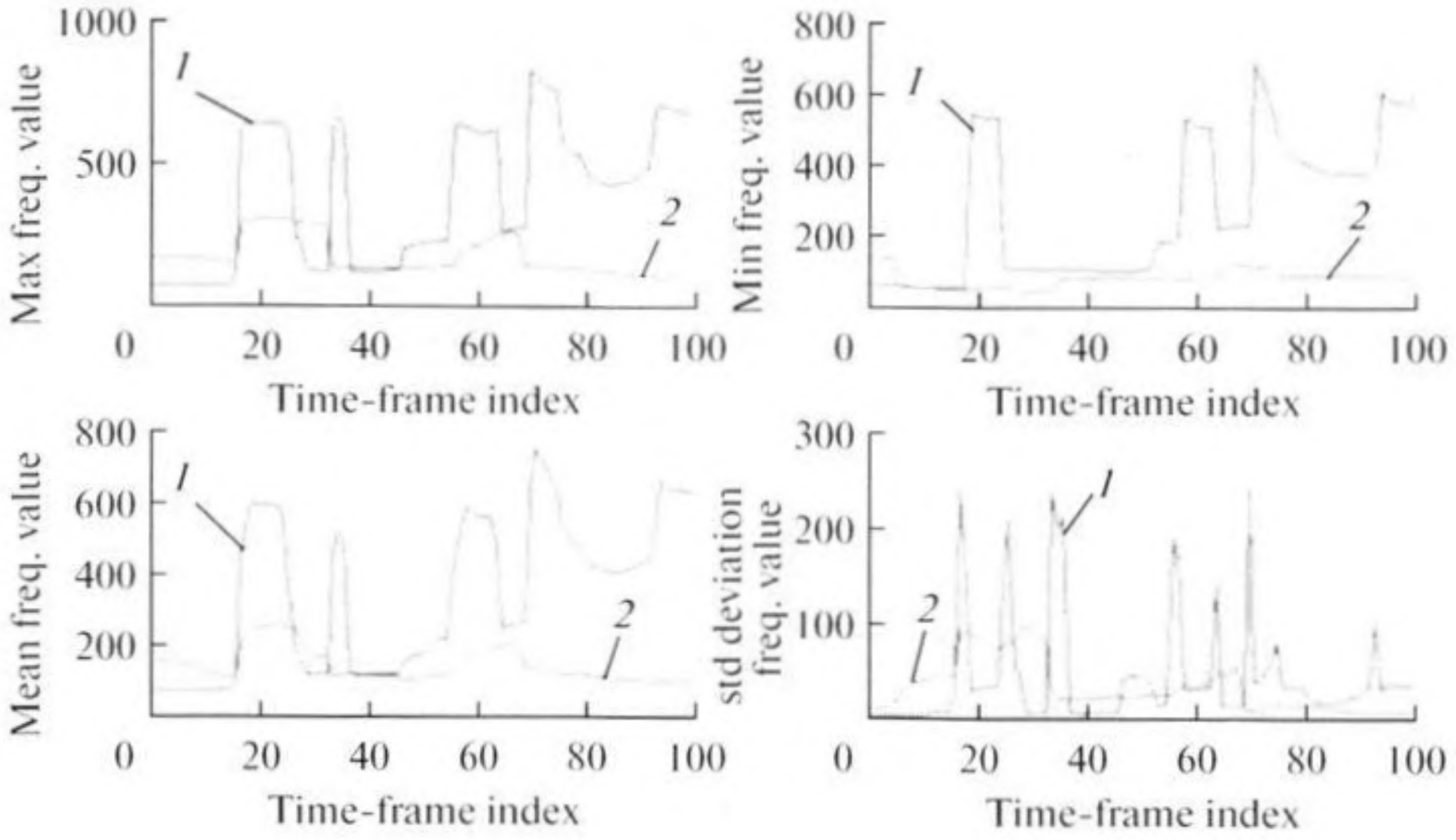

Fig. 5. Plot of feature values of train and airport background noisy speech signals corresponding to ST (*1*—train noise, *2*—airport noise).

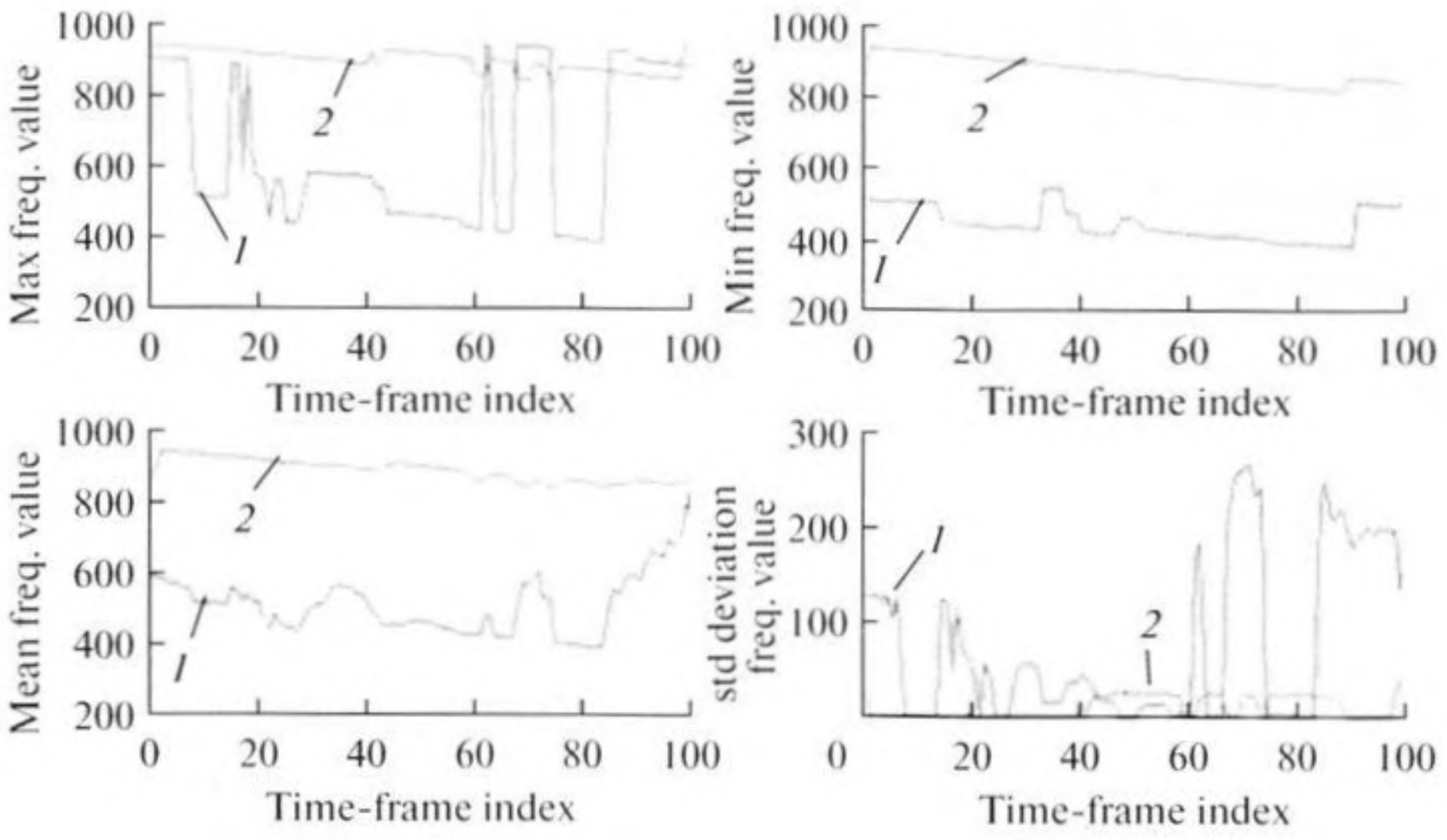

Fig. 6. Plot of feature values for train and airport background noisy speech signals corresponding to GSFT at $a = 0.5$ (*1*—train noise, *2*—airport noise).

order/fractional power, a and ST. The fractional power is already defined in section 2, equation (5). The Euclidean distance increases with maximum value at fractional power of GFST, $a = 0.5$, which is much higher as compared to the value corresponding to ST. It shows that the separation between the feature vectors corresponding to two different noise sources makes the classifier convenient and unambiguous to take the decision of classification.

The classification accuracy using k-NN classifier is also increased from 68% (in case of ST) to 85% (in case of GSFT with $a = 0.5$) for discrimination between some pairs of noise sources whose results are shown in the plot. This trend has been observed and verified in the 40 samples of the noisy speech signals for all four types of noise sources as mentioned. In the case of car–restaurant noise pair, the results are observed better for $a = 0.2$. This is the beauty of the method. A theoretical model cannot be presented for choosing the suitable value of a for discrimination of different background noise signals in the world, as this decision is absolutely based on the simulation outcomes on MatLab platform.

The Euclidean distance between feature vectors of different pairs of background noise sources is plotted in Fig. 10 for different SNR values. The reduction of Euclidean distance at $a = 0.5$ as SNR is improved (Fig. 10) from 5 to 10 dB. As SNR degrades, the effect

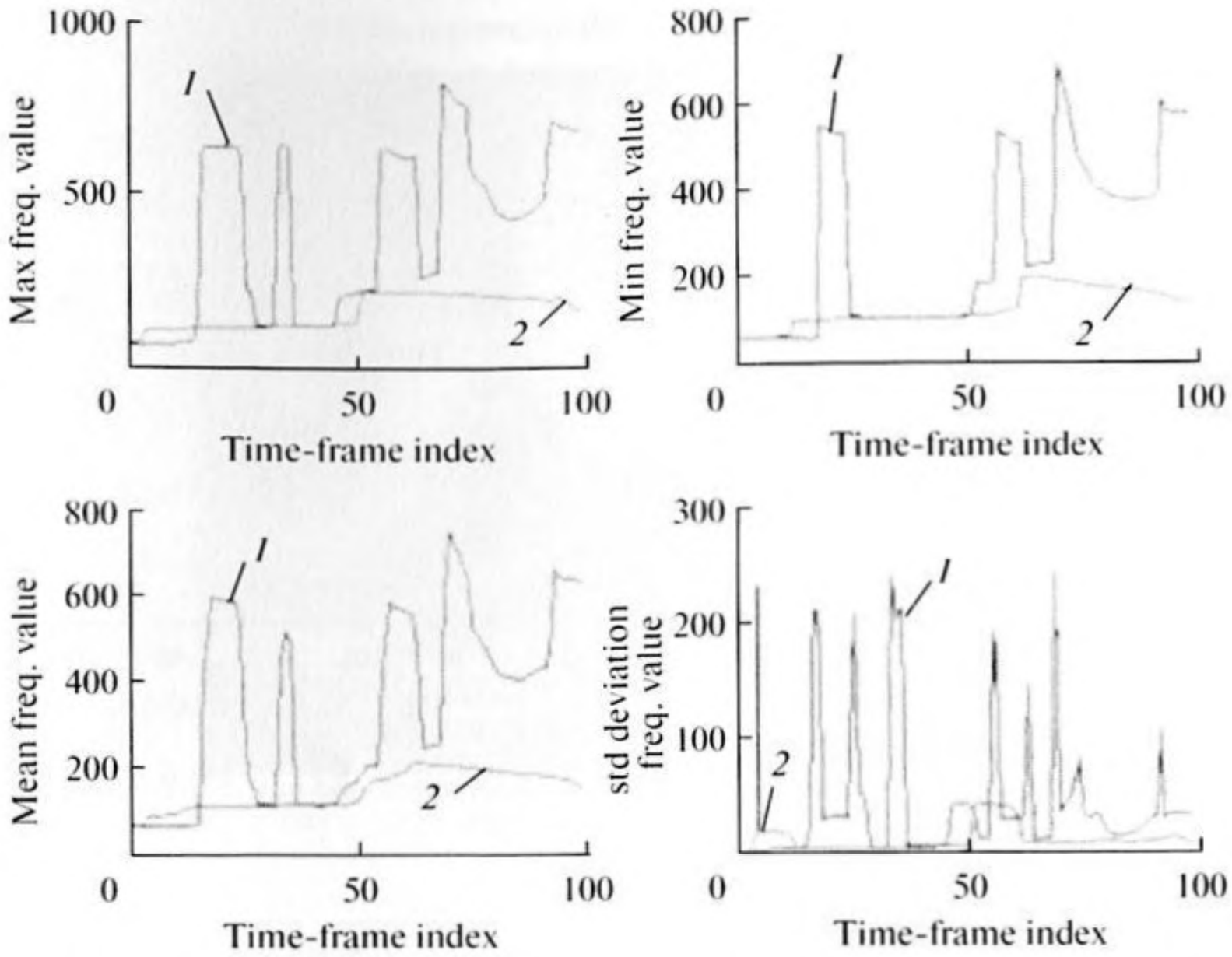

Fig. 7. Plot of feature values of train and restaurant background noisy speech signals corresponding to ST (*1*—train noise, *2*—restaurant noise).

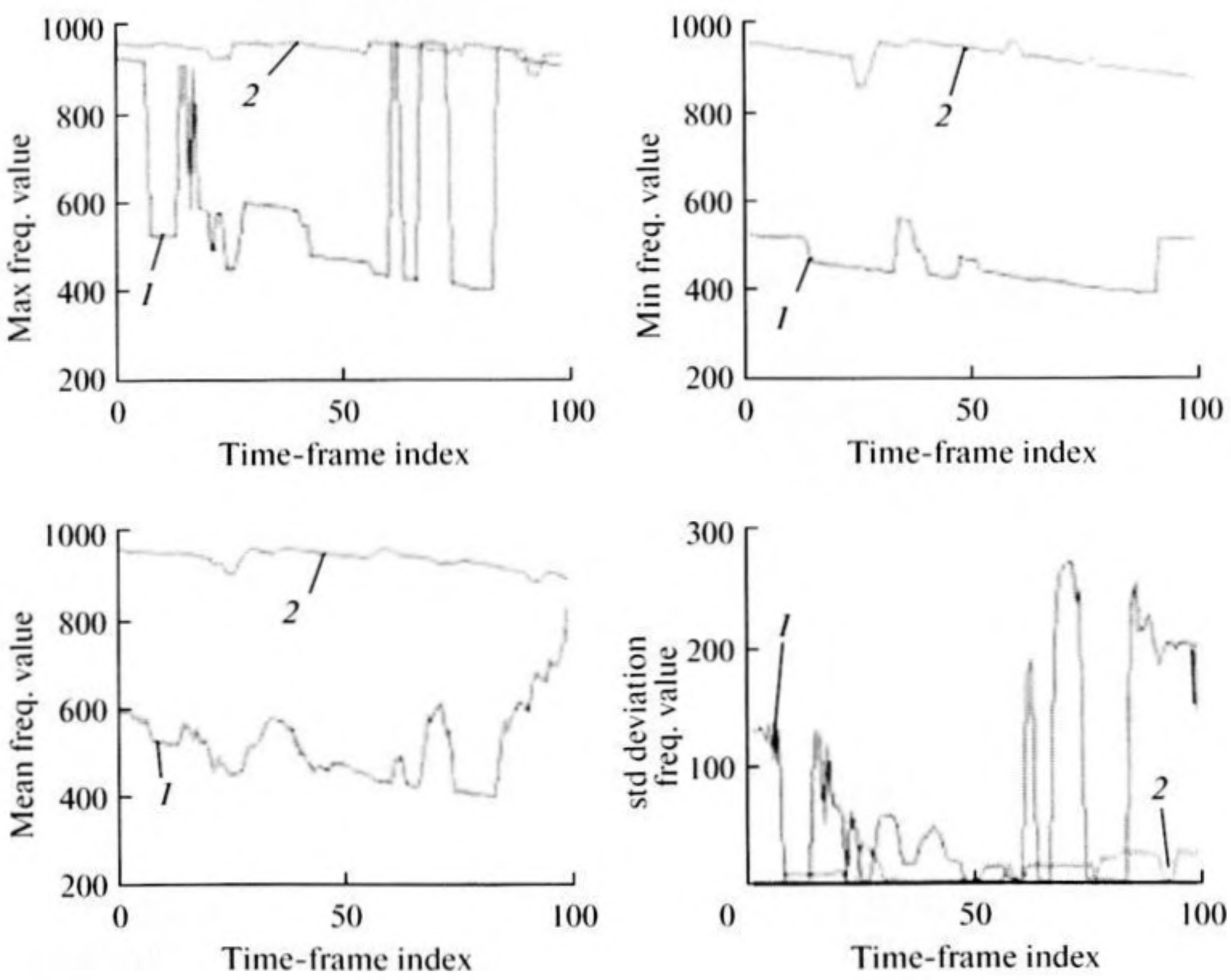

Fig. 8. Plot of feature values for train and restaurant background noisy speech signals corresponding to GSFT at $a = 0.5$ (*1*—train noise, *2*—restaurant noise).

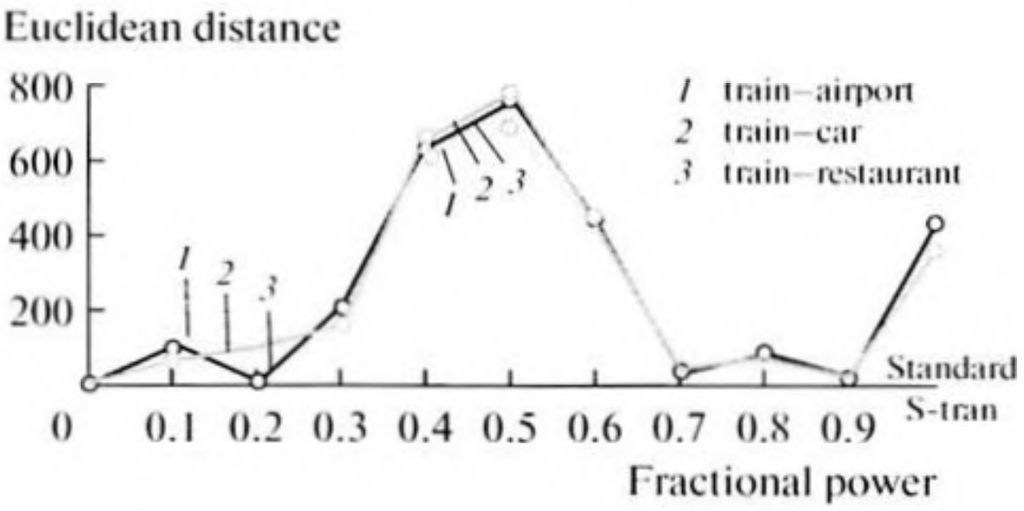

Fig. 9. Multiple plots of Euclidean distance between feature vectors of different pairs of background noise sources vs. fractional power.

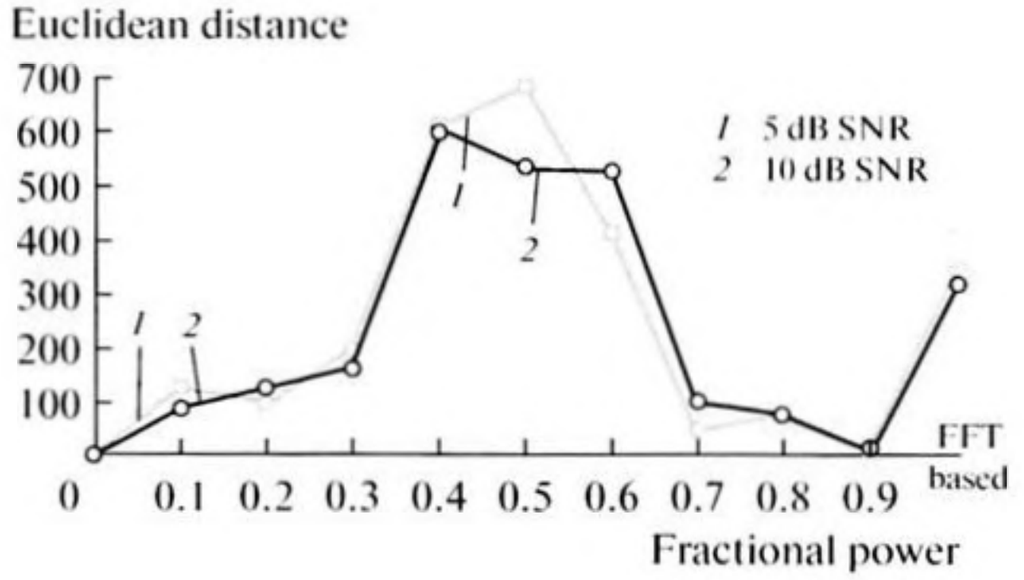

Fig. 10. Multiple plots of Euclidean distance between feature vectors of different pairs of background noise sources corresponding to SNR 5 and 10 dB vs. fractional power.

of background noise enhances over the speech signal and it is prominent at $a = 0.5$. For lower level of background noise as compared to speech content, the discrimination of different noise sources is convenient for classifiers at other fractional powers. The technique of discriminating noise sources performs better in smaller background noise also, whereas in original ST the results are independent of SNR. It proves through the experiment that the fractional power is one of the key parameter of GFST for success of classification in different cases which is absent in case of ST.

5. CONCLUSIONS

Generalized fractional ST based features are proved to be an effective tool to discriminate the background noise sources mixed with human speech signals. More features may be computed to enhance the number of noise sources for discrimination purpose. This technique may be useful in finding the location of speaker on the basis of recognizing background after analyzing the recorded signals. As shown in results, this methodology is more effective in noisy speech signals with lower SNR.

REFERENCES

1. S. Chu, S. Narayanan, and C. C. Jay Kuo, IEEE Trans. Audio, Speech Lang. Proc. **6** (17), (2009).
2. J. Yang, in *Proc. 6th Int. Conf. on Information Tech.: New Generation, 2009.*
3. M. A. Sobreira-Seoane, A. R. Molares, and J. L. Alba Castro, *Proc. Conf. "Acoustics–08," Paris, June 29–July 4,* 2008.
4. F. Hlawatsch and G. F. Bourdeaux-Bartels, IEEE Signal Proc. Mag. **9** (2), 21 (1992).
5. M. R. Portnoff, IEEE Trans. Acoust., Speech, Signal Proc. **28** (1), 55 (1980).
6. L. Cohen, Proc. IEEE **77**, 941 (1989).
7. S. Mallat, *A Wavelet Tour of Signal Processing* (Academic, London, 1998).
8. R. G. Stockwell, L. Mansinha, and P. Lowe, IEEE Trans. Signal Proc. **44**, 998 (1996).
9. R. G. Stockwell, PhD Thesis (Univ. Western Ontario, Canada, 1999).
10. C. R. Pinnegar and L. Mansinha, Geophysics **68**, 381 (2003).
11. C. R. Pinnegar and L. Mansinha, SIAM J. Sci. Comput. **24**, 1678 (2003).
12. C. R. Pinnegar and L. Mansinha, ELSEVIER J. Signal Proc. **84**, 1167 (2004).
13. I. Ya. Chebotareva, Acoust. Phys. **57**, 857 (2011).
14. A. S. Ivanenkov, A. A. Rodionov, and V. I. Turchin, Acoust. Phys. **59**, 179 (2013).
15. J. Namias, J. Inst. Math. Appl. **25**, 241 (1980).
16. A. McBride and F. Kerr, IMA J. Appl. Mat. **39**, 159 (1987).
17. L. Almeida, IEEE Trans. Signal Proc. **42**, 3084 (1994).
18. H. Ozaktas, Z. Zalevsky, and M. Kutay, in *The Fractional Fourier Transform with Applications in Optics and Signal Processing* (Wiley, Chichester, 2001), pp. 99–107.
19. H. Ozaktas, O. Arikan, M. Kutay, and G. Bozdag, IEEE Trans. Signal Proc. **44**, 2141 (1996).
20. D. Jhanwar, K. Sharma Kamlesh, S. G. Modani, in *Proc. ICCOS, Coimbatore, India, March 17–18, 2011,* pp. 689–694.
21. T. M. Cover and P. E. Hart, IEEE Trans. Inform. Theory **13**, 21 (1967).
22. V. Peltonen, MS Thesis (Tampere Univ. of Techno, Tampere, Finland, 2001).