ISSN 1063-7710, Acoustical Physics, 2015, Vol. 61, No. 5, pp. 585–596. © Pleiades Publishing, Ltd., 2015.

ACOUSTICAL SIGNAL PROCESSING AND COMPUTER SIMULATION

Discrimination of Environmental Background Noise in Presence of Speech Using Sample-Pairs Statistics Based Features[1]

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

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

[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 April 20, 2014

Abstract—A methodology to discriminate the different classes of background noise using new features based on samples of the signal is presented here. Two consecutive samples of different amplitude of the discrete-time signals are termed as sample-pair and 14 types of sample-pairs are considered here as fundamental features. Results of simulation work proves that count of some of such type of sample-pairs as well as count of few combinations of two, three and four such sample-pairs are useful to detect and discriminate the different acoustic noise mixed with speech signals. On the basis of simulation results, the performance of proposed features have proved better than other spectral features like Mel Frequency Cepstral Coefficients (MFCC), Spectral Centroid, Spectral Flux and Spectral Roll-off regarding discrimination capabilities, simplicity of extraction process and lesser dependency over speech utterances mixed with noise. These sample-pairs based features having advantage of not requiring frame-decomposition and silence period removal. Their discrimination capabilities are shown by Fisher's F-ratio as performance index. The multiclass Support Vector Machine (SVM) is used as a classifier.

Keywords: Fisher's F-ratio, Mel Frequency Cepstral Coefficients, Sample-pair, Spectral Centroid, Spectral Flux, Spectral Roll-off, Support Vector Machine

DOI: 10.1134/S1063771015050103

1. INTRODUCTION

Background environmental sound categorization 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. Human utilizes both vision and hearing information to navigate and respond to surroundings. When robotics application is employed to understand unstructured environment [1–4], the robustness and utility will get vanished, if visual information is compromised. To capture complete detail of a scene, the fusion of audio and visual information can prove to be advantageous. Audio data could be obtained at any moment when the system is functioning, in spite of challenging external conditions like low intensity of lights or visual obstruction. It is relatively easier to handle audio than visual signals. There have been recent trends in finding solutions to provide hearing information for mobile robots to enhance their context awareness with audio information [5, 6]. In the context aware applications [7, 8], 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 classroom [7] or can ring louder in the crowded place. This technique can provide information regarding the location of user [9]. Several temporal and spectral features have been used to describe audio signals. In the most audio recognition systems, the usage of MFCC feature can be widely observed. Single sound source can be characterized easily through MFCC. Environmental background noise or sound signals contain large varieties of signal for which MFCC normally fails. Other popular features for audio signals include Linear Prediction Coefficients (LPC), Band Energy Ratio, Spectral Roll-off, Spectral Centroid, Spectral Asymmetry, Spectral Bandwidth, Spectral Flatness, Zero-Crossing, and Energy [10]. The problem of using a large number of features is that there are many irrelevant features which can deteriorate the accuracy of classification.

Nowadays, research on general audio environment recognition has been obtained a little attention as compared to the structured audio such as speech or

[1] The article is published in the original.

music [11–15]. Major research in environmental sounds has been centered on recognition of specific events or sounds [16]. Due to randomness and high variance of environmental sounds, the recognition rates have been limited depending on the number of targeted classes. With the help of MFCCs and other features the recognition rates are around 92% for 5 classes [17]. The simple feature extraction technique reduces the computational complexity and running time. It is with these findings and applications that motivate us to look for a more effective approach for discriminating and classification of environmental background noise/sounds. To achieve this goal, we investigated in ways of extracting features and introduced a novel idea to extract features for unstructured sounds with simpler technique. To determine the type of mixing used in a stereo signal mix, the presented work compares two blind classification algorithms that divide individual time–frequency regions into six classes of mixing types. The results show that mixing strategies vary with the musical genre, and the classification algorithm can reduce the instability and center disintegration when up-mixing a stereo signal to a multichannel format [18]. In environmental audio analysis, individual sound events related to some activity, for example, sounds of footsteps from a walking person can be solved by matching some prototype time-frequency (TF) patterns to a TF-representation of the input signals to obtain time-varying probability functions for the prototype events. The method introduced in paper [19] is based on a small number of locally collected event patterns that are used directly to dene features for weighted cascade of weak classifiers that is trained using the AdaBoost algorithm. The results of a comparison to a traditional sound event classifier based on the mel-frequency cepstrum coefficients and a hidden Markov model are very hopeful. The increasing availability of forensic audio surveillance recordings data makes human audition not practical therefore the rationale and potential application of several techniques for high-speed computerized search is discussed in [20]. Solution of problems of estimation of direction and localization of individual sources on a ship hull is demonstrated in [21]. A series of analytical calculated models for predicting the noise in an aircraft cabin is developed and presented in [22]. A method of estimating the isotropic background sea noise level with a horizontal array in the presence of uncorrelated interference and interference with a complex spatial structure is proposed in [23].

In the presented methodology, four different environmental background noises i.e. airport, train, car and street, mixed with human speech signals are discriminated using newly proposed features and one-against-all SVM. Two consecutive samples of different amplitude of the discrete-time signals are termed as sample-pair and 14 types of such sample-pairs are considered here. Results of simulation work proves that the count of some of such type of sample-pairs as well as count of few combinations of two, three and four such sample-pairs are useful features to detect and discriminate the different acoustic noise mixed speech signals.

On the basis of simulation results, the new features have been proved better than the other spectral features like MFCCs, Spectral Centroid, Spectral Flux and Spectral Roll-off regarding discrimination capabilities, simplicity of extraction process and lesser dependency over speech utterances mixed with noise. These sample-pairs based features are simple as their extraction does not require any special pre-processing of signal including frame-decomposition and silence period removal. Their discrimination capabilities are shown by Fisher's F-ratio as performance index. The one-against-all multiclass SVM is used to discriminate different environmental background acoustic noise on the basis of proposed features. This methodology has performed well in discriminating and classifying unstructured background acoustic environment mixed with speech, achieving 70–85% classification accuracy. As a consequence, the technique has proved promising for discrimination of such type of signals.

The rest of the paper is organized as follows. Overview of multiclass SVM and Fisher's F-ratio is discussed in Section 2. The proposed features and methodology is given in Section 3. Section 4 contains MATLAB based results and analysis including performance evaluation of proposed features in comparison with other traditional features. Finally, concluding remarks are given in Section 5.

2. BACKGROUND OF MULTICLASS SVM AND FISHER'S F-RATIO

A. Multiclass SVM

One-against-all SVM is one of the effective multiclass SVM used here for discrimination purpose [24, 25]. The linear two-class SVM (hard-margin type) is the basic SVM of any category of multiclass SVM.

In linear two-class SVM, the training data are linearly separable in the input space, X. Let there be N_c—*classes* in a problem of classification, the N_c direct decision functions to be determined to discriminate one class from the remaining classes. Let the jth decision function, that separates class j from remaining classes with maximum margin be

$$D_j(X) = W_j^T X + b_j, \quad (1)$$

where $-1 < D_j(X) < 1$, decision function known as the separating hyperplane. There are an infinite number of decision functions, which are separating hyperplanes. The generalization ability depends on the location of

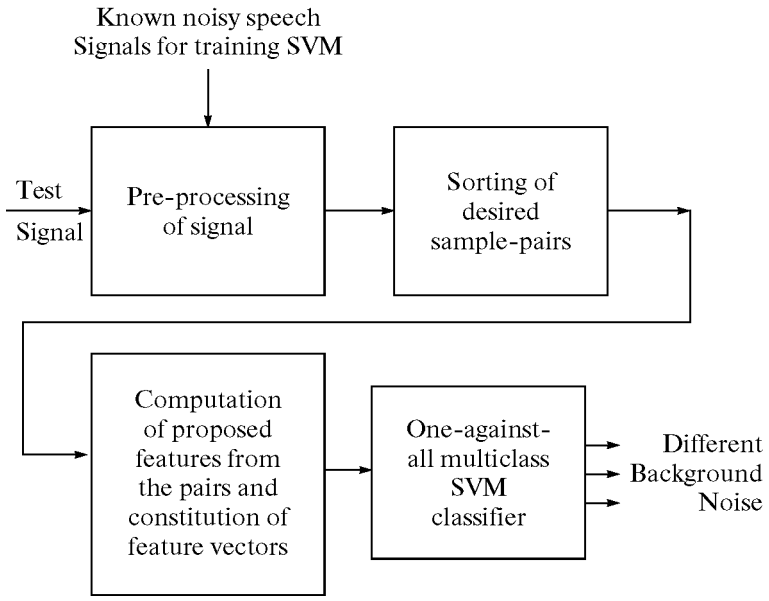

Fig. 1. Block diagram for computation of proposed features.

the separating hyperplane, and the hyperplane with the maximum margin is known as the optimal separating hyperplane [25]. W_j is the l-dim vector, b_j is the bias term and X is the original input data. If the training data are not linearly separable then it is required to map the input space into the high-dimensional feature space to enhance linear separation in the feature space. Let $m(X)$ is the mapping function which maps original input data X into the l-dim feature space then the decision function modifies as follows:

$$D_j(X) = W_j^T m(X) + b_j. \quad (2)$$

In the technique of kernel trick, function $H(X, X')$ is used in place of $m(X)$. As per Hilbert–Schmidt theory, if a symmetric function $H(X, X')$ satisfies [25]

$$\sum_{i,j=1}^{L} h_i h_j H(X, X') \geq 0, \quad (3)$$

where L takes on natural numbers and h takes on real numbers then there is a relation of $m(X)$ with the function $H(X, X')$ which is to be satisfied as follows:

$$H(X, X') = m^T(X)m(X). \quad (4)$$

There is no need to treat the high-dimensional feature space explicitly by using kernel $H(X, X')$ in place of $m(X)$. The kernel (excluding linear kernel) based SVMs are categorized under non-linear SVM [25]. The kernels such as polynomial, radial basis function, wavelet, polymetric, tensor wave, numerical, dynamic time-alignment kernels etc. may be used in SVM. For separable classification problems, the training data corresponding to class j satisfy the equation

$$D_j(X) \geq 1 \quad (5)$$

and the remaining data belonging to rest of the classes i.e. $(N_c - 1)$ classes satisfy the equation

$$D_j(X) \leq -1. \quad (6)$$

The sign of the decision function is used in classification and thus decision is discrete.

B. Fisher's F-Ratio

Fisher's F-ratio is a statistical parameter in the analysis of variance where multi-class data are available. If there are N_c number of classes, and if each class consists of N number of data points, then

$$\text{F-ratio} = \frac{\text{variance of means between the classes}}{\text{average variance within the classes}}. \quad (7)$$

If μ_j and μ_k are the mean vectors of the features of jth and kth classes $(j \neq k)$ respectively as well as σ_j^2 and σ_k^2 are corresponding intra-class variances respectively, then F-ratio is expressed as:

$$\text{F-ratio} = \sum_{j}^{N_c} \sum_{k \neq j}^{N_c} \frac{(\mu_j - \mu_k)^2}{\sigma_j^2 + \sigma_k^2}. \quad (8)$$

When the spreads of feature means of corresponding class increase, or the distributions within classes become narrower, the value of F-ratio will become larger, which means that the feature on the corresponding dimension has a larger discrimination.

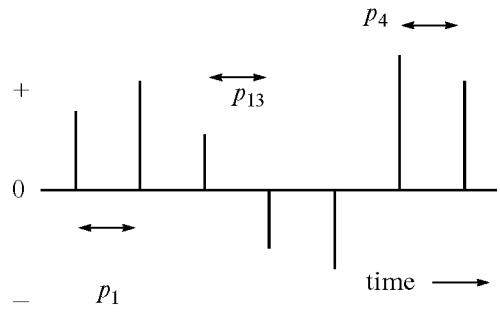

Fig. 2. Representation of sample-pairs.

3. PROPOSED FEATURES AND METHODOLOGY

The basic scheme for computation of proposed features is shown in Fig. 1. The related simulation work is done on MATLAB platform.

A. Pre-processing of Signals

The applied continuous training or testing signals are sampled at particular sampling rate f_s. In this paper, $f_s = 10.000$ samples/s is considered. The sequence of samples is cut at particular time interval from the beginning to save restricted number of samples for computation purpose. The total number of samples should be even to get integer number of sample-pairs.

Let $X[n]$ be the sample sequence of discrete-time noisy speech signal, where $1 \le n \le N$ represents the sample index at time interval of one sample period, $T_s = 1/f_s$, and N is an even number, represents the total number of samples.

B. Categorization of Sample-pairs Present in the Sequence

Defining Sample-Pairs

Let $r(i)$ be the ratio of two consecutive non-zero sample amplitude of same polarity starting from first sample of the segment, where $1 \le i \le N/2$, represents the number of sample-pair. The samples are paired as per their natural sequence of occurrence,

$$r(i) = \frac{X(2i)}{X(2i-1)}, \quad \text{if } |X(2i)| > |X(2i-1)|, \tag{9}$$

$$r(i) = \frac{X(2i-1)}{X(2i)}, \quad \text{if } |X(2i-1)| > |X(2i)|. \tag{10}$$

The ratio of the sample amplitudes in a sample-pair is not considered if any one of the following conditions are not satisfied,

$$X(2i), \quad X(2i-1) \ne 0, \tag{11}$$

$$X(2i) \ne X(2i-1), \tag{12}$$

$$X(2i)X(2i-1) > 0, \tag{13}$$

where means both should have same polarity.

Quantization of Ratio

The $r(i)$ is quantized among three bins of uniform width as follows:

$$\text{bin(1) if } 1.0 < r(i) \le 1.5, \tag{14}$$

$$\text{bin(2) if } 1.5 < r(i) \le 2.0, \tag{15}$$

$$\text{bin(3) if } 2.0 < r(i). \tag{16}$$

Sorting of Sample-Pairs

On the basis of (14–16), the sample-pairs are sorted containing samples as follows.

The sample-pairs containing respective sample amplitude as per (9), bin(1), bin(2) and bin(3) are termed as p_1, p_2 and p_3 respectively if the samples are of positive polarity and p_7, p_8 and p_9 respectively if the samples are of negative polarity.

The sample-pairs containing respective sample amplitude as per (10), bin(1), bin(2) and bin(3) are named as p_4, p_5 and p_6 respectively if the samples are of positive polarity and p_{10}, p_{11} and p_{12} respectively if the samples are of negative polarity. Two additional types of sample-pairs, irrespective of the ratio of their amplitudes, are named as p_{13} if the condition is such that $X(2i-1) > 0, X(2i) < 0$ and p_{14} if $X(2i-1) < 0$, $X(2i) > 0$.

C. Computation of Proposed Features

On the basis of sorted sample-pairs, the different features and their feature values are computed as follows.

Count of Each Type of Sample-pair

Discrete-time signal consisting of different pairs as shown in Fig. 2.

Let $np_1, np_2, np_3, \ldots, np_{14}$ are the number of sample-pairs $p_1, p_2, p_3, \ldots, p_{14}$ respectively found in the total number of samples in the corresponding signal and individually constitute as feature. The set of those features, which differs in their values considerably in signals with different background noise, constitute the feature vector. Thus elements of feature vectors may vary according to application.

$$np_1 + np_2 + np_3 + \ldots + np_{14} \approx \frac{N}{2}. \tag{17}$$

A few cases may be there in which the conditions (11), (12) and (13) are not satisfied hence the corresponding sample-pairs are not counted.

Count of Each Combination of Two Sample-pairs

The mentioned feature can be clearly defined by Fig. 3. Count of each combination of two sample-pairs is represented by $CP^2_{p,q}$ where p, q are integers and $1 \le p, q \le 14$.

Nos. of Possible Combinations, $CP^2_{p,q} = 196$. (18)

Each combination contains four samples. The combinations include those combinations also which contain the same type of sample-pairs. Each combination is counted in the total number of samples of corresponding signal from the beginning.

Count of Each Combination of Three Sample-pairs

The representation of this feature is shown in Fig. 4. Count of each combination of three sample-pairs is represented by $CP^3_{p,q,r}$, where p, q, r are integers and $1 \le p, q, r \le 14$.

Nos. of possible combinations, $CP^3_{p,q,r} = 2744$. (19)

Each combination $CP^3_{p,q,r}$ contains six consecutive samples. The feature is the count of the each combination in the total number of available samples of corresponding signal from the beginning. The feature vector includes count of those combinations which vary considerably from noise to noise irrespective of utterances.

Count of each combination of four sample-pairs

The representation of the feature is shown in Fig. 5. Count of each combination of four sample-pairs is represented by $CP^4_{p,q,r,s}$, where p, q, r, s are integers and $1 \le p, q, r, s \le 14$.

Nos. of possible combinations,

$$CP^4_{p,q,r,s} = 38.416. \quad (20)$$

Each combination $CP^4_{p,q,r,s}$ contains eight consecutive samples. The feature is the count of the each combination in the total number of available samples of corresponding signal from the beginning. As the order of combination increases, the corresponding feature value reduces.

The mentioned features are temporal features. These are simple to compute and not affected by statistical parameters. The silent parts in the acoustic signals are not influencing the results very much. These features are less affected by mixing of speech and environmental background acoustic signals. The higher order combinations are more in number, so their individual count reduces drastically for the total available samples of the signal under test.

Selection of Features

The feature vectors contain only those elements which show effective discrimination among the different noisy signals under test. The sample-pairs of unstructured background sound mixed with speech are highly uncorrelated. Therefore, the effectiveness of features for discrimination purpose can be measured in terms of Fisher's F-ratio (FR). The proposed selection method of features is shown in Fig. 6.

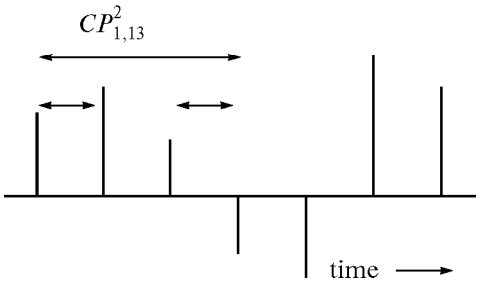

Fig. 3. Combinations of two sample-pairs.

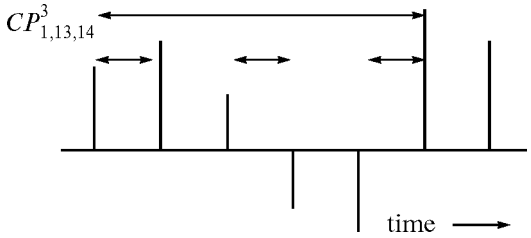

Fig. 4. Combination of three sample-pairs.

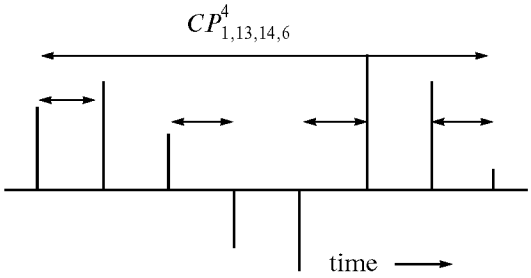

Fig. 5. Combination of four sample-pairs.

Constitution of Feature Vectors

In the proposed scheme, four types of feature vectors FV_1, FV_2, FV_3 and FV_4 are constituted as follows:

$$np_p \in NP, \quad (21)$$

$$CP^2_{p,q} \in CP^2, \quad (22)$$

$$CP^3_{p,q,r} \in CP^3, \quad (23)$$

$$CP^4_{p,q,r,s} \in CP^4, \quad (24)$$

$$FV_1 \subseteq NP, \quad (25)$$

$$FV_2 \subseteq CP^2, \quad (26)$$

$$FV_3 \subseteq CP^3, \quad (27)$$

$$FV_4 \subseteq CP^4. \quad (28)$$

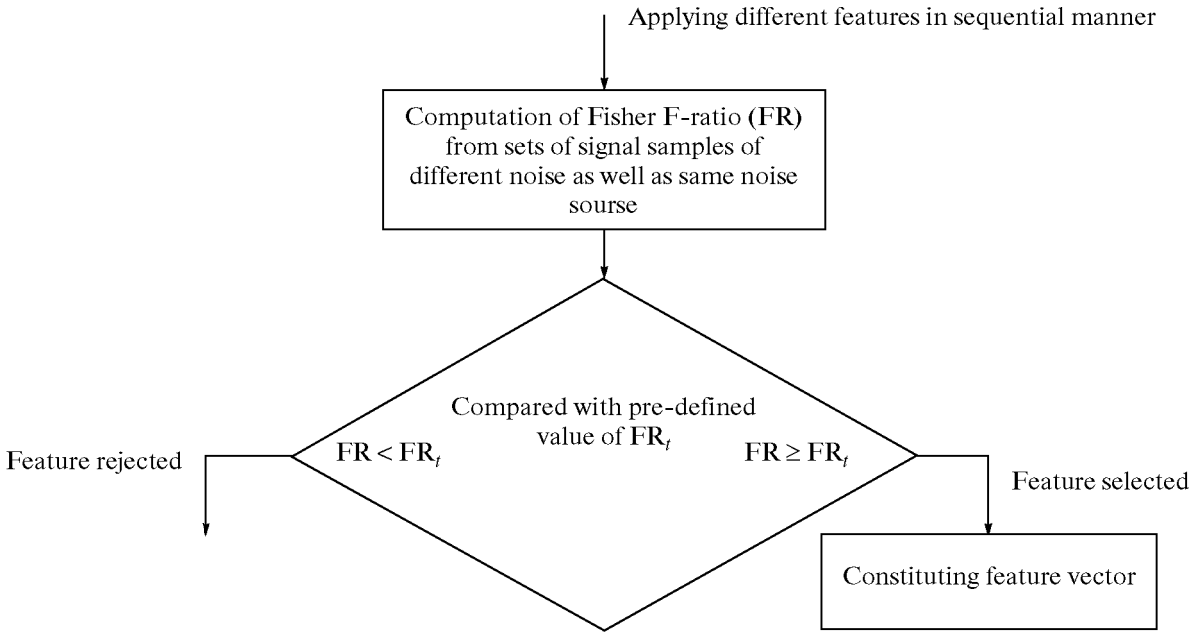

Fig. 6. Proposed feature selection method.

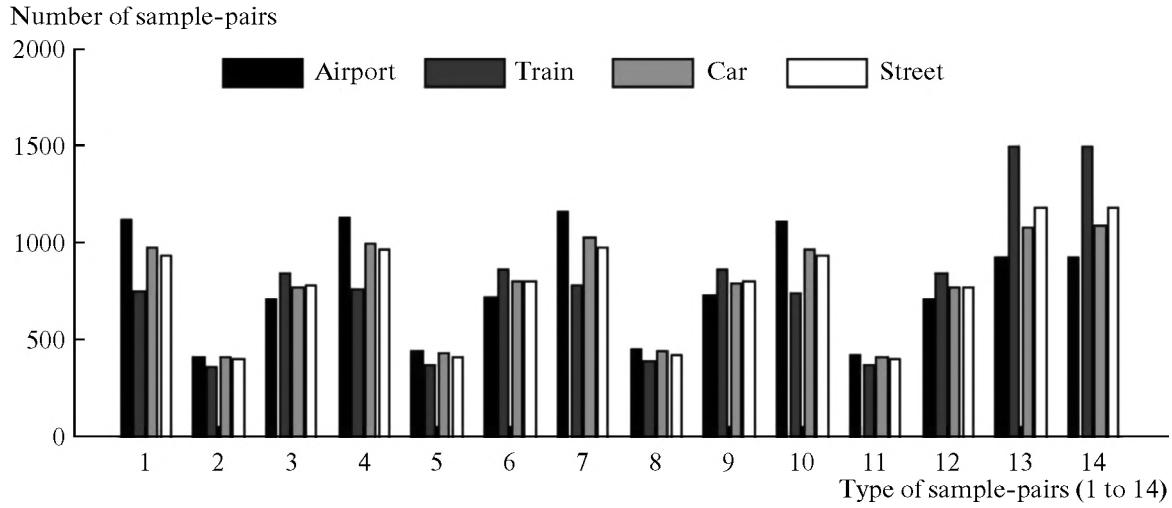

Fig. 7. Mean value of number of sample-pairs of each noise category.

4. RESULTS AND ANALYSIS

The signals for experiment are taken from the NOIZEUS noisy speech database [28]. The noisy database contains 30 IEEE sentences (produced by three male and three female speakers) corrupted by eight different real-world noises at different SNRs. The noise was taken from the AURORA database and includes suburban train noise, babble, car, exhibition hall, restaurant, street, airport, and train-station noise [28].

A. Nos. of Single Sample-pairs of Different Noise Sources

In Fig. 7 the bars show the number of sample-pairs in the first 22000 samples. This is the mean value over 20 signals of each type of noise source.

B. Performance Evaluation of Proposed Features

The following plots show the performance of discrimination of the proposed features in comparison

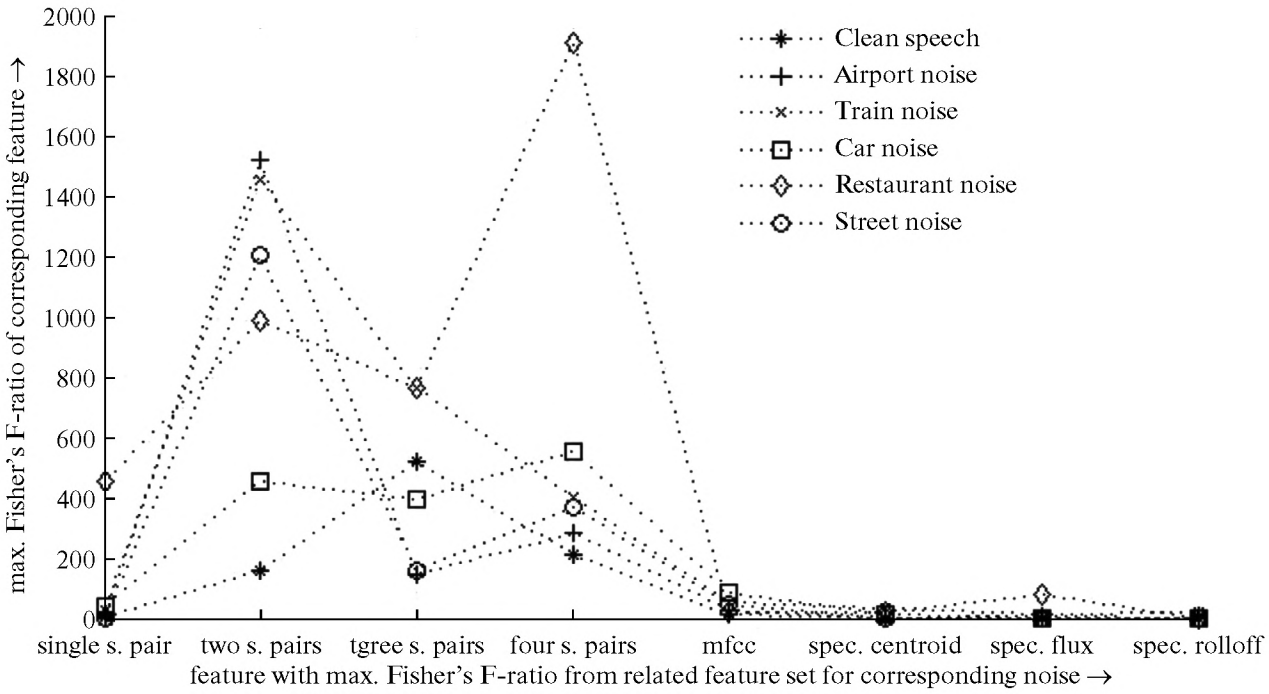

Fig. 8. Fisher's F-ratio of different feature types.

with traditional features like MFCCs, Spectral Centroid, Spectral Roll-off and Spectral Flux. 13 nos. of MFCCs are considered here. The values of different parameters are also computed for clean speech signal and restaurant noise.

Fisher's F-ratio

The maximum value of Fisher's F-ratio as parameter of effectiveness of different features is shown in Fig. 8. The plot clearly displays that the values corresponding to the proposed features are much higher as compared to the other traditional features.

Normalized Euclidean Distance Between Feature Vectors of Different Noise Sources

The normalized Euclidean distance between feature vectors of corresponding noise pairs of different features is depicted in Fig. 9. The plot shows that the distance parameter is considerable for almost all possible noise pairs in case of sample-pair based features in comparison to other traditional features.

C. Selection of Proposed Features from the Large Set of Three and Four Sample-pairs Combination based Features

In Figs. 10, 11 the Coefficient of Variance (COV) of some features displays better discriminating performance among different noise sources, and their closeness for the same type of noise sources with different speech utterances are shown. This exhibits that the new features are independent of speech utterances mixed with background signal. Although more features may be incorporated in the graph but due to limitation of the space and figure, some selected features are presented here.

In Tables 1 and 2 the different feature values are indicated as a mean value over 30 signals of each noise

Table 1. Single and double sample-pairs based discriminating feature vectors for different noises

Feature vector	Features	Airport	Train	Car	Street
		count of single sample-pairs			
FV_1	p_1	1143	823	1029	955
	p_7	1217	857	1082	1027
	p_{13}	881	1392	1033	1146
	p_{14}	879	1379	1023	1132
FV_2	$CP^2_{1,4}$	358	179	261	237
	$CP^2_{3,1}$	182	83	151	115
	$CP^2_{3,13}$	84	220	123	161
	$CP^2_{4,13}$	247	135	232	196
	$CP^2_{13,3}$	55	151	79	107
	$CP^2_{14,13}$	119	462	181	249

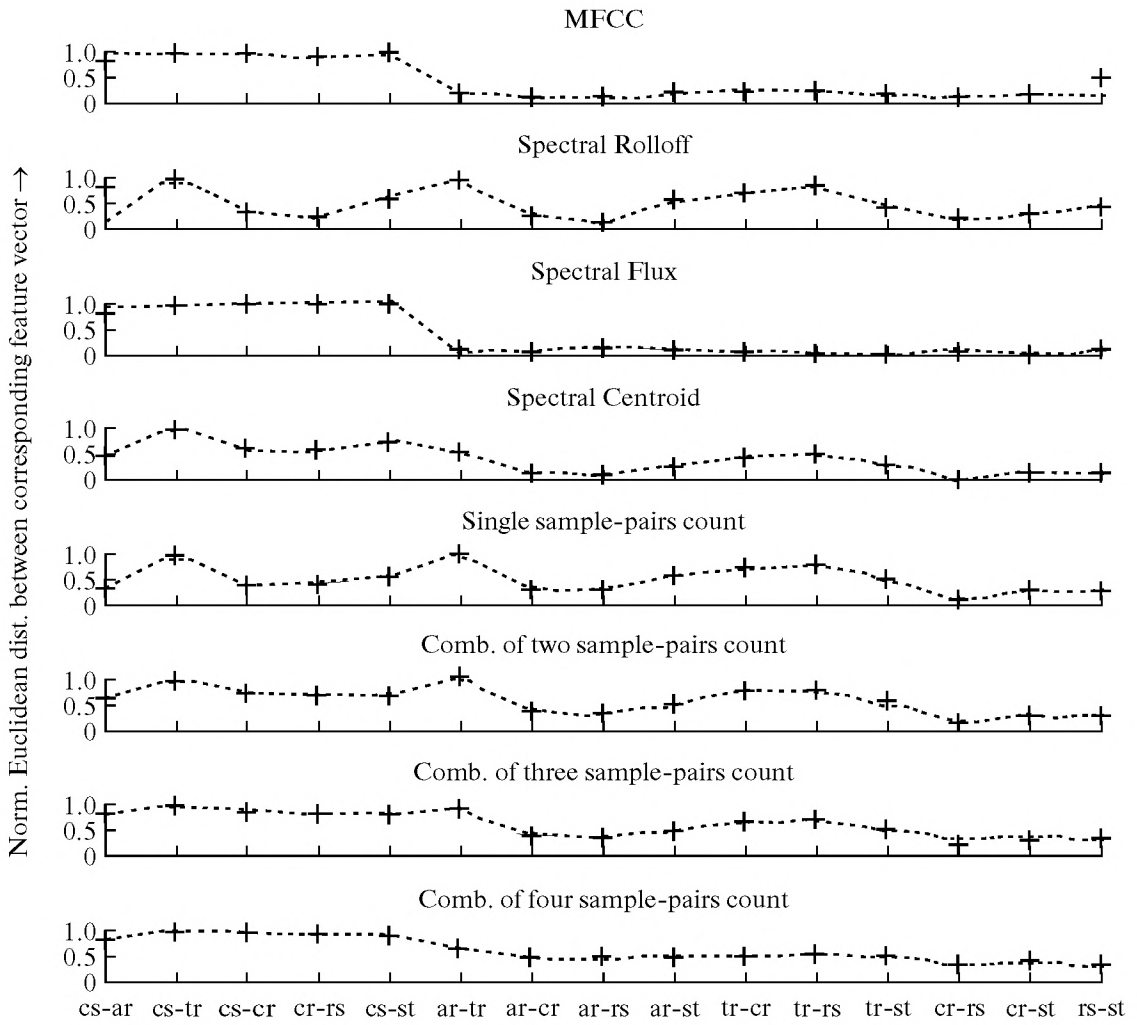

Fig. 9. Normalized Euclidean distance between feature vectors of corresponding noise pairs of different features ("cs"—clean speech, "tr"—train, "cr"—car, "ar"—airport, "rs"—restaurant and "st"—street noise).

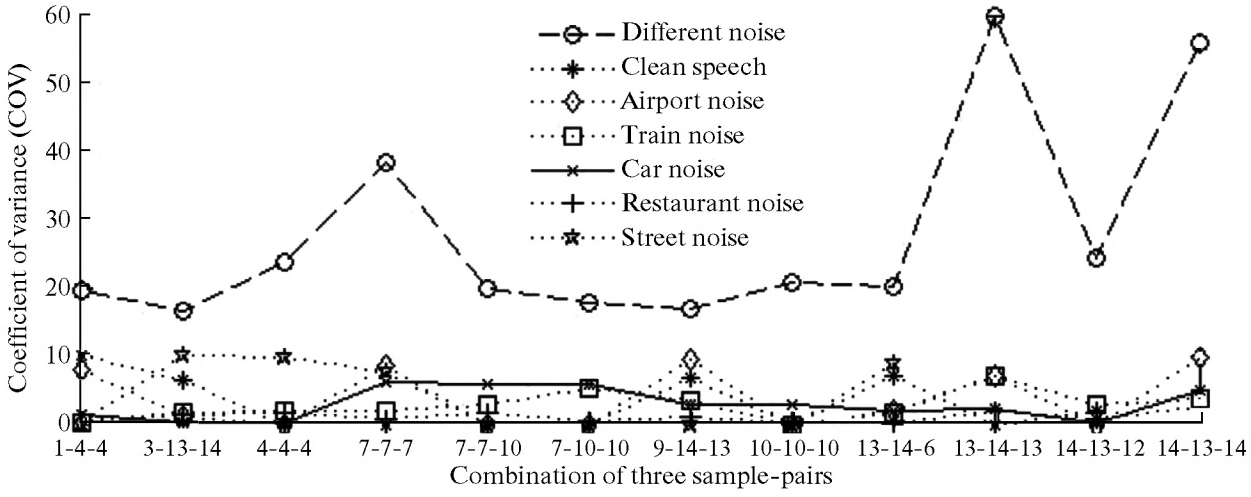

Fig. 10. COV of features based on three sample-pairs combination.

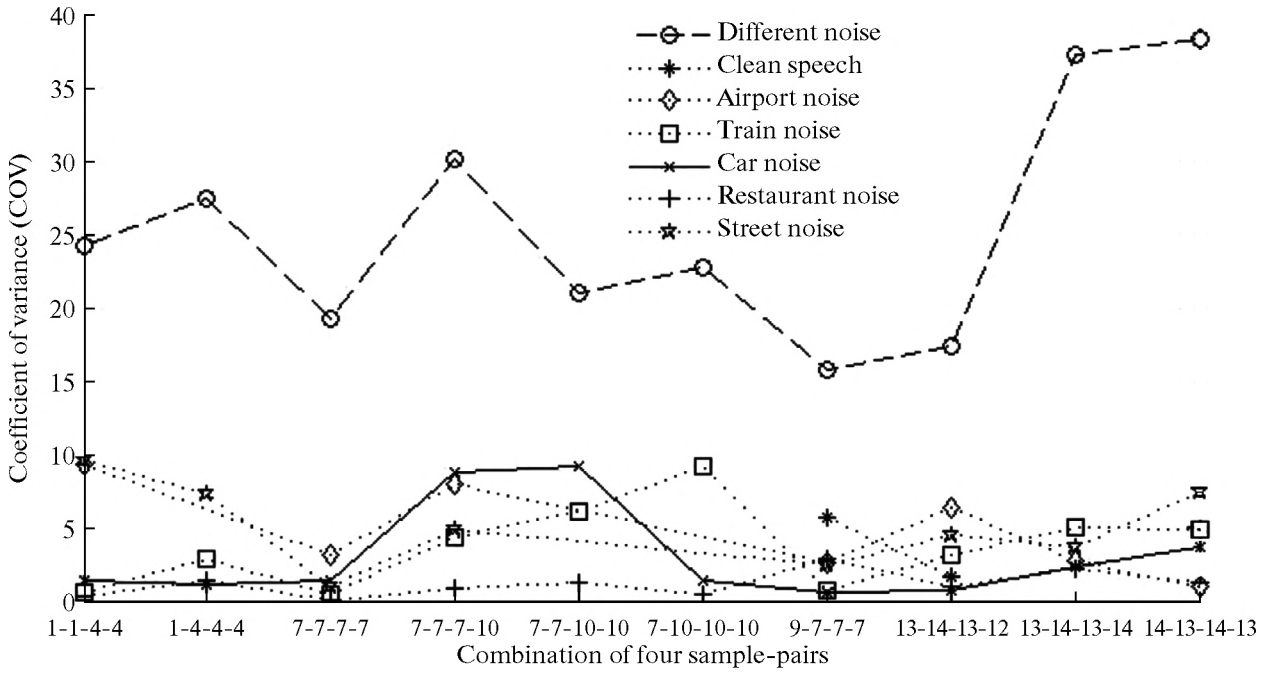

Fig. 11. COV of features based on four sample-pairs combination.

category. The pairs or combination of pairs are counted over common time-segment of first 22000 samples of corresponding discrete-time noisy speech signals. Only those features are reported here whose values are quite discriminating for the different type of background noise sources.

In Tables 3–6 the confusion matrices based on feature vectors FV_1, FV_2, FV_3 and FV_4 are shown. These are the results of one-against-all multiclass SVM classifier. All type of noise samples are applied to classifier on the basis of individual feature vector. Around 30 samples of each noise category are used for training purpose to train the classifier and more than 20 samples of each noise category are tested for discrimination. The segment of each noisy signal is of 2 s duration. The main outcome is that the train noise is very much different than any other type of noise considered here corresponding to any of the feature vector. The new feature vectors may be constituted as per (25–28) for better discrimination of noise sources applied for classification.

The experimental work is further supported by using different classifiers like k-NN ($k = 1$), k-NN ($k = 3$) and GMM for the same type of feature vector (FV_1) which have been used in case of SVM classifier. The performances of the classifiers in terms of classification accuracy are compared and shown in Table 7.

From Table 7 it is unambiguous that for this particular experiment GMM classifier is appropriate as far as overall performance is observed for discriminating any of the noise from rest of the noisy speech signals.

Table 2. Three and four sample-pairs based discriminating feature vectors for different noises

FV_3	$CP^3_{1,6,9}$	48	14	42	25
	$CP^3_{1,13,14}$	19	40	24	32
	$CP^3_{3,1,4}$	65	25	49	34
	$CP^3_{3,13,14}$	14	82	25	38
	$CP^3_{7,10,14}$	75	30	69	50
	$CP^3_{7,12,3}$	50	16	41	27
	$CP^3_{10,14,1}$	77	20	56	46
	$CP^3_{12,3,1}$	64	12	40	22
	$CP^3_{13,7,10}$	75	27	55	44
	$CP^3_{14,13,14}$	27	196	50	73
	$CP^3_{1,1,1,4}$	23	7	10	13
	$CP^3_{1,4,6,9}$	25	4	14	7
	$CP^3_{1,4,13,7}$	27	5	17	11
FV_4	$CP^4_{3,1,4,6}$	14	3	10	5
	$CP^4_{4,13,7,7}$	25	7	18	14
	$CP^4_{7,10,14,1}$	25	7	19	13
	$CP^4_{7,7,10,14}$	20	6	15	10
	$CP^4_{9,14,13,14}$	4	37	8	16
	$CP^4_{14,13,14,13}$	7	86	10	18

Table 3. Confusion matrix on the basis of FV_1

	Noise types	Predicted classes			
		airport	train	car	street
Actual Classes	Airport	72%	3%	16%	9%
	Train	1%	91%	5%	3%
	Car	8%	4%	81%	7%
	Street	8%	6%	9%	77%

Table 4. Confusion matrix on the basis of FV_2

	Noise types	Predicted classes			
		airport	train	car	street
Actual classes	Airport	73%	2%	13%	12%
	Train	1%	92%	3%	4%
	Car	6%	3%	70%	21%
	Street	6%	4%	18%	72%

Table 5. Confusion matrix on the basis of FV_3

	Noise types	Predicted classes			
		airport	train	car	street
Actual classes	Airport	65%	1%	28%	6%
	Train	2%	74%	4%	20%
	Car	22%	2%	68%	8%
	Street	2%	9%	23%	66%

Table 6. Confusion matrix on the basis of FV_4

	Noise types	Predicted classes			
		airport	train	car	street
Actual Classes	Airport	69%	3%	12%	16%
	Train	2%	84%	6%	8%
	Car	12%	4%	67%	17%
	Street	11%	4%	17%	68%

Table 7. Performance comparison of different classifiers

Type of classifier	Train noise from rest of the noisy signals	Car noise from rest of the noisy signals	Street noise from rest of the noisy signals
SVM	91.0%	81.0%	77.0%
k-NN, $k = 1$	82.0%	83.0%	70.0%
k-NN, $k = 3$	80.0%	73.0%	77.0%
GMM	69.0%	70.0%	65.0%

Table 8. Effect of change in noise level in terms of SNR on sample-pairs

Type of sample-pair	% change in the count of sample-pair corresponding to different SNR	
	train noise	airport noise
p_1 to p_3	Less than 2.0%	Less than 2.0%
p_4	Around 10.0%	Less than 2.0%
p_5 to p_6	Less than 2.0%	Less than 2.0%
p_7	10 to 12%	12 to 13%
p_8 to p_9	Less than 2.0%	Less than 2.0%
p_{10}	12 to 13%	Less than 2.0%
p_{11} to p_{14}	Less than 2.0%	Less than 2.0%

In Fig. 12 the count of sample-pair feature is extracted from noisy speech signals of airport noise and train noise with SNR of 5, 10 and 15 dB. It is shown that almost negligible change in the number of sample-pairs is observed corresponding to different values of SNR for both type of noise except a few type of sample pairs shown enclosed. In case of airport noise, the marginal change of around 12 to 13% in the number of sample pairs of p_7 type corresponding to 15 dB and 10 dB or 5 dB where as in case of train noise, a small change of same value is noticed in the number of sample-pairs of p_4, p_7 and p_{10} type corresponding to different SNRs.

A Table 8 is shown to clarify the effect of change in noise level in terms of SNR on sample-pairs. Count of only a few type of sample-pairs changes if SNR of the noisy signal varies in steps from 5 to 15 dB.

The final result is that the numbers of most of the type of sample-pairs are not affected by changing the level of noise added to the pure speech signal in terms of SNR. This is one of the important and desirable properties of the feature being used for discrimination purpose.

5. CONCLUSIONS

The aim is for developing a technique to discriminate different background noise signals by acoustic information only. This method may be applied to any type of acoustic background noise generated in industrial environment in order to predict or analyze the machine problems as well as to diagnose the diseases on the basis of background sound generated by different organs of body in medical field. The focus of the presented work is to study and investigate the utility of newly developed features for separation of background acoustic signals mixed with speech signals to be used along with one-against-all SVM classifier. The simulation results show that the classification based on FV_1 provides a recognition accuracy of 72% for airport noise and 81% for train noise. With the second, third and fourth feature vector FV_2, FV_3 and FV_4 the recognition accuracy are 92, 74 and 84% respectively by which the train noise can be distinguished from rest of all noise signals. The other noise signals are also discriminated with somewhat lower accuracy using same feature vectors. The performance of the features and its comparison with other traditional features shows its effectiveness and suitability for discrimination. In this methodology, the constitution and selection of feature vectors are important as they will vary depending upon the type of background noise to be discriminated. The results show that for certain categorization of noise signals, good recognition accuracy can be obtained by constituting new feature vectors.

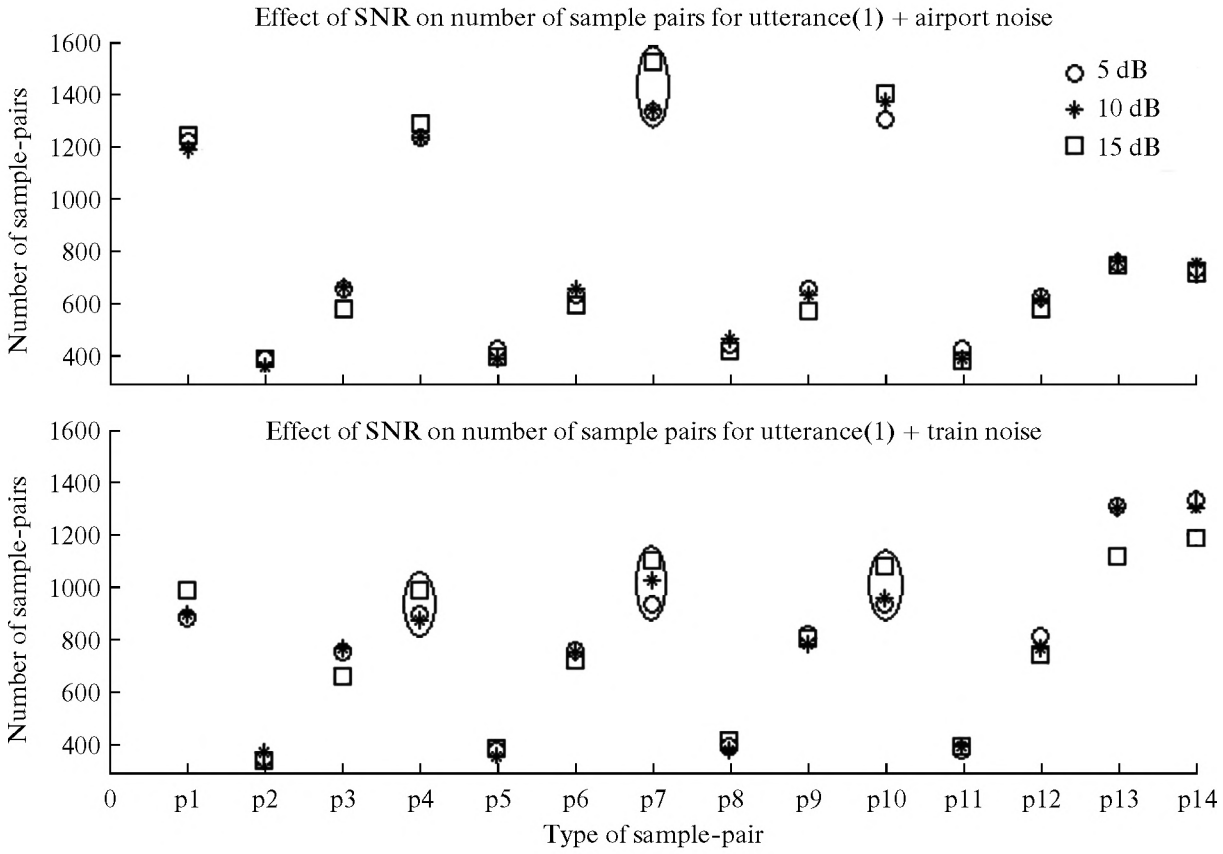

Fig. 12. Effect of SNR of the noisy signal on the number of sample-pairs.

The performance of the features and its comparison with the other traditional features shows its effectiveness and suitability for discrimination. It is further supported by the additional experiment which proves that the count of the sample-pairs is not much affected by changing the level of added background noise in terms of SNR. In the last part of the experiment, the different classifiers are tried for the same work. Their performance of classification is compared for common set of feature vectors and observed that for this particular work, the multiclass SVM has an upper edge over rest of the classifiers. A future vision is to construct an adaptive system capable of learning new environment and applying high-level knowledge in making the decisions.

REFERENCES

1. J. Pineau, M. Montemerlo, N. Roy, M. Pollack, and S. Thrun, Robot. Autonom. Syst. **42**, 271 (2003).
2. S. Thrun, M. Bennewitz, W. Burgard, A. B. Cremers, D. Fox, F. Dellaert, D. Haehnel, N. Roy, C. Rosenberg, J. Schulte, and D. Schulz, in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA), 1999.*
3. H. A. Yanco, *Lecture Notes in Artificial Intelligence: Assistive Technology and Artificial Intelligence* (Springer-Verlag, New York, 1998).
4. A. Fod, A. Howard, and M. J. Mataric, in *Proc. Int. Conf. Robot. Autom. (ICRA), 2002.*
5. S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, in *Proc. IEEE Int. Conf. Multimedia Expo (ICME), Toronto, Canada, 2006.*
6. J. Huang, in *Proc. IEEE Int. Conf. Multimedia Expo (ICME), 2002*, p. 253.
7. A. Waibel, H. Steusloff, and R. Stiefelhagen, in *Proc. Int. Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS, 2004.*
8. D. P. W. Ellis and K. Lee, in *Proc. Workshop on Continuous Archival and Retrieval of Personal Experiences, CARPE, 2004.*
9. J. Mantyjarvi, P. Huuskonen, and J. Himberg, J. IEEE Trans. Wireless Commun., **9**, 39 (2002).
10. T. Zhang and C.-C. Jay Kuo, J. IEEE Trans. Audio, Speech Lang. Proc. **9**, 441 (2001).
11. D. P. W. Ellis, *PhD Dissertation* (Cambridge, MA, 1996).
12. A. Eronen, V. Peltonen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, J. IEEE Trans. Audio, Speech Language Proc. **14**, 321 (2006).
13. R. G. Malkin and A. Waibel, in *Proc. Int. Conf. Audio, Speech and Language (ICASSP), 2005*, p. 509.

14. V. Peltonen, *MS Thesis*, (Tampere, Finland, 2001).
15. J.-J. Aucouturier, B. Defreville, and F. Pachet, J. Acoust. Soc. Am. **122**, 881 (2007).
16. R. Cai, L. Lu, A. Hanjalic, and H. Zhang, J. IEEE Trans. Audio, Speech and Language Proc. **14**, 1026 (2006).
17. A. Eronen and A. Klapuri, in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Proc. (ICASSP), Istanbull, Turkey, 2000*, p. 753.
18. Härmä and Aki, J. Am. Electr. Soc. **59**, 707 (2011).
19. Härmä and Aki, in *Proc. 45th Int. Conf: Appl. Time Freq. Proc. Audio, 2012.*
20. R. C. Maher and J. Studniarz, in *Proc. 46th Int. Conf: Audio Forensics, 2012.*
21. A. Gordienko, N. V. Krasnopistsev, V. N. Nekrasov, and V. N. Toropov, Acoust. Phys. **57**, 168 (2011).
22. V. M. Efimtsov and L. A. Lazarev, Acoust. Phys. **58**, 404 (2012).
23. A. S. Ivanenkov, A. A. Rodionov, and V. I. Turchin, Acoust. Phys. **59**, 179 (2013).
24. V. N. Vapnik, *The Nature of Statistical Learning Theory* (Springer-Verlag, New York, 1995).
25. Shigeo Abe, *Support Vector Machines for Pattern Classification* (Springer-Verlag, London, 2005).
26. J.Wolf, J. Acoust. Soc. Am. **51**, 2044 (1971).
27. G. Saha, S. Chakroborty, and S. Senapati, in *Proc. IEEE Ann. Conf. Indicon, 2004*, (2005), p. 70.
28. www.utdallas.edu/~loizou/speech/noizeus/