# Two-stage classification of respiratory sound patterns

Emin Çağatay Güler[a], Bülent Sankur[b], Yasemin P. Kahya[b,*], Sarunas Raudys[c]

[a]*Biomedical Engineering Institute, Boğaziçi University, Bebek, 34342 Istanbul, Turkey*
[b]*Electrical Engineering Department, Boğaziçi University, Bebek 34342 Istanbul, Turkey*
[c]*Institute of Mathematics and Informatics, Akademijos 4, 2600 Vinnius, Lithuania*

## Abstract

The classification problem of respiratory sound signals has been addressed by taking into account their cyclic nature, and a novel hierarchical decision fusion scheme based on the cooperation of classifiers has been developed. Respiratory signals from three different classes are partitioned into segments, which are later joined to form six different phases of the respiration cycle. Multilayer perceptron classifiers classify the parameterized segments from each phase and decision vectors obtained from different phases are combined using a nonlinear decision combination function to form a final decision on each subject. Furthermore a new regularization scheme is applied to the data to stabilize training and consultation.
© 2003 Elsevier Ltd. All rights reserved.

## 1. Introduction

Auscultation of the chest via a stethoscope provides useful information to the physician for the diagnoses of respiratory disorders. However, due to the subjectivity in auditory perception among physicians, and variability in their verbal descriptions of sound characteristics, fuzzy and qualitative nature of the diagnosis has become the major problem for this rewarding method [1–3]. In the last three decades, on the other hand, technical advances in sound measurement and signal processing techniques have opened new avenues for the auscultation based diagnosis of pulmonary disorders

* Corresponding author. Electrical Engineering Department, Boğaziçi University, Bebek, 34342 Istanbul, Turkey. Tel.: +90-212-358-1540x1851; fax: +90-212-287-2465.
*E-mail address:* kahya@boun.edu.tr (Y.P. Kahya).

[4–9]. Automatic recognition of respiratory sounds is useful in providing a computer-aided tool to auscultation and increases its potential diagnostic value [10].

There are two major difficulties in developing such a tool: (i) respiratory signals are nonstationary due to changes in lung volume and flow rate during a cycle. (ii) These sounds have a large inter-subject variability due to age, weight and physiology and considerable intra-subject differences due to the evolution state of pathology. Therefore the use of conventional classification algorithms prove inadequate, and novel approaches must take into account the problems of small sample size, diversity of sounds, and the cyclic behavior of signals.

There exist other biological signals of cyclic nature, other than respiratory sounds, such as, electromyogram signals in gait, pulsating blood flow signals, electrogastrogram, etc. These signals have in common the property that from cycle to cycle the waveforms can be assumed to be statistically identical. This means that, although the waveforms are not strictly periodic, the statistical characteristics of the process evolve in cycles. For example, if the changes of autoregressive parameters over a cycle are tracked, one can observe typical trends from the onset until the termination of the cycle. This follows from the nonstationarity of the signals, which impacts on the intra-subject variability.

In this study, we propose a new classification scheme that, on the one hand takes explicitly into account the cyclic nature of the lung sounds, and on the other hand, attempts to mitigate the feature overlaps and improve on the classification performance. The contribution of our work is the design of a novel classification scheme for cyclo-stationary signals and comparative assessment of various classification fusion methods.

Classification experiments are performed on a three-class respiratory database to test the proposed novel classification scheme considering respiratory sound signals as cyclic biological data. The respiratory pathologies taken into account are obstructive, and restrictive pulmonary diseases of which chronic bronchitis, emphysema and asthma are known as obstructive pulmonary disorders whereas fibrosing alveolitis, pneumonia, pleural diseases are the most common restrictive lung disorders. The third group used in the classification experiments is respiratory sound data collected from healthy subjects.

In the proposed classification scheme, respiratory sound signals are first partitioned into segments that are short enough to guarantee their stationarity but long enough to allow for reliable parameter estimation. A number of these segments are knotted together to form one of the six designated phases of a respiration cycle. For example, the initial inspiratory phase, defined as early inspiration when lung volume is still small but airflow rate is maximum, consists of ten observation segments. All the six phases, in order, form a whole respiratory cycle. This partitioning of the signal into segments and phases is advantageous because it (i) reduces the dimensionality of the feature space to a manageable level where "small sample size-high dimensionality" trade off [11,12] is overcome to some extent and mitigates, to some extent, the evolutionary nature of the sounds, and (ii) avails one of the possibility to zoom on the different sound production mechanisms governing each phase.

In summary we design a two-stage classifier, where in the first stage we concentrate on the time waveforms patterns in the segments, and in a second stage, on the six-dimensional phase decision patterns issued from the first stage. While the first stage does waveform classification, the second stage looks more like a consultation session among experts. In fact, we conceive each phase classifier as a separate expert proffering its opinion on the pathology. Then we fuse the expert opinions to reach a verdict for the whole respiratory cycle.
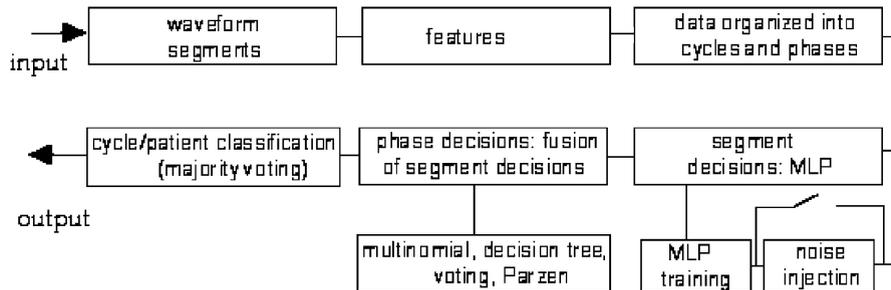
Fig. 1. Block diagram of the proposed classification method.

The steps of the proposed two-stage classification method [13–15] are given below while the block diagram of the proposed classification method is given in Fig. 1.

(1) Each segment is parametrized, that is, we extract parsimoniously linear signal model parameters from each segment. Segments of different phases are treated separately due to differing signal production mechanisms.
(2) We design separate multilayer perceptron classifiers, each for one of the six phases. Later in the testing stage, sound segments of each phase are classified using the corresponding phase classifier.
(3) A "decision feature vector" consisting of the combined segment decisions from the six phases is constructed.
(4) The final decision for each "decision vector" is obtained using a nonlinear decision combination function. A multinomial classifier, a decision tree classifier, simple voting algorithm and Parzen window classifier are the four algorithms used for the decision combination.

The proposed classification method can be equally well applied to other cyclic biological data, such as blood flow, ECG, EGG or to industrial sounds from rotating machinery, like shaft, drill sounds, or to data obtained from seasonal climatic, or oceanic measurements.

The paper is organized as follows. In Section 2, details of feature extraction, data organization and regularization methodologies employed are presented and discussed. First stage of classification, which is on classification of phases of a cycle, and various decision-making algorithms of the second stage classification used in this work are introduced in Section 3. Experimental results are given in Section 4. Conclusions are drawn in Section 5.

## 2. Preliminaries on feature extraction, data organization and regularization

### 2.1. Feature extraction

Signals in every cycle are first divided into a fixed number of segments. Therefore, each patient is represented by a single segmented cycle assuming that a cycle contains the total information requested for recognition. The segment lengths necessarily vary over cycles of patients, as the duration of each cycle is different. For respiratory sound cycles with typical durations of 1.5–2.5 s, the fixed number

of segments is chosen to be 60. The duration of segments varies typically in the range from 25 to 60 ms. These interval sizes represent a good compromise between parameter accuracy and stationarity requirements. Furthermore keeping the number of segments fixed makes the task of segment decision fusion easier.

Features are extracted from each segment of a cycle. AR (autoregressive) parameters and cepstral coefficients were chosen as features in the experiments with respiratory data. These parameters are obtained directly from linear prediction coefficients with the following recursive formula [16]:

$$c_S(1) = -a_S(1),$$

$$c_S(p) = \sum_{k=1}^{p-1} \left(1 - \frac{k}{p}\right) a_S(k) c_S(p-k) + a_S(p), \quad p = 1, \ldots, P, \tag{1}$$

where, for the $S$th segment, $c_S(p)$ and $a_S(p)$ are the $p$th cepstral parameter and linear prediction coefficient, respectively, $P$ is the order of the autoregressive (AR) model assumed for the $S$th segment. The cepstral parameters are evaluated from the AR coefficients using the nonlinear relationship in (1). Although the two feature sets are related, cepstral coefficients prove to be better for sound classification [16].

The AR model order, $P$, was chosen as six, as it was shown in [17] that such a model is adequate for classifying respiratory data. Consequently, each segment, $S = 1, 2, \ldots, 60$, of a cycle is represented either by a length six cepstral or AR feature vector, $\mathbf{f}_S$,

$$\mathbf{f}_S = [f_S(1) f_S(2) \ldots f_S(P)], \quad P = 6 \text{ (model order for a short time respiratory segment)},$$

$$f_S(p) = a_S(p) \text{ or } c_S(p). \tag{2}$$

Segmentation of a respiratory breath cycle and features extracted from segments are shown symbolically in Fig. 2.

## 2.2. Data organization: construction of the cycle feature set and its separation into phases

Feature vectors, $\mathbf{f}_S$, extracted from segments of respiratory cycles of patients are pooled into a feature set. The total number of feature vectors is the number of segments times the number of
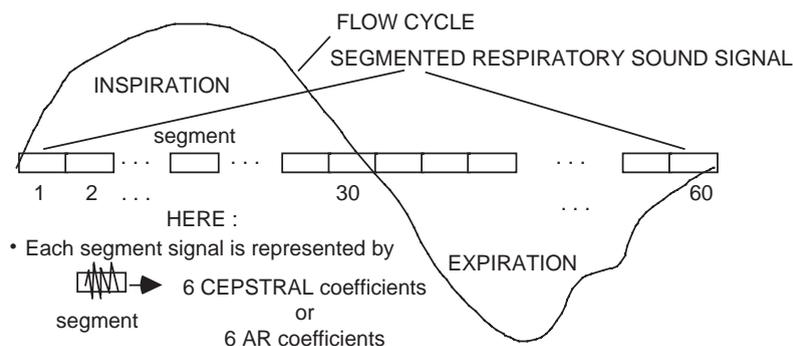


Fig. 2. Segmentation of a respiratory breath cycle and features extracted from segments.

patients in the respective classes. Thus our feature set consists of "60 (total number of segments) × 57 (total number of subjects)".

The 60 AR or Cepstral vectors in a cycle set are grouped into $\phi = 6$ phase sets, in consideration of the evolutionary nature of the cyclic data. This grouping is necessary since the class membership information in each phase may not be not identical due to differing sound production characteristics. The necessity to divide a cycle into phases rises also because of the nonstationary nature of the data within a cycle. Furthermore the use of a whole cycle in the classification process results in too large a dimensionality. It is difficult to design a classifier in such high-dimensional spaces, especially when the training and testing sample sizes are relatively small. The respiratory cycle feature set was separated into $\phi = $ six phases as (1) early expiration, (2) mid-expiration, (3) late expiration, (4) early inspiration, (5) mid-inspiration, (6) late inspiration.

Such a division of a respiratory cycle is also consistent with the auscultation terminology since these six phases are accepted as the most informative and distinctive parts of a respiratory sound signal. The duration of each phase is taken as equal, i.e., one-sixth of the whole respiratory cycle. Thus the total number of measured feature vectors available in a phase, e.g., in early inspiration, is "10 × total number of patients". After separation of a cycle feature set into phase feature sets, each phase is classified separately.

## 2.3. Regularization of the feature space

Due to the relative sparseness of the training data, it is not possible to populate adequately in a large dimensional feature space, and consequently the classifier can be adversely affected by the gaps. The test data may fall in those regions of the feature space for which no training samples were presented.

The training data set can be enriched by judicious noise injection, in other words, a set of $M$ new feature vectors, $\{\mathbf{f}_S^{\text{new}}, \ k = 1, \ldots, M\}$, can be generated for each feature vector, $\mathbf{f}_S$ by adding Gaussian noise with zero-mean, and variance $\lambda$, to its components:

$$\mathbf{f}_S^{\text{new}} = \mathbf{f}_S + \mathbf{u}_S, \quad k = 1, \ldots, M, \tag{3}$$

where $\mathbf{u}_S$ is a $1 \times P$ random vector taken from a normal distribution, $N(\mathbf{0}_{1 \times P}, \Lambda \mathbf{I}_{P \times P})$ with zero-mean, and covariance matrix $\Lambda \mathbf{I}$, here $\mathbf{I}$ is a $P \times P$ identity matrix. Consequently, while a certain phase data was represented by, say, $S$ feature vectors, after noise injection one obtains $S \times M$ feature vectors (Fig. 3). The enriched feature set constructed from the measured training set by noise injection is called the "regularized training set" in the sequel.

Training set or "learning set" is used to design classifier, while the regularized set is used for validation, that is, to compare the algorithms and to select the best information processing strategy. The justification for adding "noise to gain more information" lies in a space filling argument, that is, whenever the measurement space is not adequate to fill the multidimensional feature space. The classifier designed with a sparse set may strongly adapt itself to the training vectors by disregarding the gaps. One can therefore expect an improvement of the classification performance on the average with injection of a judicious amount of noise to the features. The noise fills the gaps between the training set vectors, and thus, in a way, regularizes the feature space. Such an enrichment of the

SUBPHASE VALIDATION SET :
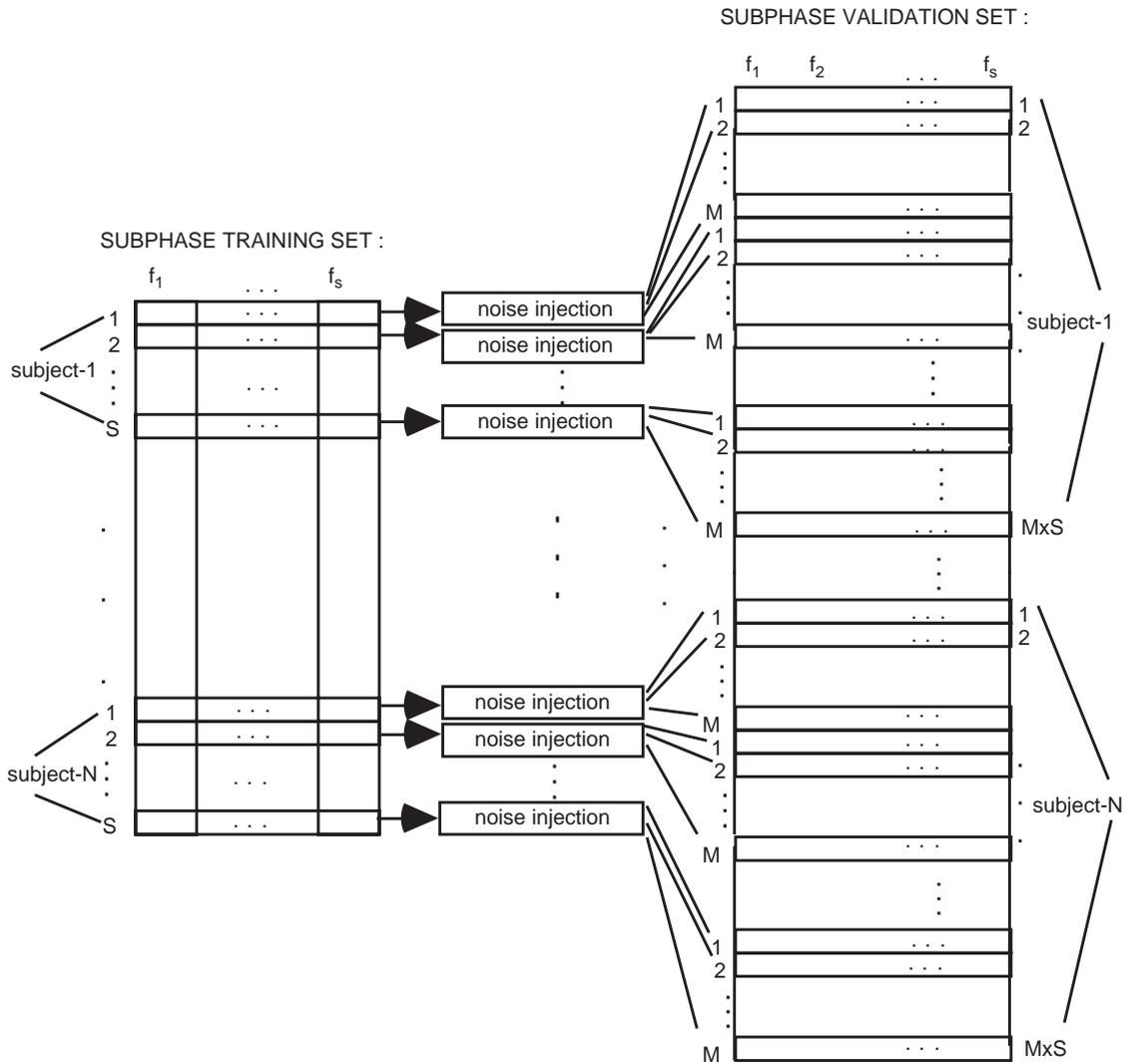
SUBPHASE TRAINING SET :

Fig. 3. Subphase regularized training set constructed from a subphase training set.

measurement set is not new, but appears in other guises such as Parzen's kernels, regularization in the training of neural networks, the ridge estimate of the class covariance matrices in regularized discriminant analysis. In principle, variance of noise could be chosen differently for each class and subphase. However, in our classification experiments, we generated $M = 20$ feature vectors for each given one, by adding zero-mean Gaussian noise with variance $\sigma^2 = 0.065$ to all phase feature sets. A two-dimensional feature space, composed of 40 vectors, and its appearance after addition of 20 vectors from a normal distribution $N(\mathbf{0}_{1\times2}, 0.0625\mathbf{I}_{2\times2})$ to each feature are shown in Fig. 4.
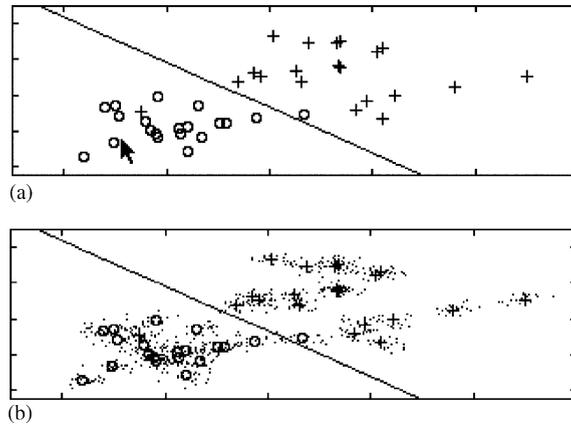
Fig. 4. (a) A two-dimensional feature space composed of two classes where each class has 20 feature vectors, (b) appearance after injection of 20 vectors from a normal distribution $N(\mathbf{0}_{1\times2}, 0.0625\mathbf{I}_{2\times2})$ to each feature vector in the space.
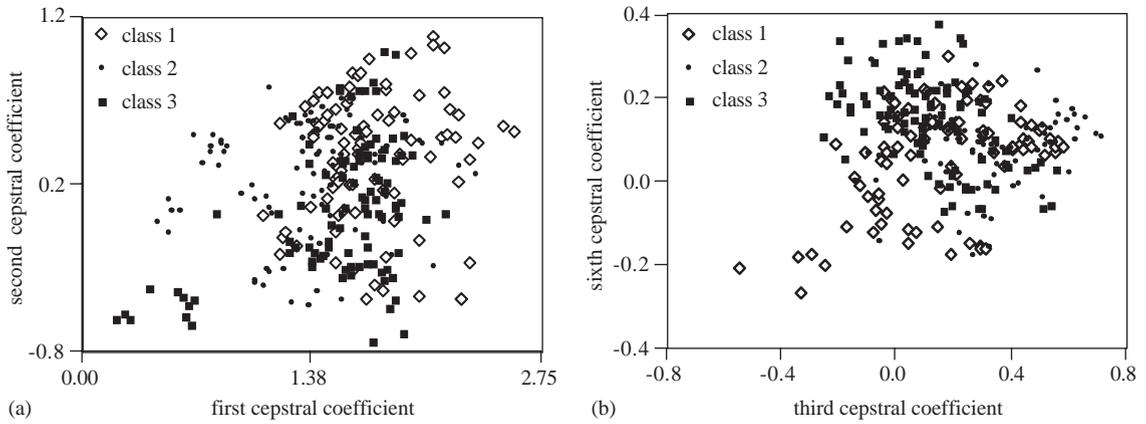


Fig. 5. Scatter diagrams of features from mid expiration training data: (a) $c_S(1)$ vs. $c_S(2)$, (b) $c_S(3)$ vs. $c_S(6)$.

## 3. Classification

First, classification is performed on individual segments using the respective phase classifier, for the three, respectively, training, regularized, and test sets. For each phase, a phase data classifier is trained to establish the three-class decision boundaries. In the following classification of the regularized set and the test set, these decision boundaries are identical.

In order to select the type of the classification algorithm to be used in the first stage of decision-making, an exploratory analysis of segment pattern vectors is performed. To give an insight into the nature of this three-class pattern recognition problem, scatter diagrams of two of the "best" pairs of the cepstral features from the mid-expiration phase are presented in Fig. 5. These scatter diagrams are more or less typical for all the data analyzed.

Scatter diagrams in Fig. 5 indicate that three classes overlap substantially, despite their separate clustering tendency. Thus there is strong evidence that a classifier is needed that allows obtaining nonlinear decision boundaries.

A nonlinear classifier can be obtained either using nonparametric statistical classifiers or resorting to an artificial neural network. We opted for the multilayer perceptron (MLP), the simplest and most popular neural network, as a phase expert is preferred [18–20]. Thus a separate MLP was designed for each phase.

Using different groupings of segment data, a $\phi$-component decision vector is obtained. More explicitly, the decisions from the $s$th segments ($s$ ranging from 1 to 10) from each phase are combined into a $\phi$-component decision pattern:

$$\mathbf{D}_S = [d_1^s \ldots d_\phi^s]. \tag{4}$$

Thus the consultation process, that is, merging of phase decisions, is executed for each of the segment collections. This scheme can be interpreted with the analogy of the phase experts expressing their opinion, and later in a "consultation" session merging them into a single decision for the patient. A "boss classifier" having the phase decision patterns as input enacts the consultation session.

There are various schemes to fusion the phase decisions, the simplest being the majority voting, and solving the ties by random choice. On the other extreme, the maximum number of bins, i.e., different decision patterns, which can occur for an $L$-class problem is $L\phi$, where $\phi$ is the number of experts and $L$ is the number of classes. In our specific problem, with $L=3$ classes and $\phi=6$ experts, there results $3^6=729$ bins. Clearly, one would need a huge data collection to estimate the conditional probability of bins for all the classes. The decision for $\mathbf{D}_S$ would correspond then to the most probable bin. This type of expert consultation will be denoted as the multinomial classifier in the sequel. In this multinomial classifier, the decision for any $\phi$-fold segment group $\mathbf{D}_S$ would correspond to the most probable bin over the classes. Thus using the class probabilities, $P_c$, $c=1,\ldots,L$, one estimates

$$\arg \max_c \{P_c P_l(c)\}, \quad c = 1, 2, \ldots, L, \tag{5}$$

$$P_l(c) = \frac{n_l(c)}{N_c} \quad \text{and} \quad \sum_{l=1}^{L^\phi} P_l(c) = 1, \tag{6}$$

where $P_l(c)$ are the class conditional bin probabilities, $n_l(c)$ and $N_c$ are the total number of $l$th bins and decision patterns, $\mathbf{D}_S$, existing in the $c$th class, respectively.

Due to the forbidding number of bins, the use of the multinomial classifier becomes impractical. In such cases, one might use the decision tree [21] or the nonparametric Parzen window classifier with a suitable distance metric. Note that in the small sample size cases, simpler classification rules are often better than the complex ones.

The two principle assumptions used to simplify the multinomial classifier are as follows:

- Many states belonging to one class can be joined together and described by smaller number of bins.
- Class membership of the bins with extremely small probabilities, $P_l(c)$, can be neglected.

To implement these simplifications Boolean algebra, fast algorithms for multivariate histograms, or decision tree classifiers can be used. The decision tree classifier is advantageous in its simplicity and tractability of the results, which is important in decision-making [21].

In the decision tree schema, the one-layer $L\phi$-branch tree is effectively substituted by a multilayer tree, where the complexity of each layer is considerably less. This simplification results in part by the partitioning of the bins between layers, and in part for a large number of bins, the class membership probability is negligible. In fact, as a result of pruning of branches and leaves, each problem might have a different tree architecture [21]. In other words, to make a solution concerning the class membership of the discrete decision vector, $\mathbf{D}_S$, the decision tree classifier uses only one element of $\mathbf{D}_S$ at a time and allocates vector $\mathbf{D}_S$ to one of the branches where it will be classified further (sent to a next branch), so that the complexity of the decision mechanism is simplified.

Parzen window classifier is another alternative consultation approach to multinomial scheme for the decision vector, $\mathbf{D}_S$. In multi-category case it performs the classification according to a maximum of the density estimate

$$g_c(\mathbf{D}_S) = P_c f(\mathbf{D}_S \mid c) = \frac{P_c}{N_c} \sum_{j=1}^{N_c} \kappa(D(\mathbf{D}_S, \mathbf{D}_{j,c}), \lambda^2),\tag{7}$$

where $\kappa(D(\mathbf{D}_S, \mathbf{D}_{j,c}), \lambda^2)$ is a discrete kernel function which determines the contribution of a single training vector from the $c$th class, $\mathbf{D}_{j,c}$, while classifying an unknown vector, $\mathbf{D}_S$, and $\lambda$ controls the degree of the contribution in accordance with distance between $\mathbf{D}_S$ and $\mathbf{D}_{j,c}$. In this work, the following kernel and distance measure are used:

$$\kappa(D) = \lambda^{1 - D(\mathbf{D}_S, \mathbf{D}_{j,c})}(1 - \lambda)^{D(\mathbf{D}_S, \mathbf{D}_{j,c})},\tag{8}$$

where $D(\mathbf{D}_S, \mathbf{D}_{j,c})$ is the distance between $\mathbf{D}_S$ and $\mathbf{D}_{j,c}$, e.g., the number of disagreements between these two vectors.

Finally, in all of the three consultation schemes discussed above, both the training and test patterns, $\mathbf{D}_S$, consist of the decision patterns obtained by classifying segments of the $\phi$ phases by the respective classifiers. Recall that in our case, "segments" are represented either by the cepstral or AR features. On the other hand, when regularized data is used, the training data is enriched $M$-fold, while the test data is still the original measured data (not enriched).

## 4. Experimental results and discussion

In this section classification results for the 57 respiratory sound signals are presented. Measurement records from 18 chronic obstructive patients (class-1), 19 restrictive lung disease patients (class-2) and 20 healthy subjects (class-3) are analyzed.

In the measurement procedure, two air-coupled electret microphones were used to record respiratory sounds from the basilars. The signal was amplified with a low noise amplifier and bandpass-filtered between 80 to 2000 Hz. The high pass section which filters out heart sounds and frictional noise was a sixth-order Bessel filter with almost linear phase characteristics so that crackles, which may be present in the signal, are minimally distorted. Similarly eight order Butterworth filter with flat passband characteristics was used as anti-aliasing filter. The gain of the amplifier-filter unit was 60 dB. The signal was sampled with a 12-bit analog to digital converter at 5000 sampling rate and inputted to the computer. In order to adjust the flow rate such that a minimum of 1 l/s was achieved and to synchronize on the inspiration–expiration cycles, the flow signal was also recorded by a Fleisch type flowmeter and digitized with the same ADC. Recordings from each site were 12.8 s in

Table 1
Segment misclassification probabilities of the six phase training sets, and test sets using cepstral and AR features

|  | Phase 1 | | Phase 2 | | Phase 3 | | Phase 4 | | Phase 5 | | Phase 6 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Feature set | Train | Test | Train | Test | Train | Test | Train | Test | Train | Test | Train | Test |
| Cepstral | 0.32 | 0.43 | 0.31 | 0.43 | 0.34 | 0.41 | 0.31 | 0.37 | 0.23 | 0.28 | 0.27 | 0.33 |
| AR | 0.38 | 0.52 | 0.38 | 0.45 | 0.40 | 0.43 | 0.33 | 0.38 | 0.25 | 0.29 | 0.36 | 0.45 |

duration, a period covering 3–4 respiration cycles. After each recording, the quality of the recorded sound was monitored via the headphones and by the computer against ambient noise. The recordings were done in the presence of a pulmonary physician in the hospital environment [17].

Recorded signal in a respiration cycle is divided into a fixed number of segments as illustrated in Fig. 2. The number of segments is chosen to be 60 for a whole respiration cycle. Each segment is characterized by six cepstral and six AR coefficients forming two feature sets, which are classified separately and comparatively. These segments in each cycle are further partitioned into six phases, namely, early, mid, late inspiration/expiration phases, each group consisting of ten consecutive segments. The segments belonging to each phase are further split into a training and a test set. More explicitly, every other segment of a phase is assigned to the training set, while the remaining segments of the same phase are used in the test stage. The choice of such interleaving test and training sets is made purely as a proving ground for any improvements that the proposed scheme could bring.

A separate multilayer perceptron classifier is designed for each phase feature set. After several preliminary classification experiments, six hidden units were chosen for each expert. In order to find the weights, a standard backpropagation MLP training algorithm is used. Results of the segment classification of the training, and test sets, using AR and cepstral coefficients are presented in Table 1.

As documented in Table 1, the classification performances of phase experts on the individual segments are rather mediocre (0.37 and 0.42 error on the average for cepstral and AR coefficients, respectively). Their performance is boosted up by the strategy of using a second stage decision on the pattern of segment decisions, which can be thought as "a consultation of experts". Furthermore, since the six cepstral features result in best classification performances for the test sets, decision feature vectors used in the decision combination algorithms are extracted from the length six cepstral vectors only.

Each decision vector, $\mathbf{D}_S$, has six elements each of which can take the value either 1 or 2 or 3, i.e., the class membership decision given by a phase expert, and a class decision is reached by one of the consultation schemes for each $\mathbf{D}_S$. In each phase there results five segment decisions for each subject (since the other five alternating segments are reserved for testing). The decisions on segments in different phases but with the same sequence number, i.e., segment-s of phase-1, segment-s of phase-2, etc., are fused via consultation. This reduces the votes for a cycle from $5 \times 6 = 30$ to just five. Finally, the classification of a cycle is based on majority voting of the consultation decisions on each of the five "fused" segments. The consultation actually takes place via a classifier operating on the pattern of class decisions from each segment. This procedure is applied five times to each of the five $\mathbf{D}_S$'s resulting from every cycle of respiration.

Table 2
Average segment (decision vector) classification errors using various consultation schemes

| Multinomial | | Decision tree | | Parzen window | | Voting | |
|---|---|---|---|---|---|---|---|
| Train | Test | Train | Test | Train | Test | Train | Test |
| 0.028 | 0.340 | 0.077 | 0.242 | 0.147 | 0.179 | 0.165 | 0.218 |
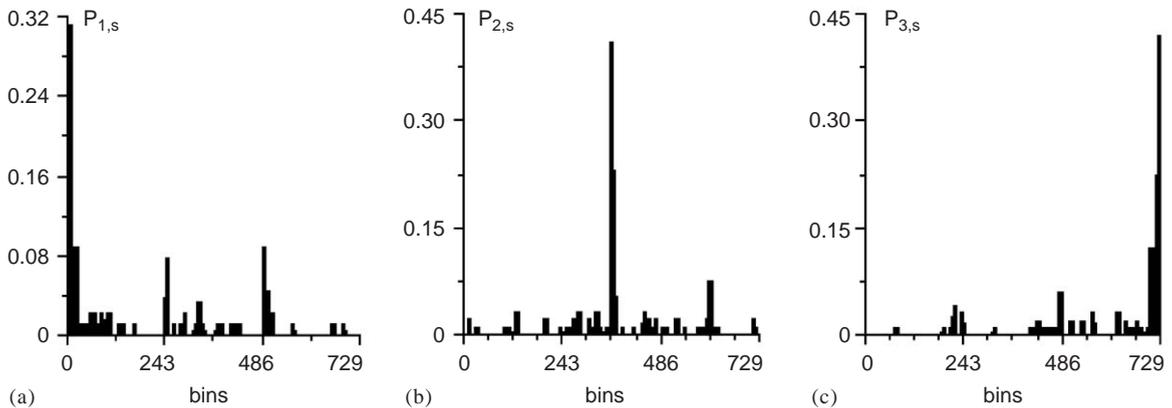


Fig. 6. Histogram probabilities: (a) $P_{1,s}$, (b) $P_{2,s}$, (c) $P_{3,s}$, of classes according to the bins, $s = 1, 2, \ldots, m = 729$.

The four decision fusion algorithms, namely, the multinomial classifier ($M$), the decision tree classifier (DT), Parzen window classifier ($P$), voting ($V$) schemes, have been compared. The improvement in the classification performance of segments (decision patterns, $\mathbf{D}_S$) by fusing expert decisions is shown in Table 2. It can be observed that all fusion schemes bring significant improvements but to varying degrees. The segment classification error of phase experts on the test sets is at an average of 0.37 (Table 1), while after the consultation, segment classification error drops down to 0.34, 0.24, 0.18, 0.22 for the $M$, DT, $P$, and $V$ schemes, respectively.

However one notices, also, a large discrepancy between training and test performances of consultation schemes. The discrepancy is most severe for the $M$ case and can be explained on the basis of the sparseness of the data. In fact the $M$ scheme has 729 states, but is trained with a set of 285 vectors overall. This means that even if all bins taken by decision vectors are different from each other, class membership probabilities of $729-285 = 444$ states are zero. Furthermore, class membership probabilities of many of the states are extremely small in this application. This can be easily noticed when the histogram probabilities and probability distributions of the three classes are observed (Fig. 6). As can be noticed from Table 2, $P$ and DT schemes give error rates of 0.18 and 0.24 on the test set, respectively, while this value is 0.34 for the $M$ scheme. In other words, as such, the multinomial decision fusion turns out to be useless. Thus simplification of the $M$ type of decision fusion is justified. Simpler consultation schemes are more robust in that they do not get overtrained. In Table 2, the DT had 49 final leaves (instead of 729 bins, thus the data is significantly compressed), and the $P$ window parameter is $\lambda = 1.6$.

A second alternative for the small sample size problem is the enrichment of the feature space by noise injection to the training data, i.e., the construction of a regularized training set. In other words, since feature data are sparse with respect to the dimensionality of the feature space, data space is enriched by adding noise to features. To this purpose, several runs of independent and identically distributed Gaussian noise, with zero mean and variance $\sigma^2$, is added to each training set vector. If $\mathbf{f}$ denotes a feature vector, it is replicated $M$ times ($M$ was taken 20) with the addition of noise to its components. A classifier designed with a sparse set may adapt itself too strongly to the training vectors due to the gaps. The added noise fills the gaps between the training vectors, and thus, in a way, regularizes and stabilizes the feature space. Noise injection to the feature space is similar to Parzen kernels, in that each measured feature points is interpreted as signaling other potential vectors in its neighborhood. Noise is added to features in each phase, with an empirical variance, $\sigma^2$. The "regularized training decision set" constructed from 20 runs of independent Gaussian noises with $\sigma^2 = 0.0645$ injected into the features of six phase training sets resulted in the best $M$ performance for the "test decision set". Results of the segment classification of the regularized training sets, using AR and cepstral coefficients are presented in Table 3 to give an insight.

The optimum value of the standard deviation of the injected noise is determined by comparing the performances of the multinomial classifiers designed based on the regularized training decision sets with various noise having different standard deviation values. Changes in the error rates of the multinomial classifiers with different standard deviation values for the regularized training and test decision sets are depicted in Fig. 7. As can be noticed, the value of the error for the test

Table 3
Segment misclassification probabilities of the six phase regularized training sets using cepstral and AR features

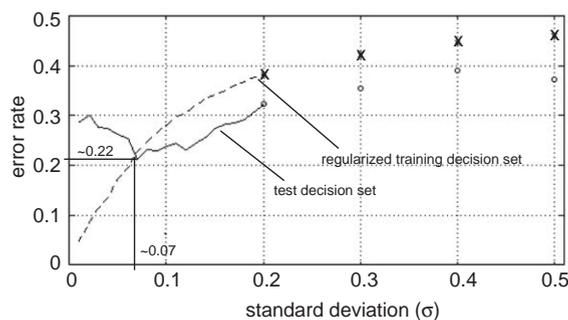|  | Phase 1 | Phase 2 | Phase 3 | Phase 4 | Phase 5 | Phase 6 |
|---|---|---|---|---|---|---|
| Feature set | Regularized training | Regularized training | Regularized training | Regularized training | Regularized training | Regularized training |
| Cepstral | 0.43 | 0.44 | 0.44 | 0.40 | 0.32 | 0.37 |
| AR | 0.49 | 0.48 | 0.51 | 0.41 | 0.38 | 0.43 |



Fig. 7. The change of the multinomial error rate with the standard deviation of the injected zero mean noise to the validation and test sets data.
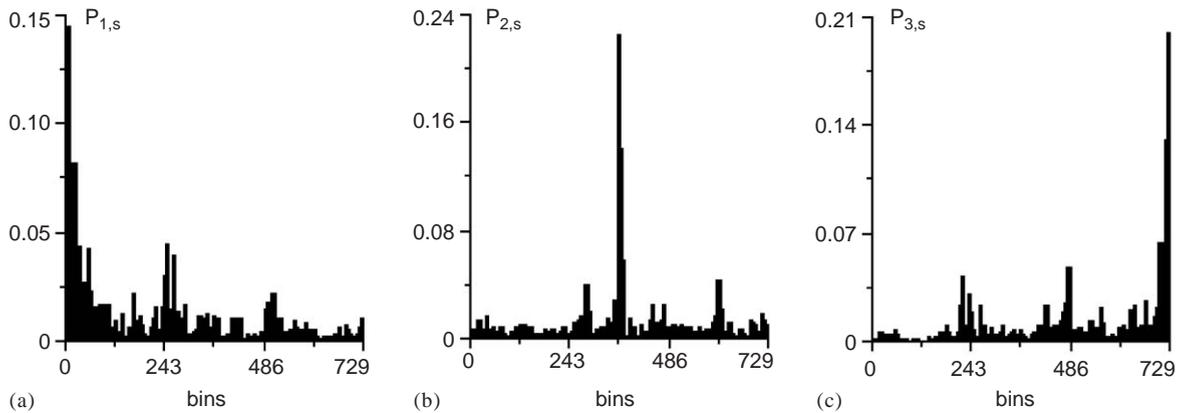
Fig. 8. Histogram probabilities, (a) $P_{1,s}$, (b) $P_{2,s}$, (c) $P_{3,s}$, of classes according to the bins, $s = 1, 2, \ldots, m = 729$, for the regularized training set.

Table 4
"Test decision set" segment classification errors for the multinomial and decision tree schemes designed from the regularized training decision set

| Multinomial | Decision tree |
| --- | --- |
| 0.218 | 0.196 |

set becomes minimum (0.22) when the standard deviation of the injected noise is 0.0645. Fig. 8 shows the histogram probabilities of the three classes in the regularized training decision sets. One can easily notice from Figs. 6 and 8 that the decision feature space is filled adequately by noise injection and that most of the bins have values, which are considerably greater than zero. Note also that, the total number of segments in each phase training set increases from 57 (total number of subjects)× 5(number of segments representing a subject in each phase)=285 to 57(total number of subjects)× 5(number of segments representing a subject in each phase)× 20(total number of noisy segments)=5700 in each phase regularized training set.

In Table 4, "test decision set" segment classification errors of $M$ and DT designed with the regularized training set are presented. Note from Tables 2 and 4 that the use of the regularized training set instead of the actual training set results 0.12 and 0.046 improvements in the performances of $M$ and DT consultations, respectively. Thus with noise injection both the multinomial and decision tree classifiers improve to a level comparable to that of the Parzen scheme.

The third stage of classification is in fact to reach a class decision for the whole cycle therefore for the subject, itself. This is achieved by merging the five consultation decisions resulting from the five $\mathbf{D}_S$'s of a cycle via majority logic. The misclassification probabilities of subjects based on majority voting of the $M$, DT, $P$ and $V$ consultation decisions on $\mathbf{D}_S$'s are presented in Table 5.

Subject misclassification probabilities of the $M$ and DT schemes designed with the regularized training set are also given in Table 5. Table 5 states that the performance of the $P$ scheme is the best with 0.14 error rate, and the DT classifier with 0.16 error rate follows $P$.

Table 5

(a) Subject classification errors based on majority voting of various consultation decisions

| Multinomial | Decision tree | Parzen window | Voting |
|---|---|---|---|
| 0.333 | 0.158 | 0.140 | 0.175 |

(b) Subject classification errors (again with majority logic) of the multinomial and the decision tree classifiers designed using the regularized training set

| Multinomial | Decision tree |
|---|---|
| 0.140 | 0.140 |

Table 6
Classification results in terms of probability of misclassification obtained from 5, 4, and 3-expert decision vectors. Numbers in parentheses show the subject classification performance based on majority voting of decision vectors

| Classification with | Multinomial (train/test) | Decision tree (train/test) | Parzen window (train/test) | Voting (train/test) |
|---|---|---|---|---|
| 5 Experts | 0.05/0.22 (0.02/0.13) | 0.10/0.24 (0.09/0.16) | 0.07/0.18 (0.02/0.12) | 0.16/0.22 (0.09/0.14) |
| 4 Experts | 0.08/0.27 (0.04/0.18) | 0.11/0.24 (0.05/0.21) | 0.09/0.22 (0.04/17.54) | 0.18/0.24 (0.11/0.13) |
| 3 Experts | 0.15/0.27 (0.02/0.18) | 0.16/0.26 (0/0.11) | 0.16/0.26 (0.07/0.19) | 0.19/0.27 (0.11/0.18) |

The performance of the DT is 0.017 better than $V$ scheme in subject classification, while $V$ was 0.024 better compared to DT in segment classification (Table 2). On the other hand, the use of the regularized training set, i.e., noise injection to the training space, in the design of $M$ and DT schemes is once more justified: Subject classification performances of $M$ and DT are improved by 0.193 and 0.018, respectively.

One of the possible solutions to overcome the "small sample size-complexity of the classifier trade-off" is to decrease the number of experts that consult on respiratory cycles. This can be achieved by neglecting the phase experts that give the lowest classification performance on the training and regularized training sets. To this purpose, classification experiments are repeated with five, four, and three experts by leaving out the third (late inspiration), first (early inspiration) and third, first, second (mid inspiration) and third experts, respectively (see Table 1), in the decision-making procedures. Classification results of the decision-making algorithms on decision vectors and further subject classification performances based on voting are given in Table 6.

It can be observed from Table 6 that the performance of the $M$ classifier on the test set in the case of smaller number of experts is much better than the six experts case both in the sense of vector classification and subject classification. On the other hand five-expert case gives the best performance with the $M$ classifier compared to three- and four-expert cases. This shows that the choice of five

best experts out of six is a good compromise between the complexity of the classifier and the sample size ($3^5 = 243$ bins to classify 285 decision vectors). On the other hand, performance of the $P$ decreases with the decreasing number of experts. The optimum value of the smoothing parameter $\lambda$ is found to be 0.6 for five, four, and three experts. The decrease in the performance may be due to the kernel and distance measure used to calculate the maximum of the density estimate, which cannot handle well the contribution of each single training vector while classifying an unknown vector in the compressed versions of feature spaces. DT classifiers used with six, five, four, and three experts result approximately in the same classification performances on decision vectors. This shows that the simplification of the $M$ classifier in the case of varying number of experts is possible to some extent for this data. It is interesting to note that the DT designed using three experts gives the worst performance (0.26) on the decision vectors, but the best performance (0.11) in the further classification of subjects via voting. Therefore the simplification of the $M$ by the use of the DT and the use of the smaller number of experts is beneficial to organize the decision vectors of a subject in a reasonable way. These experiments show that either $M$, or DT, or $P$ classifiers overcome simple voting traditionally used in the cooperation of classifiers, but simplification of the $M$ case in some reasonable way is necessary.

In summary, for the MLP classifier with six hidden neurons designed for the whole breath cycle, the use of one expert resulted in the classification error of 0.42 for the segments, and 0.33 for subjects, while the use of the cooperation of phase experts resulted in the classification error of 0.2–0.3 for segments, and 0.1–0.2 for subjects in the classification experiments. The multinomial classifier, the decision tree classifier, the Parzen window classifier, and simple voting are used as alternate decision combination algorithms. It has been shown that the "small learning set size-complexity of the multinomial classifier tradeoff" can be overcome to some extent with the use of (1) the Parzen window approach (2) the decision tree classifier, (3) regularization, and (4) smaller number of phase experts.

## 5. Conclusion

In conclusion, a new hierarchical decision-making scheme based on the cooperation of neural networks to classify respiratory sound patterns has been proposed in this paper. To this purpose each breathing cycle is divided into phases, and a separate MLP classifier, which is called, a phase expert is used for each of them. Phase decisions are then combined via a decision combination function. Furthermore a novel regularization scheme is applied to the data to stabilize training and decision-making. The MLP classifier with six hidden neurons designed for whole breath cycles, e.g., the use of one expert, resulted in the correct classification performance of 58% for the segments, and 67% for subjects, while the use of the cooperation of phase experts resulted in the correct classification of 70–80% for segments, and 80–90% for subjects.

The multinomial classifier, the decision tree classifier, the Parzen window classifier, and simple voting are used as alternate decision combination algorithms. It has been shown that the "small learning set size-complexity of the multinomial classifier tradeoff" can be overcome to some extent with the use of (1) the Parzen window approach (2) the decision tree classifier, (3) regularization, and (4) smaller number of phase experts.

## Acknowledgements

This project was sponsored by Boğaziçi University Research Fund.

## References

[1] H.C. Smyllie, L.M. Blendis, P. Armitage, Observer disagreement in physical signs of the respiratory system, Lancet (1965) 412–414.

[2] R.L. Wilkins, J.R. Dexter, R.L.H. Murphy, E.A. DelBono, Lung sound nomenclature survey, Chest 98 (1990) 886–889.

[3] D.W. Cugell, Lung sounds: classification and controversies, Semin. Respir. Med. 6 (1985) 180–182.

[4] H.S. Pasterkamp, S. Kraman, G.R. Wodicka, Respiratory sounds-advances beyond the stethoscope, Am. J. Respir. Crit. Care Med. 156 (1997) 974–987.

[5] R.P. Malmberg, K. Kallio, S. Haltsonen, T. Katila, A.R.A. Sovijarvi, Classification of lung sounds in patients with asthma, emphysema, fibrosing alveolitis and healthy lungs by using self-organizing maps, Clin. Physiol. 16 (1996) 115–129.

[6] A. Cohen, D. Landsberg, Analysis and automatic classification of breath sounds, IEEE Trans. BME 31 (1984) 585–590.

[7] J. Vanderschoot, H.J.W. Schreur, Flow and volume related AR-modeling of lung sounds, IEEE Trans. BME 13 (1991) 385–386.

[8] N. Gavriely, M. Nissan, A.H.E. Rubin, D.W. Cugell, Spectral characteristics of chest wall breath sounds in normal subjects, Thorax 50 (1995) 1292–1300.

[9] K. Anderson, S. Aitken, J.E. MacLeod, F. Moran, Lung sound transmission and reliability of the chest signs, Lancet 2 (1988) 228–229.

[10] A.R.A. Sovijarvi, J. Vandershoot, J.E. Earis (Eds.), Computerized Sound Analysis (CORSA): Recommended Standards for Terms and Techniques-ERS Task Force Report. European Respiratory Review, Vol. 10, ERS Journals Ltd, UK, 2000, Review No. 77.

[11] K. Fukunaga, Effects of small sample size in classifier design, IEEE Trans. Pattern Anal. Mach. Intell. 11 (1989) 873–883.

[12] S. Raudys, A.K. Jain, Small sample size effects in statistical pattern recognition: recommendations for practitioners, IEEE Trans. Pattern Anal. Mach. Intell. 13 (1991) 252–264.

[13] F. Fogelman-Soulie, E. Vinnet, B. Lamy, Multimodular neural network architectures: application in optical character and human face recognition, Int. J. Pattern Recog. Artif. Intell. 7 (1993) 721–755.

[14] T.K. Ho, J.J. Hull, S.N. Srihari, On multiple classifier systems for pattern recognition, Proceedings of the 11th IAPR International Conference on Pattern Recognition, Vol. 2, The Hague, The Netherlands, 1992, pp. 84–87.

[15] L.K. Hansen, P. Salamon, Neural network ensembles, IEEE Trans. Pattern Anal. Mach. Intell. 12 (1990) 993–1001.

[16] W.J. Kang, J.R. Shin, C.K. Cheng, J.S. Lai, H.W. Tsa, T.S. Kuo, The application of cepstral coefficient and maximum-likelihood method in EMG pattern recognition, IEEE Trans. BME 42 (1995) 777–785.

[17] B. Sankur, Y.P. Kahya, E.Ç. Güler, T. Engin, Comparison of AR based algorithms for respiratory sounds classification, Comput. Biol. Med. 23 (1994) 67–76.

[18] R.P. Lippmann, An introduction to computing with neural nets, IEEE ASSP Magazine, April, 1987, pp. 4–22.

[19] R. Hecht-Nielsen, Neurocomputing, Addison-Wesley, Reading, MA, 1989.

[20] R.P. Lippmann, Pattern classification using neural networks, IEEE Commun. Mag. 27 (1989) 47–64.

[21] S.R. Safavian, D. Landgrebe, A survey of decision tree classifier methodology, IEEE Trans. Systems Man Cybern. 21 (1991) 660–774.

**Emin Çağatay Güler** received his B.S. degree in Electronics and Communication Engineering from Istanbul Technical University, Istanbul, in 1989 and his M.S. and Ph.D. degrees in Biomedical Engineering from Boğaziçi University, Istanbul in 1992 and in 1998, respectively. He has worked as an R&D engineer at Arçelik A.Ş., which is a leading European household appliances company, and taken a patent (United States Patent: 6,634,191) with the application of several signal

processing and detection methods for determining the amount and type of a laundry in a washing machine. Dr. Güler currently working as a vice president at Sensormatic Electronics Security Company, Istanbul. His research interests are adaptive signal processing, spectral estimation, pattern recognition, wavelet transform theory and their applications in biomedical engineering and industry.

**Bülent Sankur** has received his B.S. degree in Electrical Engineering at Robert College, Istanbul and completed his M.Sc. and Ph.D. degrees at Rensselaer Polytechnic Institute, New York, USA. He has been active at Boğaziçi University in the Department of Electric and Electronic Engineering in establishing curricula and laboratories, and guiding research in the areas of digital signal processing, image and video compression and multimedia systems. His current research interests are biometric systems, watermarking, functional imaging and biological shape. He was the chairman of the ITC'96: International Telecommunications Conference and the technical co-chairman of ICASSP'2000. Dr. Sankur has held visiting positions at the University of Ottawa, Canada, Istanbul Technical University, Technical University of Delft, The Netherlands, and Ecole Nationale Supérieure des Telecommunications, France.

**Yasemin P. Kahya** received the B.S. degrees in Electrical Engineering and in Physics at Boğaziçi University in 1980, M.S. degree in Engineering and Applied Science at Yale University in 1980 and Ph.D. degree in Biomedical Engineering at Boğaziçi University in 1987. She has been teaching ever since in the Department of Electric and Electronic Engineering at Boğaziçi University. Her current research interests are in the areas of biomedical instrumentation, respiratory acoustics and biomedical signal processing applications. She was the program co-chair of IEEE-EMBS 2001 Conference which was held in Istanbul.

**Sarunas Raudys** received his Ph.D. degree in Technical Sciences from Institute of Mathematics and Cybernetics of Academy of Sciences of Lithuania in 1979. He has been with the Vilnius Technical University, has been a consultant to industry and has held visiting positions at Michigan State University, Baylor University, Delft University of Technology, Energy Research Center of Netherlands and University of Pierre et Marie Curie. His research interests are in the areas of statistical pattern recognition, artificial neural networks and data analysis.