Audio Signal Feature Extraction and Classification Using Local Discriminant Bases

Karthikeyan Umapathy, *Student Member, IEEE*, Sridhar Krishnan, *Senior Member, IEEE*, and Raveendra K. Rao, *Senior Member, IEEE*

***Abstract*—Audio feature extraction plays an important role in analyzing and characterizing audio content. Auditory scene analysis, content-based retrieval, indexing, and fingerprinting of audio are few of the applications that require efficient feature extraction. The key to extract strong features that characterize the complex nature of audio signals is to identify their discriminatory subspaces. In this paper, we propose an audio feature extraction and a multigroup classification scheme that focuses on identifying discriminatory time-frequency subspaces using the local discriminant bases (LDB) technique. Two dissimilarity measures were used in the process of selecting the LDB nodes and extracting features from them. The extracted features were then fed to a linear discriminant analysis-based classifier for a three-level hierarchical classification of audio signals into ten classes. In the first level, the audio signals were grouped into artificial and natural sounds. Each of the first level groups were subdivided to form the second level groups viz. instrumental, automobile, human, and nonhuman sounds. The third level was formed by subdividing the four groups of the second level into the final ten groups (drums, flute, piano, aircraft, helicopter, male, female, animals, birds and insects). A database of 213 audio signals were used in this study and an average classification accuracy of 83% for the first level (113 artificial and 100 natural sounds), 92% for the second level (73 instrumental and 40 automobile sounds; 40 human and 60 nonhuman sounds), and 89% for the third level (27 drums, 15 flute, and 31 piano sounds; 23 aircraft and 17 helicopter sounds; 20 male and 20 female speech; 20 animals, 20 birds and 20 insects sounds) were achieved. In addition to the above, a separate classification was also performed combining the LDB features with the mel-frequency cepstral coefficients. The average classification accuracies achieved using the combined features were 91% for the first level, 99% for the second level, and 95% for the third level.**

***Index Terms*—Audio classification, dissimilarity measures, feature extraction, linear discriminant analysis (LDA), local discriminant bases (LDB), wavelet packets.**

I. Introduction

AUDIO feature extraction and classification techniques have been addressed by many existing works over the years. The general methodology of audio classification involves extracting discriminatory features from the audio data and feeding them to a pattern classifier. Different approaches and various kinds of audio features were proposed with varying success rates. Audio feature extraction serves as the basis for a wide range of applications in the areas of speech processing [1], multimedia data management and distribution [2]–[5], security [6], biometrics, and bioacoustics [7]. The features can be extracted either directly from the time domain signal or from a transformation domain depending upon the choice of the signal analysis approach. Some of the audio features that have been successfully used for audio classification include mel-frequency cepstral coefficients (MFCCs) [4], [5], spectral similarity [8], timbral texture [5], band periodicity [2], linear prediction coefficient derived cepstral coefficients (LPCCs) [9], zero crossing rate [2], [9], MPEG-7 descriptors [10], entropy [11], and octaves [3]. Few techniques generate a pattern from the features and use it for classification by the degree of correlation. Few other techniques use the numerical values of the features coupled to statistical classification methods.

While most of the above-mentioned works focus on music and speech signals, there are applications which need audio processing and feature extraction on a wider range and categories of audio signals. Auditory scene analysis [12] for environment noise/sound detection is one such important application. The following are few examples where environment detection (by audio) plays an important role in our day-to-day life: people with hearing disability depend on assistive devices such as hearing aids to listen to the sounds around them. It is very important for these assistive devices to determine the environment using the auditory clues in order to build better instruments with automatic switching features [13]. This would improve the quality of life of people with disability. Some practical situations would be in adaptively changing the noise reduction strategy depending on the noise environments [14], [15], alerting the hearing-impaired listener when a fast approaching automobile is detected, and audio source localization for navigation. Audio source localization and environment detection also finds application in robot navigation [16]. Another interesting application of audio environment detection would be in wildlife conservatories. Animal, bird, and insect sounds can be used to automatically keep track of the wildlife population and their movement patterns. By using audio-triggered gadgets, it would be possible to understand and create unique audio patterns representing the life style of various species. Audio environment detection can also form as an vital educational tool (audio tutor) particularly for children. Audio environments can be synthesized and, in combination with visuals, could serve as an effective tool in teaching children about different environments without physically being there. In order to perform applications such as the one mentioned above, efficient audio feature extraction and classification schemes covering wider cat-

Manuscript received October 1, 2005; revised May 30, 2006. This work was supported the Natural Sciences and Engineering Research Council (NSERC). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dilek Hakkani-Tur.

K. Umapathy and R. K. Rao are with the Department of Electrical and Computer Engineering, University of Western Ontario, London, ON N6A 5B8, Canada (e-mail: karthi@ee.ryerson.ca).

S. Krishnan is with the Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada (e-mail: krishnan@ee.ryerson.ca).

Digital Object Identifier 10.1109/TASL.2006.885921

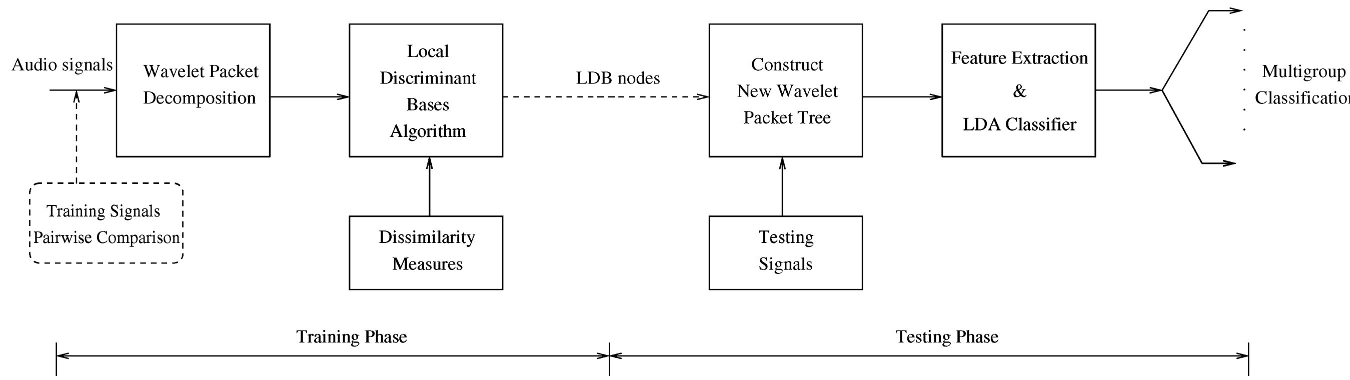

Fig. 1. Block diagram of the proposed technique.

egories of audio signals are required. To achieve this, the proposed methodology is an attempt to develop local discriminant bases (LDBs)-based automated multigroup audio classification system. The proposed methodology addresses only the sound source (type) identification part of an auditory scene analysis and not the complete auditory analysis by itself.

Audio signals are highly nonstationary in nature, and the best way to analyze them is to use a joint time-frequency (TF) approach. The previous works [3], [11] of the authors have demonstrated the success of the TF approach in music signal classification. In order to perform efficient TF analysis on the signals for feature extraction and classification purposes, it is essential to locate the subspaces on the TF plane that demonstrate high discrimination between different classes of the signals. Once the target subspaces are identified, it is easier to extract relevant features for classifications. In the proposed work, we used LDB [17] with wavelet packet bases to identify these target subspaces in the TF plane to classify the audio signals. The optimal choice of LDBs depends on the nature of the dataset and the dissimilarity measures used to distinguish between and among classes. A combination of multiple dissimilarity measures can be used to achieve high classification accuracies. The proposed work follows the nonstationary signal analysis approach similar to our previous work [3]; however, compared to our previous work, the proposed work focuses on developing a generic methodology for automatic identification of discriminatory subspaces and uses simple features to classify a wider range of audio signals. The proposed work will also be computationally less expensive as the orthonormal wavelet packet basis sets are much smaller than the overcomplete and redundant TF dictionaries used in our previous study.

The block diagram of the proposed technique is shown in Fig. 1. An audio database with 213 audio signals consisting of 113 artificial and 100 natural sounds were used in this study. The sampling frequency of all the signals were set uniformly to 44.1 kHz at a resolution of 16 bits/sample. The artificial sounds consisted of instrumental sounds (drums, flute, and piano) and automobile sounds (aircraft and helicopter). The natural sounds consisted of human (male and female speech) and nonhumans (animals, birds, and insects). A subset of signals from each of the four second level audio classes (instruments, automobile, humans, and nonhumans) were used as the training set. The audio classes were compared by taking two classes at a time (all possible pairwise combinations). The training signals for each of the class within the pair of classes were decomposed into wavelet packet trees. The corresponding nodes of the trees were compared using a set of dissimilarity measures to identify the nodes that exhibit high discriminative values between the classes. The nodes exhibiting high average discriminatory value among all the pairwise combination over a number of trials were selected as the final LDB nodes. The process can be repeated with different wavelets, and the wavelet that exhibits overall better discrimination between classes can be chosen as the best wavelet basis. Having selected the best wavelet basis and the significant LDB nodes, a new wavelet packet tree was constructed. All the signals from the database (including the training signals) were then decomposed using this new wavelet packet tree. Features were extracted from the LDB node basis vector coefficients followed by classification using a linear discriminant analysis (LDA)-based classifier. A hierarchical classification was performed starting from the first level two groups to third level ten subgroups (individual classification of each class). Fig. 2 shows the hierarchical classification of audio classes. The number in the brackets after the class labels indicate the number of signals in each of the classes used in this study. A comparative analysis of the proposed technique with MFCC-based classification scheme is also presented. This paper is organized as follows. Section II covers the methodology comprising of LDB, dissimilarity measures, LDB selection process, audio database, feature extraction, and pattern classification. Results and discussions are covered in Section III, and a conclusion is presented in Section IV.

II. Methodology

A. LDB Algorithm

The LDB [17] algorithm is a pruning algorithm that identifies the subspaces that exhibit high discrimination between signal classes using a given dissimilarity measure. LDB has been successfully applied to address real world classification problems in the areas of robotics, radar, biomedical, multimedia, and image processing [18]–[23]. The optimal choice of LDB subspaces for a given dataset is driven by the nature of the dataset and the dissimilarity measures [24] used to distinguish between classes. The dissimilarity measures indirectly control the classification accuracy achieved. When dealing with complex datasets, it is often required to project the dataset on to

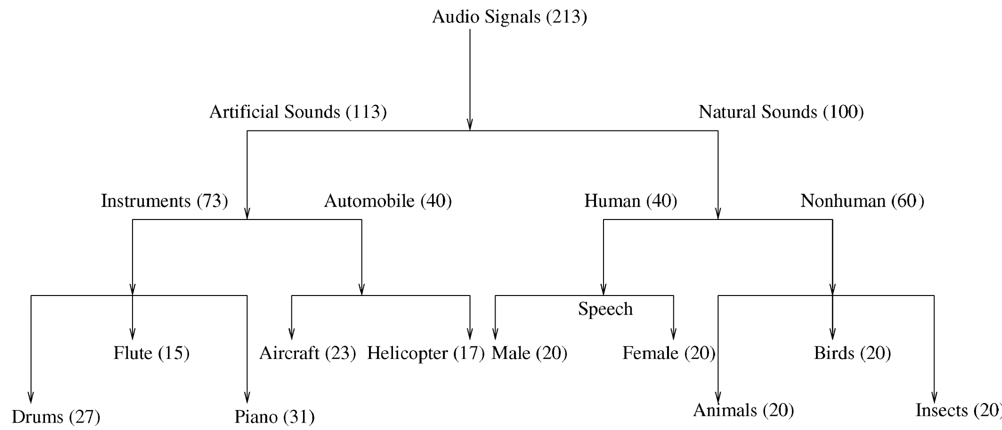

Fig. 2. Proposed hierarchical classification tree.

a multidimensional feature space to achieve efficient classification. Similarly, a combination of multiple dissimilarity measures with varying complexity could be used to select the LDB with different characteristics to achieve high classification accuracies.

In the LDB algorithm with wavelet packet bases, a set of training signals for all C classes are decomposed into full wavelet packet trees of order N. Let each of the signals in the training set be denoted as x_i^c, where the index i and c corresponds to the ith training signal in cth class. In this study, for ease of computation we restricted our analysis to binary wavelet packets trees [25].

Let $\Omega_{0,0}$ be a vector space R^n corresponding to the node 0 of the parent tree. Then at each level the vector space is split into two mutually orthogonal subspaces given by

$$\Omega_{j,k} = \Omega_{j+1,2k} \oplus \Omega_{j+1,2k+1} \quad (1)$$

where j indicates the level of the tree, and k represents the node index in level j, given by $k = 0, \ldots, 2^j - 1$. This process is repeated until level J, giving rise to 2^J mutually orthogonal subspaces. Our goal was to select the set of best subspaces that provided maximum dissimilarity information between the different classes of the signal. Each subspace $\Omega_{j,k}$ is spanned by a set of basis vectors $\{\mathbf{w}_{j,k,l}\}_{l=0}^{l=2^{u-j}-1}$, where 2^u corresponds to the length of the signal. The vector $\mathbf{w}_{j,k,l}$ represents the wavelet packet basis function that could be of length 2^j, oscillates with a center frequency in the kth frequency band, and translates over l locations in steps of 2^j.

The signal x_i can then be expressed as

$$x_i = \sum_{j,k,l} [\alpha_{j,k,l}]_i . \mathbf{w}_{j,k,l} \quad (2)$$

where in the index (j,k,l), (j,k) corresponds to the terminal (leaf) nodes and $\alpha_{j,k,l}$ are the basis vector coefficients at (j,k). The class label c is dropped from the signal x_i^c for notational convenience. In other words, the signal x_i is decomposed into 2^J subspaces with $\alpha_{j,k,l}$ coefficients in each subspace.

With the training signals decomposed into wavelet packet coefficients, we need to define a dissimilarity measure so as to identify those subspaces which had larger statistical distance between classes. In the LDB technique proposed in [17], relative entropy was used as a dissimilarity measure in identifying the LDB. The wavelet packet tree was pruned in such a way that a node was split if the cumulative discriminative measure of the children nodes were greater than that of the parent node. In other words, a node was split only if the children nodes had better discriminative power than that of the parent node. At the end of this iterative process, the tree structure contained only those terminal nodes, which contributed to maximizing the distance between different classes. Once the discriminatory nodes are identified, all the l wavelet packet coefficients from each of the terminal node of the pruned tree could be used as features. However, in order to reduce the feature dimension, a small set of dominant coefficients $(g \ll l)$ can be used as features or using the l wavelet packet coefficients a new set of feature(s) can be created by applying some mathematical operation.

B. Dissimilarity Measures

In the proposed method, we used a modified LDB approach. Instead of using a single dissimilarity measure, we used multiple dissimilarity measures to arrive at the final tree structure. Using multiple dissimilarity measures provides additional feature dimensions for classification. Especially for complex data sets such as audio signals, a single dissimilarity measure may not be able to capture all the characteristic information about its class.

Almost all existing literature on LDB deals with dissimilarity measures from a information theory point of view, i.e., the measures basically provide some numerical values without having a real physical meaning that could be related to some signal characteristics. Extracting nonabstract and meaningful features is essential in understanding the physical phenomenon that alters the signal characteristics according to the environment or conditions.

In this paper, we introduced meaningful dissimilarity measures in the process of identifying the discriminant subspaces. The choice of dissimilarity measures originated from the following facts that the main characteristics of any signal is embedded in its: 1) envelope or amplitude modulated component; 2) energy distribution over frequencies; and 3) time-varying dynamics (typically related to nonstationarity).

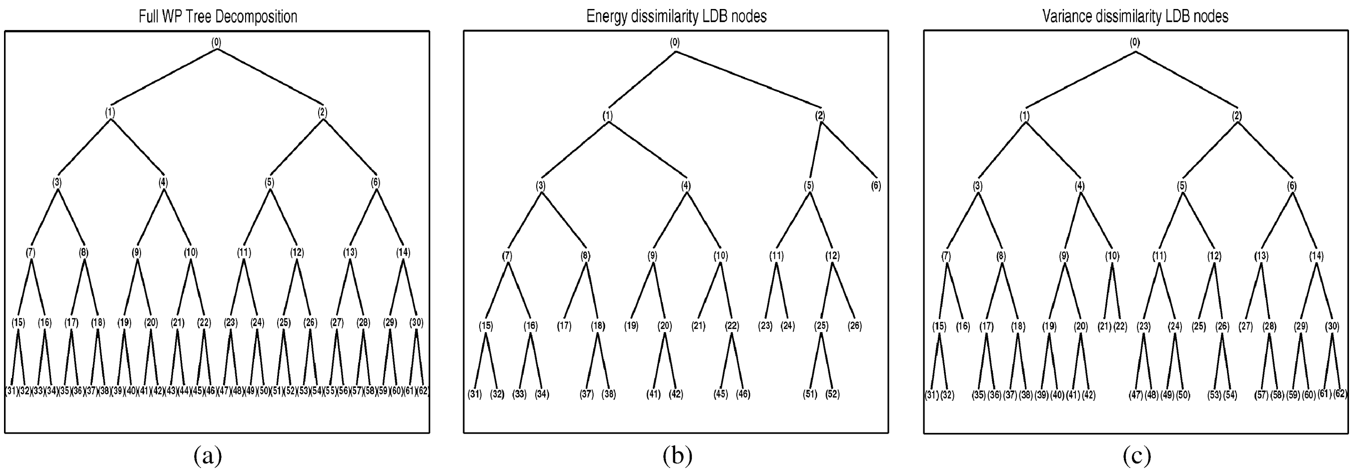

Fig. 3. (a) Sample five-level full wavelet packet tree. (b) Sample LDB tree obtained using the dissimilarity measure D_1 and db4 wavelet. (c) Sample LDB tree obtained using the dissimilarity measure D_2 and db4 wavelet.

Using the above rationale, we designed three dissimilarity measures to measure the temporal (related to envelope) characteristics, energy distribution over frequencies, and nonstationarity index. While the temporal dissimilarity measure, which was basically a normalized correlation index between the corresponding node coefficients is feasible for performing two group comparisons, it is not a computationally feasible measure for a multigroup classification such as the one presented in the study. Hence, in this study, we included only the dissimilarity measure for energy distribution over frequencies (D_1) and a measure of nonstationarity index (D_2).

The first dissimilarity measure D_1 is defined as the difference in the normalized energy between the corresponding nodes of the training signals from one of the $\binom{P}{2}$ combination of classes. P is the number of classes, and operation $\binom{P}{2}$ means that from a given number of P classes there would be $\binom{P}{2}$ number of possible combinations, if we choose to take two classes at a time from P classes. This measure is expected to reveal the energy concentration locations on the TF plane for different types of audio signals

$$D_1 = E_{j,k}^1 - E_{j,k}^2 \tag{3}$$

where $E_{j,k}^1$ and $E_{j,k}^2$ are the normalized energy of the corresponding nodes for one of the $\binom{P}{2}$ combination of signals. Fig. 3(b) shows a sample LDB tree obtained using the dissimilarity measure D_1.

The dissimilarity measure D_2 is a measure of estimating the nonstationarity of the basis vector coefficients. It is computed as the set of variances along the segments of the basis vector coefficients. Each node contains 2^{u-j} basis vector coefficients. These coefficients were segmented into L equal segments and variance for each of these L segments were computed to get a set of variances. The dissimilarity measure D_2 was then computed as the variance of the set of variances. The ratio of this variance measure between the signals from each of the $\binom{P}{2}$ combination of classes indicate the amount of deviation observed in the nonstationarity between the classes. The variability in time-varying signal structures could be represented using this dissimilarity measure D_2

$$D_2 = \frac{\mathrm{var}\left[\mathbf{v}_{(j,k)}^1\right]}{\mathrm{var}\left[\mathbf{v}_{(j,k)}^2\right]} \tag{4}$$

where $\mathbf{v}^1$ and $\mathbf{v}^2$ are vectors of length L, containing the variances of L equal segments obtained by segmenting the basis vector coefficients at node (j,k) for one of the $\binom{P}{2}$ combination of classes. An L value of 10 was used in the proposed study. The value of L decides the length of the segment that is used to compute the variance sequence. For example, considering the down sampling by two at each level, the length of the wavelet packet coefficients at the fifth level (terminal nodes) for a signal of length M would be approximately $M/(2^5)$. So the segment length for the fifth-level terminal nodes used for computing the variance sequence would be $M/(32 * L)$ samples. Depending upon the length and characteristics of a signal, the value of L could be chosen so that the segment used for variance computation is neither too short nor too long. Fig. 3(c) shows a sample LDB tree obtained using the dissimilarity measure D_2.

The dissimilarity measure D_2 is also nonadditive in nature in contrary to the conventional additive LDB dissimilarity measures. For fast pruning of the wavelet packet trees, one of the required condition is that the dissimilarity measure should be additive. If not additive, a union operation of the dissimilarity values corresponding to the child nodes is necessary to obtain a comparable measure with the parent node. This requirement of additivity limits the choice of the dissimilarity measures. The goal of pruning or selective splitting of the nodes is to avoid redundancy in the selection of the final LDBs. In order to use a nonadditive dissimilarity measure and at the same time avoid redundancy in the choice of the LDBs, in the proposed study we do not perform selective splitting of the nodes but use all the nodes in the process of the LDB selection. Later, in the LDA portion, the redundancy within the final set of LDBs were removed in the feature evaluation process.

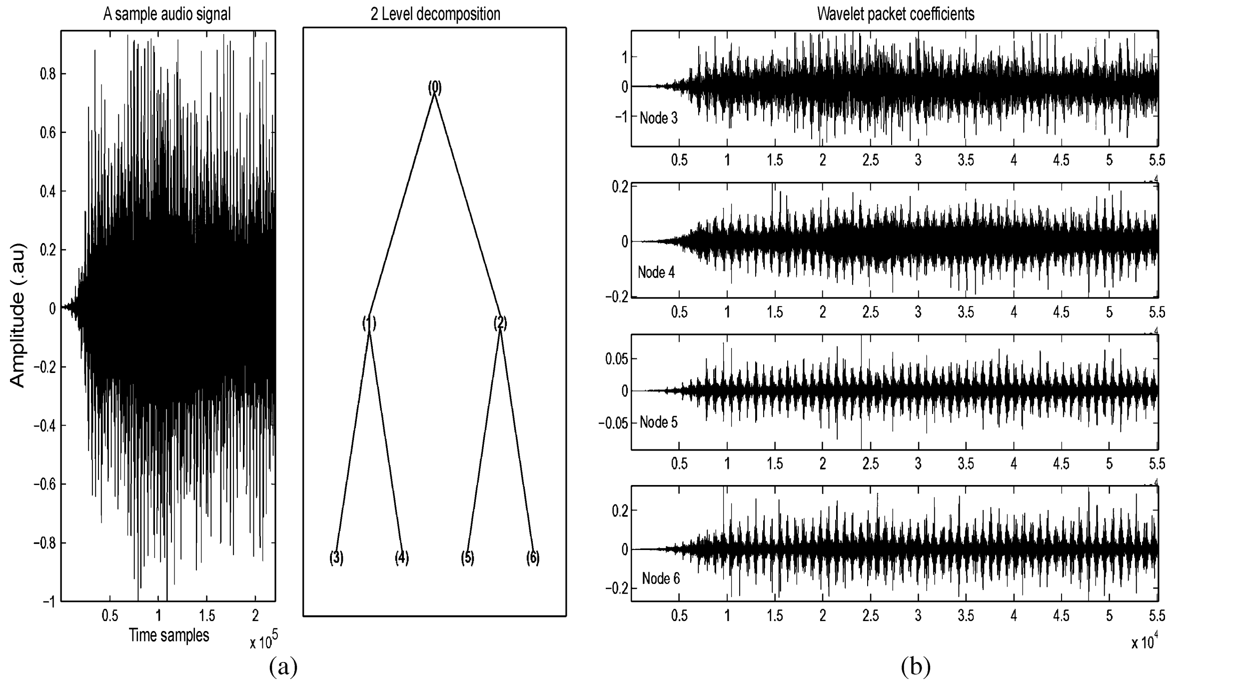

Fig. 4. (a) Sample audio signal and a two-level wavelet packet tree. (.au)—arbitrary units. (b) Wavelet packet coefficients at the four terminal nodes (3, 4, 5, 6)

C. LDB Selection Process

The goal is to identify those nodes from the full wavelet packet tree which demonstrate high discriminatory values between all the classes for a given dissimilarity measure D_n. If there are P classes, then the dissimilarity measure is computed by taking two classes at a time, i.e., $\binom{P}{2}$ combinations. Let $\beta = 1, 2, \ldots, \binom{P}{2}$ denote the index of $\binom{P}{2}$ combinations. In our case, we had a $P = 4$ classes (instrumental, automobile, human, and nonhuman sounds). For each of the $\binom{P}{2}$ combinations (taking in pairs such as instrumental and automobile, instrumental and human, instrumental, and nonhuman and so on), one audio signal for each of the classes in a combination was selected randomly and decomposed using a full wavelet packet tree.

The decomposition was performed following the (2). Fig. 4 shows a sample two-level wavelet packet decomposition of an audio signal. The left panel shows the audio signal in time domain and its two-level wavelet packet tree. The right panel shows the wavelet packet coefficients at the four terminal nodes. Once both the audio signals from each of $\binom{P}{2}$ combinations were decomposed into full wavelet packet trees, the corresponding nodes of these trees were compared using the dissimilarity measure D_n. The discriminatory values obtained were stored for each of the nodes. In our case, we used a five-level wavelet packet decomposition, so leaving the parent node 0, we would have $z = (1, 2, \ldots, 62)$ nodes in total.

Likewise, we performed T trials for each of the $\binom{P}{2}$ combinations. In each trial, audio signals for the combinations were randomly chosen from the database. In the proposed work, we used 15 ($t = 1, 2, \ldots, T$ where $T = 15$) trial runs, i.e., 15 pairs of audio signals for each of the $\binom{P}{2}$ combinations were used. The trials were chosen to be 15 using the following rational. From Fig. 2, we could observe that the largest split at any level was into three groups, i.e., 1) drums, flute, and piano and 2) animals, birds, and insects. In order to distribute the third-level groups evenly in their parent training groups (instrumental, automobile, human, and nonhuman) and at the same time not using up all the signals of the smallest class (1) male and female, 2) aircraft and helicopter) for training, we arrived at 15 trials. Fifteen trials will work out to be five signals per class in the three group category and seven to eight signals per class in the two group category. The idea was to make sure that the LDB training groups (instrumental, automobile, human, and nonhuman sounds) contained a balanced population of its subclasses, and at the same time all the four groups have equal number of signals for pairwise comparison. The more number of trials, the better would be the average performance of the selected LDB nodes. However, depending on the classification accuracy, database size, database distribution, and processing time requirements, the number of trials can be chosen. In our case, at the end of all trials, we would have a matrix of 62 discriminatory values over 15 trials for each of the $\binom{P}{2}$ combinations for a particular dissimilarity measure D_n. Let us denote the matrix as $\boldsymbol{Cl}_{(\beta,n)}$

$$\boldsymbol{Cl}_{(\beta,n)} = \left\{ \begin{array}{cccc} D_n^{(z=1,t=1)} & D_n^{(z=1,t=2)} & .\ . & D_n^{(z=1,t=T)} \\ D_n^{(z=2,t=1)} & D_n^{(z=2,t=2)} & .\ . & D_n^{(z=2,t=T)} \\ . & . & .\ . & . \\ . & . & .\ . & . \\ D_n^{(z=62,t=1)} & D_n^{(z=62,t=2)} & .\ . & D_n^{(z=62,t=T)} \end{array} \right\}_{\beta} . \tag{5}$$

Each row in the above matrix corresponds to the discriminatory values of a particular node over all the trials. Each column corresponds to the discriminatory values of a particular trial over all nodes. In order to identify if a particular node was discriminatory and consistent in performance, the following procedure was performed.

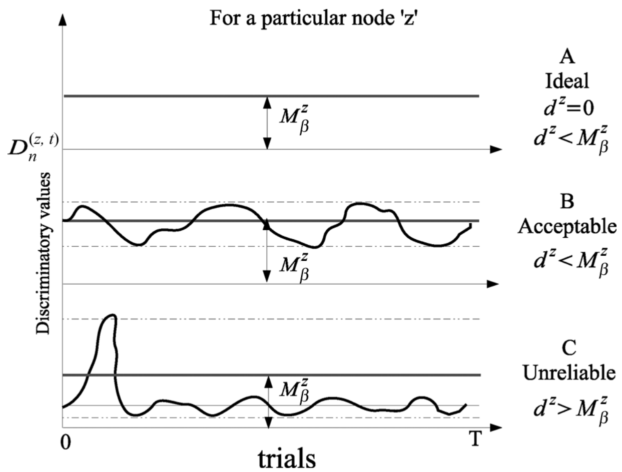

Fig. 5. Sample discriminatory values for a particular node over all trials.

- For a given β and D_n, the mean of the dissimilarity values over all the trials for each of the z nodes were computed using (6), as follows:

$$M_\beta^z = \left\{ \frac{1}{T} \sum_t D_n^{(z,t)} \right\}_\beta . \tag{6}$$

- The z nodes were arranged in the decreasing order of their M_β^z values.
- Starting from the node with the highest M_β^z value, each node was verified for its consistency over the trials.
- A high value for M_β^z indicates that the node is a good discriminator, but it is also possible to have high M_β^z value due to few spurious high discriminatory values over the trials in which case the node is not a consistent performer. Especially when the number of trials are small, using the M_β^z values alone may not be a good indicator of discriminatory node.
- As shown in Fig. 5, there could be three situations (A, B, C) with a possible high M_β^z value. The top panel of the figures shows the ideal condition where the discriminatory value of the node is constant through the trials. This is a desirable but often an unrealistic situation. The second panel shows an oscillatory pattern around the M_β^z value over the trials. This is a practical and acceptable situation as long as the oscillations are close to the M_β^z value. The last panel in the figure shows the inconsistent performance of the node where the discriminatory value is low for major part of the trials but spuriously high in one or few trials. This might show a high average value, but the inconsistency makes the choice of this node undesirable.
- In order to test if the oscillation around the M_β^z value is within acceptable range, we compute the first derivative $d^z(t')$ of the discriminatory values over the trials after sorting them in decreasing order of their amplitude. $d^z(t')$ is given by (7), where t' is the index of the trials that were obtained by sorting the D_n^z values in descending order over the T trials. By doing so, we compute the amplitude swing between subsequent maxima. If the swing between two maxima is higher than M_β^z, it denotes a transient or spurious value which makes the discriminatory value of the node unreliable. Ideally for a reliable discriminatory node, (8) should be satisfied. However, depending upon the accuracy required, the number of maxima used to compute the derivative could be controlled. In the proposed technique, the condition is set that at least the first 50% of the amplitude sorted trials (i.e., the first 50% of the subsequent maximas) should satisfy the condition in (8), as follows:

$$d^z(t') = D_n^{(z,t')} - D_n^{(z,t'-1)} \tag{7}$$
$$\forall t' : d^z(t') < M_\beta^z. \tag{8}$$

The above inequality holds good when the discriminatory value oscillates around M_β^z without huge discontinuities. One of the desirable conditions that satisfies (8) and thereby guarantees a reliable discriminating node is given by

$$\left| D_n^{(z,t)}{}_{max} - D_n^{(z,t)}{}_{min} \right| \ll M_\beta^z. \tag{9}$$

- This is repeated over the z nodes until Q LDB nodes were chosen for further processing.
- If there are less than Q nodes that satisfy the above conditions, the remainder of Q nodes were filled with zeros.

The above steps are repeated for every β combination and using D_n dissimilarity measures. So for each dissimilarity measure, we would obtain $\binom{P}{2} \times Q$ LDB nodes. Among this pool of $\binom{P}{2} \times Q$ LDB nodes, we analyze the frequency of occurrence of each of the z nodes. From this pool, H highly occurring nodes are chosen to be the final set of LDB nodes for each of the dissimilarity measure.

The following are the numerical values used in the proposed study.

- For $P = 4$ classes, there were six ($\beta = 1$ to 6) combinations of classes taken two at a time.
- The number of trials $T = 15$.
- For each β, $Q = 15$ (max) LDBs were selected. Theoretically, any maximum value can be set for choosing Q. Irrespective of the set maximum value for Q, only the number of nodes that satisfy (8) will be chosen. If there are less than Q nodes that satisfy (8), the remaining nodes are filled with zeros. Considering the wide range of audio signals to be classified and at the same time not to include less discriminatory (noisy) nodes into the analysis, a maximum Q value of 15 was chosen. So for $P = 4$ and $\beta = 1$–6, there would be a pool of 6 × 15 LDB nodes, i.e., the 6 × 15 nodes are formed by many occurrences of z nodes in each of the $\binom{P}{2}$ combinations. The histogram of the pool of LDBs ($\binom{P}{2} \times Q$) for both the dissimilarity measures are show in Fig. 6.
- Among these pool of LDB nodes, the highly occurring nodes $H = 15$ were selected as the final LDBs. When few of the nodes occur frequently, we might not have enough nodes to choose 15 out of them. In which case, we limit H to the maximum possible number of nodes (<15).
- For two dissimilarity measures (D_1 and D_2) we finally will have a maximum of $2 \times H$ LDB nodes, which would be 30

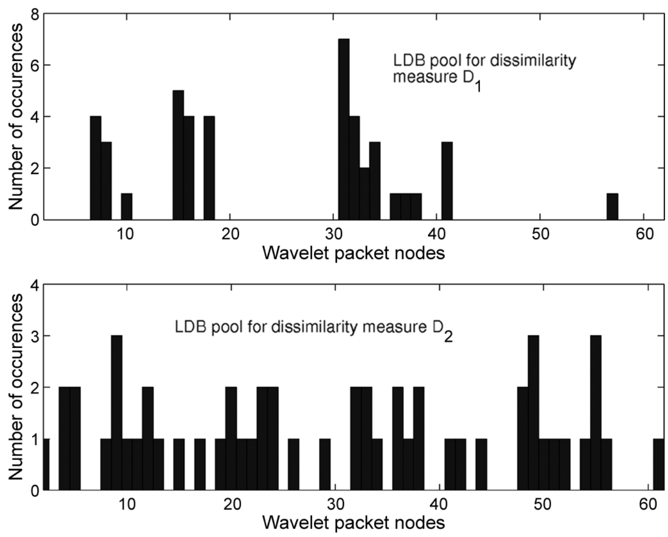

Fig. 6. Sample LDB pool for six pairwise class combinations.

LDB nodes. Some of the nodes may be common for both dissimilarity measures.

At the end of this selection process, we will have a maximum of 15 LDB nodes in the wavelet packet tree for each of the dissimilarity measure D_n that demonstrate relatively high discriminatory behavior among all the combination of P classes. In other words, these nodes demonstrate high statistical distance between all the P classes. The chosen 30 LDBs (some nodes may be common for both D_n) can be used to construct a composite wavelet packet tree which is used to decompose the testing set and extract features as will be explained in Section II-E. We analyzed the database using the above LDB selection process with few variations of Daubechies, Symlet, and Coiflet wavelets. The difference in the performance for the few tested wavelet basis were significantly small to arrive at a conclusion of the best wavelet for the given conditions of the application in hand. Hence, in this study, we choose to present the results using only the db4 (the Daubechies wavelets generated from the four tap conjugate quadrature filter) wavelet packet decompositions.

D. Audio Database

A database consisting of 213 audio signals were used to evaluate the proposed technique. Each audio signal is a segment of 5-s duration with a sampling rate of 44.1 kHz and a resolution of 16 bits/sample. The 213 audio signals consisted of 27 drums, 15 flute, 31 piano, 23 aircraft, 17 helicopter, 20 male speech, 20 female speech, 20 animal, 20 bird, and 20 insect sounds. The instrumental sounds (drums, flute, and piano) were solo. The speech signals were the following sentence: "When the sunlight strikes rain drops in the air they act like a prism and form a rainbow" as spoken by normal male and female subjects [26]. Nonhuman sounds consisted of sounds from various animals, birds, and insects such as a lion, elephant, donkey, horse, dog, monkey, sheep, cat, cow, eagle, hummingbird, sparrow, bluejay, chicken, duck, cricket, bee, wasp, etc. Most of the music samples were collected from the Internet and suitably processed to have uniform sampling frequency and duration.

E. Feature Extraction

LDB Features: All the 213 signals from the audio database were decomposed using the composite wavelet packet tree with LDB nodes as explained in Section II-C. The basis vector coefficients from each of the LDB nodes from this wavelet packet tree could be directly used as features. However, considering the dimensions of the basis vectors, we extracted the normalized node energy $(E_{j,k})$ and the variance of set of variances $(\mathbf{v}_{j,k})$ using the same dissimilarity measures (D_1 and D_2) on the LDB nodes and used them as features. Here, it should be noted that since we are only interested in extracting $E_{j,k}$ and $\mathbf{v}_{(j,k)}$ as features from the individual signals, we do not need a second signal in computing D_1 and D_2. We just replace the $E_{j,k}$ of the second signal to 0 and $\mathbf{v}_{(j,k)}$ to 1. Using each of the dissimilarity measures this way, 15 features were extracted for each audio signal. So, for two dissimilarity measures, there were 30 features for each signal. The combination of these 30 features were evaluated for their significance in the class separability. Redundant features were removed to generate a compact and highly discriminative feature set. A step-wise feature selection procedure using the SPSS [27] software was used to choose a nonredundant feature set that provided minimum classification error. The resulting set of features were then fed to a LDA-based classifier using the SPSS software [27].

MFCC Features: In order to evaluate the relative performance of the proposed LDB work, we compared it with the well-known MFCC features. MFCCs are short-term spectral features and are widely used in the area of audio and speech processing. Some of the related audio classification work using MFCC features are [4], [5], [28], and [29]. To obtain MFCCs, the audio signals were segmented and windowed into short frames of 256 samples. Magnitude spectrum was computed for each of these frames using fast Fourier transform (FFT) and converted into a set of mel scale filter bank outputs. Logarithm was applied to the the filter bank outputs followed by discrete cosine transformation to obtain the MFCCs. Typically, the first 13 MFCCs are used as features. In our case, we computed the mean and variance of the first 13 MFCCs over all the segments of the entire length of the audio signals and used them as our features, i.e., for each audio signal we arrived at 26 features, 13 features were from the mean of the segment MFCCs and the remaining 13 were the variance of the segment MFCCs. These 26 features were computed for all the 213 signals and fed to a LDA-based classifier for classification.

F. Pattern Classification

In linear discriminant analysis, the feature vector comprising of all the features or a subset of the features were transformed into canonical discriminant functions such as

$$f = u_1 b_1 + u_2 b_2 + \cdots + u_q b_q + a \tag{10}$$

where $\{\mathbf{u}\}$ is the set of features extracted from the wavelet packet coefficients at each of the LDB nodes and $\{\mathbf{b}\}$ and a are the coefficients and constant, respectively. The feature dimension q represents the number of features used in the analysis. Using the discriminant scores and the prior probability values of each group, the posterior probabilities of each sample occurring in each of the groups were computed. The sample was then assigned to the group with the highest posterior probability [27].

The classification accuracy was estimated using the leave-one-out method which is known to provide a least bias estimate [30]. In the leave-one-out method, one sample is excluded from the dataset, and the classifier is trained with the remaining samples. Then, the excluded signal is used as the test data and the classification accuracy is determined. This is repeated for all samples of the dataset. Since each signal is excluded from the training set in turn, the independence between the test and the training set are maintained.

III. Results and Discussions

Following the feature extraction procedure as explained in the previous Sections II-C and E, the classification accuracies were computed using a LDA-based classifier and verified using the leave-one-out method. For the proposed LDB approach, the classification accuracies were computed for all the splits (as shown in Fig. 2) in each level, i.e., even though we trained the LDB selection process with four groups, we started one level up from the four groups and verified if the features extracted for four groups fall into their respective parent groups (artificial and natural sounds). This was followed by the two 2 group classifications in second level and then the third-level subgroup classification. Each of the classification of the splits were performed individually every time including the original number of signals in each group or subgroups. The MFCC-based classification followed the same hierarchical classification scheme as we did for the LDB features.

The classification results are presented in three stages, which are: 1) using MFCC features alone; 2) using the proposed LDB features alone; and 3) using a combination of both MFCC and LDB features. As there are totally seven different individual classifications performed, for easy visualization we decided to show them in a figure (Figs. 7–9) rather than in multiple tables. In Figs. 7–9, the number in the brackets after the class label represents the total number of original signals present in that particular class. The correct classification accuracy in percentage for each of the classes is presented using two methods, which are: 1) regular LDA (resubstitution method) and 2) leave-one-out-based LDA (marked with a "*," which are given just above the splits. In the regular LDA method, the LDA-based classifier was trained with all available data and tested with all available data. This is called the resubstitution method of estimating classification accuracy [31]. Since all the data is used for training and testing, this method is optimistically biased and, hence, the results can be considered as the upper bound of the achievable accuracy. On the other hand, the leave-one-out-based approach is known to provide least bias and, hence, the results obtained using this approach can be considered as the lower bound of the achievable accuracy. A combination of these two results (i.e., resubstitution and leave-one-out) would be more appropriate for presenting the range of the achievable classification accuracies [32].

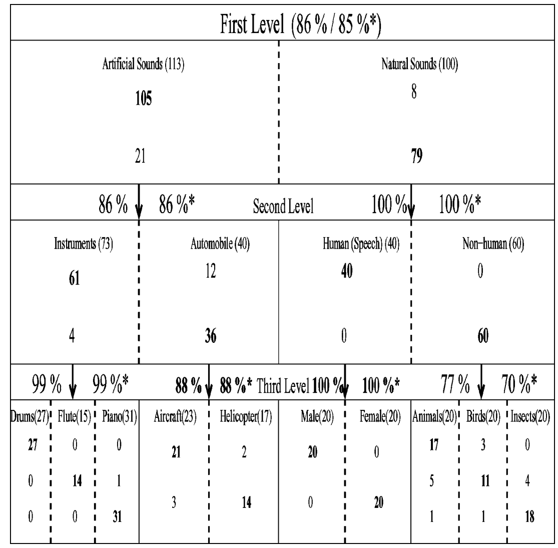

Fig. 7. MFCC features—results for seven individual classifications over three levels. The "*" indicates the percentage classification obtained using leave-one-out-based LDA.

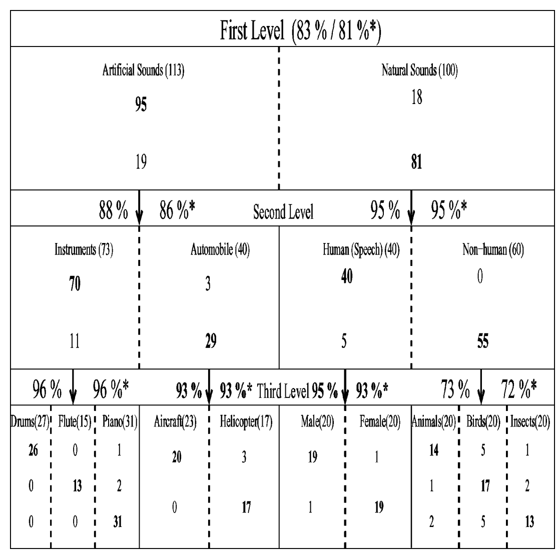

Fig. 8. LDB features—results for seven individual classifications over three levels. The "*" indicates the percentage classification obtained using leave-one-out-based LDA.

Below the percentage, a set of numbers are given vertically. The first number in bold font corresponds to the number of signals correctly classified to its original group and the next number(s) corresponds to the number of signals from other groups misclassified into this group. From top to bottom, each of this number corresponds to the next class label from left to right. For more clarity, let us explain the classification accuracy numbers using the third-level split for animals, birds, and insects in Fig. 7. The overall classification accuracy is shown just above the split which is 77% using regular LDA and 70% for the leave-one-out-based LDA. The column "Animals(20)" shows there are 20 animal signals in the original animal group. Out of this 20 animal signals, 17 were correctly classified, indicated as the bold number in the first row. The second row shows that there were five signals from the Birds group that

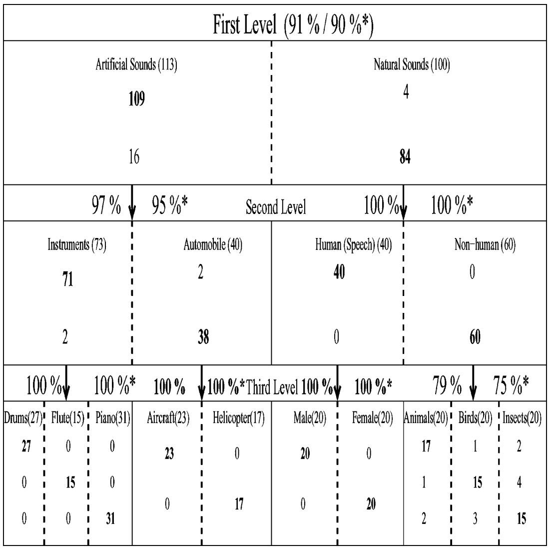

Fig. 9. MFCC and LDB features—results for seven individual classifications over three levels. The "*" indicates the percentage classification obtained using leave-one-out-based LDA.

were misclassified as animals. The third row shows that there was one insect signals that were misclassified as animal signals; likewise for the other groups. The diagonal bold numbers from top left corner to the bottom right corner of any split gives the correct number of signals classified into their respective original groups. These numbers shown in the figure representing the classification accuracy were obtained for the regular LDA method.

A. MFCC Features

The classification results for the MFCC features are shown in Fig. 7. As explained in Section II-E, we used 26 MFCC features for each audio signal. In each of the classifications, the total number of features (26 features) to start with were the same; however, the final optimal set was decided by a step-wise feature selection procedure [27]. On average, we observed six features chosen from the 26 features for each of the classifications. From the results, we observe that the overall classification accuracies ranged from 77% to 100% for the seven individual classification splits over three levels. The average classification accuracy (using the regular LDA results) achieved over each level are: 86% for the first level (artificial and natural sounds), 93% for the second level (instrumental and automobile; human and nonhuman), and 91% for the third-level (drums, flute, and piano; aircraft and helicopter; male and female speech; animals, birds, and insects). It is interesting to note that a classification accuracy of 99%–100% were achieved for the classes in level two "[human, nonhuman]," level three "[drums, flute, piano]," and level three "[speech: male, female]." This was an expected result, as MFCC is known to perform well with music and especially speech signals (natural signals). The lowest classification accuracy of 77% was achieved for the level three "[animal, bird, and insect]" classification.

B. LDB Features

The classification results for the LDB features are shown in Fig. 8. As explained in Section II-E, we used 30 LDB features for each signal. In each of the classifications, the total number of features (30 features) to start with were the same; however, the final optimal set was decided by a step-wise feature selection procedure [27]. On average, we observed five features chosen out of the 30 features for each of the classifications. From the results, we observe that the overall classification accuracies ranging from 73% to 100% for the seven individual classification splits over three levels. The average classification accuracy (using the regular LDA results) achieved over each level are as follows: 83% for the first level (artificial and natural sounds), 92% for the second level (instrumental and automobile; human and nonhuman), and 89% for the third level (drums, flute, and piano; aircraft and helicopter; male and female speech; animals, birds, and insects). Comparing with the results obtained from MFCC, we see that MFCC performs slightly better than the proposed LDB approach. However, we see the proposed LDB approach is slightly better in classifying artificial sounds. In the artificial sound category, drums, piano, and helicopter share common characteristics. In these signals, the sounds occur at some rhythmic intervals with a sharp onset (transient) followed by a smooth decay like damped sinusoids. These asymmetric signal structures match well with the asymmetric daubechies wavelet (db4) used in the proposed work. This could be the reason for the proposed work to perform better in classifying the artificial sounds. This inference indicates the flexibility and the ability of the proposed work in possibly choosing optimal basis functions for each class of audio signals.

It should also be noted that the training of the LDB selection stage was performed only using the four second level groups. By using each of the seven splits in the training of the LDB selection stage and identifying optimal basis function for each of the audio class, it would be possible to achieve better results. However, as the focus of this work is to demonstrate the applicability of the proposed LDB-based methodology for audio classification and also considering the minimal difference observed in comparative results with MFCC, we limit our analysis to the above.

C. MFCC and LDB Features

Comparing the MFCC and LDB features, we observed that the MFCC features were performing better for natural sounds, whereas the LDB features were performing slightly better for artificial sounds. This indicates that the combination of LDB and MFCC features would form a better feature fusion than the standalone MFCC or LDB feature sets. The combined features are expected to ensure that the unique discriminatory clues in the respective domains of MFCC and LDB are well represented and captured. So, a third set of classification accuracies were computed for the combined set of MFCC and LDB features. The total of 56 features (30 from LDB and 26 from MFCC) were fed to the LDA-based classifier, and like before, a step-wise feature selection method was used. On average out of the 56 features, eight features were used for each of the classifications. The classification results for the combined features is shown

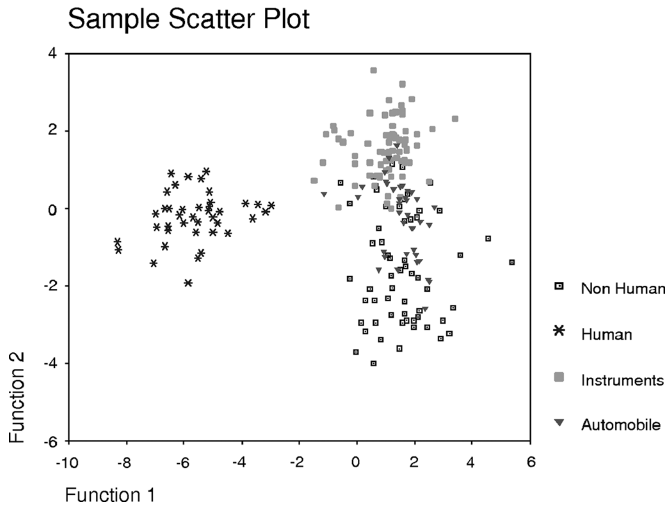

Fig. 10. Sample four group (automobile, instrumental, human, and nonhuman) scatter plot using the first two dominant discriminant functions derived from combined LDB and MFCC features.

in Fig. 9. This combined features resulted in very high classification accuracies scoring 100% or close to 100% in four out of the seven classifications. The average classification accuracy achieved over each level are as follows: 91% for the first level (artificial and natural sounds), 99% for the second level (instrumental and automobile; human and nonhuman), and 95% for the third level (drums, flute, and piano; aircraft and helicopter; male and female speech; animals, birds, and insects). These were excellent results compared to the individual results from MFCC and LDB features. Also on an average only eight (combined) features were used to achieve the reported classification accuracies. A sample scatter plot using the first two dominant discriminant function for the four group (instrumental, automobile, human, and nonhuman) classification using the combined features is shown in Fig. 10.

From the results presented for the MFCC, LDB, and combined MFCC and LDB, we could infer that the proposed LDB approach performs well and is comparable to the results of the MFCC. The combination of LDB features with MFCC features improves the classification accuracy, indicating that they both complement each other in classifying a complex multiclass problem as presented in this paper.

IV. Conclusion

A novel LDB-based audio classification scheme covering a wide range of audio signals was presented. High classification accuracies were achieved using the proposed methodology. Simple dissimilarity measures like node energy and nonstationarity index performed well in identifying the discriminatory nodes between the audio classes. The presented results suggest significant potential for LDB-based audio classification in auditory scene analysis or environment detection. A comparative study with the well-known MFCC features was presented. The proposed LDB approach performs comparably to the MFCC results with added benefits in the choice of basis functions and the inherent ability to handle nonstationarity of audio signals. The combination of MFCC and LDB features shows promising benefits and could be an interesting area for further analysis. The lower computational expense, automated detection of discriminant subspaces, and thereby extraction of discriminant features makes the proposed methodology attractive for multigroup audio classification applications. Future work involves improving the LDB selection process, arriving at an optimal number of LDBs for a given classification problem and include more dissimilarity measures for audio classification.

References

[1] J. P. Campbell, Jr., “Speaker recognition: A tutorial,” *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

[2] L. Lu and H.-J. Zhang, “Content analysis for audio classification and segmentation,” *IEEE Trans. Speech Audio Process.*, vol. 10, no. 7, pp. 504–516, Oct. 2002.

[3] K. Umapathy, S. Krishnan, and S. Jimaa, “Multi-group classification of audio signals using time-frequency parameters,” *IEEE Trans. Multimedia*, vol. 7, no. 2, pp. 308–315, Apr. 2005.

[4] G. Guo and S. Z. Li, “Content-based audio classification and retrieval by support vector machines,” *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 209–215, Jan. 2003.

[5] G. Tzanetakis and P. Cook, “Music genre classification of audio signals,” *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.

[6] C. J. Burges, J. C. Platt, and S. Jana, “Distortion discriminant analysis for audio fingerprinting,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 165–174, May 2003.

[7] J. L. Dugelay, J. C. Junqua, C. Kotropoulos, R. Kuhn, F. Perronnin, and I. Pitas, “Recent advances in biometric person authentication,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2002, vol. 4, pp. 4060–4063.

[8] M. Cooper and J. Foote, “Summarizing popular music via structural similarity analysis,” in *IEEE Workshop Applicat. Signal Process. Audio Acoust.*, 2003, pp. 127–130.

[9] C. Xu, N. C. Maddage, and X. Shao, “Automatic music classification and summarization,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 441–450, May 2005.

[10] H. G. Kim, N. Moreau, and T. Sikora, “Audio classification based on MPEG-7 spectral basis representations,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 716–725, May 2004.

[11] S. Esmaili, S. Krishnan, and K. Raahemifar, “Content based audio classification and retrieval using joint time-frequency analysis,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2004, pp. V 665–V 668.

[12] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, “Computational auditory scene recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2002, vol. 2, pp. 1941–1944.

[13] S. Ravindran and D. V. Anderson, “Audio classification and scene recognition and for hearing aids,” in *IEEE Int. Symp. Circuits Syst. (ISCAS)* , May 2005, pp. 860–863.

[14] L. Ma, D. Smith, and B. Milner, “Context awareness using environmental noise classification,” in *Proc. Eurospeech*, 2003, pp. 2237–2240.

[15] M. Akbacak and J. H. L. Hansen, “Environmental sniffing: Noise knowledge estimation for robust speech systems,” in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP)*, Apr. 2003, vol. 2, pp. 113–116.

[16] H. G. Okuno, T. Ogata, K. Komatani, and K. Nakadai, “Computational auditory scene analysis and its application to robot audition,” in *Proc. Int. Conf. Informatics Research Development of Knowledge Society Infrastructure (ICKS)*, Mar. 2004, pp. 73–80.

[17] N. Saito and R. R. Coifmann, “Local discriminant bases and their applications,” *J. Math. Imaging Vision*, vol. 5, no. 4, pp. 337–358, 1995.

[18] C. M. Spooner, “Application of local discriminant bases to HRR-based ATR,” in *Proc. Signals, Syst., Comput., 35th Asilomar Conf.*, Nov. 2001, pp. 1067–1073.

[19] B. Christian, “Local discriminant bases and optimized wavelet to classify ultrasonic echoes: Application to indoor mobile robotics,” *Proc. IEEE Sens.*, vol. 2, no. 12–14, pp. 1654–1659, Jun. 2002.

[20] D. J. Strauss, W. Delb, J. Jung, and P. K. Plinker, “Adapted filter banks in machine learning: Applications in biomedical signal processing,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2003, pp. VI 425–VI 428.

[21] K. Englehart, B. Hudgins, and P. A. Parker, "A wavelet-based continuous classification scheme for multifunction myoelectric control," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 3, pp. 302–311, Mar. 2001.

[22] K. Umapathy, S. Krishnan, and R. K. Rao, "Audio signal feature extraction and classification using local discriminant bases," in *Proc. Int. Conf. Signal Process. Commun. (SPCOM)*, Dec. 2004, pp. 457–461.

[23] K. Umapathy, S. Krishnan, and A. Das, "Sub-dictionary selection using local discriminant bases algorithm for signal classification," in *Proc. IEEE Can. Conf. Elect. Comput. Eng.*, Niagara Falls, ON, Canada, May 2004, pp. 2001–2004.

[24] A. Web, *Statistical Pattern Recognition*. West Sussex, U.K.: Wiley, 2002.

[25] S. Mallat, *A Wavelet Tour of Signal Processing*. San Diego, CA: Academic, 1998.

[26] M. Eye and E. Infirmary, Voice Disorders Database, version 1.03 (CDROM). Kay Elemetrics Corporation, Lincoln Park, NJ, 1994.

[27] *"SPSS Advanced Statistics User's Guide,"* SPSS, Inc., Chicago, IL, 1990.

[28] S. Essid, G. Richard, and B. David, "Instrument recognition in polyphonic music based on automatic taxonomies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 68–80, Jan. 2006.

[29] A. J. Eronen *et al.*, "Audio-based context recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 321–329, Jan. 2006.

[30] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. San Diego, CA: Academic, 1990.

[31] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[32] K. Fukunaga and R. R. Hayes, "Estimation of classifier performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 10, pp. 1087–1101, Oct. 1989.

Karthikeyan Umapathy (S'02) received the B.E degree in electronics and communication engineering from Bharathiyar University, Coimbatore, India, in 1992 and the M.Phil. degree in electronic, communication, and electrical engineering from the University of Hertfordshire, Hatfield, U.K., in 2002, in collaboration with Ryerson University, Toronto, ON, Canada. He is currently pursuing the Ph.D. degree in electrical and computer engineering at the University of Western Ontario, London.

From 1993 to 2000, he was with Philips Medical Systems India in his last position as the Area Manager/Country Specialist—Magnetic Resonance Imaging for both technical and clinical applications.

Mr. Umpathy holds the Canada Graduate Scholarship (CGS) from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Sridhar Krishnan (S'94–M'99–SM'05) received the B.E. degree in electronics and communication engineering from Anna University, Chennai, India, in 1993, and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Calgary, Calgary, AB, Canada, in 1996 and 1999, respectively.

He joined Ryerson University, Toronto, ON, Canada, in 1999 as an Assistant Professor in Electrical and Computer Engineering and was promoted to Associate Professor in 2003. His research interests include adaptive signal processing, biomedical signal/image analysis, and multimedia processing and communications.

Dr. Krishnan is the Chair of the IEEE Signals and Applications Chapter, Toronto Section, and is a Registered Professional Engineer (P.Eng.) in the Province of Ontario.

Raveendra K. Rao (S'83–M'88–SM'01) received the M.Tech. and Ph.D degrees from the Indian Institute of Technology, Delhi, India, in 1982 and 1989, respectively.

From 1988 to 1999, he was a tenured member of the Delhi Institute of Technology. He served the Institute in various capacities including Chair of the Electrical and Computer Engineering Department and Member of the Board of Governors. From 1996 to 1998, he was Chair of Electronic Department and Deputy Principal of Caledonian College of Engineering (Affiliate of Glasgow Caledonian College of Engineering, Glasgow, U.K.), Muscat, Oman. Also, he has served on the engineering faculties of Delhi, Mysore, and Roorkee Universities in India. Since 1999, he has been with the Department of Electrical and Computer Engineering, University of Western Ontario (UWO), Toronto, ON, Canada. He founded the Communication and Signal Processing (CASP) Laboratory, Delhi Institute of Technology, and was in charge of the Laboratory from 1989 to 1999. He played major roles in the establishment and development of Electronic and Communication Engineering Laboratories, Caledonian College of Engineering. Currently, he is in charge of the Bell Center for Information Engineering in the Department of Electrical and Computer Engineering, UWO. His research interests include digital communications, wireless networks, statistical signal processing, and multicarrier and multiuser communication systems. He has authored more than 40 technical papers in journals and international conferences. He has supervised more than 25 M.S. and Ph.D. students.

Dr. Rao received the IEEE AES/COM Chapter India Award in 1985 for contributions in the area of communications, the Young Scientist of International Union of Radio Science (URSI) award in 1987, and the Student Journal Award of IETE (India) for best paper published in 1994. He was nominated for the prestigious Vikas Jyothi Award, India, in 1993. He is a Fellow the IETE and a registered Professional Engineer in the Province of Ontario, Canada. He is listed in Asia's *Who's Who of Men and Women of Achievement*.