Environmental Sound Recognition: A Survey

Sachin Chachada and C.-C. Jay Kuo
Ming Hsieh Department of Electrical Engineering
University of Southern California, Los Angeles, CA 90089, USA
E-mails: chachada@usc.edu; cckuo@sipi.usc.edu

***Abstract*—Although research in audio recognition has traditionally focused on speech and music signals, the problem of environmental sound recognition (ESR) has received more attention in recent years. Research on ESR has significantly increased in the past decade. Recent work has focused on the appraisal of non-stationary aspects of environmental sounds, and several new features predicated on non-stationary characteristics have been proposed. These features strive to maximize their information content pertaining to signal's temporal and spectral characteristics. Furthermore, sequential learning methods have been used to capture the long-term variation of environmental sounds. In this survey, we will offer a qualitative and elucidatory survey on recent developments. It includes three parts: i) basic environmental sound processing schemes, ii) stationary ESR techniques and iii) non-stationary ESR techniques. Finally, concluding remarks and future research and development trends in the ESR field will be given.**

I. Introduction

A considerable amount of research has been made towards modeling and recognition of environmental sounds over the past decade. By environmental sounds, we refer to various quotidian sounds, both natural and artificial (*i.e.* sounds one encounters in daily life other than speech and music). Environmental sound recognition (ESR) plays a pivotal part in recent efforts to perfect machine audition.

With a growing demand on example-based search such as content-based image and video search, ESR can be instrumental in efficient audio search applications [36]. ESR can be used in automatic tagging of audio files with descriptors for keyword-based audio retrieval [9]. Robot navigation can be improved by incorporating ESR in the system [5], [42]. ESR can be adopted in a home-monitoring environment, be it for assisting elderly people living alone in their own home [3], [32] or for a smart home [37]. ESR, along with image and video analysis, find applications in surveillance [7], [24]. ESR can also be tailored for recognition of animal and bird species by their distinctive sounds [1], [39].

Among various types of audio signals, speech and music are two categories that have been extensively studied. In its infancy, ESR algorithms were a mere reflection of speech and music recognition paradigms. However, on account of considerably non-stationary characteristics of environmental sounds, these algorithms proved to be ineffective for large-scale databases. For example, the speech recognition task often exploits the phonetic structure that can be viewed as a basic building block of speech. It allows us to model complicated spoken words by breaking them down into elementary phonemes that can be modeled by the Hidden Markov Model (HMM) [22]. In contrast, general environmental sounds, such as that of a thunder or a storm, do not have any apparent substructures like phonemes. Even if we were able to identify and learn a dictionary of basic *units* (analogous to phonemes in speech) of these events, it would be difficult to model their variation in time with HMM as their temporal occurrences would be more random as against preordained sequence of phonemes in speech. Similarly, as compared to music signals, environmental sounds do not exhibit meaningful stationary patterns such as melody and rhythm [23]. To the best of our knowledge, there was only one survey article on the comparison of various ESR techniques done by Cowling and Sitte [6] about a decade ago.

Research on ESR has significantly increased in the last decade. Recent work has focused on the appraisal of non-stationary aspects of environmental sounds, and several new features predicated on non-stationary characteristics have been proposed. These features, in essence, strive to maximize their information content pertaining to signal's temporal and spectral characteristics as bounded by the uncertainty principle. For most real life sounds, even these features exhibit non-stationarity when observed over a long period of time. To capture these long-term variations, sequential learning methods have been applied.

It becomes evident that ESR methods not only have to model non-stationary characteristics of sounds, but also have to be scalable and robust as there are numerous categories of environmental sounds in real life situations. Despite increased interest in the field, there is no single consolidated database for ESR, which often hinders benchmarking of these new algorithms.

In this work, we will present an updated survey on recent developments and point out the future research and development trends in the ESR field. In particular, we will elaborate on non-stationary ESR techniques. The rest of this paper is organized as follows. We will first discuss three commonly used schemes for environmental sound processing in Section II. Then, we will conduct a survey on stationary and non-stationary ESR techniques in Sections III and IV, respectively. Finally, concluding remarks and future research trends will be given in Section V.

II. Environmental Sound Processing Schemes

Before delving into the details of various ESR techniques, we first describe three commonly used environmental sound processing schemes in this section.

1) Framing-based processing: Audio signals to be classified are first divided into frames, often using a Haaning or a Hamming window. Features are extracted from each frame and this set of features is used as one instance of training or testing. A classification decision is made for each frame and, hence, consecutive frames may belong to different classes. A major drawback of this processing scheme is that there is no way of selecting an optimal framing-window length suited for all classes. Some sound events are short-lived (*e.g.* gun-shot) as compared to other longer events (*e.g.* thunder). If the window length is too small, the long-term variations in the signal would not be well captured by the extracted features, and the framing method might chop events into multiple frames. On the other hand, if the window length is too large, it becomes difficult to locate segmental boundaries between consecutive events and there might be multiple sound events in a single frame. Also, one has to rely on features to extract non-stationary attributes of the signal since such a model does not allow the use of sequential learning methods.

2) Sub-framing-based processing: Each frame is further segmented into smaller sub-frames, usually with overlap, and features are extracted from each sub-frame. In order to learn a classifier, features extracted from sub-frames are either concatenated to form a large feature vector or averaged so as to represent a single frame. Another possibility is to learn a classifier for each sub-frame and make a collective decision for the frame based on class labels of all sub-frames (*e.g.*, a majority voting rule). This model allows the use of both non-stationary features and sequential classifiers. Even with a non-sequential classifier, this processing scheme can represent each frame better as the collective distribution over all sub-frames allows one to model intra-frame characteristics with greater accuracy. This method offers more flexibility in segmenting consecutive sound events based on class labels of sub-frames.

3) Sequential processing: Audio signals are still divided into smaller units (called a segment), which is typically of 20-30 ms long with 50% overlap. The classifier makes decisions on class labels and segmentation both based on features extracted from these segments. As compared to the above two methods, this method is unique in its objective to capture the inter-segment correlation and the long-term variations of the underlying environment sound. This can be achieved using a sequential signal model such as the Hidden Markov Models (HMM).

Any ESR algorithm basically follows one of the above three processing schemes with minor variations in its preprocessing and feature selection/reduction schemes. For example, a pre-emphasis filter can be used to boost the high frequency content or an A-weight filter can be used for equalized loudness. For feature selection/reduction, there is an arsenal of tools to choose from [16], [20], [35]. We will not pay attention to these minor differences in later sections.

III. Stationary ESR Techniques

Features developed for speech/music based applications have been traditionally used in stationary ESR techniques. These features are often based on psychoacoustic properties of sounds such as loudness, pitch, timbre, etc. A detailed description of features used in audio processing was given in [17], where a novel taxonomy based on the properties of audio features was provided (see Fig. 1).

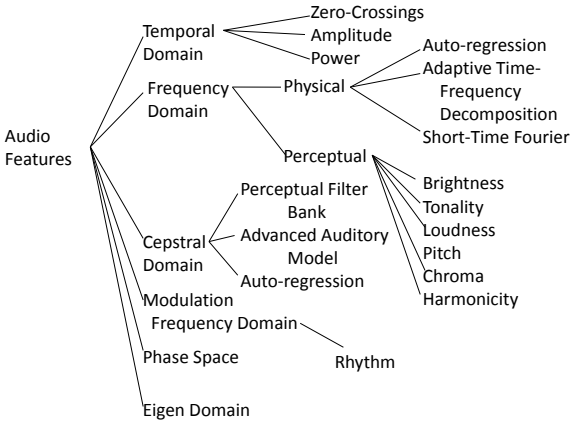

Fig. 1. Taxonomy for audio features as proposed in [17].

Features such as Zero-Crossing Rate (ZCR), Short-Time Energy (STE), Sub-band Energy Ratio, Spectral Flux, etc. are easy to compute and used frequently along with other refined set of features. These features provide rough measures about temporal and spectral content properties of an audio signal. For more details on basic features, we refer to [8], [17], [19], [21].

Cepstral features are widely used features. They include: Mel-Frequency Cepstral Coefficients (MFCC) and their first and second derivatives (ΔMFCC and $\Delta\Delta$MFCC), Homomorphic Cepstral Coefficients (HCC), Bark-Frequency Cepstral Coefficients (BFCC), etc. MFCC were developed to resemble the human auditory system and have been successfully used in speech and music applications. As mentioned before, due to lack of a standard ESR database, MFCC are often used by researchers for benchmarking their work. A common practice is to concatenate MFCC features with newly developed features to enhance the performance of a system.

MPEG-7 based features are also popular for speech and music applications. They demand low computational complexity and encompass psychoacoustic (or perceptual-based) audio properties. Wang *et al.* [38] proposed to use low-level audio descriptors such as Audio Spectrum Centroid and Audio Spectrum Flatness with a hybrid classifier constituted of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN). They converted the classifier outputs from SVM and KNN into probabilistic scores and fused them to improve classification accuracy. Muhammad *et al.* [18] combined several low-level MPEG-7 descriptors and MFCC and used Fisher's Discriminant Ratio (F-Ratio) to discard irrelevant features. Although MPEG-7 features perform better than MFCC, MFCC and MPEG-7 descriptors are shown to be complementary to each other and, when used together, the classification accuracy

can be improved.

Auto-regression based features, in particular, Linear Prediction Coefficients (LPC), have been prevalent in speech processing applications. Linear Prediction Cepstrum Coefficients (LPCC), which are an alternate representation of LPC, are also commonly used. However, LPC and LPCC embody the source-filter model for speech and, hence, they are not useful for ESR. Tsau *et al.* [30] proposed the use of the Code Excited Linear Prediction (CELP) based features along with the LPC, pitch and pitch gain features. Since CELP uses a fixed codebook for excitation of a source-filter model, it is more robust than LPC. Tsau *et al.* [30] reported improved performance over MFCC. CELP and MFCC together further increase the classification accuracy, specially noticeable for classes like rain, stream and thunder which are difficult to recognize.

ESR algorithms relying on the sub-framing processing scheme usually learn signal-models in each sub-frame and, thus, do not utilize the temporal structure. One variation to exploit the temporal structure is when a signal-model is learned based on features from all ordered sub-frames such as HMM. Another example was recently proposed by Karbasi *et al.* [14], which attempted to capture the temporal variation among sub-frames in a new set of features called "Spectral Dynamic Features (SDF)" as detailed below.

Let $x_{sb}(i)$ denote the i^{th} sub-frame, with $i \in [1, N]$. From each sub-frame $x_{sb}(i)$, MFCC and other features are extracted in a vector y_i with dimension $L \times 1$. Let $Y = [y_1, \ldots, y_N]$ be a matrix with columns y_i of feature vectors for N sub-frames. For each row of Y, the N-point FFT is applied followed by the logarithmic filter bank, and then followed by the N-point DCT to yield the final set of features. This method essentially extracts cepstral features (MFCC-like features) considering each row of Y as a time series. For example, if 13 MFCC coefficients are extracted from each sub-frame, then we end up with 13 time series, one for each dimension. The cepstral features are evaluated for each of these time series by capturing the dynamic variation of sub-frame features over the entire frame. The superior performance of SDF against several conventional features such as ZCR, LPC, MFCC under three classifiers (*i.e.*, KNN, GMM and SVM) was demonstrated. It was shown in [14] that the combined features of MFCC and ΔMFCC give the performance bound of *static* features, which is not improved by adding more conventional features. A system with a feature vector consisting of ZCR, Band-Energy, LPC, LPCC, MFCC and ΔMFCC, performs poorly as compared to that with only MFCC and ΔMFCC under the SVM or GMM classifiers. In contrast, the *dynamic* feature set, SDF, achieves an improvement of $10 - 15\%$ over the *static* bound.

Filter-banks are often used to extract features local to smaller bands, encapsulating spectral properties effectively. On the other hand, the auto-correlation function (ACF) represents the time-evolution and has an intimate relationship with the power spectral density (PSD) of the underlying signal. Valero and Alias [33] proposed a new set of features called the Narrow-Band Auto Correlation Function features (NB-ACF). The extraction of NB-ACF features can be explained using Fig. 2. First, a signal is passed through a filter bank with $N = 48$ bands whose center frequencies being tuned to the Mel-scale. Then, the sample ACF of the filtered signal in the ith band is calculated, which is denoted by $\Phi_i(\tau)$. One can calculate four NB-ACF features based on each ACF as follows.

1) $\Phi_i(0)$: Energy at lag $\tau = 0$. It is a measure of the perceived sound pressure at the i^{th} band.
2) τ_{i_1}: Delay of the first positive peak which represents the dominant frequency in the i^{th} band.
3) $\Phi_{i_1}(\tau_{i_1})$: Normalized ACF of the first positive peak. It is related to the periodicity of the signal and, hence, gives a sense of pitch of the filtered signal at the i^{th} band.
4) τ_{i_e}: Effective duration of the envelope of normalized ACF. It is defined as the time taken by normalized ACF to decay 10 dB from its maximum value, and it is a measure of reverberation of the filtered signal at the i^{th} band.

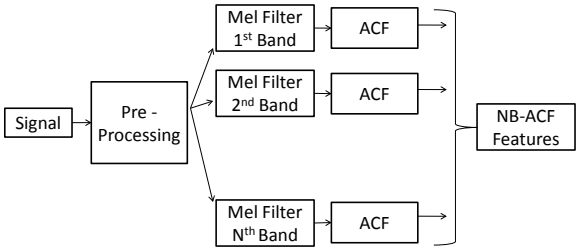

Fig. 2. Illustration of the NB-ACF feature extraction process.

As a rule of thumb, the sample ACF is meaningful up to lag τ_{opt} if and only if the signal length is at least four times the lag length. This demands a sub-frame length to be much larger than that used in sub-framing processing. It is recommended in [33] that a be sub-frame of size 500ms with an overlap of 400 ms in a frame of 4 seconds. Finally, KNN and SVM classifiers are used for decision making in each sub-frame. The performance of NB-ACF features was compared with MFCC and Discrete Wavelet Transform (DWT) coefficients with a data-set consisting of 15 environmental scenes. Dynamically changing scenes, such as *office*, *library* and *classroom*, pronouncedly benefited from the new NB-ACF features. It is well known that ACF is instrumental in the design of linear predictors for time-series because they capture the temporal similarity/dissimilarity well. As a result, the NB-ACF features offer better performance for wide-sense-stationary (WSS) signals than most static features discussed in this section.

IV. Non-stationary ESR Techniques

A. Wavelet-Based Methods

The performance of commonly employed features for audio recognition, including Mel-Frequency Cepstral Coefficients (MFCC), Homomorphic Cepstral Coefficients (HCC), time-frequency features derived using Short-Term Fourier Trans-

form (STFT), Discrete Wavelet Transform (DWT) and Continuous Wavelet Transform (CWT) was compared by Cowling and Sitte in [6], where the Learning Vector Quantization (LVQ), Artificial Neural Networks, Dynamic Time Warping (DTW) and Gaussian Mixture Models (GMM) were used as classifiers. The experiments were conducted on three types of data – speech, music and environmental sounds. For the environmental sound, the data set consisted of 8 classes, and the framing-based processing scheme was adopted. It was reported that the best performance for ESR was achieved with CWT features with the DTW classifier, which was comparable to that of MFCC features with the DTW classifier. It is surprising that CWT, which is a time-frequency representation, and MFCC gave very similar results while DWT and STFT did not give good performance. It was noted in [6] that the dataset was too small to make any meaningful comparison between MFCC and CWT. Given other factors being equal, MFCC features can be more favored than CWT features because of their lower computational complexity. DTW was clearly the best classifier in the test, yet the claim should be further verified by a larger environmental sound database.

Han and Hwang [13] used the Discrete Chirplet Transform (DChT) and the Discrete Curvelet Transform (DCuT) along with several other common features such as MFCC, ZCR, etc. When compared, all features gave similar performance, yet significant improvement was observed when they were used together.

Valero and Alias [34] adapted the Gammatone mother function to meet wavelet admissibility conditions, used the squared sum of Gammatone representations of signal as features, called the Gammatone wavelet features, and adopted the SVM classifier. A comparable performance was observed between Gammatone wavelet features and DWT. When both features were used together, classification accuracy was improved even in noisy conditions. Gammatone features perform well in classes such as footsteps and gunshots due to their capability in characterizing transient sounds.

Umapathy *et al.* [31] proposed a new set of features based on the binary wavelet packet tree (WPT) decomposition. More recently, Su *et al.* [28] used a similar approach to recognize sound events in an environmental scene consisting of many sound events. This ESR algorithm was conducted with the framing-based processing scheme. The signal in the ith frame, x_i, is first transformed to a binary WPT representation denoted by $\Omega_{j,k}$, where j is the depth of the tree and k is the node index at level j. Each subspace $\Omega_{j,k}$ is spanned by a set of basis vectors $\{\mathbf{w}_{j,k,l}\}_{l=0}^{2^u-1}$, where 2^u is the length of x_i. Then, we have

$$x_i = \sum_{j,k,l} [\alpha_{j,k,l}]_i \ \mathbf{w}_{j,k,l}, \tag{1}$$

where $\alpha_{j,k,l}$ is the projection coefficient at node (j,k). Once all training samples are decomposed to a binary WPT, the Local Discriminant Bases (LDB) algorithm is used to identify the most discriminant nodes of the WPT. The LDB algorithm can be simply described below. For each pair of classes in the data set, one can determine a set of Q discriminatory nodes based on a dissimilarity measure. Two dissimilarity measures were proposed in [28]:

1) the difference of normalized energy
$$D_1 = E_1^{(j,k)} - E_2^{(j,k)}$$
of the two sound classes at the same node (j,k);
2) the ratio of the variances of projection coefficients of the two sound classes at node (j,k),
$$D_2 = var[\mathbf{v}_1^{(j,k)}]/var[\mathbf{v}_2^{(j,k)}],$$
where $\mathbf{v}_i^{(j,k)}$ is the vector of variance of locally grouped coefficients at node (j,k).

Strictly speaking, none of these two dissimilarity measures are distance metrics. The selected Q nodes should be consistent. It was recommended to conduct multiple trials with randomly selected training samples from two classes, and consistent nodes should be selected from these random trials. The above process should be repeated among all possible class pairs. Finally, we select H nodes that occur most frequently among the Q nodes for each pair, and use coefficients and/or dissimilarity measure quantities at these H nodes as features.

The LDA-based classifier was used in [31] while the KNN and HMM were used in [28]. It was observed in [31] that WPT-LDB and MFCC features gave similar performance, yet much better performance was achieved when the two were combined together. It was reported in [28] that MFCC performed better than WPT-LDA, and a significant improvement could be obtained by combining the two features. Note that the classification performance in [28] was given for environmental scenes rather than individual events.

Despite being time-frequency features, the performance of wavelet features is not better than that of MFCC features but at a comparable level. When being combined with MFCC, the performance does improve yet the required complexity overhead to extract wavelet features might not always justify the gain in classification accuracy except for the Gammatone features. The Gammatone features are proved to be complementary to MFCC owing to their strong capability in representing impulsive signal classes such as footsteps and gun-shots.

B. Sparse-Representation-Based Methods

Chu *et al.* [4] proposed to use the Matching Pursuit (MP) based features for ESR. The basis MP (BMP) is a greedy algorithm used to obtain a sparse representation of signals based on atoms in an over-complete dictionary. Given signal x and an over-complete dictionary $D = [d_1, d_2, ...]$, BMP obtains the sparse representation of x on D as follows.

1) Initialize the residue at the 0^{th} iteration as $R^0 x = x$
2) For $t = 1$ to T
    a) Select the atom with the largest inner product with the residue via
$$d_t = \max_i < R^{t-1}x, d_i > .$$

b) Update the residue via

$$R^t x = R^{t-1} x - \alpha_t d_t,$$

where $\alpha_t =< R^{t-1}x, d_t >$ is the projection coefficient of $R^{t-1}x$ on d_t.

3) The BMP projection of x on D is given by

$$\hat{x} = \sum_{i=1}^{T} \alpha_i d_i$$

One stopping criterion for this algorithm is a fixed number of iterations (atoms), T. Another one is to use the energy of the residual signal, i.e. decomposition stops at t when $||R^{t-1}x||^2 <$ Threshold.

An over-complete Gabor dictionary consisting of frequency modulated Gaussian functions (called Gabor atoms) was used in [4]:

$$g_{s,u,\omega,\theta} = \frac{K_{s,u,\omega,\theta}}{\sqrt{s}} e^{-\pi(n-u)^2/s^2} \cos\left[2\pi\omega(n-u)+\theta\right], \quad (2)$$

where s, u, ω and θ are atom's scale, location, frequency and phase, respectively, and $K_{s,u,\omega,\theta}$ is a normalization constant so that $||g_{s,u,\omega,\theta}||^2 = 1$.

The following parameters were chosen: $s = 2^p (1 \leq p \leq 8)$, $u = \{0, 64, 128, 192\}$, $\omega = 0.5 \times 35^{-}2.6\ i^{2.6}$ $(0 \leq i \leq 35)$, and $\theta = 0$, with each atom of size $N = 256$ given signal sub-frames of size 256 at a sampling frequency of 22.05 kHz. The classification accuracy is not affected much for $T > 5$ so that the first $T = 5$ atoms in the MP algorithm is used. The selected features are the mean and the variance of scale and frequency parameters of the 5 selected atoms, *i.e.*, $[\mu_s, \mu_w, \sigma_s, \sigma_w]$, which are referred to as the MP-Gabor features. The location and phase parameters are ignored. It adopts the sub-framing processing scheme with a frame of 4 seconds and a sub-frame of 0.11 ms with 50% overlap. For classification, KNN and GMM classifiers were tested. The MP-Gabor features perform marginally better than MFCC, and the classification accuracy is further improved when used together with MFCC. Sound classes with broad-spectrum fare well with the MP-Gabor features, but classes with highly non-stationary characteristics such as thunder sounds have poorer recognition accuracy.

To improve the performance of MP-Gabor features, Sivasankaran and Prabhu [25] proposed several modifications. First, they construct a signal-dependent over-complete dictionary (rather than using a fixed dictionary) for signals. The normalized frequency scale is divided into N sub-bands, and the normalized energy present in each sub-band is calculated using DFT coefficients. Suppose that a total of N_f frequency points are to be used in the dictionary. The number of frequency points in each sub-band is proportional to its normalized energy and equally-spaced frequency points in each sub-band are used. Second, the Orthogonal Matching Pursuit (OMP), which is a variant of BMP, was used. At each iteration, OMP computes the orthogonal projection matrix using previously selected atoms and calculates projection coefficients using this projection matrix. Third, the weighted sample mean and variance are used. They achieved high classification accuracy by using modified MP-Gabor features and MFCC yet without performance benchmarking with other methods. The modified MP-Gabor and MFCC features together perform well for most sound classes including thunder. The only two classes with lower classification accuracy were ocean and rain. They are actually quite similar when heard for a small duration of time.

Yamakawa *et al.* [41] compared the Haar, Fourier and Gabor bases with the HMM classifier using the sequential processing scheme. Instead of using the mean and the standard deviation of scale and frequency parameters of MP-Gabor atoms, they concatenated them to construct a feature vector. Since MP is a greedy algorithm, one may not expect ordered atoms to offer an accurate approximation to non-stationary signals. Due to the use of the HMM classifier, results for Gabor features are still good when the sound classes were restricted to impulsive sounds. The classification accuracy of Haar wavelets was low in the experiment, which is counter-intuitive since the Haar basis matches the impulse-like structure well in the time domain. This work does show that HMM can better capture the variations in features when 6 mixtures are used in GMM to model hidden states. Also, the performance of time-frequency Gabor features and stationary Fourier features are comparable.

To conclude, MP-based features that are capable of extracting the information of high time-frequency resolution improve the performance of an ESR system when used together with the popular MFCC. Moreover, classification accuracy can be further improved using sequential learning methods such as HMM.

C. Power-Spectrum-Based Methods

The spectrogram provides useful information about signal's energy in a well localized time and frequency region. It is an intuitive tool to extract transient and variational characteristics of environmental sounds. However, it is not easy to use the spectrogram features in learning models for ESR for a small database due to its higher dimensionality.

Khunarsal *et al.* [15] used the sub-framing processing scheme to calculate the spectrogram as the concatenation of the Fourier Spectrum of sub-frames and adopted the Feed-Forward Neural Network and KNN for classification. Extensive study was done on the selection of spectrogram size parameters, the audio signal length, the sampling rate and other model parameters needed for accurate classification. The features were compared with MFCC and LPC and MP-Gabor features. The spectrogram features perform consistently better against MFCC and LPC and give comparable results against MP-Gabor features. Although a combination of the spectrogram, LPC and MP-Gabor features gives the best results, classification results with other feature combination are comparable to the best one. This implies that there is redundancy in these features.

Recently, Ghoraani and Krishnan [10] proposed a novel feature extraction method based on the spectrogram using the framing processing scheme. First, the MP representation for

a signal is achieved with the Gabor dictionary that has fine granularity in scale, frequency, position and phase. To render a good approximation of the signal, the stopping criterion is set to $T = 1000$ iterations. Let $x(t)$ be the signal and $g_{\gamma_i}(t)$ be the Gabor atom with $\gamma_i = \{s, u, \omega, \theta\}$ as parameters in Eq. (2). After T iterations, we have

$$x(x) = \sum_{i=1}^{T} \alpha_i \; g_{\gamma_i}(t) + R^T x. \tag{3}$$

The Time-Frequency Matrix (TFM) representation of $x(t)$ can be written as

$$V(t, f) = \sum_{i=1}^{T} \alpha_i \; WVG_{\gamma_i}(t), \tag{4}$$

where WVG_{γ_i} is the WignerVille distribution (WVD) of Gabor atom $g_{\gamma_i}(t)$. The WVD is a quadratic time-frequency representation in form of

$$W(t, f) = \frac{1}{2\pi} \int x(t - \tau/2) x^*(t + \tau/2) e^{-jf\tau} \; d\tau. \tag{5}$$

If signal $x(t)$ has more than one time-frequency component, its WVD will have cross-terms. However, given the decomposition of $x(t)$ in terms of Gabor atoms which consist of a single time-frequency component, $WVG_{\gamma_i}(t)$ in (4) will not have a cross-term interference. As a result, TFM $V(t, f)$, can be considered as an accurate representation of the spectrogram of the signal. Since only first T atoms are used, less significant time-frequency components are filtered out and the desired structural property of the energy distribution is captured in $V(t, f)$. Then, the Non-Negative Matrix Factorization (NFM) is applied to $V(t, f)$ to obtain a more compact representation in terms of time and frequency:

$$V = WH, \tag{6}$$

where W and H capture the frequency and temporal structures of each component, respectively. One can reduce the redundant information in $V(t, f)$ by decomposing it into fewer components. Finally, the following four features are extracted.

1) *Joint TF moments.* The p^{th} temporal and q^{th} spectral moments are defined as

$$MO_{h_j}^{(p)} = \log_{10} \sum_n (n - \mu_{h_j})^p h_j(n), \tag{7}$$

$$MO_{w_j}^{(q)} = \log_{10} \sum_n (n - \mu_{w_j})^q w_j(n). \tag{8}$$

2) *Sparsity.* The measure of sparseness of temporal and spectral structures help in distinguishing between transient and continuous components. They are defined as

$$S_{h_j} = \log_{10} \frac{\sqrt{N} - \Big(\sum_n h_j(n)\Big) / \sqrt{\sum_n h_j^2(n)}}{\sqrt{N} - 1} \tag{9}$$

$$S_{w_j} = \log_{10} \frac{\sqrt{N} - \Big(\sum_n w_j(n)\Big) / \sqrt{\sum_n w_j^2(n)}}{\sqrt{N} - 1} \tag{10}$$

3) *Discontinuity.* The abrupt changes in the structure of temporal and spectral components are measured by the following parameters:

$$D_{h_j} = \log_{10} \sum_n h'^2_j(n), \tag{11}$$

$$D_{w_j} = \log_{10} \sum_n w'^2_j(n), \tag{12}$$

where $h'_j(n)$ and $w'_j(n)$ are the first order derivatives of temporal and spectral components, respectively.

4) *Coherency.* The coherency of the MP decomposition of a given signal, $x(t)$, can be evaluated as

$$CMP = \log_{10} \frac{\sum_{t=2}^{T} \alpha_t - \alpha_{t-1}}{E_x}, \tag{13}$$

where E_x is the total energy of signal $x(t)$.

Finally, LDA is used for classification.

There are justifications to the approach proposed in [10]. First, the WVD is a quadratic representation and so is energy (and in turn the spectrogram). By using the WVD of a single component, one obtains a cross-term free estimate of the spectrogram by retaining all useful properties of the WVD while leaving out its drawback. Second, the NMF yields a compact pair of vectors which contain important time-frequency components in the signal. Hence, features derived from these components tend to be characteristics of the underlying signal. When compared to MP-Gabor features, the first and second order moments estimated with this method are more reliable.

On the other hand, there are several weaknesses in this approach. First, there might be a problem with the discontinuity measure. The NMF results in non-unique decomposition. An intuitive initialization based on signal properties was adopted in [10]. However, it is not guaranteed that the discontinuity measure would be stable for signals of the same class as the order of of spectral and temporal components in vectors W and H affect this measure. It would be better to sort the components before taking the first derivative of these quantities. Second, its computational complexity is way too high. One needs to perform the MP decomposition of a 3-second signal sampled at $F_s = 22.05$ kHz up to 1000 iterations. Moreover, all possible discrete points of scale, frequency, location and orientation parameters are needed. Given these conditions, each iteration would require about $(6F_s + 1)M$ operations, where M is the number of atoms in the Gabor dictionary. The length of a 3-second signal $x(t)$ is $3F_s$, and an over-complete dictionary with at least $M = 4 \times 3F_s$ should be used. As a result, the total number of operations needed at each iteration would be about $72F_s^2 \approx 1.58$ million operations. It is desirable to implement the algorithm using the sub-framing processing scheme, yet this will result in a distorted estimate of long-term variations.

In [11], Ghoraani and Krishnan applied a nonlinear classifier called the Discriminant Cluster Selection (DSS) to the time-frequency features in [10]. The DSS uses both unsupervised

and supervised clustering methods. First, all features, irrespective of their true classes, undergo an unsupervised clustering scheme. Resulting clusters are subsequently categorized as *discriminant* or *common* clusters. Discriminant clusters are dominated with majority membership from one single class while common clusters house features from all or multiple classes with no obvious champion-class. For a test signal, all features are first extracted from the signal. Then, each feature's membership is determined. Features belonging to common clusters are ignored. The final decision for a test signal is made based on the labels of discriminant clusters. Two schemes; namely, hard and soft/fuzzy clustering, are used in the last step. The crux of this algorithm is that it determines discriminant sub-spaces in the entire feature space. Each discriminant region is assigned to a single class. Given that a single test signal is represented by multiple features, its final labeling is done based on the cluster-membership relationship of its discriminant features.

The spectrogram offers a tool for visually analyzing the time-frequency distribution of an audio signal. This has inspired the development of visual features derived from the spectrogram of music signals [12], [43], [44]. The original application in [43] was texture classification, yet the plausible use for music instrument classification was mentioned. Souli and Lachiri subsequently used this method for ESR in [26]. They also proposed another set of nonlinear features in [27]. In [27], non-linear visual features are extracted from the log-Gabor filtered spectrogram. The log-Gabor filtering is often used in image feature extraction. One polar representation of the log-Gabor function in the frequency domain is given by

$$G(r,\theta) = G_{radial}(r)G_{angular}(\theta), \tag{14}$$

where

$$G_{radial}(r) = e^{-\log(r/f_0)^2/2\sigma_r^2} \tag{15}$$

$$G_{angular}(\theta) = e^{-(\theta/\theta_0)^2/2\sigma_\theta^2} \tag{16}$$

are frequency responses for the radial and the angular components, respectively, f_0 is the center frequency of the filter, θ_0 is the orientation angle of the filter, and σ_r^2 and σ_θ^2 are the scale and the angular bandwidths, respectively. This method extracted features from the log-Gabor filtered spectrogram (instead of the raw spectrogram). Since no performance comparison was made between features obtained from the log-Gabor filtered spectrogram and the raw spectrogram in [26], the advantages and shortcomings of this approach need to be explored furthermore.

V. Conclusion and Future Work

We did an in-depth survey on recent developments in the ESR field in this paper. Existing ESR methods can be categorized into two types: stationary and non-stationary ESR techniques. The stationary ESR techniques are dominated by spectral features. While these features are easy to compute, there are limitations in the modeling of non-stationary sounds. The non-stationary ESR techniques obtain features derived from the wavelet transform, the sparse representation and the spectrogram. Wavelet based methods give results comparable to stationary methods. Sparse representation and spectrogram based methods in general perform better. MFCC features are often combined with one or more features to boost classification accuracy furthermore. While the non-stationary methods give improved performance, they are often computationally expensive.

Finally, we would like to point out the following three future research and development directions.

- Database Construction and Performance Benchmarking
  One major problem in the ESR field is the lack of a universal database. Each paper in this field presents its results with its own dataset consisting of an arbitrary number of environmental sound classes from various sources collected from the Internet. In the absence of a standard database, it is difficult to conduct a quantitative comparison of various approaches. Baseline classifiers with Mel-Frequency Cepstral Coefficients (MFCC) are often used to benchmark the performance of a new algorithm. However, due to significant differences in datasets of any two papers, such a performance benchmarking is futile. In solving real world problems, application-specific databases are more relevant. Also, there are numerous kinds of environmental sounds yet there is no standard taxonomy for them. Potamitis and Ganchev [21] made an effort to classify sounds to various categories from the application perspective. However, this problem is far from completion.
- Features Analysis
  In this paper, we attempted to give a qualitative overview of the recently developed ESR features. However, a deeper analysis of these features along with experimentation is needed. It is crucial to examine the performance of each of these features against a common dataset, thereby brining to light their ability to represent well-defined or even latent characteristics of sounds. For example, timbre, pitch, loudness etc. are well defined properties of an audio signal. On the other hand, often there are unidentified signal properties shared among several classes which might not be immediately apparent like those between sounds of machine-gun, foot-steps, typewriter, etc. Rigorous analysis of multiple features over a large data-set can be used to identify sub-groups of sounds which share one or more of such latent phenomenon, thereby empowering us with better feature selection strategies.
- Ensemble-based ESR
  A set of features with simplicity of stationary methods and accuracy of non-stationary methods is still a puzzle piece. Moreover, considering the numerous types of environmental sounds, it is hard to fathom a single set of features suitable for all sounds. Another problem with using a single set of features is that different features need different processing schemes, and hence several meaningful combination of features, that would

be otherwise functionally complementary to each other, are incompatible in practice. This school of though, and success stories of [2], [29], [40] directly leads us to ensemble learning methods. Instead of learning/training a classifier for single set of features, we can use multiple classifiers (experts) targeting different aspects of signal characteristics by using a set of complementary features. Unfortunately, there is no best way to design an ensemble framework. Hence, considerable effort is needed on this front.

REFERENCES

[1] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K.-H. Tauchert, and K.-H. Frommolt, "Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1524–1534, 2010.

[2] R. M. Bell, Y. Koren, and C. Volinsky, "The bellkor solution to the Netflix prize," *KorBell Teams Report to Netflix*, 2007.

[3] J. Chen, A. H. Kam, J. Zhang, N. Liu, and L. Shue, "Bathroom activity monitoring based on sound," in *Pervasive Computing*. Springer, 2005, pp. 47–61.

[4] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1142–1158, 2009.

[5] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, "Where am I? Scene recognition for mobile robots using audio features," in *Multimedia and Expo, 2006 IEEE International Conference on*. IEEE, 2006, pp. 885–888.

[6] M. Cowling and R. Sitte, "Comparison of techniques for environmental sound recognition," *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2895–2907, 2003.

[7] M. Cristani, M. Bicego, and V. Murino, "Audio-visual event recognition in surveillance video sequences," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 257–267, 2007.

[8] J. D. Deng, C. Simmermacher, and S. Cranefield, "A study on feature analysis for musical instrument classification," *IEEE Trans. Syst., Man, Cybern. B*, vol. 38, no. 2, pp. 429–438, 2008.

[9] S. Duan, J. Zhang, P. Roe, and M. Towsey, "A survey of tagging techniques for music, speech and environmental sound," *Artificial Intelligence Review*, pp. 1–25, 2012.

[10] B. Ghoraani and S. Krishnan, "Time–frequency matrix feature extraction and classification of environmental audio signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2197–2209, 2011.

[11] ——, "Discriminant non-stationary signal features' clustering using hard and fuzzy cluster labeling," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, p. 250, 2012.

[12] A. Ghosal, R. Chakraborty, B. C. Dhara, and S. K. Saha, "Song / instrumental classification using spectrogram based contextual features," in *Proceedings of the CUBE International Information Technology Conference*. ACM, 2012, pp. 21–25.

[13] B.-j. Han and E. Hwang, "Environmental sound classification based on feature collaboration," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*. IEEE, 2009, pp. 542–545.

[14] M. Karbasi, S. Ahadi, and M. Bahmanian, "Environmental sound classification using spectral dynamic features," in *Information, Communications and Signal Processing (ICICS) 2011 8th International Conference on*. IEEE, 2011, pp. 1–5.

[15] P. Khunarsal, C. Lursinsap, and T. Raicharoen, "Very short time environmental sound classification based on spectrogram pattern matching," 2013, (in press). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025513003113

[16] H. Liu, H. Motoda, R. Setiono, and Z. Zhao, "Feature selection: An ever evolving frontier in data mining," in *Proc. The Fourth Workshop on Feature Selection in Data Mining*, vol. 4, 2010, pp. 4–13.

[17] D. Mitrović, M. Zeppelzauer, and C. Breiteneder, "Features for content-based audio retrieval," *Advances in computers*, vol. 78, pp. 71–150, 2010.

[18] G. Muhammad, Y. A. Alotaibi, M. Alsulaiman, and M. N. Huda, "Environment recognition using selected MPEG-7 audio features and Mel-Frequency Cepstral Coefficients," in *Digital Telecommunications (ICDT), 2010 Fifth International Conference on*. IEEE, 2010, pp. 11–16.

[19] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 2. IEEE, 2002, pp. II–1941.

[20] J. Pickens, "A survey of feature selection techniques for music information retrieval," 2001.

[21] I. Potamitis and T. Ganchev, "Generalized recognition of sound events: Approaches and Applications," in *Multimedia Services in Intelligent Environments*. Springer, 2008, pp. 41–79.

[22] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[23] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: a survey," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 133–141, 2006.

[24] R. Sitte and L. Willets, "Non-speech environmental sound identification for surveillance using self-organizing-maps," in *Proceedings of the Fourth conference on IASTED International Conference: Signal Processing, Pattern Recognition, and Applications*, ser. SPPR'07. Anaheim, CA, USA: ACTA Press, 2007, pp. 281–286. [Online]. Available: http://dl.acm.org/citation.cfm?id=1331978.1332027

[25] S. Sivasankaran and K. Prabhu, "Robust features for environmental sound classification," in *Electronics, Computing and Communication Technologies (CONECCT), 2013 IEEE International Conference on*, 2013, pp. 1–6.

[26] S. Souli and Z. Lachiri, "Environmental sounds classification based on visual features," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer, 2011, pp. 459–466.

[27] ——, "Environmental sounds spectrogram classification using log-gabor filters and multiclass support vector machines," *arXiv preprint arXiv:1209.5756*, 2012.

[28] F. Su, L. Yang, T. Lu, and G. Wang, "Environmental sound classification for scene recognition using local discriminant bases and HMM," in *Proceedings of the 19th ACM international conference on Multimedia*. ACM, 2011, pp. 1389–1392.

[29] A. Töscher, M. Jahrer, and R. M. Bell, "The bigchaos solution to the Netflix grand prize," *Netflix prize documentation*, 2009.

[30] E. Tsau, S.-H. Kim, and C.-C. J. Kuo, "Environmental sound recognition with CELP-based features," in *Signals, Circuits and Systems (ISSCS), 2011 10th International Symposium on*. IEEE, 2011, pp. 1–4.

[31] K. Umapathy, S. Krishnan, and R. K. Rao, "Audio signal feature extraction and classification using local discriminant bases," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1236–1246, 2007.

[32] M. Vacher, F. Portet, A. Fleury, and N. Noury, "Challenges in the processing of audio channels for ambient assisted living," in *e-Health Networking Applications and Services (Healthcom), 2010 12th IEEE International Conference on*. IEEE, 2010, pp. 330–337.

[33] X. Valero and F. Alías, "Classification of audio scenes using narrow-band autocorrelation features," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*. IEEE, 2012.

[34] ——, "Gammatone wavelet features for sound classification in surveillance applications," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*. IEEE, 2012, pp. 1658–1662.

[35] L. Van der Maaten, E. Postma, and H. Van den Herik, "Dimensionality reduction: A comparative review," *Journal of Machine Learning Research*, vol. 10, pp. 1–41, 2009.

[36] T. Virtanen and M. Helén, "Probabilistic model based similarity measures for audio query-by-example," in *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*. IEEE, 2007, pp. 82–85.

[37] J.-C. Wang, H.-P. Lee, J.-F. Wang, and C.-B. Lin, "Robust environmental sound recognition for home automation," *IEEE Trans. Autom. Sci. Eng.*, vol. 5, no. 1, pp. 25–31, 2008.

[38] J.-C. Wang, J.-F. Wang, K. W. He, and C.-S. Hsu, "Environmental sound classification using hybrid SVM/KNN classifier and MPEG-7 audio low-level descriptor," in *Neural Networks, 2006. IJCNN'06. International Joint Conference on*. IEEE, 2006, pp. 1731–1735.

[39] F. Weninger and B. Schuller, "Audio recognition in the wild: static and dynamic classification on a real-world database of animal vocalizations,"

in *acoustics, speech and signal processing (ICASSP), 2011 IEEE international conference on*. IEEE, 2011, pp. 337–340.

[40] M. Wu, "Collaborative filtering via ensembles of matrix factorizations," in *Proceedings of KDD Cup and Workshop*, vol. 2007, 2007.

[41] N. Yamakawa, T. Kitahara, T. Takahashi, K. Komatani, T. Ogata, and H. G. Okuno, "Effects of modelling within-and between-frame temporal variations in power spectra on non-verbal sound recognition," in *Proc. 2010 International Conference on Spoken Language Processing, Makuhari*. Citeseer, 2010, pp. 2342–2345.

[42] N. Yamakawa, T. Takahashi, T. Kitahara, T. Ogata, and H. G. Okuno, "Environmental sound recognition for robot audition using Matching-Pursuit," in *Modern Approaches in Applied Intelligence*. Springer, 2011, pp. 1–10.

[43] G. Yu and J.-J. Slotine, "Fast wavelet-based visual classification," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, 2008, pp. 1–5.

[44] ——, "Audio classification from time-frequency texture," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 1677–1680.