Reduced-Rank Spectra and Minimum-Entropy Priors as Consistent and Reliable Cues for Generalized Sound Recognition

Michael A. Casey

MERL, Cambridge Research Laboratory
casey@merl.com

Abstract

We propose a generalized sound recognition system that uses reduced-dimension log-spectral features and a minimum entropy hidden Markov model classifier. The proposed system addresses the major challenges of generalized sound recognition—namely, selecting robust acoustic features and finding models that perform well across diverse sound types. To test the generality of the methods, we sought sound classes consisting of time-localized events, sequences, textures and mixed scenes. In other words, no assumptions on signal composition were imposed on the corpus.

Comparison between the proposed system and conventional maximum likelihood training showed that minimum entropy models yielded superior performance in a 20-class recognition experiment. The experiment tested discrimination between speech, non-speech utterances, environmental sounds, general sound effects, animal sounds, musical instruments and commercial music recordings.

1. Introduction

1.1. Generalized Sound Recognition

There are many uses for generalized sound recognition in audio applications. For example, robust speech / non-speech classifiers may be used to enhance the performance of automatic speech recognition systems, and classifiers that recognize ambient acoustic sources may provide signal-to-noise ratio estimates for missing-feature methods. Additionally, audio and video recordings may be indexed and searched using the classifiers and model state-variables utilized for fast query-by-example retrieval tasks from large general audio databases.

1.2. Previous Work

With each type of classifier comes the task of finding robust features that yield classifications with high accuracy on novel data sets. Previous work on non-speech audio classification has addressed recognition of audio sources using ad-hoc collections of features that are tested and fine-tuned to a specific classification task.

Such audio classification systems generally employ front-end processing to encode salient acoustic information; such as fundamental frequency, attack time and spectral centroid. These features are often subjected to further analysis to find an optimal set for a given task such as speech/music discrimination, musical instrument identification and sound effects recognition, [1][2][3]. Whilst each of these systems performs satisfactorily in their own right, they do not generalize beyond their intended applications due to the prior assumptions on the structure and composition of the input signals. For example, fundamental frequency assumes periodicity and, along with the spectral centroid, also assumes that the observable signal was produced by a single source.

Here, we are concerned with general methods that can be uniformly applied to diverse source classification tasks with accurate performance, which is the goal of generalized sound recognition (GSR). An acceptable criterion for GSR performance is >90% recognition for a multi-way classifier tested on novel data.

2. Maximally Informative Features

Machine learning systems are dependent upon the choice of representation of the input data. A common starting point for audio analysis is frequency-domain conversion using basis functions. The complex exponentials used by the Fourier transform form such a basis and yield complete representations of spectral magnitude information. The advantage of this complete spectral basis approach is that no assumptions are made on signal composition.

However, this representation consists of many dimensions and yields a high degree of correlation in the data. This renders much of the data redundant and therefore requires greater effort to be expended in parameter inference for statistical models. In many cases the redundancy also creates problems with numerical stability during training and adversely affects model performance during recognition.

To understand why such representations are problematic consider that higher dimensional populations of samples are more sparsely distributed across each dimension. This encourages over-fitting of the available data points, thus decreasing the reliability of density estimates. In contrast, a low dimensional representation of the same population yields a more densely sampled distribution from which parameters are more accurately inferred.

2.1.1. Independent Subspace Analysis

To address the problems of dimensionality and redundancy, whilst keeping the benefits of complete spectral representations, we use projection to low-dimensional subspaces via reduced-rank spectral basis functions. It is assumed that much of the information in the data occupies a subspace, or manifold, that is embedded in the larger spectral data space. A number of methods exist that yield maximally informative subspaces multivariate data; such as, local-linear embedding, non-linear principal components analysis, projection pursuit and independent component analysis. It has been shown that these algorithms form a family of closely

related algorithms that use information maximization to find salient components of multivariate data, [4][5][6].

We use independent subspace analysis (ISA) for extracting statistically independent reduced-rank features from spectral coefficients. ISA has previously been used for scene analysis and source separation from single-channel mixtures, [7]. The singular value decomposition (SVD) is used to estimate a new basis for the data, and the right singular basis functions are cropped to yield fewer basis functions that are then passed to independent component analysis (ICA). The SVD transformation produces de-correlated, reduced-rank features and the ICA transformation imposes the additional constraint of minimum mutual information between the marginal components of the output features. The resulting representation consists of the complete data projected onto a lower-dimensional subspace with marginal distributions that are as statistically independent as possible.

2.1.2. Independent Subspace Extraction

To extract reduced-rank spectral features a log-ferquency power spectrum was initially computed with a hamming window of length 20ms advanced at 10ms intervals. Frequency channels were logarithmically spaced in ¼-octave bands spaced between 62.5Hz and 8kHz. The resulting log-frequency power spectrum was converted to a decibel scale and each spectral vector was constrained to unit L2-norm, thus yielding spectral shape coefficients. The full-rank features for each frame *l*, consisted of both the L2-norm gain value $r^{(l)}$ and unit spectral shape vector $\hat{\mathbf{x}}^{(l)}$:

$$r^{(l)} = \sqrt{\sum_{k=1}^{N} \left(10 \log_{10} \left\{x_k{}^{(l)}\right\}\right)^2}, \qquad (1)$$

and

$$\hat{\mathbf{x}}^{(l)} = \frac{10 \log_{10} \left\{\mathbf{x}^{(l)}\right\}}{\left\|r^{(l)}\right\|}, \quad 1 \le l \le M \qquad (2)$$

where *N* was the number of spectral coefficients and *M* was the frame count. The next step was to extract a subspace using the singular value decomposition. To yield a statistically independent basis we used a reduced-rank set of SVD basis functions and applied a linear transformation $\mathbf{W}$ obtained by independent component analysis, see Figure 1. The resulting features were the product of the full-rank observation matrix $\mathbf{X}$, the dimension-reduced SVD basis functions $\mathbf{V}_\rho$ and the ICA transformation matrix $\mathbf{W}$

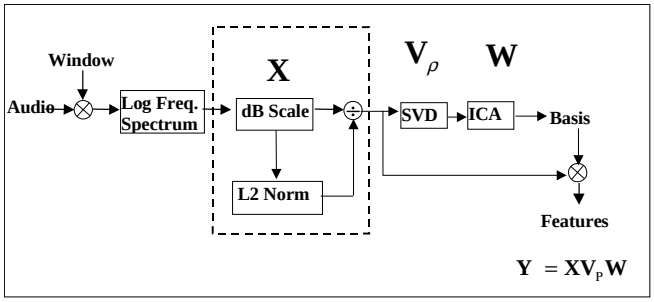
$$\mathbf{Y} = \mathbf{X}\mathbf{V}_\rho\mathbf{W}. \qquad (3)$$

The proportion of information retained for reduced-rank features of dimension ρ is given by:

$$I_\rho = \frac{1}{\sum_i \sigma_i} \sum_{j=1}^{\rho} \sigma_j, \qquad (4)$$

where *N* is the total number of SVD basis functions and σ_i are the singular values.

Figure 1. Block diagram of independent subspace feature extraction.



To see how the representation affects classification we trained two HMM classifiers; one using the complete spectral information as given by Equations (1) and (2), and the other using the reduced-rank form of Equation (3). Table 1 shows results for the classifiers tested on musical instrument classification. The HMMs trained with the reduced rank form of spectral data performed significantly better than the HMMs trained using full-rank spectra. Analysis of variance indicates that the results are significant with $p < .00018$. The details of classifier training and testing are discussed in the rest of the paper.

Table 1 Performance statistics of 7-class classifier trained on direct spectra and reduced-rank spectra.

	Direct Spectra		**Reduced-Rank Spectra**	
Class	**# Hit**	**# Miss**	**# Hit**	**# Miss**
Flute	1	3	4	0
Piano	5	0	5	0
Cello	5	1	5	1
Cor Anglais	1	3	4	0
Guitar	0	3	3	0
Trumpet	4	1	5	0
Violin	4	2	6	0
Totals	**20**	**13**	**32**	**1**
Performance	**60.61%**		**92.65%**	

3. Minimum Entropy HMMs

3.1. HMM Classifiers

Hidden Markov models consist of three components; an initial state distribution $\pi_i = P(q_1 = i)$ with $q_t \in \{1 \ldots N\}$, a state transition matrix $A_{ij} = P(q_t = j \mid q_{t-1} = i)$ and the observation density function $b_j(\mathbf{y}) = P(\mathbf{y} \mid q_t = j)$ for each state. Continuous HMMs set $b_j(\mathbf{y})$ to a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_j$ and covariance matrix $\mathbf{K}_j$, giving $B_j = \{\boldsymbol{\mu}_j, \mathbf{K}_j\}$ for each state. We use the

notation $\theta_j = \{A_j, B_j, \pi_j\}$ to identify the parameters for hidden Markov model *j*.

Classification proceeds using the viterbi algorithm which estimates the most likely state sequence $Q = \{q_1, q_2, \cdots, q_M\}$ given observed data $Y = \{y_1, y_2, \cdots, y_M\}$ and model parameters θ_j. The output is a sequence of states and a likelihood that measures the probability of each model given the observed data. The HMM classifier chooses the model with the maximum likelihood score, amongst *L* competing models

$$L^* \equiv \arg\left\{ \max_{1 \le l \le L} \left[P(Y, Q \mid \theta_l) \right] \right\} \tag{5}$$

3.2. HMM Parameter Inference

To infer a model from data, we seek parameters that maximize the probability of the posterior distribution given the training data for a class. To do this we invoke Baye's rule

$$P(\theta \mid y) = \frac{P(y|\theta)P(\theta)}{P(y)} \tag{6}$$

where the likelihood $P(y|\theta)$ measures the accuracy in modeling the data, and the prior $P(\theta)$ measures the consistency of the model with our background knowledge. Conventional HMMs use maximum likelihood estimation with uniform priors. However, we often possess knowledge that should lead us to prefer some model configurations to others, thus we assume a prior on the likely form of the model. But what is this background knowledge and how can it be quantified?

3.3. Maximum Structure Priors

Brand [8,9] proposed an elegant way to include background knowledge about structured data. The goal is to maximize the information content of the model such that those parameters with low probability are forced to zero in favor of parameters with high probabilities. In Bayesean terms, parameters that do not reduce the uncertainty are improbable, thus only those parameters that reduce the entropy of the model should be considered.

For multinomial parameters consisting of *N* conditional probabilities $\boldsymbol{\theta} = \{\theta_1, \cdots \theta_N\}$ the entropic prior is

$$P_e(\theta) = e^{-H(\theta)} = \exp\left[\sum_{i=1}^{N} \theta_i \log \theta_i \right] = \prod_{i=1}^{N} \theta_i^{\theta_i} \tag{7}$$

substituting the entropic prior into Equation (6) yields

$$P_e(\boldsymbol{\theta} \mid y) \propto \frac{\prod_{i=1}^{N} \theta_i^{\theta_i} \prod_{i=1}^{N} \theta_i^{y_i}}{P(\mathbf{y})} \propto \prod_{i=1}^{N} \theta_i^{\theta_i + y_i} \tag{8}$$

The maximum *a-posteriori* estimate for the parameters is obtained by setting the derivative of the log-posterior term to zero which yields

$$0 = 1 + \frac{\mathbf{y}_i}{\theta_i} + \log \theta_i + \lambda \tag{9}$$

where λ is a Lagrange multiplier to ensure that $\sum_i^N \theta_i = 1$. This equation corresponds to the E-step in *expectation maximization* (EM). The M-step is given by

$$\theta_i = \frac{-\mathbf{y}_i}{\mathrm{W}\left(-\mathbf{y}_i e^{1+\lambda}\right)} \tag{10}$$

where W is the Lambert W function, a multi-valued inverse function satisfying $W(x)e^{W(x)} = x$. Each row of the transition matrix of an HMM is a multinomial, as is the initial state distribution, thus they can be estimated with the above formulae by iterating between the E-step and the M-step.

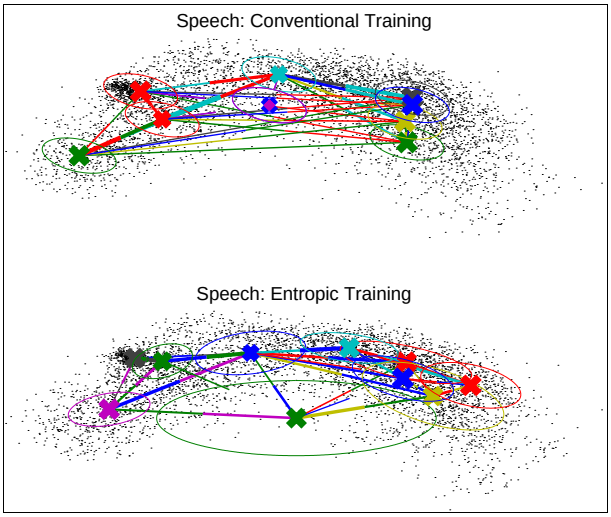
The entropic prior for Gaussian distributions is $P_e(\boldsymbol{\mu}, \mathbf{K}) \propto |\mathbf{K}|^{1/2}$. The maximum *a-posteriori* (MAP) estimate of these parameters is obtained by setting the gradient of the log posterior to zero and an approximate solution is derived

$$\hat{\mathbf{K}} = \frac{\sum_n^N \mathbf{y}_n \mathbf{y}_n^T}{N + Z} \tag{11}$$

where $Z = T - T0$ is a function of the temperature for deterministic annealing. T is the current temperature and T0 the initial temperature in the annealing schedule.

To illustrate the utility of minimum entropy priors, Figure 2 shows the difference between conventional and minimum entropy models trained on the same set of speech data. The structure of the entropic model appears somewhat sparser and easier to interpret than that of the conventionally trained model.

Figure 2. Conventional and minimum entropy HMMs viewed from two dimensions of speech data. Gaussian states are represented by ellipses and transition probabilities are represented by line thickness.



4. Results

We trained a collection of 20 hidden Markov models for diverse sound classes using both conventional maximum likelihood estimation and maximum *a-posteriori* training with minimum entropy priors. The data corpus consisted of 1000 sounds taken from various sources including the "TIMIT" speech database, the "Pro Sonus" musical instrument sample library, the "Sound Ideas" general sound effects library and several hours of music recordings from commercial compact discs.

The duration of the sound sequences was between 1 and 60 seconds with no restrictions on segmentation boundaries. The data was divided by a 70%/30% split into training and testing sets. For some sound classes the split was random. For others, such as speech classes, care was taken to make sure that no speaker appeared in both the training and testing sets.

For each class, a reduced-rank basis was extracted from log-spectral gain and shape coefficients as described in §2.1.2. The full-rank training data matrix was projected against the basis functions to yield a 10-dimensional observation matrix that was used for HMM training. For conventional training, the state means were initialized using the *k-means* algorithm with randomly generated full covariance matrices. For minimum entropy training, the state means were initialized by distributing them uniformly across the data and the covariances were initialized with the full data covariance.

To test the classifiers, the novel data was presented to each HMM and the model with the highest likelihood was selected using Equation 5. The results of classification for the two sets of HMMs are shown in Table 2.

Correct classifications are counted as *Hits*, and incorrect classifications are indicated as *Misses*. The performance for each classifier was measured as the percentage of correct classifications for the entire set of 275 test sequences. The results indicate that both models performed well for such a diverse set of classes; however, the models trained with minimum entropy priors yielded significantly better results overall; analysis of variance indicated that the results are significant with $p < 0.0388$.

5. Conclusions

We have shown that reduced-rank spectral data and HMM classifiers with minimum entropy priors yield multi-way classifiers that perform significantly better than conventionally trained HMM classifiers for generalized sound recognition problems.

In future work, we will evaluate the effects of hierarchical and factorial model combinations and investigate applications of these architectures to mixtures of sound classes. The results reported herein indicate some promise for such extensions.

Table 2 Recognition results for conventional and minimum entropy 20-way HMM classifies.

	Conventional Training		**Entropic Training**	
Class	**# Hit**	**# Miss**	**# Hit**	**# Miss**
Speech:Female	39	1	39	1
Speech:Male	69	1	68	2
Music	12	0	12	0
Bird Calls	9	6	12	3
Applause	6	0	5	1
DogBarks	14	1	15	0
Explosions	6	1	7	0
FootSteps	11	0	10	1
Glass Smash	8	5	12	1
Gunshots	13	0	12	1
Shoe Squeaks	2	2	4	0
Laughter	13	5	17	1
Telephones	14	4	12	6
Flute	2	2	4	0
Piano	3	2	5	0
Cello	5	1	6	0
Cor Anglais	2	2	4	0
Guitar	1	2	3	0
Trumpet	3	2	4	1
Violin	6	0	5	1
Totals	**238**	**37**	**256**	**19**
Performance	**86.55%**		**93.09%**	

6. References

[1] Wold, E., Blum, T., Keislar, D., and Wheaton, J., Content-based classification, search and retrieval of audio. *IEEE Multimedia*, pp.27-36, Fall 1996.

[2] Martin, K. D. and Kim, Y. E., Musical instrument identification: a pattern-recognition approach. In *Proc. 136th Meeting of the Acoustical Society of America*, VA, 1998.

[3] Zhang, T. and Kuo, C., Content-based classification and retrieval of audio. *Proc. Conference on Advanced Signal Processing Algorithms, Architectures and Implementations VIII*, San Diego, CA, 1998.

[4] Bell, A. J. and Sejnowski, T.J. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129-1159, 1995.

[5] Cardoso, J.F. and Laheld, B.H., Equivariant adaptive source separation. *IEEE Trans. On Signal Processing*, 4:112-114, 1996.

[6] Hyvarinen, A., Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. On Neural Networks*, 10(3):626-634, 1999.

[7] Casey, M.A., and Westner, A., Separation of mixed audio sources by independent subspace analysis. *Proceedings of the International Computer Music Conference*, ICMA, Berlin, 2000.

[8] Brand, M., Pattern discovery via entropy minimization. In *Proceedings, Uncertainty'99.* Society of Artificial intelligence and Statistics #7, FL, 1999.

[9] Brand, M., Structure discovery in conditional probability models via an entropic prior and parameter extinction. *Neural Comput.*, vol. 11, no. 5, pp. 1155-1183, 1999.