ANCHOR SPACE FOR CLASSIFICATION AND SIMILARITY MEASUREMENT OF MUSIC

*Adam Berenzweig and Daniel P.W. Ellis**

Columbia University
LabROSA, Electrical Engineering
New York, NY, U.S.A.

Steve Lawrence

NEC Research Institute, Inc.
Princeton, NJ, U.S.A.

ABSTRACT

This paper describes a method of mapping music into a semantic space that can be used for similarity measurement, classification, and music information retrieval. The value along each dimension of this *anchor space* is computed as the output from a pattern classifier which is trained to measure a particular semantic feature. In anchor space, distributions that represent objects such as artists or songs are modeled with Gaussian Mixture Models, and several similarity measures are defined by computing approximations to the Kullback-Leibler divergence between distributions. Similarity measures are evaluated against human similarity judgements. The models are also used for artist classification to achieve 62% accuracy on a 25-artist set, and 38% on a 404-artist set (random guessing achieves 0.25%). Finally, we describe a music similarity browsing application that makes use of the fact that anchor space dimensions are meaningful to users.

1. INTRODUCTION

A natural way of describing unfamiliar music is to relate it to musical categories that are well-known, such as genres or a single artist's work. For example, the All Music Guide describes Jeff Buckley as "Van Morrison meets Led Zeppellin" but more "folkie" and influenced by "lounge jazz"[1]. While we find it difficult to describe music in words, it is often easier to draw parallels to familiar *anchors*. Analogously, researchers working on content-based music retrieval have not yet reached consensus on how to automatically extract high-level, musically-relevant descriptive features from low-level audio features. This paper explores a way to automate the folk wisdom approach of describing things as combinations of well-known categories.

A high-level representation of music is desirable to combat the "curse of dimensionality" that plagues machine learning techniques. Presumably, information that humans use to describe music has already been filtered by some intermediate perceptual processes, and is quite succinct.

In the work described here, we train pattern classifiers to recognize musically-relevant classes such as genre or artist label, or more general categories such as male vs. female vocalists. Training data is labeled by hand or by using information gathered from the Internet. We then collect the output (posterior probabilities) from several such classifiers into a feature vector, the dimensions of which represent soft membership in one of the musically-meaningful anchor classes. The feature space, which we call *anchor space*, can be used for similarity measurement, classification, and clustering of music.

From a machine learning perspective, the anchor classifiers can be seen as nonlinear feature extractors, where the nonlinear function is obtained by machine learning techniques. Seen this way, this work is related to the work of Bollacker and Ghosh [1] on "supra-classifiers" and knowledge reuse. Slaney [2] uses a similar technique for content-based audio retrieval.

In the remainder of the paper, we describe anchor space and how it is constructed, use it to compute a music similarity measure, and use it for an artist classification task. Finally, in section 5 we show how it can be used in a music similarity browsing application.

2. ANCHOR SPACE

Anchor space is an M-dimensional Euclidean space in which each dimension represents soft membership in one of the M anchor classes. Points in anchor space are mapped from vectors of low-level audio features (here, we use mel-frequency cepstral coefficients) taken over a short time window. The nonlinear mapping is learned by a pattern classifier; in this work, we use neural networks.

The M neural networks are trained to recognize membership in the M anchor classes, which could represent genres, tempos, timbres, or any other musically meaningful category.

Points in anchor space are vectors of posterior probabil-

*Support from NEC Research Institute

[1] All Music Guide, http://www.allmusic.com

ities of membership in the anchor classes, given the input:

$$\big(p(\omega_i|x), .., p(\omega_M|x)\big),$$

where ω_i represents the i^{th} anchor class, and x is a vector of spectral coefficients, described below. In some of the experiments we use the neural net outputs before application of the typical softmax non-linearity, so that anchor space is not restricted to a [0,1] hypercube.

Because feature vectors are computed from short time segments of audio, an entire song induces a cloud of points in feature space. The cloud can be thought of as samples from a distribution that characterizes the song, and we can attempt to model that distribution using standard statistical techniques. Extending this idea, we can conceive of a distribution in feature space that characterizes an entire album, an artist's repertoire, a genre, or any other set of music. Modeling and comparing these distributions is the subject of section 3.

2.1. Training and using anchor models

We experimented with two different configurations of neural networks. In one configuration, each anchor class is recognized by a separate "one-vs-many" network. In the other configuration, we train a single M-way network. The difference is that the M-way network is trained discriminatively (strong activation of one anchor will suppress activation in all others), while with M separate classifiers, each dimension acts independently.

Note that these are merely two ways of handling what is essentially a multi-class machine learning problem. Rifkin and Whitman [3] explore a related problem.

For each anchor, a two-class neural network was trained to discriminate the anchor from all others. The first 20 mel-frequency cepstral coefficients (MFCCs) were computed on 32ms frames, overlapped by 16ms. The input vector to the neural net is a 200-dimensional vector formed by the 20 MFCCs plus 20 first-order deltas (the difference between the coefficients at time t and time $t + 1$), for 5 consecutive frames of context. The hidden layer size was set to 20 units for one-vs-many networks, and 100 units for M-way networks, based on initial experiments. The relatively small hidden unit size is meant to ensure that the networks are slightly undertrained, to encourage output activity even when the input only loosely matches the training data.

2.2. Choosing anchors

We would like anchor classes that provide full coverage of "music space" but are relatively uncorrelated. For these experiments, we simply hand-picked "canonical" artists and genres with this criterion in mind. From a database of popular music, which consists of over 1000 albums from about 400 artists, we chose one set of 24 artist anchors, and several sets of genre anchors (10 and 12 genres), plus two supplemental anchors (male/female vocalist and high/low fidelity) which were added to the 12 genres to make an augmented 14-anchor set.

The training set for the artist classes consists of one full album by that artist, usually two or more albums. To train genre anchors, we first selected several artists that, in our opinion, represent the genre particularly well.

This training process depends on our subjective choice of training examples. A more principled way to choose anchors, for example by using an information-theoretic criteria to minimize the correlation between anchor dimensions, is under investigation. One step we took in that direction was to prune out highly correlated dimensions: for example, in an early anchor set, Hard Rock and Punk were highly correlated, and so were merged into a single model.

3. SIMILARITY MEASURES: COMPARING CLOUDS

We can measure similarity between objects in anchor space (e.g., songs, artists, or albums) in order to classify them into categories, or for music information retrieval. Recall that an entire song induces a cloud of points in anchor space, which we think of as samples from a distribution. Therefore, it is natural to use models of probability distributions to model objects in anchor space. We choose Gaussian Mixture Models (GMMs) because of their ability to model complex multi-modal distributions and the ease with which they are trained using the Expectation-Maximization (EM) algorithm.

To measure similarity, we need a way of comparing clouds of points in anchor space. We would like to use the Kullback-Leibler divergence, which is the natural way to define distance between probability distributions. However, we run into difficulty because no closed form for the KL-divergence between GMMs is known [4].

We tried approximating the KL-divergence by sampling points from distribution P and then computing the likelihood of model Q given the samples. However, in high dimensions, the number of samples required to adequately characterize the distribution is large, and it becomes computationally prohibitive to use enough points to make this work in practice.

Instead, we reduce clouds to a single point (the centroid) and then take the Euclidean distance. This approach, though overly simple, is tractable and gives decent results. Further work using the earth-mover's distance [5] is in progress.

3.1. Evaluating Similarity Measures

Evaluation is difficult in this domain because there is no clear ground truth for music similarity [6]. We are forced to rely on subjective judgment and user tests.

We evaluate the measures by comparing them with data gathered from human subjects in a survey about artist similarity. We presented subjects a list of 10 artists $(a_1, ..a_{10})$, and a single target artist a_t, and asked "Which of these artists is most similar to the target artist?" We interpret each response to mean that the chosen artist a_c is more similar to the target artist a_t than any of the other artists in the list $(a_1, ..a_{10})$, *if* the artists are known to the subject. More details are available in [6].

To evaluate an experimental similarity metric, we check how often the metric agrees with the human subjects. For each list $(a_1, ..a_{10})$ that was presented to a subject, we order it by similarity to the target a_t under the experimental metric. We then find the rank of the artist a_c chosen by the subject. For example, if the experimental metric agrees perfectly with the human subject, then the ranking of a_c will be 1 in every case, and a random ordering of the artists would produce an average ranking of 5.5. In practice, the ideal score of 1.0 is not possible because different users do not always agree about artist similarity; therefore, the ceiling is the single, consistent metric that best matched the survey data. For our data, this was computed to be 1.98.

First we compare several sets of anchor models: a set of 24 artist anchors, a set of 12 genre anchors, and a set of 14 anchors (consisting of the 12 genres plus two supplemental anchors: male/female voice and high/lo-fidelity). To determine the effect of dimensionality reduction alone, we also include a set of 12 "meaningless" anchors trained on randomly chosen songs.

The results are summarized in Table 1. As expected, the "meaningless" anchors perform almost as poorly as random guessing, showing that that similarity measures are better defined in terms of semantically meaningful features.

ank12-rand	ank12-g	ank14-g+	ank24-a
5.2	4.02	3.97	4.13

Table 1. Survey-based evaluation of anchor sets using D_{centroid}. The random baseline is 5.42, and the optimal ceiling is 1.98.

In [6], we describe several similarity measures based on other sources of human opinion such as preference data, community metadata (text from websites describing the artists, such as fan sites and music reviews), and expert opinion. Table 2 partly lists the results, with the addition of D_{centroid} computed on the 14-anchor set. The audio-based anchor measure outperforms randomness, WebText, and Pref, however it does not perform as well as the measure based on expert opinion.

D_{centroid}	Expert	Pref	WebText	Random	Ceiling
3.97	3.83	4.05	4.53	5.42	1.98

Table 2. Survey-based evaluation of D_{centroid} on ank14-g+ vs. measures derived from human opinion: expert opinion, user collections (Pref), and descriptive text found on the web. Details of the survey and measures are in [6].

4. CLASSIFICATION

Having defined anchor space and a similarity measure on it, we use it for an artist classification task. For each artist, we fit a Gaussian Mixture Model to the anchor space points induced by the music. The number of Gaussians was chosen as 5 (sensitivity to this parameter was not extensively examined). For each of 10 test query songs witheld from the training set, the total likelihood of the points in anchor space induced by the query song is computed for each candidate model, and the model with highest likelihood is declared the winner. The (log) likelihood of a GMM with parameters θ, given the observed set of samples $\mathcal{X} = (x_1, .., x_T)$ is

$$\begin{aligned} l(\theta|\mathcal{X}) &= \sum_{t=1}^{T} \log p(x_t|\theta) \\ &= \sum_{t=1}^{T} \log \sum_{k=1}^{K} \pi_{ik} \mathcal{N}(x_t|\mu_{ik}, \Sigma_{ik}), \end{aligned} \tag{1}$$

where $\theta = (\pi_1, .., \pi_K, \mu_1, .., \mu_K, \Sigma_1, .., \Sigma_K)$ are respectively the priors, means, and covariance matrices of the K mixture components, and $\mathcal{N}(x|\mu, \Sigma)$ is the Normal distribution.

There were several experimental conditions to examine the effect of preprocessing: the softmax non-linearity applied to the output of the anchor nets, and an online normalization step. In [7], we noted that classifiers trained on only a single album tend to learn overall spectral characteristics of the album (perhaps an artifact of the equalization applied during mastering). Online normalization attempts to remove constant channel effects by normalizing with a running mean and variance before presentation to the neural net.

Several tasks were used: a medium-size task of 404 artist classes; the set of 21 artists used in [7] and [8], for comparison of results (they achieved 65% and 50%, respectively); and a set of 25 artists for which we had at least three albums available, to explore the effect of online normalization to reduce the "album effect". Table 3 presents the results in terms of classification accuracy (percent of correctly classified songs). Set25 has two conditions: in set25-matched, the training set contains songs from all of

the albums in the test set; in set25-mismatched, one entire album is left out of the training set for testing. Online normalization boosts the accuracy on the mismatched set from 19.6% to 46.6% (130% relative improvement).

The best preprocessing settings (online normalization with linear network outputs) were used for the 404-artist task, resulting in 38% accuracy (random guessing would only achieve 0.25% accuracy). Note that the entire query song is used for classification. An experiment using only 30% of each song resulted in 32.9% accuracy.

	Anchor set		
test set	**linear**	**norm**	**norm+lin**
set21	23.9	47.1	48.5
set25-matched	53.8 (1.41)	53.0 (3.11)	62.6 (0.28)
set25-mismatched	19.6 (3.04)	40.4 (1.84)	46.0 (2.47)
set404			38.0
set404, 30%			32.9

Table 3. Classification accuracy by % songs correct. Numbers in parentheses are the standard deviations for cases where multiple validations sets were used.

5. MUSIC SIMILARITY BROWSING

One important property of anchor space is that the feature extractors produce meaningful dimensions that can be understood and manipulated by human users. This enables applications such as a music-space browsing tool. A web site has been constructed to demonstrate this application[2].

Users browse through anchor space by moving sliders that represent each of the dimensions, and the system then displays a list of artists or songs in the anchor space neighborhood.

The system currently contains music from the 400 artists in the evaluation set, as well as about 17,000 songs from new bands who have made their music available for free on the Internet. This demonstrates an important advantage of audio-based similarity methods, namely the ability to include new music for which no metadata yet exists.

In addition to demonstrating the usefulness of anchor models, the website is a way to gather more evaluation data from human subjects. Users are asked to give feedback about whether or not they agree with the system's similarity results. This data will be used to evaluate different similarity measures and choices of anchors.

6. CONCLUSION

We have presented a method for mapping perceptual space into a semantic attribute space, similarity measures in that space, and results from a classification task using the space. The similarity metric was evaluated against human similarity judgments, and shown to be comparable with a metric based on the opinion of human experts. For artist classification on a set of 404 artists (which is a significantly larger set than that of any published results we are aware of), accuracy is 38%. Online normalization improved results by 130% (relative) in cases where the test set is taken from different albums than the training set.

There is plenty of room for further development. The centroid-based similarity measure is overly simplistic, and better results may be obtained by using something more sophisticated like the earth-mover's distance. We also wish to personalize the similarity measure by finding transformations of attribute space that best account for a user's collection.

We have also used anchor models to construct a music similarity browsing system, in which the user can explore anchor space by moving sliders that represent the dimensions, that will be the subject of future research.

[2] http://www.playola.org

7. REFERENCES

[1] Kurt D. Bollacker and Joydeep Ghosh, "A supra-classifier architecture for scalable knowledge reuse," in *Proc. 15th Int. Conf. on Machine Learning*, 1998.

[2] M. Slaney, "Mixtures of probability experts for audio retrieval and indexing," in *ICME*. 2002.

[3] Brian Whitman and Ryan Rifkin, "Musical query-by-description as a multiclass learning problem," in *Proc. IEEE Multimedia Signal Processing Conf. (MMSP)*, December 2002.

[4] Nuno Vasconcelos, "On the complexity of probabilistic image retrieval," in *ICCV'01*. Vancouver, 2001.

[5] Y. Rubner, C. Tomasi, and L. Guibas, "A metric for distributions with applications to image databases," in *Proc. ICCV*, 1998.

[6] Daniel P.W. Ellis, Brian Whitman, Adam Berenzweig, and Steve Lawrence, "The quest for ground truth in musical artist similarity," 2002.

[7] Adam Berenzweig, Dan Ellis, and Steve Lawrence, "Using voice segments to improve artist classification of music," in *AES 22nd International Conference*, Espoo, Finland, 2002.

[8] Brian Whitman, Gary Flake, and Steve Lawrence, "Artist detection in music with minnowmatch," in *Proc. of the 2001 IEEE Workshop on Neural Networks for Signal Processing*. Falmouth, Massachusetts, 2001.