Learning Word Meanings and Descriptive Parameter Spaces from Music

Brian Whitman
MIT Media Lab
Music, Mind and Machine
Cambridge, MA USA
bwhitman@media.mit.edu

Deb Roy
MIT Media Lab
Cognitive Machines
Cambridge, MA USA
dkroy@media.mit.edu

Barry Vercoe
MIT Media Lab
Music, Mind and Machine
Cambridge, MA USA
bv@media.mit.edu

Abstract

The audio bitstream in music encodes a high amount of statistical, acoustic, emotional and cultural information. But music also has an important linguistic accessory; most musical artists are described in great detail in record reviews, fan sites and news items. We highlight current and ongoing research into extracting relevant features from audio and simultaneously learning language features linked to the music. We show results in a "query-by-description" task in which we learn the perceptual meaning of automatically-discovered single-term descriptive components, as well as a method of automatically uncovering 'semantically attached' terms (terms that have perceptual grounding.) We then show recent work in 'semantic basis functions' – parameter spaces of description (such as *fast ... slow* or *male ... female*) that encode the highest descriptive variance in a semantic space.

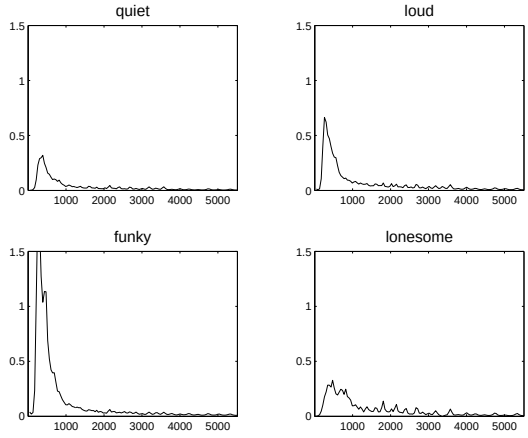

Figure 1: Mean spectral characteristics of four different terms uncovered by the spectral frame-based single term attachment system. Magnitude of frequency on the y-axis, frequency in Hz on the x-axis.

1 Introduction

What can you learn by listening to the radio all day? If the DJ was wordy enough, we argue that you can gain enough knowledge of the language of perception, as well as the grammar of description and the grammar of music.

Here we develop a system that uncovers descriptive parameters of perception completely autonomously. Relations between English adjectives and audio features are learned using a new 'severe multi-class' algorithm based on the support vector machine. Training data consists of music reviews from the Internet correlated echnology and Entertainment Media: Rights and Responsibilities with acoustic recordings of the reviewed music. Once trained, we obtain a perceptually-grounded lexicon of adjectives that may be used to automatically label new music. The predictive accuracy of the perceptual models are evaluated on unseen test music-review data samples. We consider terms with high predictive accuracy (i.e., that agree with word usage of musical reviews not used during training) to be well grounded. We extend our prior work by introducing a 'linguistic expert,' in the form of a lexical knowledge base that provides human-encoded symbolic knowledge about lexical relations. We apply lexical relations to well grounded adjectives to determine the perceptual correlates of opposition. This enables us to move from isolated word groundings to a gradation system by discovering the perceptual basis underlying lexical opposition of adjective pairs (*fast ... slow*, *hard ... soft*, etc.). Once we have uncovered these gradations, we effectively obtain a set of "semantic basis functions" which can be used to characterize music samples based on their perceptual projections onto these lexically determined basis functions.

Term	Precision	Term	Precision
acoustic	23.2%	annoying	0.0%
classical	27.4%	dangerous	0.0%
clean	38.9%	gorgeous	0.0%
dark	17.1%	hilarious	0.0%
electronic	11.7%	lyrical	0.0%
female	32.9%	sexy	1.5%
happy	13.8%	troubled	0.0%
romantic	23.1%	typical	0.0%
upbeat	21.0%	wicked	0.0%
vocal	18.6%	worldwide	2.8%

Table 1: Selected adjective terms and their weighted precision in predicting a description of as-yet 'unheard' music in the frame-based single term attachment system. The very low baseline and noisy ground truth contribute to low overall scores, but the difference between 'ungroundable' and high-scoring terms are significant– for example, the system cannot find a spectral definition of 'sexy.'

2 Background

In the general audio domain, work has recently been done (Slaney, 2002) that links sound samples to description using the labeled descriptions on the sample sets. In the visual domain, some work has been undertaken attempting to learn a link between language and multimedia. The lexicon-learning aspects in (Duygulu et al., 2002) study a set of fixed words applied to an image database and use a method similar to EM (expectation-maximization) to discover where in the image the terms (nouns) appear. (Barnard and Forsyth, 2000) outlines similar work. Regier has studied the visual grounding of spatial terms across languages, finding subtle effects that depend on the relative shape, size, and orientation of objects (Regier, 1996). Work on motion verb semantics include both procedural (action) based representations building on Petri Net formalisms (Bailey, 1997; Narayanan, 1997) and encodings of salient perceptual features (Siskind, 2001). In (Roy, 1999), we explored aspects of learning shape and color terms, and took first steps in perceptually-grounded grammar acquisition.

We refer to a word as "grounded" if we are able to determine reliable perceptual or procedural associations of the word that agree with normal usage. However, encoding single terms in isolation is only a first step in sensory-motor grounding. Lexicographers have traditionally studied lexical semantics in terms of lexical relations such as opposition, hyponymy, and meronymy (Cruse, 1986). We have made initial investigations into the *perceptual grounding of lexical relations*. We argue that *gradations* or linguistic parameter spaces (such as *fast ... slow* or *big ... small*) are necessary to describe high-dimensional perceptual input.

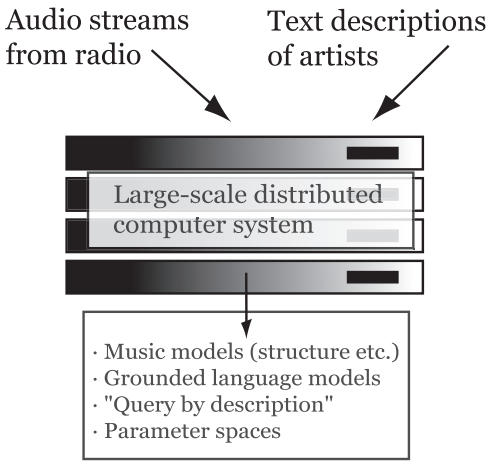

Figure 2: The "Radio, Radio" platform for autonomously learning language and music models. A bank of systems (with a distributed computing back-end connecting them) listens to multiple genres of radio streams and hones an acoustic model. When a new artist is detected from the metadata, our cultural representation crawler extracts language used to describe the artist and adds to our language model. Concurrently, we learn relations between the music and language models to ground language terms in perception.

Our first approach to this problem was in (Whitman and Rifkin, 2002), in which we learned the descriptions of music by a combination of automated web crawls for artist description and analysis of the spectral content of their music. The results for that work, which appear in Figure 1 and Table 1, show that we can accurately predict (well above an impossibly low baseline) a label on a held-out test set of music. We also see encouraging results in the set of terms that were accurately predicted. In effect we can draw an imaginary line in the form of a confidence threshold around our results and assign certain types of terms 'grounded' while others are 'ungroundable.' In Table 1 above, we note that terms like 'electronic' and 'vocal' that would appear in the underlying perceptual feature space get high scores while more culturally-influenced terms like 'gorgeous' and 'sexy' do not do as well. We have recently extended this work (Whitman et al., 2003) by learning parameters in the same manner. Just because we know the spectral shape of 'quiet' and 'loud' (as in Figure 1) we cannot infer any sort of connecting space between them unless we know that they are antonyms. In this work, we infer such gradation spaces through the use of a lexical knowledge base, 'grounding' such parameters through perception. As well, to capture important time-aware gradations such as *'fast...slow'* we introduce a new machine listening

np Term	Score	adj Term	Score
beth gibbons	0.1648	cynical	0.2997
trip hop	0.1581	produced	0.1143
dummy	0.1153	smooth	0.0792
goosebumps	0.0756	dark	0.0583
soulful melodies	0.0608	particular	0.0571
rounder records	0.0499	loud	0.0558
dante	0.0499	amazing	0.0457
may 1997	0.0499	vocal	0.0391
sbk	0.0499	unique	0.0362
grace	0.0499	simple	0.0354

Table 2: Top 10 terms (noun phrase and adjective sets) for the musical group 'Portishead' from community metadata.

representation that allows for far more perceptual generality in the time domain than our previous work's single frame-based power spectral density. Our current platform for retrieving audio and description is shown in Figure 2.

We acknowledge previous work on the computational study of adjectival scales as in (Hatzivassiloglou and McKeown, 1993), where a system could group gradation scales using a clustering algorithm. The polar representation of adjectives discussed in (Miller, 1990) also influenced our system.

3 Automatically Uncovering Description

We propose an unsupervised model of language feature collection that is based on *description by observation*, that is, learning target classifications by reading about the musical artists in reviews and discussions.

3.1 Community Metadata

Our model is called *community metadata* (Whitman and Lawrence, 2002) and has been successfully used in style detection (Whitman and Smaragdis, 2002) and artist similarity prediction (Ellis et al., 2002). It creates a machine understandable representation of artist description by searching the Internet for the artist name and performing light natural language processing on the retrieved pages. We split the returned documents into classes encompassing n-grams (terms of word length n), adjectives (using a part-of-speech tagger (Brill, 1992)) and noun phrases (using a lexical chunker (Ramshaw and Marcus, 1995).) Each pair $\{artist, term\}$ retrieved is given an associated salience weight, which indicates the relative importance of $term$ as associated to $artist$. These saliences are computed using a variant of the popular TF-IDF measure gaussian weighted to avoid highly specific and highly general terms. (See Table 2 for an example.) One important feature of community metadata is its time-sensitivity; terms can be crawled once a week and we can take into account trajectories of community-level opinion about certain artists.

Although tempting, we are reticent to make the claim that the community metadata vectors computationally approach the "linguistic division of labor" proposed in (Putnam, 1987) as each (albeit unaware) member of the networked community is providing a small bit of information and description about the artist in question. We feel that the heavily biased opinion extracted from the Internet is best treated as an approximation of a 'ground truth description.' Factorizing the Internet community into relatively coherent smaller communities to obtain sharpened lexical groundings is part of future work. However, we do in fact find that the huge amount of information we retrieve from these crawls average out to a good general idea of the artists.

4 Time-Aware Machine Listening

We aim for a representation of audio content that captures as much perceptual content as possible and ask the system to find patterns on its own. Our representation is based on the MPEG-7 (Casey, 2001) standard for content understanding and metadata organization.[1] The result of an MPEG-7 encoding is a discrete state number l ($l = [1...n]$) for each $\frac{1}{100}th$ of a second of input audio. We histogram the state visits into counts for each n-second piece of audio.

5 Relating Audio to Description

Given an audio and text model, we next discuss how to discover relationships between them. The approach we use is the same as our previous work, where we place the problem as a multi-class classification problem. Our input observations are the audio-derived features, and in training, each audio feature is associated with some salience weight of each of the 200,000 possible terms that our community metadata crawler discovered. In a recent test, training 703 separate SVMs on a small adjective set in the frame-based single term system took over 10 days. In most machine learning classifiers, time is dependent on the number of classes. As well, due to the unsupervised and automatic nature of the description classes, many are incorrect (such as when an artist is wrongly described) or unimportant (as in the case of terms such as 'talented' or 'cool'– meaningless to the audio domain.) Lastly, because the decision space over the entire artist space is so large, most class outputs are negative. This creates a bias problem for most machine learning algorithms. We next show our attempt at solving these sorts of problems using a new classifier technique based on the support vector machine.

[1] Our audio representation is fully described in (Whitman et al., 2003).

5.1 Regularized Least-Squares Classification

Regularized Least-Squares Classification (Rifkin, 2002) allows us to solve 'severe multi-class' problems where there are a great number of target classes and a fixed set of source observations. It is related to the Support Vector Machine (Vapnik, 1998) in that they are both instances of Tikhonov regularization (Evgeniou et al., 2000), but whereas training a Support Vector Machine requires the solution of a constrained quadratic programming problem, training RLSC only requires solving a single system of linear equations. Recent work (Fung and Mangasarian, 2001), (Rifkin, 2002) has shown that the accuracy of RLSC is essentially identical to that of SVMs.

We arrange our observations in a Gram matrix K, where $K_{ij} \equiv K_f(x_i, x_j)$ using the *kernel function* K_f. $K_f(x_1, x_2)$ is a generalized dot product (in a Reproducing Kernel Hilbert Space (Aronszajn, 1950)) between $\mathbf{x_i}$ and $\mathbf{x_j}$. We use the Gaussian kernel

$$K_f(x_1, x_2) = e^{-\frac{(|x_1 - x_2|)^2}{\sigma^2}} \quad (1)$$

where σ is a parameter we keep at 0.5.

Then, training an RLSC system consists of solving the system of linear equations

$$(K + \frac{I}{C})\mathbf{c} = \mathbf{y}, \quad (2)$$

where C is a user-supplied *regularization constant*. The resulting real-valued classification function f is

$$f(x) = \sum_{i=1}^{\ell} c_i K(x, x_i). \quad (3)$$

The crucial property of RLSC is that if we store the inverse matrix $(K + \frac{I}{C})^{-1}$, then for a new right-hand side $\mathbf{y}$, we can compute the new $\mathbf{c}$ via a simple matrix multiplication. This allows us to compute new classifiers (after arranging the data and storing it in memory) on the fly with simple matrix multiplications.

5.2 Evaluation for a "Query-by-Description" Task

To evaluate our connection-finding system, we compute the *weighted precision* $P(a_t)$ of predicting the label t for audio derived features of artist a. We train a new $\mathbf{c_t}$ for each term t against the training set. $f_t(x)$ for the test set is computed over each audio-derived observation frame x and term t. If the sign of $f_t(x)$ is the same as our supposed 'ground truth' for that $\{artist, t\}$, (i.e. did the audio frame for an artist correctly resolve to a known descriptive term?) we consider the prediction successful. Due to the bias problem mentioned earlier, the evaluation is then computed on the test set by computing a 'weighted precision': where $P(a_p)$ indicates overall positive accuracy (given an audio-derived observation, the probability that a positive association to a term is predicted) and

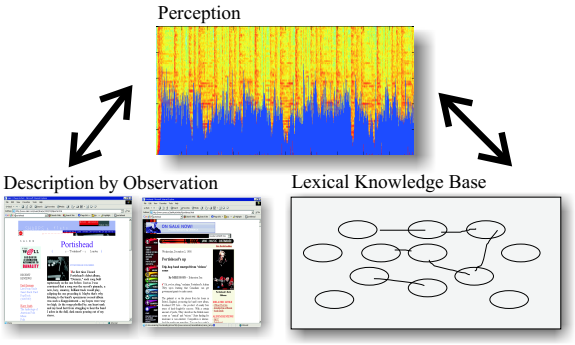

Figure 3: Overview of our parameter grounding method. Semantically attached terms are discovered by finding strong connections to perception. We then ask a 'professional' in the form of a lexical knowledge base about antonymial relations. We use those relations to infer gradations in perception.

$P(a_n)$ indicates overall negative accuracy, $P(a)$ is defined as $P(a_p)P(a_n)$, which should remain significant even in the face of extreme negative output class bias.

Now we sort the list of $P(a_t)$ and set an arbitrary threshold ε. In our implementation, we use ε = 0.1. Any $P(a_t)$ greater than ε is considered 'grounded.' In this manner we can use training accuracy to throw away badly scoring classes and then figure out which were incorrect or unimportant.

6 Linguistic Experts for Parameter Discovery

Given a set of 'grounded' single terms, we now discuss our method for uncovering parameter spaces among those terms and learning the knobs to vary their gradation. Our model states that certain knowledge is not inferred from sensory input or intrinsic knowledge but rather by querying a 'linguistic expert.' If we hear 'loud' audio and we hear 'quiet' audio, we would need to know that those terms are antonymially related before inferring the gradation space between them.

6.1 WordNet

WordNet (Miller, 1990) is a lexical database hand-developed by lexicographers. Its main organization is the 'synset', a group of synonymous words that may replace each other in some linguistic context. The meaning of a synset is captured by its lexical relations, such as hyponymy, meronymy, or antonymy, to other synsets. WordNet has a large community of users and various APIs for accessing the information automatically. Adjectives in WordNet are organized in two polar clusters of synsets, which each focal synset (the head adjective) linking to some antonym adjective. The intended belief is that

northern - southern	playful - serious
unlimited - limited	naive - sophisticated
foreign - native	consistent - inconsistent
outdoor - indoor	foreign - domestic
dissonant - musical	physical - mental
opposite - alternate	censored - uncensored
unforgettable - forgettable	comfortable - uncomfortable
concrete - abstract	untamed - tame
partial - fair	empirical - theoretical
atomic - conventional	curved - straight
lean - rich	lean - fat

Table 3: Example *synant* relations.

descriptive relations are stored as polar gradation spaces, implying that we can't fully understand 'loud' without also understanding 'quiet.' We use these antonymial relations to build up a new relation that encodes as much antonymial expressivity as possible, which we describe below.

6.2 Synant Sets

We define a set of lexical relations called *synants*, which consist of every antonym of a source term along with every antonym of each synonym and every synonym of each antonym. In effect, we recurse through WordNet's tree one extra level to uncover as many antonymial relations as possible. For example, "quiet"'s anchor antonym is "noisy," but "noisy" has other synonyms such as "clangorous" and "thundering." By uncovering these second-order antonyms in the synant set, we hope to uncover as much gradation expressivity as possible. Some example synants are shown in Table 3.

The obvious downside of computing the synant set is that they can quickly lose synonymy– following from the example above, we can go from "quiet" to its synonym "untroubled," which leads to an synantonymial relation of "infested." We also expect problems due to our lack of sense tagging: "quiet" to its fourth sense synonym "restrained" to its antonym "demonstrative," for example, probably has little to do with sound. But in both cases we rely again on the sheer size of our example space; with so many possible adjective descriptors and the large potential size of the synant set, we expect our connection-finding machines to do the hard work of throwing away the mistakes.

7 Innate Dimensionality of Parameters

Now that we have a set of grounded antonymial adjectives pairs, we would like to investigate the mapping in perceptual space between each pair. We can do this with a multidimensional scaling (MDS) algorithm. Let us call all acoustically derived data associated with one adjective as $X1$ and all data associated with the syn-antonym $X2$. An MDS algorithm can be used to find a multidimensional embedding of the data based on pairwise similarity distances between data points. The similarity distances between music samples is based on the representations described in the previous section. Consider first only the data from $X1$. The perceptual diversity of this data will reflect the fact that it represents numerous artists and songs. Overall, however, we would predict that a low dimensional space can embed $X1$ with low stress (i.e., good fit to the data) since all samples of $X1$ share a descriptive label that is well grounded. Now consider the embedding of the combined data set of $X1$ and $X2$. In this case, the additional dimensions needed to accomodate the joint data will reflect the *r*elation between the two datasets. Our hypothesis was that the additional perceptual variance of datasets formed by combining pairs of datasets on the basis of adjective pairs which are (1) well grounded, and (2) synants, would small compared to combinations in which either of these two combinations did not hold. Following are intial results supporting this hypothesis.

7.1 Nonlinear Dimensionality Reduction

Classical dimensional scaling systems such as MDS or PCA can efficiently learn a low-dimensional weighting but can only use euclidean or tangent distances between observations to do so. In complex data sets, the distances might be better represented as a nonlinear function to capture inherent structure in the dimensions. Especially in the case of music, time variances among adjacent observations could be encoded as distances and used in the scaling. We use the Isomap algorithm from (Tenenbaum et al., 2000) to capture this inherent nonlinearity and structure of the audio features. Isomap scales dimensions given a NxN matrix of distances between every observation in N. It roughly computes global geodesic distance by adding up a number of short 'neighbor hops' (where the number of neighbors is a tunable parameter, here we use $k = 20$) to get between two arbitrarily far points in input space. Schemes like PCA or MDS would simply use the euclidean distance to do this, where Isomap operates on prior knowledge of the structure within the data. For our purposes, we use the same gaussian kernel function as we do for RLSC (Equation 1) for a distance metric, which has proved to work well for most music classification tasks.

Isomap can embed in a set of dimensions beyond the target dimension to find the best fit. By studying the residual variance of each embedding, we can look for the "elbow" (the point at which the variance falls off to the minimum)– and treat that embedding as the innate one. We use this variance to show that our highly-grounded parameter spaces can be embedded in less dimensions than ungrounded ones.

8 Experiments and Results

In the following section we describe our experiments using the aforementioned models and show how we can automatically uncover the perceptual parameter spaces underlying adjective oppositions.

8.1 Audio dataset

We use audio from the NECI Minnowmatch testbed (Whitman et al., 2001). The testbed includes on average ten songs from each of 1,000 albums from roughly 500 artists. The album list was chosen from the most popular songs on OpenNap, a popular peer-to-peer music sharing service, in August of 2001. We do not separate audio-derived features among separate songs since our connections in language are at the artist level (community metadata refers to an artist, not an album or song.) Therefore, each artist a is represented as a concatenated matrix of F_a computed from each song performed by that artist.

F_a contains N rows of 40-dimensional data. Each observation represents 10 seconds of audio data. We choose a random sampling of artists for both training and testing (25 artists each, 5 songs for a total of N observations for testing and training) from the Minnowmatch testbed.

8.2 RLSC for Audio to Term Relation

Each artist in the testbed has previously been crawled for community metadata vectors, which we associate with the audio vectors as a y_t truth vector. In this experiment, we limit our results to adjective terms only. The entire community metadata space of 500 artists ended up with roughly 2,000 unique adjectives, which provide a good sense of musical description. The other term types (n-grams and noun phrases) are more useful in text retrieval tasks, as they contain more specific information such as band members, equipment or song titles. Each audio observation in N is associated with an artist a, which in turn is related to the set of adjectives with pre-defined salience. (Salience is zero if the term is not related, unbounded if related.) We are treating this problem as classification, not regression, so we assign not-related terms a value of -1 and positively related terms are regularized to 1.

We compute a c_t for each adjective term t on the training set after computing the stored kernel. We use a C of 10. After all the c_ts are stored to disk we then bring out the held-out test set and compute relative adjective weighted prediction accuracy $P(a)$ for each term. The results (in Table 4) are similar to our previous work but we note that our new representation allows us to capture more time- and structure-oriented terms. We see that the time-aware MPEG-7 representation creates a far better sense of perceptual salience than our prior frame-based power spectral density estimation, which threw away all short- and mid-time features.

Term	Precision	Term	Precision
busy	42.2%	artistic	0.0%
steady	41.5%	homeless	0.0%
funky	39.2%	hungry	0.0%
intense	38.4%	great	0.0%
acoustic	36.6%	awful	0.0%
african	35.3%	warped	0.0%
melodic	27.8%	illegal	0.0%
romantic	23.1%	cruel	0.0%
slow	21.6%	notorious	0.0%
wild	25.5%	good	0.0%
young	17.5%	okay	0.0%

Table 4: Select adjective terms discovered by the time-aware adjective grounding system. Overall, the attached term list is more musical due to the increased time-aware information in the representation.

Parameter	Precision
big - little	30.3%
present - past	29.3%
unusual - familiar	28.7%
low - high	27.0%
male - female	22.3%
hard - soft	21.9%
loud - soft	19.8%
smooth - rough	14.6%
clean - dirty	14.0%
vocal - instrumental	10.5%
minor - major	10.2%

Table 5: Select automatically discovered parameter spaces and their weighted precision. The top are the most semantically significant description spaces for music understanding uncovered autonomously by our system.

8.3 Finding Parameter Spaces using WordNet Lexical Relations

We now take our new single-term results and ask our professional for help in finding parameters. For all adjectives over our predefined ε we retrieve a restricted synant set. This restricted set only retrieves synants that are in our community metadata space: i.e. we would not return 'soft' as a synant to 'loud' if we did not have community-derived 'soft' audio. The point here is to only find synantonymial relations that we have perceptual data to 'ground' with. We rank our synant space by the mean of the $P(a)$ of each polar term. For example, $P(a_{soft})$ was 0.12 and we found a synant 'loud' in our space with a $P(a_{loud})$ of 0.26, so our $P(a_{loud...soft})$ would be 0.19. This allows us to sort our parameter spaces by the maximum semantic attachment. We see results of this process in Table 5.

We consider this result our major finding: from listening to a set of albums and reading about the artists, a computational system has *automatically derived the opti-*

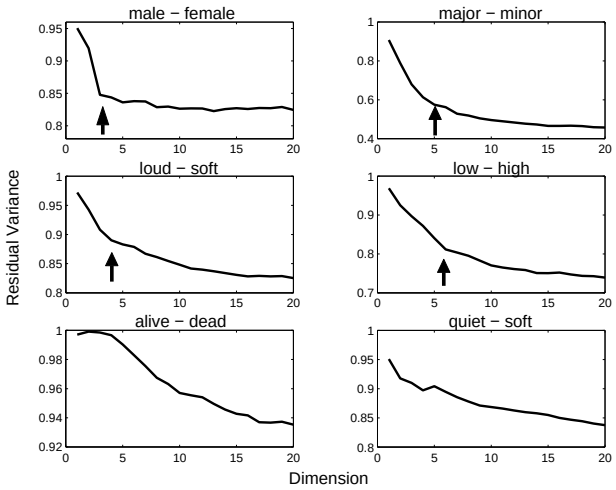

Figure 4: Residual variance elbows (marked by arrows) for different parameter spaces. Note the clear elbows for grounded parameter spaces, while less audio-derived spaces such as "alive - dead" maintain a high variance throughout. Bad antonym relations such as "quiet - soft" also have no inherent dimensionality.

mal (strongest connection to perception) *semantic gradation spaces to describe the incoming observation*. These are not the most statistically significant bases but rather the most *semantically significant* bases for understanding and retrieval.

8.4 Making Knobs and Uncovering Dimensionality

We would like to show the results of such understanding at work in a classification or retrieval interface, so we then have another algorithm learn the d-dimensional mapping of the two polar adjectives in each of the top n parameter spaces. We also use this algorithm to uncover the natural dimensionality of the parameter space.

For each parameter space a_1 ... a_2, we take all observations automatically labeled by the test pass of RLSC as a_1 and all as a_2 and separate them from the rest of the observations. The observations F_{a_1} are concatenated together with F_{a_2} serially, and we choose an equal number of observations from both to eliminate bias. We take this subset of observation $F_{a_{12}}$ and embed it into a distance matrix D with the gaussian kernel in Equation 1. We feed D to Isomap and ask for a one-dimensional embedding of the space. The result is a weighting that we can feed completely new unlabeled audio into and retrieve scalar values for each of these parameters. We would like to propose that the set of responses from each of our new 'semantic experts' (weight matrices to determine parameter values) define the most expressive semantic representation possible for music.

By studying the residual variances of Isomap as in Figure 4, we can see that Isomap finds inherent dimensionality for our top grounded parameter spaces. But for 'ungrounded' parameters or non-antonymial spaces, there is less of a clear 'elbow' in the variances indicating a natural embedding. For example, we see from Figure 4 that the "male - female" parameter (which we construe as gender of artist or vocalist) has a lower inherent dimensionality than the more complex "low - high" parameter and is lower yet than the ungroundable (in audio) "alive - dead." These results allow us to evaluate our parameter discovery system (in which we show that groundable terms have clearer elbows) but also provide an interesting window into the nature of descriptions of perception.

9 Conclusions

We show that we can derive the most semantically significant description spaces automatically, and also form them into a knob for future classification, retrieval and even synthesis. Our next steps involve user studies of music description, an attempt to discover if the meaning derived by community metadata matches up with individual description, and a way to extract a user model from language to specify results based on prior experience.

We are also currently working on new automatic lexical relation discovery techniques. For example, from the set of audio observations, we can infer antonymial relations without the use of an expert by finding optimally statistically separable observations. As well, meronymy, hyponymy and synonymy can be inferred by studying artificial combinations of observation (the mixture of 'loud' and 'peaceful' might not resolve but the mixture of 'sexy' and 'romantic' might.)

From the perspective of computational linguistics, we see a rich area of future exploration at the boundary of perceptual computing and lexical semantics. We have drawn upon WordNet to strengthen our perceptual representations, but we believe the converse is also true. These experiments are a step towards grounding WordNet in machine perception.

References

N. Aronszajn. 1950. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404.

D. Bailey. 1997. *When push comes to shove: A computational model of the role of motor control in the acquisition of action verbs*. Ph.D. thesis, University of California at Berkeley.

K. Barnard and D. Forsyth. 2000. Learning the semantics of words and pictures.

Eric Brill. 1992. A simple rule-based part-of-speech tagger. In *Proc. ANLP-92, 3rd Conference on Applied Natural Language Processing*, pages 152–155, Trento, IT.

Michael Casey. 2001. General sound recognition and similarity tools. In *MPEG-7 Audio Workshop W-6 at the AES 110th Convention*, May.

D.A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press.

P. Duygulu, K. Barnard, J.F.G. De Freitas, and D.A. Forsyth. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary.

Dan Ellis, Brian Whitman, Adam Berezweig, and Steve Lawrence. 2002. The quest for ground truth in musical artist similarity. In *Proc. International Symposium on Music Information Retrieval ISMIR-2002*.

Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. 2000. Regularization networks and support vector machines. *Advanced In Computational Mathematics*, 13(1):1–50.

Glenn Fung and O. L. Mangasarian. 2001. Proximal support vector classifiers. In Provost and Srikant, editors, *Proc. Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 77–86. ACM.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1993. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st Annual Meeting of the ACL*.

G.A. Miller. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–312.

S. Narayanan. 1997. *Knowledge-based Action Representations for Metaphor and Aspect (KARMA)*. Ph.D. thesis, University of California at Berkeley.

H. Putnam. 1987. *Representation and Reality*. MIT Press.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In David Yarovsky and Kenneth Church, editors, *Proc. Third Workshop on Very Large Corpora*, pages 82–94, Somerset, New Jersey. Association for Computational Linguistics.

T. Regier. 1996. *The human semantic potential*. MIT Press, Cambridge, MA.

Ryan M. Rifkin. 2002. *Everything Old Is New Again: A Fresh Look at Historical Approaches to Machine Learning*. Ph.D. thesis, Massachusetts Institute of Technology.

D. Roy. 1999. *Learning Words from Sights and Sounds: A Computational Model*. Ph.D. thesis, Massachusetts Institute of Technology.

J. Siskind. 2001. Grounding the Lexical Semantics of Verbs in Visual Perception using Force Dynamics and Event Logic. *Journal of Artificial Intelligence Research*, 15:31–90.

Malcolm Slaney. 2002. Semantic-audio retrieval. In *Proc. 2002 IEEE International Conference on Acoustics, Speech and Signal Processing*, May.

J.B. Tenenbaum, V. de Silva, and J.C. Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323.

Vladimir N. Vapnik. 1998. *Statistical Learning Theory*. John Wiley & Sons.

Brian Whitman and S. Lawrence. 2002. Inferring descriptions and similarity for music from community metadata. In *Proc. Int. Computer Music Conference 2002 (ICMC)*, pages 591–598, September.

Brian Whitman and Ryan Rifkin. 2002. Musical query-by-description as a multi-class learning problem. In *Proc. IEEE Multimedia Signal Processing Conference (MMSP)*, December.

Brian Whitman and Paris Smaragdis. 2002. Combining musical and cultural features for intelligent style detection. In *Proc. Int. Symposium on Music Inform. Retriev. (ISMIR)*, pages 47–52, October.

Brian Whitman, Gary Flake, and Steve Lawrence. 2001. Artist detection in music with minnowmatch. In *Proc. 2001 IEEE Workshop on Neural Networks for Signal Processing*, pages 559–568. Falmouth, Massachusetts, September 10–12.

Brian Whitman, Deb Roy, and Barry Vercoe. 2003. Grounding a lexicon and lexical relations from machine perception of music. *submitted*.