# Semantic–Audio Retrieval

*Malcolm Slaney*
IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120
malcolm@almaden.ibm.com

## ABSTRACT

This paper describes a system for connecting sounds and words in linked multi-dimensional vector spaces. The acoustic space is represented using anchor models and partitioned using agglomerative clustering. The semantic space is modeled by a hierarchical multinomial clustering model. Nodes in one space are linked by probabilistic models to the other space. With these linked models, users retrieve sounds with natural language, and the system describes new sounds with words.

## 1. THE PROBLEM

This paper describes a method of connecting sounds to words, and words to sounds. Given a description of a sound, the system finds the audio signals that best fit the words. Thus, a user might make a request with the description "the sound of a galloping horse," and the system would respond by presenting recordings of a horse running on different surfaces, and possibly of musical pieces that sound like a horse galloping. Conversely, given a sound recording, the system describes the sound or the environment in which the recording was made. Thus, given a recording made outdoors, the system says confidently that the recording was made at a horse farm where several dogs reside.

We build a system that has these functions, called **SAR** (semantic–audio retrieval), by learning the connection between a semantic space and an auditory space. **Semantic space** maps words into a high-dimensional probabilistic space. **Acoustic space** describes sounds by a multidimensional vector. In general, the connection between these two spaces will be many to many. Horse sounds, for example, might include footsteps and neighs.

Figure 1 shows one half of SAR; how to retrieve sounds from words. Annotations that describe sounds are clustered within a hierarchical semantic model that uses multinomial models. The sound files, or acoustic documents, that correspond to each node in the semantic hierarchy are modeled with **Gaussian mixture models** (GMMs). Given a semantic request, SAR identifies the portion of the semantic space that best fits the request, and then measures the likelihood that each sound in the database fits the GMM linked to this portion of the semantic space. The most likely sounds are returned to satisfy the user's semantic request.

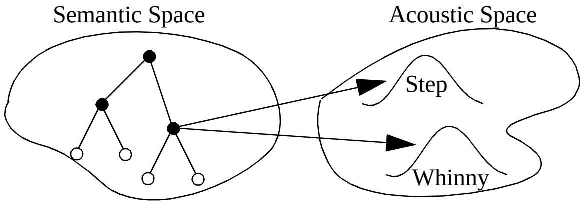

Figure 1: SAR models all of semantic space with a hierarchical collection of multinomial models, each portion in the semantic model is linked to equivalent sound documents in acoustic space with a GMM.

Figure 2 shows the second half of SAR; how to generate words to describe a sound. SAR analyzes the collection of sounds and builds models for arbitrary sounds called anchors. All sounds in the database are described by how well they are modeled by these anchor sounds. This approach gives us a multidimensional representation of any sound, and a distance metric that permits agglomerative clustering in the acoustic space. Given an acoustic request, SAR identifies the portion of the acoustic space that best fits the request. Each portion of the acoustic space has an associated multinomial word model, and from this model SAR generates words to describe the query sound.

The SAR algorithm is illustrated with a closed set of acoustic and semantic documents about animals. Six CDs (#12, 30, 34, 35, 37, 38) from the BBC Sound Effects Library contained 215 separate tracks, with 330 minutes of audio recordings of animal sounds. In addition, the concatenated name of the CD (e.g., "Horses I") and track description (e.g., "One horse eating hay and moving around") form a unique semantic label for each track. The audio from the CD track and the liner notes form a pair of acoustic and semantic documents used to train the SAR system.

## 2. THE EXISTING SYSTEMS

There are many multimedia retrieval systems that use a combination of words and/or examples to retrieve audio (and video) for users.

An effective way to find an image of the space shuttle is to enter the words "space shuttle jpg" into a text-based web search engine. The original Google system did not know about images, but, fortunately, many people created web pages with the phrase "space shuttle" that contained a JPEG image of the shuttle. More recently, both Google and AltaVista for images, and Compusonics for audio, have built systems that automate these searches. They allow people to look for images and sound based on nearby words. The SAR work expands those search techniques by considering the acoustic and semantic similarity of sounds to allow

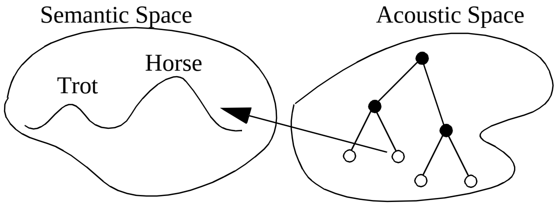

Figure 2: SAR describes with words an audio query by partitioning the audio space with a set of hierarchical acoustic models and then linking each set of audio files (or documents) to a probability model in semantic space.

users to retrieve sounds without running searches on the exact words used on the web page.

Many existing image- and audio-retrieval systems perform query by example [3]. Given an example of a sunset, these systems can find other images that have similar properties. These systems are difficult to use unless the user formulates the query using exactly the same features used to describe the original image. The queries often fail because the underlying feature space does not fit human expectations: Humans do not think about images in terms of their quantitative texture metrics.

Barnard suggested a system [1] that is closest to SAR. He used Hofmann's hierarchical clustering algorithm [2] to build a model that combined words and image features to create a single hierarchical model that spanned both semantic and image features. He demonstrated the effectiveness of coupled clustering for an information-retrieval task and argued that the words written by a human annotator describing an image (e.g., "a rose") often provide information that complements the obvious information in the image (it is red).

SAR improves on two aspects of Barnard's approach. First, the semantic and image features do not have the same probability distributions. Hofmann's algorithm assumes that image features can be described by a multinomial distribution, while a Gaussian is probably more appropriate. Second, and perhaps more important, there is nothing in Hoffman's algorithm that guarantees that the features used to build each stage of the model include both semantic and image features. Thus, the algorithm is free to build a model that completely ignores the image features and clusters the "documents" based on only semantic features.

The audio-retrieval problem is easier than image retrieval for two reasons. First, we do not have to solve the foreground–background problem (yet). People who want a picture of a tiger usually do not care whether the tiger is surrounded by grass, sand, water, or even a sunset. Fortunately, in the audio world, we can often assume that only one sound is present at a time. Second, it is difficult to know which are the features that best characterize an image.

## 3. THE FEATURE SPACES

The key to SAR is choosing a set of features that allows us to talk about mathematical spaces for sounds and semantics. In Section 3.1, we consider what semantic features are, how we describe the position of a concept in semantic space, and how we measure the distance between two concepts. Section 3.2 describes the same procedure for sounds. In both cases, we have a rich foundation of prior work on which to build.

### 3.1 The Semantic Space

SAR uses multinomial models to represent and cluster a collection of semantic documents. The likelihood that a document matches a given multinomial model is described by $L = \Pi p_i^{n_i}$, where $p_i$ is the probability that the word $i$ occurs in this type of document, and $n_i$ is the number of times that word $i$ is found in this document. The set of probabilities, $p_i$, is different for different types of documents. Thus, a model for documents about cows will have a relatively high probability for containing "cow" and "moo," whereas a model for documents that describe birds with have a high probability of containing "feather."

SAR uses the same cluster abstraction technique used in Barnard's work, but only applies it to the semantic data. Cluster abstraction builds a hierarchical collection of multinomial models to describe a set of documents. This hierarchical model (Figure 3) assumes that an unknown process generates documents using two hidden variables. The first hidden variable assigns each document to one of K clusters at the leaves of the tree, where each cluster describes one type of document (e.g., "galloping"). The second hidden variable describes how words are generated by a hierarchy of word-probability models. Common words, such as "the" and "a," are generated with the same probability distribution for all documents in the collection and are represented by a single distribution at the top of the tree. The words that distinguish the broad classes in the collection are described by probability distributions slightly lower in the tree. These different probability distributions allow us to distinguish the different types of documents and to assign different documents to different points in semantic space. Finally, the lowest level of the hierarchy represents words that are specialized to just a few documents in the collection. The process generates any one document in the collection by picking a cluster (which determines the semantic content), and then generates each word by picking a level in the hierarchy and then picking the specific word from the corresponding probability distribution at that level and above. SAR uses these models to partition and recognize portions of the semantic space.

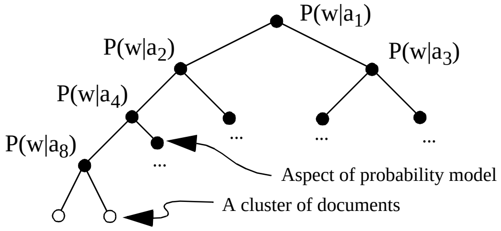

Figure 3: The cluster abstraction model calculates a hierarchical set of multinomial models ($a_x$)to generate clusters of documents.

The text used to describe one track of a CD is called a document. SAR uses the PORTER stemmer [4] to remove common suffixes from the words, and deletes common words on the SMART list [5] before further processing. After this preprocessing, there were 414 unique stemmed words. The 215 documents were grouped into 32 clusters, and there were 63 nodes in the tree that described the probability of a word given a position in the hierarchy. In effect, a 414-dimensional vector (the multinomial coefficients) describes a point in semantic space, and SAR partitions the space into a hierarchy of 63 overlapping regions.

Hofmann's approach clusters all documents in the training collection as leaf nodes. All nodes in the hierarchy above a leaf contribute probability distributions for words in a training document. This structure is appropriate, since all of the sound labels that SAR uses to build the hierarchical model are specific. User queries are not required to be so specific. Thus, if the user requests "horse" sounds, many clusters of documents satisfy this request and a document described by only high-level nodes in the tree is the best match for the query. This work extends Hofmann's work by allowing less specific documents to be generated using only the higher-level nodes in the tree.

### 3. 2 The Acoustic Space

Sound is difficult to analyze because it is dynamic. The sound that we describe as that of a horse galloping is constantly changing at time scales in the hundreds of milliseconds; a hoofstep is

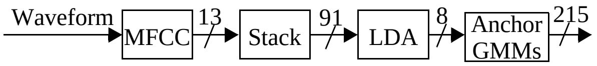

Figure 4: The acoustic signal processing chain. Arrows are marked with the signal's dimensionality. All but the last are sampled at 50Hz. The final output is sampled once per sound.

followed by silence, and then by another hoofstep. Yet we would like a means to transform the sound of a galloping horse into a single point in an acoustic space. This section describes acoustic features that allow us to describe each sound as a single point in acoustic space, and to cluster related sounds.

Conventional acoustic features for speech recognition and for sound identification use a short-term spectral slice to characterize the sound at 20-ms intervals. A combination of signal-processing and machine-learning calculations endeavors to capture the sound of a horse as a point in auditory space.

I measured the efficacy of a number of different features by testing them in a 10-way discrimination test. I manually separated many of the BBC sounds into 10 classes that were each clearly related to one animal. For each class, a GMM was trained to model the probability of that classes' features. The abilities of the GMMs to learn the class feature probabilities and to predict the class of test sounds are measures of how well the features discriminate among the 10 classes of sound.

The best feature in this test combined three signal-processing steps (Figure 4). Mel-frequency cepstral coefficients (MFCC) [6] decompose each signal into a broad spectral channels and compress the loudness of the signal. Then seven frames of data—three before the current frame, the current frame, and the three frames following the current frame—are stacked together. Finally, linear discriminant analysis (LDA) [6] uses the intra- and inter-class scatter matrices for the labeled data to project the data on the optimum dimensions for linear separability. GMM recognizers (10 element, diagonal covariance) gave a 35% error on the 20% of the data that was held out for testing.

New features, designed to represent the long-term temporal properties of the sounds, were not successful at discriminating the 10 classes of data. These new features—which used histograms, correlation and small-order all-pole models after the MFCC calculation—do not change when the feature calculation is shifted 20 ms. This property is clearly not true for the MFCC-LDA features that provided the best discrimination. The MFCC-LDA feature during a footstep is different from its value in the silence between steps.

The success of the MFCC-LDA representation is due to the multi-centroid GMM that formed the discrimination system. One of the Gaussians might capture the start of the footstep, a second captures the steady-state portion, a third captures the footstep's decay, and, finally, a fourth captures the silence between footsteps. The temporal order is ignored by a GMM model.

SAR converts the MFCC-LDA plus GMM recognition system into an auditory space by using model likelihood scores to measure the closeness of a sound to pretrained acoustic models. These known sounds and their GMMs are called anchor models [7]. The negative log-likelihood that a sound fits a model is a measure of the distance of the new sound from the test model. By this method, we can use the distances to the GMMs for 215 prototype sounds as the measure of the position of the sound in a 215-dimensional acoustic space.

I built a 10-center (diagonal covariance) GMM to model each anchor sound. Each of the 215 sounds was then evaluated with each of the 215 acoustic models. Figure 5 shows the resulting log-likelihood per frame (*L*). Each model did a good job of matching the sound on which it was trained (the diagonal). The sounds that came from the same CD, and were often from the same animal, are the larger, light squares along the diagonal with lower error.

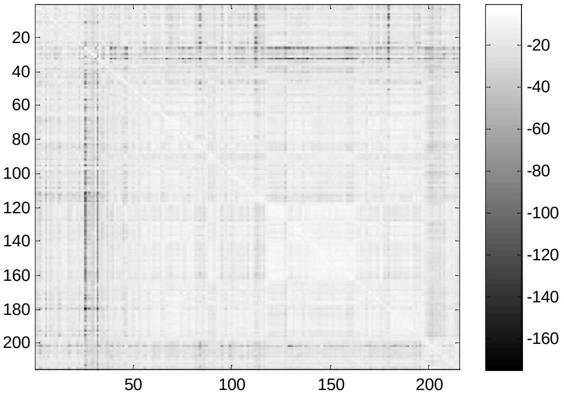

Figure 5: Log-likelihood of all sounds given their GMM models. Sound tracks are in numerical order.

SAR uses agglomerative clustering [8] to group nearby sounds in acoustic space into larger clusters. It computes the distance between each pair of training sounds -[*L*(model a, sound b) + *L*(model b, sound a)]/2. At each step, agglomerative clustering grows another layer of a hierarchical model by merging the two remaining clusters that have between them the smallest distance. SAR uses "complete" linkage, which uses the maximum distance between the points that form the two clusters, to decide which clusters should be combined.

# 4. THE MODEL LINKAGE

Linking the acoustic and semantic spaces is straightforward given the hierarchical representations of the spaces.

Each node of the acoustic hierarchy represents multiple similar sounds. SAR represents the collection of sounds below a node using a GMM, forming a new super-anchor model. In addition, each sound that contributes to a node has a semantic document that describes the sound. We can use this collection of documents to build a (single mixture) multinomial model that predicts the most likely words used to describe this cluster of sounds. The probability of word *i* is given by $P(w|a_c) = n_i / N_d$, where $n_i$ is the number of times the word is seen in this collection of documents, and $N_d$ is the number of words in the cluster.

Given a sound, SAR consults the ensemble of anchor models that partition the acoustic hierarchy. Sometimes, one of the leaf nodes is the best match for the sound. SAR simply returns the words that describe the prototype.

More often, the sound will match an intermediate node in the hierarchy: The sound is a mixture of two different sounds, or somehow falls in between two or more acoustic models. The sound is best described by the most likely words from the combined multinomial model at that level in the hierarchy.

In the semantic space, each leaf and all intermediate nodes define a set of multinomial models that describe documents at that level. At each node in the semantic hierarchy, there is a well-defined set of documents that contribute words to that aspect of the semantic model. We collect all those documents, then train a 10-element GMM with all the corresponding audio CD tracks.

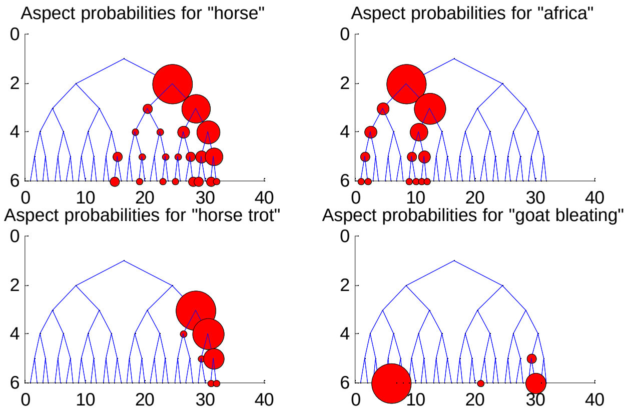

Figure 6: Probabilities of four semantic queries. The size of the dot shows the relative probability for different aspects of the cluster abstraction model.

Given a semantic query, SAR queries each multinomial model and determines which model is most likely to have generated that set of words. Then it uses the corresponding GMM to query all of the sounds in the acoustic database, and it returns the most likely sounds.

## 5. THE RESULTS

It is instructive to look at the results of several queries. A search in the semantic space for the words "horse trot" is translated by the semantic model into a query for aspect 7 in the third level from the top of the hierarchy (see Figure 6). This aspect generates a large number of words, of which the most common words are (words as they are stored in the database, after stemming): hors (p=0.65), walk (p=0.10), trot (p=0.07), track (p=0.066), tarmac (p=0.03), approach (p=0.02), canter (p=0.02), pass (p=0.01).

Each aspect of the semantic model defines an acoustic GMM that describes sounds that fit those words. All sounds in the database were evaluated with the GMM connected to aspect 7; the most likely sound tracks for the query are shown in Table 1.

**Table 1: Acoustic results from semantic query "horse trot"**

| L | Track | Label |
|---|---|---|
| -5.26 | 37.37 | Horses I One horse approaches at walk |
| -5.33 | 37.22 | Horses I One horse walks past through grass |
| -5.37 | 37.12 | Horses I One horse walks past on tarmac |
| -5.46 | 37.13 | Horses I One horse canters past on grass |
| -5.49 | 37.36 | Horses I One horse approaches at walk, on rough track |
| -5.55 | 37.38 | Horses I One horse walks past on rough track |
| -5.62 | 37.40 | Horses I One horse trots past on rough track |

It is instructive to look at how the system does when translating from audio to semantic and then back to audio again. This back-and-forth test is often used with machine-translation systems to show how difficult it is to generate reasonable results from language-specific idioms. I started with sound track 34.21 which is described as "Livestock I Cattle passing on tarmac, about 30–50." A semantic query garners the response that aspect 7 is most likely to generate this semantic document. The most common words for this semantic aspect are cattl (p=0.25), livestock (p=0.25), pass (p=0.25), tarmac (p=0.25). Returning to the acoustic domain, this semantic aspect is most likely to generate the CD tracks shown in Table 2.

**Table 2: Result of back-and forth query to track 34.21**

| L | Track | Description |
|---|---|---|
| -4.9 | 34.21 | Livestock I Cattle passing on tarmac, about 30–50 |
| -5.2 | 38.16 | Horses II One horse working in an indoor school |
| -5.7 | 35.17 | Livestock II Cattle TB tested in yard |
| -6 | 37.16 | Horses I One horse galloping through field |
| -6.1 | 37.13 | Horses I One horse canters past on grass |
| -6.1 | 37.14 | Horses I One horse trots past on grass |

## 6. THE CONCLUSIONS

This paper has described a system for converting sounds and words into points in separate multi-dimensional vector spaces. Using hierarchical clustering ideas, SAR group the points and build models that link one domain to the other.

There are two shortcomings of the present work. First, this paper provides only a proof of concept. We need to test SAR with a larger data set—with separate training and testing data—to quantify the efficacy of this approach. Second, choosing only acoustic and semantic clusters with the largest score limits query results to well-defined locations. A means of interpolating between nodes will allow results that fit the user's query more precisely.

## ACKNOWLEDGMENTS

I appreciate the assistance that I received from Jay Kadis, Dulce Ponceleon, Byron Dom, Arnon Amir, Myron Flickner, Chalapathy Neti, Michele Covell, John Fisher and Lyn Dupré. I used Ian Nabney's NETLAB software to calculate the GMMs; Roger Jang provided the agglomerative clustering code.

## REFERENCES

[1] Korbus Barnard and David Forsyth. "Learning the semantics of words and pictures." *Proceedings of the 2001 International Conference on Computer Vision,* Vol. 2, pp. 408–415, 2001.

[2] Thomas Hofmann. "The cluster-abstraction model: Unsupervised learning of topic hierarchies." *Proceedings of the Int. Joint Conf. on Artificial Intelligence,* pp. 682–687, 1999.

[3] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, P. Yanker. "The QBIC project: Query images by content using color, texture and shape." *SPIE Storage and Retrieval of Image and Video Database,* pp. 173–181, 1993.

[4] M. F. Porter. "An algorithm for suffix stripping." *Program,* 14(3), pp. 130–137, 1980.

[5] G. Salton. *The SMART Retrieval System.* Prentice Hall, Englewood Cliffs, NJ, 1971.

[6] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon. *Spoken Language Processing.* Prentice Hall, Upper Saddle River, NJ, 2001.

[7] D. Sturim, D. Reynolds, E. Singer, J. Campbell. "Speaker indexing in large audio databases using anchor models." *Proc. of ICASSP,* vol. I, pp. 429–433, 2001.

[8] Anil K. Jain, Richard Dubes. *Algorithms for Clustering Data.* Prentice Hall, Englewood Cliffs, NJ, 1988.