Presented at the IEEE International Conference on Multimedia and Expo, Lausanne, Switzerland, August 26–29, 2002.

Mixtures of Probability Experts for Audio Retrieval and Indexing

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 non-speech sounds and words using linked multi-dimensional vector spaces. An approach based on mixture of experts learns the mapping between one space and the other. This paper describes the conversion of audio and semantic data into their respective vector spaces. Two different mixture-of-probability-expert models are trained to learn the association between acoustic queries and the corresponding semantic explanation, and visa versa. Test results are presented based on commercial sound effects CDs.

1. THE APPROACH

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 responds 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.

A system that has these functions, called **MPESAR** (mixtures of probability experts for semantic–audio retrieval), learns the connections between a semantic space and an acoustic 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 MPESAR: how to retrieve sounds from words. Annotations that describe sounds are clustered and represented with multinomial models. The sound files, or acoustic documents, that correspond to each node in the semantic space are modeled with **Gaussian mixture models** (GMMs). Given a semantic request, MPESAR 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.

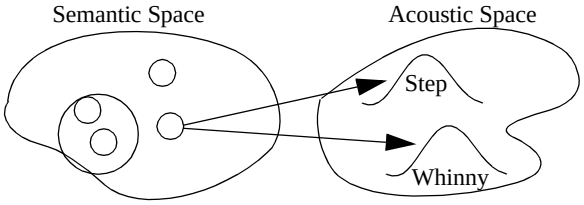

Figure 1: MPESAR models all of semantic space with overlapping multinomial clusters, each portion in the semantic model is linked to equivalent sound documents in acoustic space with a GMM.

Figure 2 shows the other half of MPESAR: how to generate words to describe a sound. MPESAR analyzes the collection of sounds and builds models for arbitrary sounds. This approach gives us a multi-dimensional representation of any sound, and a distance metric that permits agglomerative clustering in the acoustic space. Given an acoustic request, MPESAR 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 MPESAR generates words to describe the query sound.

In general, sounds that are close in acoustic space might correspond to many different points in semantic space, and vice versa. Thus, MPESAR builds two completely separate sets of models: one connecting audio to semantic space and the other connecting semantic to audio space.

2. THE EXISTING SYSTEMS

There are many multimedia retrieval systems that use a combination of words 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" and 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 words on a web page near the picture. The MPESAR work expands those search techniques by considering the acoustic and semantic similarity of sounds to allow 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

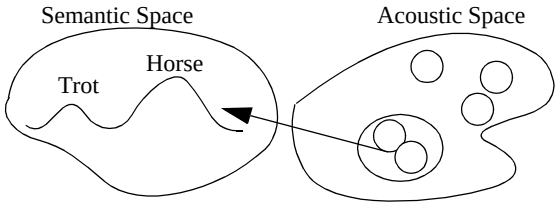

Figure 2: MPESAR describes with words an audio query by partitioning the audio space with a set of acoustic models and then linking each cluster of audio files (or documents) to a probability model in semantic space.

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 [1] used a hierarchical clustering algorithm 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).

My previous solution to this problem, SAR [7], built *separate* hierarchical models for each space and then learned an association for each node to link the two spaces together. SAR improved on two aspects of Barnard's approaches. First, the semantic and image features do not have the same probability distributions. Barnard's algorithm assumes that image features can be described by a multinomial distribution, while a Gaussian is probably more appropriate. Second, and perhaps most important, there is nothing in Barnard'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.

MPESAR improves on SAR by interpolating between models. SAR assigned each document to a single cluster and used a single model (winner-take-all) to map to the opposite domain. On the other hand, MPESAR calculates the probability that each cluster generates the query and then calculates a *weighted* average of models based on the cluster probabilities.

The MPE algorithm is appropriate for mapping one type of media to another. I illustrate the idea here using audio and semantic documents because audio retrieval is a simple problem [7].

3. THE ALGORITHM

3.1 Mixture of Probability Experts

MPESAR uses a mixture of experts approach [8] to link semantic and audio spaces. A mixture of experts approach uses a different expert for different regions of an input space. Thus, one expert might be responsible for horse sounds while another is responsible for bird sounds.

Mathematically, a mixture of probability experts for semantic to audio retrieval is summarized by the following equation

$$P(a|q) = \sum_c P(c|q)P(a|c)$$

Here $P(c|q)$ represents the probability that a semantic query (q) matches a cluster (c). The probability that a particular portion of acoustic space is associated with an expert or cluster (c) is given by $P(a|c)$. To find the overall probability of a point in audio space given the query, $P(a|q)$, I sum over all possible clusters, essentially interpolating the different expert's opinions to arrive at the final probability estimate.

We want to calculate the probability of a cluster given a query. I group semantic documents into clusters and then estimate $P(q|c)$. Using Bayes' rule, $P(c|q) = P(c)P(q|c)/P(q)$. The P(c) and $P(q|c)$ terms will be calculated using clustering algorithms described in Sections 3.4 and 3.5. Since the query is given, we can ignore the $P(q)$ term. The same formalism is used for the audio to semantic problem.

3.2 Semantic Features

MPESAR 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 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."

A semantic document contains the text used to describe an audio clip. MPESAR uses the PORTER stemmer to remove common suffixes from the words, and deletes common words on the SMART list before further processing [7]. In effect, a 705-dimensional vector (the multinomial coefficients) describes a point in semantic space, and MPESAR partitions the space into overlapping clusters of regions.

Smoothing is used in statistical language modeling to compensate for a paucity of data. It is called smoothing because the probability associated with likely events is reduced and distributed to events that were not seen in the training data. The most successful methods [2] use a backoff method, where data from simpler language models are used to set the probability of rare events. MPESAR uses a unigram word model, so the backoff model suggests a uniform low probability for all words.

3.3 Acoustic Features

Sound is difficult to analyze because it is dynamic. The sound of a horse galloping is constantly changing at time scales in the hundreds of milliseconds; a hoofstep is 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 10-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.

The process of converting a waveform into a point in acoustic space is shown in Figure 3. Mel-frequency cepstral coefficients (MFCC) [6] decompose each signal into broad spectral channels and compress the loudness of the signal. RASTA filtering [6] is used on the MFCC coefficients to remove long-term spectral characteristics that often occur due to the different recording environments. 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 a hand-labeled set of classes to project the data onto the optimum dimensions for linear separability.

The long-term temporal characteristics of each sound are captured using a GMM. One of the Gaussians might capture the

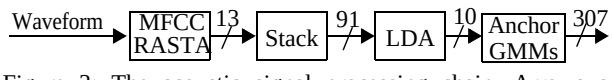

Figure 3: The acoustic signal processing chain. Arrows are marked with the signal's dimensionality. All but the last are sampled at 100Hz. The final output is sampled once per sound.

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 GMM measures the probability that a vector sequence fits a probabilistic model learned from the training sounds. Unlike hidden-Markov models (HMMs), a GMM ignores temporal order.

MPESAR converts the MFCC-RASTA-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. The negative log-likelihood that a sound fits a model is a measure of the distance of the new sound from the test model.

3.4 Acoustic to Semantic Lookup

Given representations of acoustic and semantic spaces, we can now build models to link the two spaces together. The overall algorithm for both acoustic to semantic and semantic to acoustic lookup is shown in Figure 4.

Acoustic space is clustered into regions using agglomerative clustering [4]. I compute the distance between each pair of training sounds $L(\text{model a}|\text{sound b}) + L(\text{model b}|\text{sound a})/2$ where $L(\text{model a}|\text{sound b})$ represents the likelihood that sound b is generated by model a. 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. MPESAR uses "complete" linkage, which uses the maximum distance between the points that form the two clusters, to decide which clusters should be combined. While agglomerative clustering generates a hierarchy, MPESAR only uses the information about which sounds are clustered. Leaves at the bottom of the tree are considered clusters containing a single document.

Each acoustic cluster is composed of a number of audio tracks and their associated descriptive text. A new 10-element GMM with diagonal covariance models all the sounds in this cluster and estimates the probability density for acoustic frames in this cluster, $P(a|c)$. Given a new sound, MPESAR uses this model to estimate the probability that a new sound belongs to this cluster. The text associated with each acoustic sample in the cluster is used to estimate the semantic model associated with this cluster. This is written as a simple multinomial model; there is not enough text in this study to form a richer model.

Given a new waveform, MPESAR queries all acoustic GMMs to find the probability that each possible cluster generated this query. Each cluster comes with an associated semantic model. MPESAR uses a weighted average of all the semantic models, based on cluster probabilities, to estimate the semantic model that describes the test sound. The words that describe the test sound are entries in the semantic multinomial model with the highest probabilities.

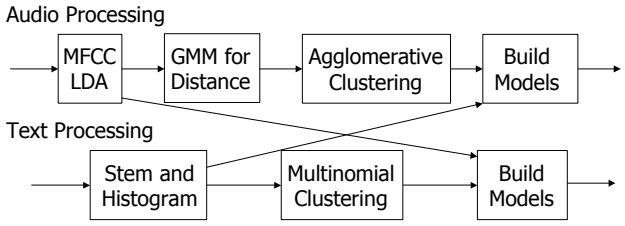

Figure 4: A schematic showing the process of building the MPESAR models. The top line shows the construction of the audio to semantic model and the bottom line shows the construction of the semantic to audio model.

3.5 Semantic to Acoustic Lookup

A similar procedure is used for semantic to acoustic lookup. A document's point in semantic space is described by the coefficients of a unigram multinomial model. Semantic space is clustered into regions using a multinomial clustering algorithm which uses an iterative expectation-maximization algorithm [5] to group documents with similar (multidimensional) models. In this work, I assign each document to its own cluster, and then split the entire corpus into a number of arbitrary-sized clusters (32, 64, 128 and 256 clusters for the corpus).

Each text cluster is composed of a number of text documents and their associated audio tracks. All the text associated with each cluster is used to form a unigram multinomial model of the text documents. All of the audio associated with a cluster is used to form a 10-element GMM to describe the link to audio space. (Note there are three sets of GMMs used in this work: the GMMs used to compute the distances as part of audio clustering, the GMMs used to model each audio cluster, and the GMMs used to model the sounds associated with each semantic cluster.)

Given a text query, MPESAR finds the probability that each semantic cluster generated the query. Then the acoustic models are averaged (weighted by the cluster probabilities) to find the probability that any one sound fits the query.

4. TESTING

This section describes several tests performed using the algorithms described above.

4.1 Data

The animal sounds from two sets of sound effect CDs were used as training and testing material. Seven CDs from the BBC Sound Effects Library (#6, 12, 30, 34, 35, 37, 38) contained 261 separate tracks and 390 minutes of animal sounds. Two CDs from the General 6000 Sound Effect library (all tracks from CD6003 and tracks 18 to 40 of CD6023.) totaled 122 tracks and 110 minutes of animal sounds.

The concatenated name of the CD (e.g., "Horses I") and track description (e.g., "One horse eating hay and moving around") forms a 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 MPESAR system.

The system training and testing described in this paper were performed on distinct sets of data. 80% of the tracks (307) from both sets of CDs were randomly assigned as training data in the procedure shown in Figure 4. The remaining 20% of the tracks (93) were reserved for testing.

Mixing the data obtained from the two sets of CDs is important for several reasons. First, the acoustic environments of the two data sets are different; RASTA reduces these effects. Second, the words and description are different because the sounds are labeled by different organizations with different needs. For example, the BBC describes the sound of a cat's vocalization as miaow and the General Sound Effects CD uses meow. Finally, the two sets of audio data do not contain the same sounds: There are many sounds in the General set which are not represented in the BBC training set.

4.2 Acoustic Feature Reduction and Language Smoothing

The audio-feature reduction using LDA was computed using portions of the audio data from both sets of CDs. I chose ten broad classes of distinct sound types (baboon, bird, cat, cattle, dog,

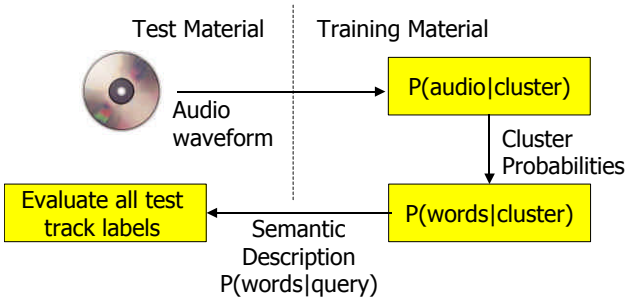

Figure 5: A schematic of the audio to semantic testing procedure.

fowl, goat, horse, lion, pig, sheep). The stacked features from only those audio tracks that fit these classes were used as input to the LDA algorithm. This computation produced a matrix that reduced the 91-dimensional data to the 10-dimensional subspace that best discriminates between these 10 classes. This dimensionality reduction was fixed for all experiments.

A simple test was used to set the amount of smoothing in the language models. Without smoothing, the semantic lookup results were poor because many of the General sounds were labeled with the word "animal," which was seldom used in the BBC labels. The results here were generated using a backoff method that added a small constant probability ($1/N_w$, where N_w is the number of words in the vocabulary) to each word model.

4.3 Labeling Tests

Figure 5 shows the test procedure for the acoustic-to-semantic task (a similar procedure is used to test semantic-to-audio labeling.) Audio from each test track is applied as an acoustic query to the system. The MPESAR system calculates the probability of each cluster given this acoustic query. These cluster probabilities are used to weight the semantic models associated with each cluster. The result is a multinomial probability distribution that represents the probabilities that each word in the dictionary describes the acoustic test track. The likelihoods that each test-track description fit the query's semantic description were sorted and the rank of the true test label was recorded.

Figures 6 and 7 show histograms of the true test ranks for both directions of the MPESAR algorithm. Figure 6 shows the acoustic-to-semantic results and the median rank of the true result over all the test tracks is 17.5. Figure 7 shows the semantic-acoustic results and the median rank of the true result for this direction is 9. At this point I do not understand the difference in performance between these two directions.

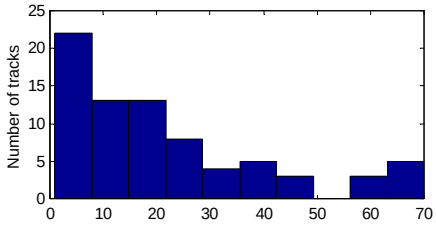

Figure 6: Histogram of true label ranks based on likelihoods from audio to semantic tests.

5. CONCLUSIONS

This paper described a system that uses mixture of probability experts to learn the connection between an audio and a semantic space, and the reverse. It describes the conversion of sound and text into acoustic and semantic spaces and the process of creating the mixture of probability experts. The system was tested using commercial sound-effect CDs and is effective at labeling acoustic queries with the most appropriate words, and for finding sounds that fit a semantic query.

There are several improvements to this system that are worth pursuing. First, an algorithm that integrates the clustering and the MPE training will improve the system's models. Second, a richer acoustic description, perhaps replacing the GMMs with hidden Markov models, will provide more discrimination power. Finally, larger training sets will improve the system's knowledge.

ACKNOWLEDGEMENTS

I appreciate the assistance that I received from Byron Dom, Dulce Ponceleon, Arnon Amir, Myron Flickner, John Fisher, Clemens Drews and Michele Covell. I used Ian Nabney's NETLAB software to calculate the GMMs; Roger Jang provided the 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] Stanley F. Chen and Joshua Goodman. "An Empirical Study of Smoothing Techniques for Language Modeling." TR-10-98, Center for Research in Computing Technology, Harvard University, August 1998.

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

[5] Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, Tom M. Mitchell. "Learning to classify text from labeled and unlabeled documents." *Proceedings of AAAI-98,* Madison, US, pp. 792–799, 1998.

[6] Thomas F. Quatieri. *Discrete-Time Speech Signal Processing: Principles and Practice*. Prentice Hall, 2002.

[7] Malcolm Slaney. "Semantic–Audio Retrieval." *Proceedings of the 2002 IEEE ICASSP*, Orlando, FL, 2002.

[8] Steven Waterhouse. "Classification and Regression using Mixtures of Experts." Ph.D Thesis, Department of Engineering University of Cambridge, 1997.

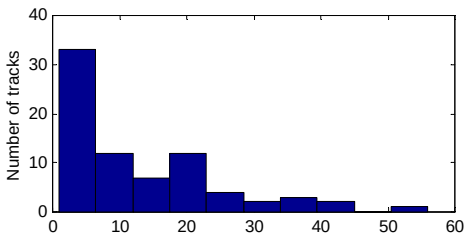

Figure 7: Histogram of true label ranks based on likelihoods from semantic-to-audio tests