Proposed Technique for Content Based Sound Analysis and Ordering Using CASA and PAMIR Algorithm

Senthil Kumar T.K [1] Dheepak.G [2] Rajalingam.S [3]

[1] ***Department of Electronics and Communication, Anna University, Rajalakshmi Institute of Technology***
Chennai, Tamilnadu, India

[2] ***Department of Electronics and Communication, Anna University, Rajalakshmi Institute of Technology***
Chennai, Tamilnadu, India

[3] ***Department of Electronics and Communication, Anna University, Rajalakshmi Institute of Technology***
Chennai, Tamilnadu, India

Abstract

Making the machine to hear as a human is one of the emerging technology of current technical world. If we can make machines to hear as humans, then we can use them to easily distinguish speech from music and background noises, to separate out the speech and music for special treatment, to know from direction sounds are coming, to learn which noises are typical and which are noteworthy. These machines should be able to listen and react in real time, to take appropriate action on hearing noteworthy events, to participate in ongoing activities, whether in factories, in musical performances, or in phone conversations. The existing auditory models for automatic speech recognition (ASR) has not been entirely successful, due to the highly evolved state of ASR system technologies , which are finely tuned to existing representations and to how phonetic properties of speech are manifest in those representations.

One particularly promising area of machine hearing research is computational auditory scene analysis (CASA). To the extent that we can analyze sound scenes into separate meaningful components, we can achieve an advantage in tasks involving processing of those components separately. Separating speech from interference is one such application. This paper deals with the retrieval of the sound from the text queries using CASA and PAMIR algorithms with Pole-Zero filter cascade peripheral model. This paper work on content-based sound ranking system that uses biologically inspired auditory features and successfully learns a matching between acoustics and known text**.**

Keyword *: PAMIR, PZFC, Sparse Coding, AGC*

1. INTRODUCTION

Machine Hearing is a field aiming to develop systems that can process, identify and classify the full set of sounds that people are exposed to. Like machine vision, machine hearing involves multiple problems: from auditory scene analysis, through "auditory object" recognition to speech processing and recognition. While considerable effort has been devoted to speech and music related research, the wide range of sounds that people – and machines – may encounter in their everyday life has been far less studied.

Such sounds cover a wide variety of objects, actions, events, and communications: from natural ambient sounds, through animal and human vocalizations, to artificial sounds that are abundant in today's environment. Building an artificial system that processes and classifies many types of sounds poses two major challenges. First, we need to develop efficient algorithms that can learn to classify or rank a large set of different sound categories. Recent developments in machine learning and particularly progress in large scale methods provide several efficient algorithms for this task. Second, and sometimes more challenging, we need to develop a representation of sounds that captures the full range of auditory features that humans use to discriminate and identify different sounds, so that machines have a chance to do so as well. To evaluate and compare auditory representations, we use a real-world task of content-based ranking sound documents given text queries. In this application, a user enters a textual search query, and in response is presented with an ordered list of sound documents, ranked by relevance to the query. For instance, a user typing "dog" will receive an ordered set of files, where the top ones should contain sounds of barking dogs. Importantly, ordering the sound documents is based solely on acoustic content: no text annotations or other metadata are used at retrieval time. Rather, at training time, a set of annotated sound documents (sound files with textual tags) is used, allowing the system to learn to match the acoustic features of a dog bark to the text tag "dog", and similarly for a large set of potential sound-related text queries. In this way, a small labeled set can be used to enable content-based retrieval from a much larger, unlabeled set. Several previous studies have addressed the problem of content-based sound retrieval, focusing mostly on the machine-learning and information-retrieval aspects of that task, using standard acoustic representations. Here we focus on the complementary problem, of finding a good representation of sounds using a given learning algorithm. The current paper proposes a representation of sounds that is based on models of the mammalian auditory system. Unlike many commonly used representations, it emphasizes fine timing relations rather than spectral analysis. We test this representation in a quantitative task: ranking sounds in response to text queries.

2. MODELING SOUNDS

In this paper we focus on a class of representations that is partially based on models of the auditory system **[1].** The Auditory Image Model (AIM) is a time-domain model of human auditory processing to represent how sounds are processed from the inner ear to the auditory nerve and cochlear nucleus **[2]**. In comparison to the frequency-domain representation of sound using spectrogram, which is widely used today as a simple and crude approximation to human hearing, the stabilized auditory image generated from AIM more accurately simulates our perceived auditory image of complex sounds. The motivation for using auditory models follows from the observation that the auditory system is very effective at identifying many sounds, and this may be partially attributed to the acoustic features that are extracted at the early stages of auditory processing. We extract features with a four-step process, illustrated in Fig. 1: (1) A nonlinear filter bank with half-wave rectified output. (2) Strobed temporal integration, that yields a stabilized auditory image (SAI). (3) Sparse coding using vector quantization. (4) Aggregate all frames features to represent the full audio document. The first two steps, filter bank and strobed temporal integration, are firmly rooted in auditory physiology and psychoacoustics The third processing step, sparse coding, is in accordance with some properties of neural coding and has significant computational benefits that allow us to train large scale models. The fourth step takes a "bag of features" approach which is common in machine vision and information retrieval. The remainder of this section describes these three steps in detail.

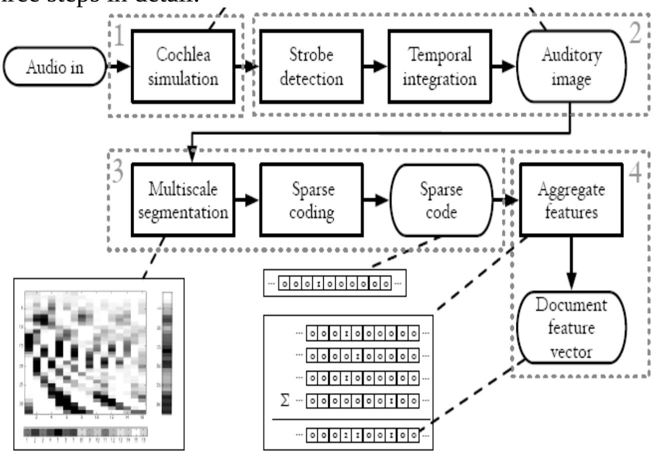

Figure 1

2.1 Cochlear model filter bank

The first processing step is a cascade filter bank inspired by cochlear dynamics **[2]**, known as the *pole–zero filter cascade (PZFC)* (Figure 2). It produces a bank of band pass-filtered, half-wave rectified output signals that simulate the output of the inner hair cells along the length of the cochlea. The PZFC can be viewed as approximating the auditory nerve's instantaneous firing rate as a function of cochlear place, modeling both the frequency filtering and the compressive or *automatic gain control* characteristics of the human cochlea**.[2]** The PZFC also models the adaptive and frequency dependent gain that is observed in the human cochlea, thereby making an *automatic gain control* (AGC) system.

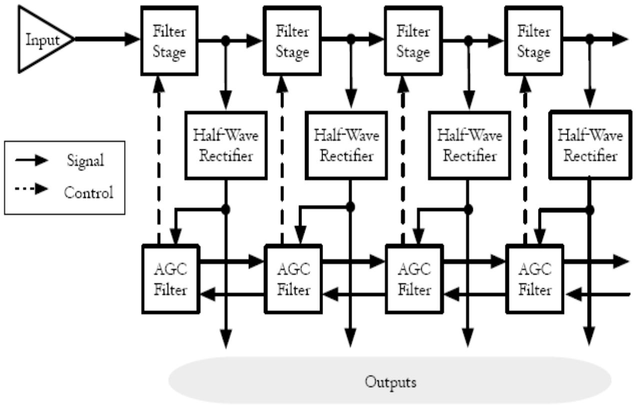

Figure 2

2.2 Strobe finding and image stabilization

The second processing step, strobed temporal integration (STI), is based on human perception of sounds, rather than purely on the physiology of the auditory system. In this step, PZFC output is passed through a strobe-finding process, which determines the position of "important" peaks in the output in each channel. These strobe points are used to initiate temporal integration processes in each channel, adding another dimension to represent time delay from the strobe, or trigger, points. Intuitively, this step "stabilizes" the signal, in the same way that the trigger mechanism in an oscilloscope makes a stable picture from an ongoing time-domain waveform. The end result of this processing is a series of two-dimensional frames of real-valued data (a "movie"), known as a "stabilized auditory image" (SAI) **[3].** Each frame in this "movie" is indexed by cochlear channel number on the vertical axis and lags relative to identified strobe times on the horizontal axis. Examples of such frames are illustrated in Fig. 3 and Fig.4.

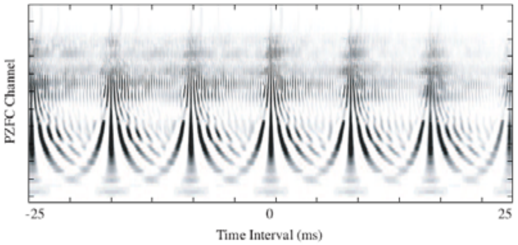

Figure 3

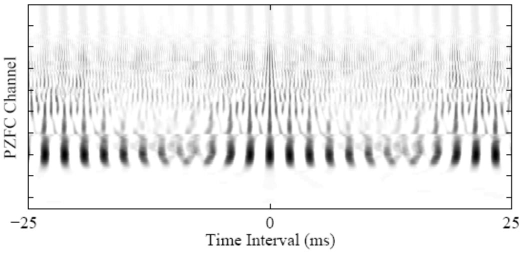

Figure 4

2.3 Sparse coding of an SAI

The third processing step transforms the content of SAI **[3]** frames into a *sparse code* that captures repeating local patterns in each SA image. Sparse codes have become prevalent in the characterization of neural sensory systems. As such it provides a powerful representation that can capture complex structures in data, while providing computational efficiency. Specifically, sparse codes can focus on typical patterns that frequently occur in the data, and use their presence to represent the data efficiently.

3. RANKING SOUNDS GIVEN TEXT QURIES

We now address the problem of ranking sound documents by their relevance to a text query. Practical uses of such a system include searching for sound files or specific moments in the sound track of a movie. For instance, a user may be interested to find vocalizations of monkeys to be included in a presentation about the rain-forest, or to locate the specific scene in a video where a breaking glass can be heard. A similar task is “musical query-by-description”, in which a relation is learned between audio documents and words. We solve the ranking task in two steps. In the first step, sound documents are represented as sparse vectors. In the second step, we train a machine learning system to rank the documents using the extracted features. In this study we use the *PAMIR* method as a learning algorithm. PAMIR uses a fast and robust training procedure to optimize a simple linear mapping from features to query terms, given training data with known tags. The query is represented as a sparse vector of terms in the tag vocabulary (about 3,000 words), and each sound file is given a score respect to a query via a linear matrix product of features times matrix times query. The matrix is trained to optimize a ranking criterion, such that it attempts to rank "relevant" documents higher, by giving them a higher score, than "nonrelevant" ones, in the training set, for a large number of training queries that include multiword queries formed from the tag vocabulary.

4. EXPERIMENTS

We planned to evaluate the auditory representation in a quantitative ranking task using a large set of audio recordings that cover a wide variety of sounds. We compare sound retrieval based on the SAI with standard MFCC features. In what follows we describe the dataset and the experimental setup

4.1 The dataset

We planned to collect a few thousands of sound effect from multiple sources. We planned to collect those from commercially available sound effect collections, BBC sound effects library and through a variety of web sites; *www.findsounds.com,acoustica.com, ilovewavs.com, simplythebest.net, wav-sounds.com*, *wavsource.com*, and *wavlist.com*. We plan to manually label all of the sound effects by listening to them and typing in a handful of tags for each sound. This was used for adding tags to existing tags (from*www.findsounds.com*) and to tag the non-labeled files from other sources. When labeling, the original file name was displayed, so the labeling decision was influenced by the description given by the original author of the sound effect. We like to restrict our tags to a somewhat limited set of terms. We also added high level tags to each file. For instance, files with tags such as ‘rain’, ‘thunder’ and ‘wind’ were also given the tags ‘ambient’ and ‘nature’. Files tagged ‘cat’, ‘dog’, and ‘monkey’ were augmented with tags of ‘mammal’ and ‘animal’. These higher level terms assist in retrieval by inducing structure over the label space. All terms are stemmed, using the Porter stemmer for English. After stemming, we planned to have around 3000 tags.

4.2 The Experimental Setup

We planned to use standard cross validation **[5]** to estimate performance of the learned ranker. Specifically, we like to split the set of audio documents in three equal parts, using two thirds for training and the remaining third for testing. Training and testing was repeated for all three splits of the data, such that we obtained an estimate of the performance on all the documents. We will remove from the training and the test set queries that had fewer than $k = 5$ documents in either the training set or the test set, and removed the corresponding documents if these contained no other tag. We will use a second level of cross validation to determine the values of the hyper parameters: the aggressiveness parameter C, and the number of training iterations. In general performance was good as long as C

was not too high, and lower C values required longer training. We selected a value of C = 0.1, which was also found to work well in other applications and 10M iterations. From our study the system is not very sensitive to the value of these parameters. To evaluate the quality of the ranking obtained by the learned model we can use the precision (fraction of positives) within the top k audio documents from the test set as ranked for each query.

4.3 SAI and sparse coding parameters

The process of transformation of SAI frames into sparse codes **[6]** has several parameters which can be varied. We plan to define a default parameter set and then performed experiments in which one or a few parameters were varied from this default set. The default parameters cut the SAI into rectangles starting with the smallest size of 16 lags by 32 channels, leading to a total of 49 rectangles. All the rectangles were reduced to 48 marginal values each, and for each box a codebook of size 256, for a total of 49 × 256 = 12544 feature dimensions. Using this default experiment as a baseline for comparisons, we can make systematic variations to several parameters and studied their effect on the retrieval precision. First,we modify two parameters that determine the shape of the PZFC filter: Pdamp and Zdamp.**[6]** Then, we modified the smallest rectangle size used for sparse segmentation and by limiting the maximum number of rectangles used for the sparse segmentation. Further variants used systematic variation of the codebook sizes used in sparse coding.

5 CONCLUSION

We described a content-based sound ranking system that uses biologically inspired auditory features and successfully learns a matching between acoustics and known text labels. We described PAMIR to study systematically many alternative sparse-feature representations ("front ends"). Our analysis support the hypothesis a front end that mimics several aspects of the human auditory system provides an effective representation for machine hearing. These aspects include a realistic nonlinear adaptive filter bank and a stage that exploits temporal fine structure at the filter bank output (modeling the cochlear nerve) via the concept of the stabilized auditory image. Importantly however, the auditory model described in this paper may not be always optimal, and future work on characterizing the optimal parameters and architecture of auditory models is expected to further improve the precision, depending on the task at hand. One approach to feature construction would have been to manually construct features that are expected to discriminate well between specific classes of sounds. For instance, periodicity could be a good discriminator between wind in the trees and a howling wolf. However, as number of classes grows, such careful design of discriminative features may become infeasible. Here we take an opposite approach, assuming that perceptual differences rely on lower level cochlear feature extraction, we proposed the models inspired by cochlear processing to obtain a very high dimensional representation, and let the learning algorithm identify the features that are most discriminative. Since our system currently uses only features from short windows, we envision future work to incorporate more dynamics of the sound over longer times, either as a bag-of-patterns using patterns that represent more temporal context, or through other methods.

References

[1] R. F. Lyon, A. C. Kat siam is, and E. M. Drakakis,"History and future of auditory filter models," in *Proc. IEEE Int.* Conf Circuits and Systems, 2010,pp.3809-3812.

[2] R. F. Lyon, "Filter cascades as analogs of the cochlea," in Neuromorphic Systems Engineering: Neural Networks in Silicon , T. S. Lande, Ed. Norwell, MA: I(luwer, 1998, pp. 3-18.

[3] R. D. Patterson, K. Robinson, J. Holdsworth, D. McI(eown , C. Zhang, and M. Aller hand , "Complex sounds and auditory images," in *Proc.* 9th Int. Symp. Hearing, Auditory Physiology and Perception, Y. Cazals, L. Demany, and Ie Horner,Eds. Oxford: Pergamon, 1992, pp. 429-446.

[4] M. Slaney and R. F. Lyon, "On the importanceof time-A temporal representation of time," in Visual Representations of Speech Signals, M. Cooke,S. Beet, and M. Crawford, Eds. New York: Wiley,1993, pp. 95-116.

[5] D. Crangier and S. Bengio, "A neural network to retrieve images from text queries," in Proc. Artificial Neural Networks-ICANN 2006, 2006, pp. 24-34.

[6] M. Rehn, R. F. Lyon, S. Bengio, T. C. Walters,and C. Chechik, "Sound ranking using auditory sparse-code representations," in ICML Workshop Sparse Methods for Music Audio, 2009.

Senthil Kumar T.K is a Lecturer in the department of Electronics and Communication at Rajalakshmi Institute of Technology, Chennai, India. He has three years of teaching experience. He received his Bachelor of Engineering in the stream of Electrical and Electronics from Bhajrang Engineering College and Master of Engineering in Applied Electronics from College of Engineering, Guindy, Anna University. He is a member of IACSIT. He presented some papers in national level conferences. His area of interest includes Digital Image-Signal Processing and Neural Networks.

Dheepak G is a Senior Lecturer in the department of Electronics and Communication at Rajalakshmi Institute of Technology, Chennai, India. He has six years of teaching experience. He received his Bachelor of Engineering in the stream of Electronics and Communication Engineering from Mahendra Engineering College and Master of Engineering in Applied Electronics from K.S.R College of Technology. He is a member of IACSIT and IAENG. His area of interest includes Digital Signal and Image Processing and Computer Networks.

Rajalingam S is a Lecturer in the department of Electronics and Communication at Rajalakshmi Institute of Technology, Chennai, India. He has three years of teaching experience. He received his Bachelor of Engineering in the stream of Electronics and Communication from SACS MAVMM Engineering College and Master of Engineering in VLSI DESIGN from Kongu Engineering College, Perundurai,. He is a member of IACSIT. He received wipro Mission10X trainer Certified. He presented some National conference. His area of interest includes Testing of Vlsi, LowPower Vlsi,Mobile Communication