AUTOMATIC SOUND ANNOTATION

Pedro Cano and Markus Koppenberger
Music Technology Group, Institut de l'Audiovisual
Universitat Pompeu Fabra, 08003 Barcelona, Spain
Phone: +34 935 422 101
Fax: +34 935 422 202
E-mail: pcano, koppi@iua.upf.es
Web: http://www.iua.upf.es/mtg

Abstract.

Sound engineers need to access vast collections of sound effects for their film and video productions. Sound effects providers rely on text-retrieval techniques to offer their collections. Currently, annotation of audio content is done manually, which is an arduous task. Automatic annotation methods, normally fine-tuned to reduced domains such as musical instruments or reduced sound effects taxonomies, are not mature enough for labeling with great detail any possible sound. A general sound recognition tool would require: first, a taxonomy that represents the world and, second, thousands of classifiers, each specialized in distinguishing little details. We report experimental results on a general sound annotator. To tackle the taxonomy definition problem we use WordNet, a semantic network that organizes real world knowledge. In order to overcome the need of a huge number of classifiers to distinguish many different sound classes, we use a nearest-neighbor classifier with a database of isolated sounds unambiguously linked to WordNet concepts. A 30% concept prediction is achieved on a database of over 50.000 sounds and over 1600 concepts.

INTRODUCTION

Post-production audio studios add the sound that accompanies the image in movies. Audio can create the illusion of reality and immersion and is fundamental component of the artistic creation. Sounds are also needed in other audiovisual productions, such as in computer games or web pages. Sometimes sounds are recorded for the occasion. Many times, sound engineers access already compiled sound effects libraries. Main sound effects providers

Table 1: The classifier assigns the metadata of the sounds of the second column to the sounds of the first.

Query Sound Caption	Nearest-neighbor Caption
Mini Cooper Door Closes Interior Persp.	Trabant Car Door Close
Waterfall Medium Constant	Extremely Heavy Rain Storm Short Loop
M-domestic Cat- Harsh Meow	A1v:Solo violin (looped)
Auto Pull Up Shut Off Oldsmobile	Ferrari - Hard Take Off Away - Fast
Animal-dog-snarl-growl-bark-vicious	Dinosaur Monster Growl Roar

employ standard text retrieval techniques to offer their collections.[1][2] The manual annotation is a labor-intensive and error-prone task. According to staff from the Sound Effects Library, it would take 60 years for a librarian to tag a collection of 2 million sounds. There are attempts towards metadata generation by automatic classification. State of the art of audio classification methods are not mature enough to provide the level of detail and coverage needed in a sound effect management system, e.g: "fast female footsteps on wood", "violin pizzicato with natural open strings" or "car door closes - interior perspective". In audio classification, researchers normally assume the existence of a well defined hierarchical classification scheme of a few categories (less than a hundred categories). On-line sound effects and music sample providers have several thousand categories. This makes the idea of generating a model for each category quite difficult as we would need several thousand classifiers.

In this context, we present experiments on an all-purpose sound recognition system based on nearest-neighbor classification rule [2]. A sound sample will be labeled with the descriptions from the similar sounding examples of an annotated database (see Table 1). Besides, the terms of the closest match will be unambiguous thanks to the use of WordNet[3] [11] as the taxonomy back-end. With unambiguous tagging, we refer to assigning concepts and not just terms to sounds. For instance, the sound of a "bar" is ambiguous, it could be "bar" as "rigid piece of metal or wood" or as "establishment where alcoholic drinks are served" where each concept has a unique identifier.

The system has been developed with the aim of easing the task of librarians in audio asset management systems [1]. However, other uses exist, for instance, it can support image-based video recognition systems. Audio can provide reliable cues that complement those of video and image. The rest of the paper is organized as follows: In Section 2 we briefly enumerate some approaches to the problem of automatic sound classification. In Section 3, we present a real-world size taxonomy for sound effect description. From Section 4 to 7 we describe the system setup as well as some results. We end-up with possible continuations of the approach.

[1] http://www.sound-effects-library.com
[2] http://www.sonomic.com
[3] http://www.cogsci.princeton.edu/~wn/

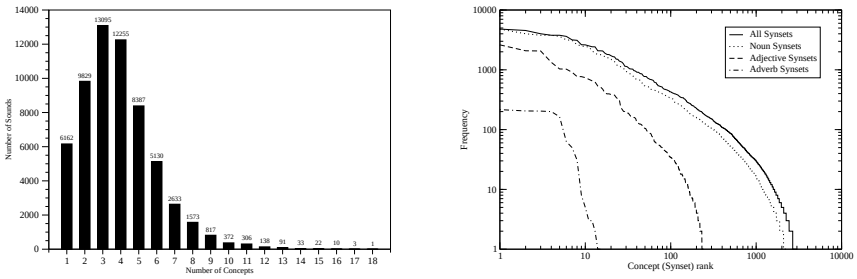
RELATED WORK

Existing classification methods are normally finely tuned to small domains, such as musical instrument classification [6, 8] or simplified sound effects taxonomies [13, 14]. Independently of the feature extraction and selection method and the type of classifier, content-based classification systems need a set of classes and a large number (e.g: 30 or more) of audio samples for each class to train the system.

In musical instrument classification, the taxonomies more or less follow perceptual-related hierarchical structures [9]. Accordingly, in such problems one can devise hierarchical classification approaches such as [10, 12] in which the system distinguishes in a first level between sustained and non-sustained sounds, and in a second level among strings, woodwinds and so on. In everyday sound classification, there is not such a parallelism between semantic and perceptual categories. On the contrary one can find hissing sounds in categories of "tea boilers", "snakes". Sound engineers exploit this ambiguity and create the illusion of "crackling fire" by recording "twisting cellophane". Alternatively, a "cat" can produce "hissing", "purring" or "miaow" sounds. Dubnov and Ben-Shalom [4] point out that one of the main problems faced by natural sounds and sound effects classifiers is the lack of clear taxonomy. It is very difficult to design and implement a taxonomy that organizes a huge amount of concepts—basically any that may appear in the real world. This is the reason why we have chosen an existing semantic network, WordNet.

TAXONOMY MANAGEMENT

WordNet is a lexical network designed following psycholinguistic theories of human lexical memory. Standard dictionary organize words alphabetically. WordNet organizes concepts in synonym sets, called *synsets*, with links between the concepts. It knows for instance that the word piano, as a noun, has two senses, the musical attribute that refers to "low loudness" and the "musical instrument". It also encodes the information that a "grand piano" is a type of "piano", and that it has parts such as a keyboard, a loud pedal and so on. Such a knowledge system is useful for retrieval. It can for instance display the results of a query "car" in types of cars, parts of car, actions of a car (approaching, departing, turning off). Although WordNet organizes over 150.000 terms and 100.000 concepts, there are some concepts missing, e.g. It knows "jaguar" as a feline but not as a British manufactured car. Fortunately, the lexical network can be extended. For further details on the implementation and evaluation of WordNet as backbone for audio taxonomy management, we refer to [3].



(a) Histogram

(b) Synset frequency rank

Figure 1: (a) Histogram of number of synsets (concepts) per sound. (b) Number of sounds described per synset as a function of the synset rank. The frequency rank is plotted for the different parts of speech: noun, verb, adjective and adverbs.

EXPERIMENTAL SETUP

Our database consists of 54.799 sounds from over 30 different libraries of sound effects, music and music samples. These sounds have been unambiguously tagged with concepts of an enhanced WordNet. Thus a cougar sound with the following caption:"cougar short snarl" has been linked to the following synsets (the numbers indicate the unique WordNet identifiers):

- 01792447%n cougar, puma, catamount, mountain lion, painter, panther, Felis concolor – (large American feline resembling a lion)
- 01399940%a short – (primarily temporal sense; indicating or being or seeming to be limited in duration((of instruments in the violin family) to be plucked with the finger)
- 05949121%n snarl – (a vicious angry growl)

In Figure 1a, we show a histogram with the number of synsets the sounds have been labeled with after disambiguation. It should be clear that the higher the number of synsets, the better a sound is described. In average, a sound is labeled with 3.88 synsets. In Figure 1b we plot the frequency of synset appearance as a function of the rank. The plot is repeated for various parts of speech, specifically: noun, verb, adjective and adverb. The distribution of 3028 synsets with respect its syntactic function is as follows: 2381 nouns, 380 verbs, 251 adjectives and 16 adverbs. The number of synsets for which there are ten or more examples sounds is 1645.

The classifier uses a set of 89 features and a nearest-neighbor classifier against the database of sounds with WordNet as taxonomy backbone.

FEATURES EXTRACTION

Every audio sample is converted to 22.05 KHz mono format and goes through a noise gate in order to set its beginning and its end. After a frame-by-frame analysis, we extract features belonging to three different groups: a first group gathering spectral as well as temporal descriptors included in the MPEG-7 standard; a second one built on Bark Bands perceptual division of the acoustic spectrum and which outputs the mean and variance of relative energies for each band; and, finally a third one, composed of Mel-Frequency Cepstrum Coefficients and their corresponding variances (For details on the implementation of the feature extraction module, see [2]).

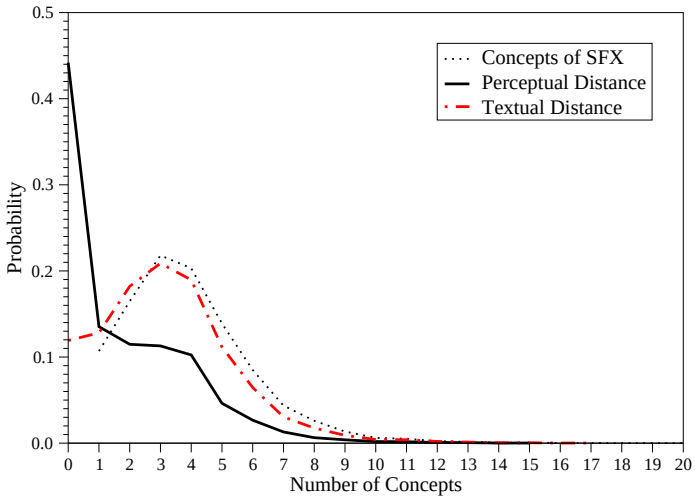


Figure 2: Histogram of correctly identified synsets. For each sound we count the intersection of concepts of the query and the retrieved nearest neighbor. The Concepts per SFX shows the perfect score. The perceptual distance prediction plot indicates the prediction accuracy using the perceptual similarity distance. The textual distance line indicates the prediction using the textual captions and a cosine distance.

NEAREST-NEIGHBOR CLASSIFIER

We use the one-nearest neighbor decision rule (1-NN)[7] for classification. The choice of a memory-based nearest neighbor classifier avoids the design and training of every possible class of sound (the order of several thousands). Besides, it does not need redesign or training whenever a new class of sounds

Table 2: Percussive instruments confusion matrix where SN:Snare, To:Tom, HH:Hihat, CR:Crash, KI:Kick, RI:Ride

	SN	TO	HH	CR	KI	RI
SN	150	1	2	2	1	20
TO	1	148	2	0	19	0
HH	5	7	153	0	1	4
CR	21	0	2	45	0	12
KI	1	17	0	0	182	0
RI	15	0	5	4	0	135

is added to the system. The NN classifier needs a database of labeled instances and a similarity distance to compare them. An unknown sample will borrow the metadata associated with the most similar registered sample. The similarity measure of the system is a normalized Manhattan distance of the above enumerated features:

$$d(x,y) = \sum_{k=1}^{N} \frac{|x_k - y_k|}{(max_k - min_k)}$$

where x and y are the vectors of features, N the dimensionality of the feature space, and max_k and min_k the maximum and minimum values of the kth feature.

In some of our experiments, the standard deviation-normalized Euclidean distance does not perform well. Specially harmful is the normalization with standard deviation. Changing the normalization from the standard deviation to the difference between maximum and minimum boosted classification. For example the percussive instrument classification (see Table 2 and Section 7) raises from 64% to 82% correct identification. Changing the distance from Euclidean to Manhattan gives us an extra 3%.

EXPERIMENTAL RESULTS

The first experiment consisted in finding a best-match for all the sounds in the database. Table 1 shows some examples: on the left column the original caption of the sound and on the right the caption of the nearest neighbor. The caption on the right would be assigned to the query sound in an automatic annotation system.

As can be inferred from Table 1, it is not trivial to quantitatively evaluate the performance of the system. An intersection on the terms of the captions would not yield a reasonable evaluation metric. The WordNet based ontology can inform us that both “Trabant” and “Mini Cooper” are narrow terms for the concept “car, automobile”. Thus, the comparison of number of common synsets on both query and nearest-neighbor could be used as a better evaluation. As was shown in previous work [2], the intersection of synsets between query and best-match is 1.5 in average, while 50% of the times the

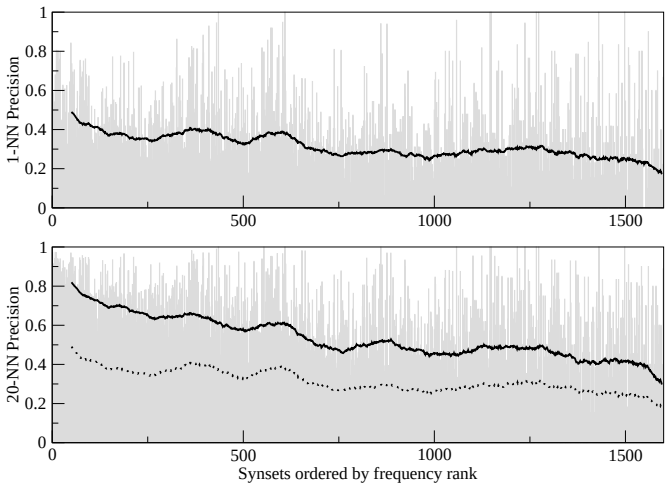


Figure 3: Synset accuracy using the 1-NN perceptual distance. The X axis corresponds to the synsets ordered by its frequency rank. The graph at the top shows the accuracy on the 1-NN. The bottom graph displays the how often at least one of the 20 best retrieved sounds was labeled by a synset. The plots have been smoothed with an average filter. The dotted line of the bottom graph reproduces the precision on the 1-NN of the top graph.

best-match did not share a single common synset (see Figure 2). The intersection of source descriptions can be zero for very similar sounding sounds. The closest-match for a "paper bag" turns out to be a "eating toast". These sounds are semantically different but perceptually similar.

Another experiment is the prediction of synsets, that is, how well a particular concept, say "cat miaow", will retrieve "miaow" sounds. The methodology is as follows. For each synset, we retrieve the sounds that have been labeled with that particular synset. For each sound its nearest-neighbor is calculated. We finally compute how many best-matching sounds are also labeled with that synset. From the total of 3028 synsets we restricted the experiment to the ones that had been attached to 10 or more sounds. This leaves us with 1645 synsets. Figure 3 displays the results. The top figure shows how often a synset retrieved sounds whose best-matches were also labeled with that synset. The bottom figure, on the other hand, shows the the probability that at least one the best 20 retrieved sounds has been labeled with the particular synset. The ordering of synsets on the x-axis corresponds to their frequency rank as displayed in Figure 3. It is interesting to see that there is not a strong correlation between the synset frequency and the precision. On a random guess one would expect some synsets predicted much better only because they are very frequent.

Table 3: Harmonic instruments confusion matrix where AF:AltoFlute, AS:AltoSax, BF:BassFlute, BT:BassTrombone, BA:Bassoon, BC:BbClarinet, CE:Cello, DB:DoubleBass, EC:EbClarinet, FL:Flute, HO:Horn, OB:Oboe, PI:Piano, SS:SopranoSax, TT:TenorTrombone.

	AF	AS	BF	BT	BA	BC	CE	DB	EC	FL	HO	OB	PI	SS	TT
AF	7	0	3	0	0	0	0	0	0	1	0	0	0	0	0
AS	0	18	0	0	0	1	0	0	0	0	0	0	0	0	0
BF	0	0	9	0	0	0	0	0	0	0	0	0	0	1	0
BT	0	0	0	9	0	0	0	0	0	0	0	0	0	0	1
BA	0	0	0	0	14	0	0	0	0	0	0	0	0	1	0
BC	0	0	0	1	0	10	1	1	0	0	0	1	0	0	0
CE	0	1	0	0	0	1	74	3	0	0	0	0	0	0	0
DB	0	0	0	0	0	0	0	72	0	0	0	0	0	0	0
EC	0	1	1	0	0	2	0	0	5	1	0	1	0	2	1
FL	1	2	0	3	0	1	0	0	0	11	0	4	0	0	0
HO	0	0	0	0	2	0	0	0	0	0	10	0	0	0	0
OB	0	1	0	0	0	0	2	0	0	0	1	7	0	0	1
PI	0	0	0	0	0	0	0	0	0	0	0	0	87	0	0
SS	0	0	0	0	0	1	0	0	0	0	0	0	0	24	0
TT	0	0	0	2	0	0	0	0	0	0	0	0	0	0	7

In another experiment, we tested the general approach in reduced domain classification regime mode, specifically percussive and harmonic instrument sound, and we achieve acceptable performances. The assumption is that there is a parallelism between semantic and perceptual taxonomies in musical instruments. Psychoacoustic studies [9] revealed groupings based on the similarities in the physical structure of instruments. We have therefore evaluated the similarity with classification on the musical instruments space, a subspace of the universe of sounds.

Table 3 depicts the confusion matrix of a 15 class classification which corresponds with a 91% (261 audio files). In the 6 class percussive instrument classification an 85% Recognition (955 audio files) using 10 fold validation (see Table 2). The particularity of those results, comparable to state of the art, is that they are achieved with a classifier that has not been fine-tuned to musical instruments nor there are discriminative or feature selection methods employed to improve the classification.

The last experiment is the robustness of the NN classification framework to audio distortions. The harmonic instruments samples of the experiments of Table 3 have been transcoded and resampled into WAV PCM format and Ogg format[4]. The results are depicted in Table 4. The percentages indicate the classification accuracy using different audio qualities. The columns are the audio qualities used as reference.

[4] `http://www.vorbis.com`

Table 4: Retrieval consistency on different distortions on the harmonic instruments classification. The columns indicate the reference audio quality and the rows the performance with the different distortions. Wav: PCM Microsoft WAV format, Ogg: Ogg Vorbis encoding, #kHz: Sampling rate

	Wav 44kHz	Ogg 44kHz	Ogg 11kHz
Wav 44kHz	91.5%	92.0%	75.0%
Wav 22kHz	86.4%	85.6%	82.0%
Wav 11kHz	71.8%	73.1%	89.3%
Ogg 44kHz	90.3%	91.5%	76.0%
Ogg 11kHz	74.0%	74.8%	91.5%

SUMMARY

A major issue when building sound classification systems is the need of a taxonomy that organizes concepts and terms unambiguously. If the task is classifying any possible sound, the taxonomy design becomes a daunting task. We need a taxonomy or classification scheme that encodes the common sense knowledge of the world. WordNet can be be used as a starting taxonomy. Normally, in identification a classifier is build to identify certain concepts: "cars" , "laughs", "piano". Sound samples are gathered and are tagged with those concepts and a classifier is trained to learn that concept. The number of concepts and its possible combinations in the real world makes this approach unfeasible, one would need to train tens of thousands of classifiers and new ones would have to be trained for new concepts. We have presented an alternative approach that uses an unambiguously labelled big audio database. The classifier applies nearest-neighbor rule and a database of sounds with WordNet as taxonomy backbone. As a results the list of possible sources is presented to the user: this sound could be a "paper bag" or "toast"+"eating". Information from text or images can be used to disambiguate the possibilities.

We acknowledge that the use of a single set of features and a single distance for all possible sound classes is rather primitive. However, as Figure 3 indicates, there is room for improvement. The NN rule can be combined with other classifiers: If the system returns that a particular sound could be a violin pizzicato or a guitar, we can then retrieve pizzicato violin and guitar sounds of the same pitch and train a classifier to decide which is more likely. Another example is "car approaches", we can look for other "cars" and other "motor vehicle" "approaches" or "departs" to decide which is the right action. This same thinking applies to adjective type of modifiers, something can be described as "loud", "bright" or "fast". The concept "fast" means something different if we talk of "footsteps" or "typing".

The system can be publicy accessed and tested through a web interface which allows users to upload sounds at `http://www.audioclas.org`.

ACKNOWLEDGMENTS

We thank the staff from the Tape Gallery for all the support, discussion and feedback. This work is partially funded by the AUDIOCLAS Project E! 2668 Eureka. We thank Nicolas Wack for Nearest-neighbor client-server implemention and Emilia Gómez, Fabien Gouyon and Perfecto Herrera for their feedback.

REFERENCES

[1] P. Cano, M. Koppenberger, S. L. Groux, P. Herrera and N. Wack, "Perceptual and Semantic Management of Sound Effects with a WordNet-based Taxonomy," in **Proc. of the ICETE**, Setúbal, Portugal, 2004.

[2] P. Cano, M. Koppenberger, S. L. Groux, J. Ricard, P. Herrera and N. Wack, "Nearest-neighbor Generic Sound Classification with a WordNet-based taxonomy," in **Proc.116th AES Convention**, Berlin, Germany, 2004.

[3] P. Cano, M. Koppenberger, P. Herrera and O. Celma, "Sound Effects Taxonomy Management in Production Environments," in **Proc. AES 25th Int. Conf.**, London, UK, 2004.

[4] S. Dubnov and A. Ben-Shalom, "Review of ICA and HOS Methods for Retrieval of Natural Sounds and Sound Effects," in **4th International Symposium on Independent Component Analysis and Blind Signal Separation**, Japan, 2003.

[5] B. Gygi, **Factors in the identification of environmental sounds**, Ph.D. Thesis, Indiana University, 2001.

[6] P. Herrera, G. Peeters and S. Dubnov, "Automatic Classification of Musical Instrument Sounds," **Journal of New Music Research**, vol. 32, no. 1, 2003.

[7] A. K. Jain, R. P. Duin and J. Mao, "Statistical Pattern Recognition: A Review," **IEEE Transactions on Pattern Analysis and Machine Intelligence**, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[8] B. Kostek and A. Czyzewski, "Representing Musical Instrument Sounds for Their Automatic Classification," **J. Audio Eng. Soc.**, vol. 49, no. 9, pp. 768–785, 2001.

[9] S. Lakatos, "A common perceptual space for harmonic and percussive timbres," **Perception & Psychoacoustics**, , no. 62, pp. 1426–1439, 2000.

[10] K. D. Martin, **Sound-Source Recognition: A Theory and Computational Model**, Ph.D. Thesis, M.I.T., 1999.

[11] G. A. Miller, "WordNet: A Lexical Database for English," **Communications of the ACM**, pp. 39–45, November 1995.

[12] G. Peeters and X. Rodet, "Hierarchical Gaussian Tree with Inertia Ratio Maximization for the Classification of Large Musical Instruments Databases," in **Proc. of the 6th Int. Conf. on Digital Audio Effects**, London, 2003.

[13] E. Wold, T. Blum, D. Keislar and J. Wheaton, "Content-Based Classification, Search, and Retrieval of Audio," **IEEE Multimedia**, vol. 3, no. 3, pp. 27–36, 1996.

[14] T. Zhang and C.-C. J. Kuo, "Classification and Retrieval of Sound Effects in Audiovisual Data Management," in **Proceedings of the 33rd Asilomar Conference on Signals, Systems and Computers**, 1999.