Vector-based Representation and Clustering of Audio Using Onomatopoeia Words

Shiva Sundaram and **Shrikanth Narayanan**
Speech Analysis and Interpretation Laboratory (SAIL),
Dept. Electrical Enginneering-Systems, University of Southern California,
3740 McClintock Ave, EEB400, Los Angeles, CA 90089. USA
sssundara@usc.edu, shri@sipi.usc.edu

Abstract

We present results on organization of audio data based on their descriptions using onomatopoeia words. Onomatopoeia words are imitative of sounds that directly describe and represent different types of sound sources through their perceived properties. For instance, the word *pop* aptly describes the sound of opening a champagne bottle. We first establish this type of audio-to-word relationship by manually tagging a variety of audio clips from a sound effects library with onomatopoeia words. Using principal component analysis (PCA) and a newly proposed distance metric for word-level clustering, we cluster the audio data representing the clips. Due to the distance metric and the audio-to-word relationship, the resulting clusters of clips have similar acoustic properties. We found that as language level units, the onomatopoeic descriptions are able to represent perceived properties of audio signals. We believe that this form of description can be useful in relating higher-level descriptions of events in a scene by providing an intermediate perceptual understanding of the acoustic event.

Introduction

Automatic techniques are required to interpret and manage the ever-increasing multimedia data that is acquired, stored and delivered in a wide variety of forms. In interactive environments, involving humans and/or robots, data is available in the form of video/images, audio and a variety of sensors depending on the nature of the application. Each of these represent different forms of communication and a variety of expressions. To utilize and manage them effectively, such as reason with them in a human-robot interaction, it is desirable to organize, index and label these forms according to their content. Language (rather textual) description or annotation is a concise representation of an event that is useful in this respect. It makes the audio and video data more presentable and accessible for reasoning, and or search/retrieval. This also aids in developing machine listening systems that can use aural information for decision making tasks. The work we present here mainly deals with ontological representation and characterization of different audio events. While the recorded data is stored in signal feature space (such as in terms of frequency components or energy etc.) for automatic processing, text annotation represents the audio clip in the semantic space. The underlying representations of an audio clip in the signal feature space and in semantic space are different. This is because the feature vectors represent signal level properties (frequency components, energy etc.) while in the semantic space the definition is based on human perception and context information. This semantic definition is often represented using natural language in textual form, since words directly represent 'meaning'. Therefore natural language representation of audio properties and events are important for semantic understanding of audio, and it is the focus of this present paper.

In what we call as content-based processing, natural language representations are typically established by a naive labeling scheme where the audio data is mapped onto a set of pre-specified classes. The resulting mapped clusters are used to train a pattern classifier and eventually used to identify the correct class for a given test data. Examples of such systems are in (Guo & Li 2003; L. Liu & Jiang 2002; T. Zhang 2001). While such an approach yields high classification accuracy, they have limited scope in characterizing generic audio scenes, save for situations where the expected audio classes are known previously. Other techniques for retrieval that better exploit semantic relations in language is implemented in (P. Cano, Herrera, & Wack 2004). Here the authors have used WordNet (Fellbaum 1998) to generate word tags for a given audio clip using acoustic feature similarities, and also retrieve clips that are similar to the initial tags. While such semantic relations in language are important in building audio ontologies, they are still sufficiently insulated from signal level properties that directly affect the perception of sources.

In our work, however, we present an approach to use semantic information that are closer to signal level properties. This is implemented using onomatopoeia words present in the English language. These are words that are imitative of sounds (as defined by the Oxford English Dictionary). We believe that such a description will help tackle the potential disambiguity in generic linguistic characterizations of audio. The presentation of the idea is as follows. We first represent the onomatopoeia words as vectors in a 'meaning space'. This is implemented using the proposed inter-word distance metric. We then tag (offline) various clips of acoustic sources from a general sound effects library with appropriate onomatopoeia words. These words are the descriptions of the acoustic properties of the corresponding audio clip. Using the tags of each clip, and the vector representation of each word, we represent and cluster the audio clips in the meaning space. Using an unsupervised clustering algorithm and a model fit measure, the clips are then clustered according to their representation in this space. The resulting clusters are both semantically relevant and share similar per-

bang	bark	bash	beep	biff	blah	blare	blat	bleep
blip	boo	boom	bump	burr	buzz	caw	chink	chuck
clang	clank	clap	clatter	click	cluck	coo	crackle	crash
creak	cuckoo	ding	dong	fizz	flump	gabble	gurgle	hiss
honk	hoot	huff	hum	hush	meow	moo	murmur	pitapat
plunk	pluck	pop	purr	ring	rip	roar	rustle	screech
scrunch	sizzle	splash	splat	squeak	tap-tap	thud	thump	thwack
tick	ting	toot	twang	tweet	whack	wham	wheeze	whiff
whip	whir	whiz	whomp	whoop	whoosh	wow	yak	yawp
yip	yowl	zap	zing	zip	zoom			

Table 1: Complete list of Onomatopoeia Words used in this work.

ceived acoustic properties. We also present some examples of the resulting clusters. Next, we briefly discuss the motivation for this research.

Motivation: Describing sounds with words. Humans are able to express and convey a wide variety of acoustic events using language. This is achieved by using words that express the properties of a particular acoustic event. For example, if one attempts to describe the event "*knocking on the door*" , the words "*tap-tap-tap*" describe the acoustic properties well. Communicating acoustic events in such a manner is possible because of a two way mapping between the acoustic space and language or semantic space. Existence of such a mapping is a result of common understanding of familiar acoustic events. The person communicating the acoustic aspect of the event "*knocking on the door*" may use the words "*tap*" to describe it. That individual is aware of a provision in language (the onomatopoeia word "*tap*") that would best describe it to another. The person who hears the word is also familiar with acoustic properties associated with the word "*tap*". Here, it is important to point out the following issues: (1) There is a difference in the language descriptions "*knocking on the door*" and "*tap-tap*". The former is an original lexical description of the event and the later is closer to the description of the acoustic properties of the knocking event. (2) Since the words such as "*tap*" describe the acoustic properties, they can also represent multiple events (for example, knocking a door, horse hooves on tarmac etc.). Other relevant examples of such descriptions using onomatopoeia words of familiar sounds are as follows:

- In case of sounds of birds: A hen *clucks*, a sparrow *tweets*, a crow or raven *caws*, and an owl *hoots*.
- Example of sounds from everyday life: A door close is described as a *thud* and/or *thump*. A door can *creak* or *squeak* while opening or closing. A clock *ticks*. A doorbell is described with the words *ding* and/or *dong* or even *toot*.

In general, onomatopoeic description of such sounds is not restricted to single word expressions. One usually uses multiple words to paint an appropriate acoustic picture. The above examples also provide the rationale for using onomatopoeic descriptions. For example, by their onomatopoeic descriptions, the sound of door bell is closer to an owl hooting whereas their lexical descriptions (that semantically represents the events using the sound sources "door bell" and "owl") are entirely different. It is also possible to draw a higher level of inference from the onomatopoeic description of an audio event. Given the scene of a thicket or a barn, the acoustic features of the sample clip with *hoot* as its description is likely to be an owl than a door bell. However, given the scene of a living room, the same acoustic features are more likely to represent a door bell. Based on such ideas, it can be seen that descriptions with onomatopoeia words automatically provide a flexible framework for recognition or classification of general auditory scenes. In the next sections, we start with the implementation of the analysis in this work.

Implementation

Distance metric in lexical meaning space

The onomatopoeia words are represented as vectors using a semantic word based similarity/distance metric and Principal Component Analysis (PCA). The details of this method follows:

A set $\{L_i\}$ consisting of l_i words is generated by a Thesaurus for each word O_i in the list of onomatopoeia words. Then the similarity between the j^{th} and k^{th} word can be defined to be:

$$s(j,k) \quad = \quad \frac{c_{j,k}}{l^d_{j,k}}, \tag{1}$$

resulting in a distance measure :

$$d(j,k) \quad = \quad 1 - s(j,k) \tag{2}$$

Here $c_{j,k}$ is the number of common words in the set $\{L_j\}$ and $\{L_k\}$ and $l^d_{j,k}$ is the total number of words in the union of $\{L_j\}$ and $\{L_k\}$. By this definition it can be seen that

$$0 \leq d(j,k) \leq 1 \tag{3}$$

$$d(j,k) = d(k,j) \tag{4}$$

$$d(k,k) = 0 \tag{5}$$

Except for the triangular inequality, it is a valid distance metric. It is also semantically relevant because the words in the set $\{L_j\}$ and $\{L_k\}$ generated by the Thesaurus have some *meaning* associated with the words O_j and O_k in the language. The similarity between two words depends on the number of common words (a measure of *sameness* in meaning). Therefore for a set of W words, using this distance metric, we get a symmetric $W \times W$ distance matrix where the $(j,k)^{th}$ element is the distance between the j^{th} and k^{th} word. Note that the j^{th} row of the matrix is a vector representation of the j^{th} word in terms of other words present in the set. We perform principal component analysis (PCA) (R. O. Duda & Stork 2000) on this set of *feature* vectors, and represent each word as a point in a smaller dimensional space $\mathcal{O}^d$ with $d < W$. In our implementation the squared sum of the first eight ordered eigenvalues covered more than 95% of the total squared sum of all the eigenvalues. Therefore $d = 8$ was selected for reduced dimension representation and $W = 83$. Thus these points (or vectors) are representation of the onomatopoeic words in meaning space.

Table 1 lists all the onomatopoeia words used in this work. By studying the words it can be seen that many have overlapping meanings (eg. *clang* and *clank*), some words are 'closer' in meaning to each other with respect to other

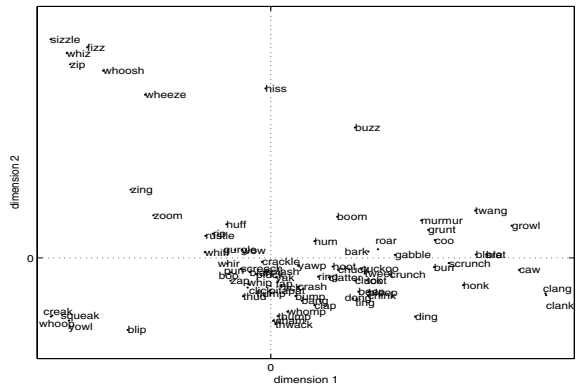

Figure 1: Arrangement of some onomatopoeia words in 2 dimensional 'meaning space'. Note that words such as *clang* and *clank* are closer to each other, but they are farther away from words such as *fizz* and *sizzle*

words (eg. *fizz* is close to *sizzle*, *bark* is close to *roar*, but *(fizz/sizzle)* and *(bark/roar)* are far from each other). These observations can also be made from Figure 1 that illustrates the arrangement of the words in a $d = 2$ dimensional space. Observe that the words *growl* and *twang* are close to each other. This is mainly because the words are represented in a low dimensional space ($d = 2$) in the figure.

Once we have the tags for each audio clip, the clips can also be represented as vectors in the meaning space. Next, we discuss the tagging procedure that results in onomatopoeic descriptions of each audio clip. Later, vector representation based on these tags is discussed.

Tagging the Audio Clips with onomatopoeia words

A set of 236 audio clips were selected from the BBC sound effects Library (http://www.sound ideas.com 2006). The clips were chosen to represent a wide variety of recordings belonging to categories such as: *animals, birds, footsteps, transportation, construction work, fireworks etc.* Four subjects (with English as their first language) volunteered to tag this initial set of clips with onomatopoeia words. A Graphical User Interface (GUI) based software tool was designed to play each clip in stereo over a pair of headphones. All the clips were edited to be about 10-14 seconds in duration. The GUI also had the complete list of the words. The volunteers were asked to choose words that best described the audio by clicking on them. The clips were randomly divided into 4 sets, so that the volunteers spent only 20-25 minutes at a time tagging the clips in each set. The chosen words were recorded as the onomatopoeia tags for the corresponding clip. The tags of all the volunteers recorded for each clip were counted. The tags that had a count of two or more were retained and the rest discarded. This results in tags that are common to all the responses of the volunteers. This tagging method is illustrated in Figure 2. Note that the resulting tags are basically onomatopoeic descriptions that best represent the perceived audio signal. The tags for this initial set of words were then transposed to other clips with similar original lexical description. For example,

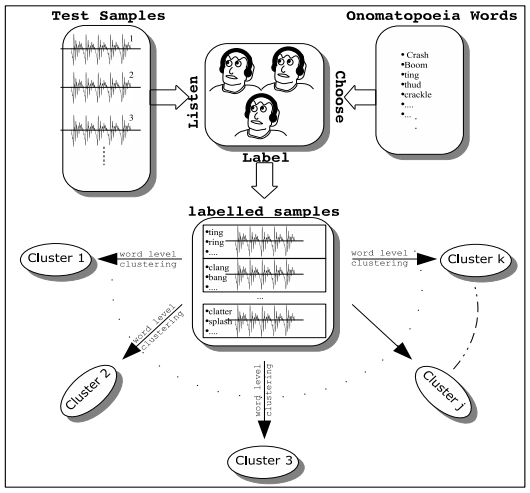

Figure 2: Tagging and clustering the audio clips with onomatopoeia words.

the clip with the lexical name BRITISH_SAANEN_GOAT_1_BB received the tags {*blah, blat, boo, yip, yowl*} and this same set of words were used to tag the file BRITISH_SAANEN_GOAT_2_BB. Similarly, the audio clip BIG_BEN_10TH_STRIKE_12_BB received the tags {*clang, ding, dong*}. These tags were also used for the file BIG_BEN_2ND_STRIKE_12_BB After transposing the tags, a total of 1014 clips was available. Next, we represent each tagged audio clip in the meaning space.

Vector representation of audio clips in meaning space

The vector representation of the tagged audio clips in two dimensions is illustrated in Figure 3. The vectors for each audio clip is simply the sum of the vectors that correspond to the onomatopoeic tags. Let the clip HORSE_VARIOUS_SURFACES_BB have the onomatopoeic description tags {*pitpat, clatter*}. Now the tags *pitpat* and *clatter* are already represented as vectors in meaning space. Performing a vector sum of the vectors that correspond to these tags (*pitpat*, *clatter*), i.e, the sum of the vectors of the points 1 and 2 shown in Figure 3. This results in the point 3. Therefore, the vector of point 3 is taken to be the vector of the clip HORSE_VARIOUS_SURFACES_BB . The implicit properties of this representation technique is as follows:

- If two or more audio clips have the same tags then the resulting vectors of the clips would be the same.
- If two clips have similar meaning tags (not the same tags) then the resulting points of the vectors of the clips would be in close proximity to each other. For example, let clips A and B have tags {*sizzle, whiz*} and {*fizz, whoosh*} respectively. Since these tags are already close to each other in the meaning space (refer to Figure 1), because of the vector sum, the resulting points of the vectors of clips A and B would also be in close proximity to each other. In contrast, if the clips have tags that are entirely different from each other, then the vector sum would result in points that are relatively far from each other. Sub-

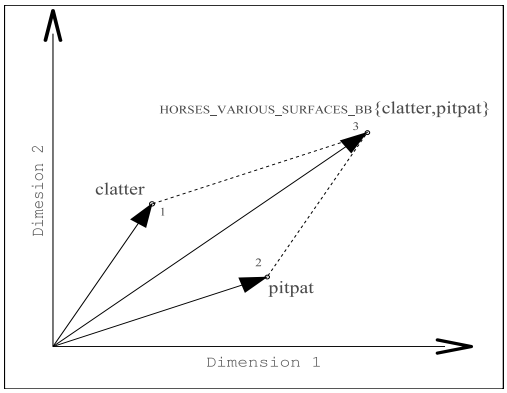

Figure 3: Vector representation of the audio clip HORSE_VARIOUS_SURFACES_BB with tags {*clatter, pitpat*}

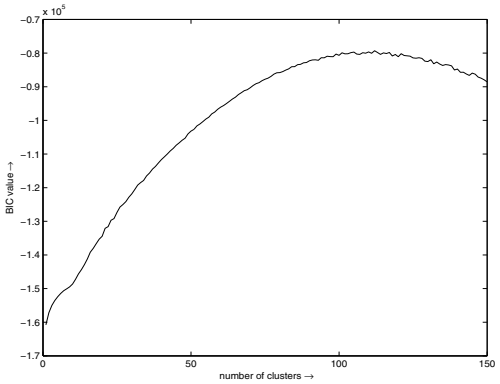

Figure 4: BIC as a function of number of clusters k in model M_k. The maximum value is obtained for $k = 112$.

sequently, using clustering algorithms in this space, audio clips that have similar acoustic and/or semantic properties can be grouped together.

Thus the audio clips can be represented as vectors in the proposed meaning space. This allows us to use conventional pattern recognition algorithms. In this work, we group clips with similar onomatopoeic descriptions (and hence similar acoustic properties) using the unsupervised k-means clustering algorithm. The complete summary of the tagging and clustering the clips is illustrated in Figure 2. The Clustering procedure is discussed in the next section.

Experiments and Results

Unsupervised Clustering of audio clips in meaning space

The Bayesian Information Criterion (BIC) (Schwarz 1978) has been used as a criteria for model selection in unsupervised learning. It is widely used for choosing the appropriate number of clusters in unsupervised clustering (Zhou & Hansen 2000; Chen & Gopalakrishnan). It works by penalizing a selected model in terms of the complexity of the model fit to the observed data. For a model fit M for an observation set $\mathcal{X}$, it is defined as (Schwarz 1978; Zhou & Hansen 2000):

$$BIC(M) = log(P(\mathcal{X}|M)) - \frac{1}{2} \cdot r_M \cdot log(R_{\mathcal{X}}), \qquad (6)$$

where $R_{\mathcal{X}}$ is the number of observations in the set $\mathcal{X}$ and r_M is the number of independent parameters in the model M. For a set of competing models $\{M_1, M_2, \ldots, M_i\}$ we choose the model that maximizes the BIC. For the case where each cluster in M_k (with k clusters) is modelled as a multivariate Gaussian distribution we get the following expression for the BIC:

$$BIC(M_k) = \qquad (7)$$
$$\sum_{j=1}^{j=k} \left(-\frac{1}{2} \cdot n_j \cdot log(|\Sigma_j|) \right) - \frac{1}{2} \cdot r_M \cdot log(R_{\mathcal{X}})$$

Here, Σ_j is the sample covariance matrix for the j^{th} cluster, k is the number of clusters in the model and n_j is the number of samples in each cluster. We use this criterion to choose k for the k-means algorithm for clustering the audio clips in the meaning space. Figure 4 is a plot of the BIC as a function of number of clusters k estimated using equation (7). It can be seen that the maximum value is obtained for $k = 112$.

Clustering Results

Some of the resulting clusters using the presented method are shown in Table 2. The table lists some of the significant audio clips in each of the clusters. Only five out of $k = 112$ clusters are shown for illustration. As mentioned previously, audio clips with similar onomatopoeic descriptions are clustered together. As a result, the clips in the clusters share similar perceived acoustic properties. For example, the clips SML_NAILS_DROP_ON_BENCH_B2.wav and DOORBELL_DING_DING_DONG_MULTI_BB.wav in cluster 5 listed in the table. From their respective onomatopoeic descriptions and an understanding of the properties of the sound generated by a doorbell and a nail dropping on a bench, a relationship can be made between them. The relationship is established by the vector representation of the audio clips in meaning space according to their onomatopoeic descriptions.

Conclusion

In this paper we represent descriptions of audio clips with onomatopoeia words and cluster them according to their vector representation in the linguistic (lexical) meaning space. Onomatopoeia words are imitative of sounds and provide a means to represent perceived audio characteristics with language level units. This form of representation essentially bridges the gap between signal level acoustic properties and higher-level audio class labels.

First, using the proposed distance/similarity metric we establish a vector representation of the words in a 'meaning space'. We then provide onomatopoeic descriptions (onomatopoeia words that best describe the sound in an audio clip) by manually tagging them with relevant words. Then, the audio clips are represented in the meaning space as the sum of the vectors of its corresponding onomatopoeia words. Using unsupervised k-means clustering algorithm, and the Bayesian Information Criterion (BIC), we cluster

Cluster #	Clip Name & Onomatopoeic Descriptions
Cluster 1	CAR_FERRY_ENGINE_ROOM_BB {*buzz,fizz,hiss*} WASHING_MACHINE_DRAIN_BB {*buzz,hiss,woosh*} PROP_AIRLINER_LAND_TAXI_BB {*buzz,hiss,whir* }
Cluster 2	GOLF_CHIP_SHOT_01_BB.wav {*thump, thwack*} 81MM_MED_MORTAR_FIRING_5_BB.wav {*bang, thud, thump*} THUNDERFLASH_BANG_BB.wav {*bang, thud, wham*} TRAIN_ELEC_DOOR_SLAM_01_B2.wav { *thud, thump, whomp* }
Cluster 3	PARTICLE_BEAM_DEVICE_01_BB.wav {*buzz, hum*} BUILDING_SITE_AERATOR.wav {*burr, hum, murmur, whir*} PULSATING_HARMONIC_BASS_BB.wav {*burr, hum, murmur*} . . .
Cluster 4	HUNT_KENNELS_FEED_BB.wav {*bark, blat, yip, yowl*} PIGS_FARROWING_PENS_1_BB.wav {*blare, boo, screech, squeak, yip*} SMALL_DOG_THREATENING_BB.wav {*bark, blare*} . . .
Cluster 5	DOORBELL_DING_DING_DONG_MULTI_BB.wav {*ding, dong, ring*} SIGNAL_EQUIPMENT_WARN_B2.wav {*ding, ring, ting*} SML_NAILS_DROP_ON_BENCH_B2.wav {*chink, clank*} . . .

Table 2: Results of unsupervised clustering of audio clips using the proposed vector representation method.

the clips into meaningful groups. The clustering results presented in this work indicate that the clips within each cluster are well represented by their onomatopoeic descriptions. These descriptions effectively capture the relationship between the audio clips based on their acoustic properties.

Discussion and Future Work

Onomatopoeia words are useful in representing signal properties of acoustic events. They are a useful provision in language to describe and convey acoustic events. They are especially useful to convey the underlying audio in media that cannot represent audio. For example, comic books frequently use words such as *bang* to represent the acoustic properties of an explosion in the illustrations. As mentioned previously, this is a result of common understanding of the words that convey specific audio properties of the acoustic events. This is a desirable trait in language level units making them suitable for automatic annotation and processing of audio. This form of representation is useful in developing machine listening systems that can exploit both semantic information and similarities in acoustic properties for aural detection and decision making tasks. As a part of our future work, we wish to explore the clustering and vector representation of audio clips directly based on their lexical labels and then relate it to the underlying properties of the acoustic sources using onomatopoeic descriptions and signal level features. For this, we would like to develop techniques based on pattern recognition algorithms that can automatically identify acoustic properties and build relationships amongst various audio events.

Acknowledgments

We would like to express our gratitude to the volunteers who took time to listen to each audio clip and tag them. We would especially like to thank Abe Kazemzadeh, Matt Black, Joe Tepperman, and Murtaza Bulut for their time and help. We gratefully acknowledge support from NSF, DARPA, the U.S. Army and ONR.

References

Chen, S. S., and Gopalakrishnan, P. S. "Clustering via the Bayesian Information Criterion with applications in Speech Recognition". *In Proc. of the International Conference on Acoustic Speech and Signal Processing (ICASSP)* Vol.2:12–15.

Fellbaum, C. D. 1998. "WordNet: An electronic lexical database" edited by. *The MIT Press* ISBN:026206197X.

Guo, G., and Li, S. Z. 2003. "Content-Based Audio Classification and Retrieval by Support Vector Machines". *IEEE Trans. on Neural Networks* 14(1).

http://www.sound ideas.com. 2006. "The BBC Sound Effects Library- Original Series.".

L. Liu, H. Z., and Jiang, H. 2002. "Content Analysis for Audio Classification and Segmentation". *IEEE Trans. on Speech and Audio Processing* 10(7).

P. Cano, M. Koppenberger, S. L. G. J. R.; Herrera, P.; and Wack, N. 2004. "Nearest-Neighbor generic Sound Classification with a WordNet-based Taxonomy". *In Proc.* 116^{th} *Audio Engineering Society (AES) Convention, Berlin, Germany.*

R. O. Duda, P. E. H., and Stork, D. 2000. "Pattern Classification". *Wiley-Interscience* 2nd edition.

Schwarz, G. 1978. "Estimating the Dimension of a Model". *The Annals of Statistics* Vol.6(2):461–464.

T. Zhang, J. K. 2001. Audio Content Analysis for Online Audiovisual data Segmentation and Classification. *IEEE Trans. on Speech and Audio Processing* 9(4).

Zhou, B., and Hansen, J. H. L. 2000. "Unsupervised Audio Stream Segmentation and Clustering Via the Bayesian Information Criterion". *In Proc. of the International Conference on Speech and Language Processing (ICSLP), Beijing, China.*