ANALYSIS OF AUDIO CLUSTERING USING WORD DESCRIPTIONS

Shiva Sundaram and Shrikanth Narayanan

Speech Analysis and Interpretation Laboratory (SAIL)
Department of Electrical Engineering-Systems
University of Southern California (USC),
3740 McClintock Ave., EEB 400, Los Angeles, CA 90089. USA.
email:ssundara@usc.edu, shri@sipi.usc.edu

ABSTRACT

We present an analysis of clustering audio clips using word descriptions that are imitative of sounds. These onomatopoeia words describe the acoustic properties of sources, and they can be useful in annotating a medium that cannot embed audio (e.g. text). First, an audio-to-word relationship is established by manually tagging a variety of audio clips (from a sound effects library) with onomatopoeia words. Using a newly proposed distance metric for word-level similarities, the feature vectors from the audio are clustered according to their tags, resulting in clusters with similarities in their onomatopoeic descriptions. By discriminant analysis of the clusters at the feature level, we present results on separability of these clusters. Our results indicate that by just using onomatopoeic descriptions, meaningful clusters with similar acoustic properties can be formed. However, in terms of audio feature level representation, clusters formed by some word groups such as *buzz, fizz* etc are better represented by signal features than percussive sounds such as *clang, clank, tap*.

Index Terms— audio ontology, audio information retrieval, analysis of audio clusters, onomatopoeia based audio descriptions

1. INTRODUCTION

Text, audio and video/images are different modalities of digital media. They represent various communication forms and expressions. To automatically process them, it is necessary to organize, and index them according to content. For this, it is desirable to compute both using lexical labels or words (for user queries) and signal level measures for (automatic) retrieval. This paper focuses on description of audio using natural language. Specifically, the work deals with ontological representation and characterization of audio and its language-level onomatopoeic descriptions.

Audio data is processed and stored in the signal feature space through a variety of time-frequency measures. On the other hand, the content and event in an audio clip, based on perception and context, are typically represented by natural language descriptions in the lexical semantic space using words. In content-based retrieval [1, 2, 3], the relationship between word-level linguistic descriptions and acoustic features is typically established by a naive, manual labeling scheme where the audio data is mapped onto a set of pre-specified classes. The resulting clusters (each belonging to a class) in the feature space are used to train a pattern classifier and eventually used to identify the correct class mapping for a given test data. Although this approach yields good performance, especially if the number of classes is small, it inherently allows for mismatch and ambiguity between the semantic information present in the linguistic labels used to describe the audio event and the signal level features that physically characterize it. For instance, consider the phrase `Nail Hammered` as a label (an example from the BBC sound effects Library [15]). It represents that the audio clip is the sound of the nail being hammered, but does not describe the acoustic properties of the event. However, the underlying automatic processing is based on similarities in the acoustic properties. This inherent ambiguity needs to be reconciled by the automatic audio classification system.

There has been some excellent work in the community on bridging this gap between the language-level semantic space and the acoustic properties. In [5] the author improves on the naive labeling scheme by creating a mapping from each node of a hierarchical model in the abstract semantic space to the acoustic space. The nodes in the hierarchical model (represented probabilistically as words) are mapped onto their corresponding acoustic models. Other techniques for retrieval using semantic relations in language include [6]. Here the authors have used WordNet [13] to generate word tags for a given audio clip using acoustic feature similarities, and also retrieve clips that are similar to the tags. While such systems which incorporate semantic language-level relations exist, they are still sufficiently insulated from signal level properties that directly determine the perception of sources.

In this paper, we present an approach to use linguistic descriptions that are closer to signal level properties. We present analysis of representing signal level measures extracted from audio clips with onomatopoeic word descriptions. These are words that are imitative of sounds[12]. The rationale being these descriptions would provide a more intuitive (based on perception) but less ambiguous lexical descriptions to aid automatic classification. For example, the audio clip of `Nail hammered` can be better described by `tap-tap` which provides more direct information about the acoustic properties of the event. Our experiments in this paper have two objectives:

1. To develop a distance metric to analyze the relationship amongst onomatopoeia words and thus cluster them. The ability to cluster these words in a quantitative space makes them useful as a meta-level representation for

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 here.

computing using words for audio retrieval.

2. To measure the effectiveness of common acoustic signal features to represent the resulting clusters. For this, we cluster the features extracted from a selection of audio clips by two methods: (a) using information from onomatopoeia word clusters as mentioned above (b) by unsupervised clustering on the whole collection of extracted feature vectors.

The two methods of clustering are compared using a Gaussian maximum a posteriori (GMAP) classifier after multiple discriminant analysis (MDA) [11]. The clustering in (a) is according to a meta-level understanding of the onomatopoeic words. These words, subsequently, are descriptive of the underlying acoustic properties of the clips. Data for the experiments were collected by having volunteers listen to an assorted set of clips from a sound effects library [15] and have them tag each clip with relevant onomatopoeia words. Next the implementation and experiments are described.

2. IMPLEMENTATION

2.1. Feature Extraction:

A total of 3.34 hours of audio from 1014 clips was used for the experiments. The files were first converted from 2 to 1-channel tracks (16 bits, 44.1kHz uncompressed PCM). A total of 29 features (commonly used in audio classification [10]) were extracted every 10 ms for each 20 ms frame of audio: *Short-term Average Energy (E) Spectral Centroid (SC)*, *Bandwidth (BW)*, *Mel-frequency Cepstral Coefficients (MFCC) (26 order)*.

2.2. Clustering words using word-level relations

The relationship between onomatopoeia words is developed and they are subsequently clustered into groups using a semantic word based similarity metric [4]. The details of this method are discussed below.

A set $\{L_i\}$ consisting of l_i words is generated by a thesaurus [14] for each onomatopoeia word O_i. Then the similarity between the j^{th} and k^{th} word can be defined to be:

$$\text{(similarity) } s(j,k) = \frac{c_{j,k}}{l^d_{j,k}}, \quad \text{(distance) } d(j,k) = 1 - s(j,k) \quad (1)$$

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

$$0 \le d(j,k) \le 1, \quad d(j,k) = d(k,j), \quad d(k,k) = 0 \qquad (2)$$

Except for the triangular inequality, it is a valid distance metric. Similar to the way humans use other words to establish

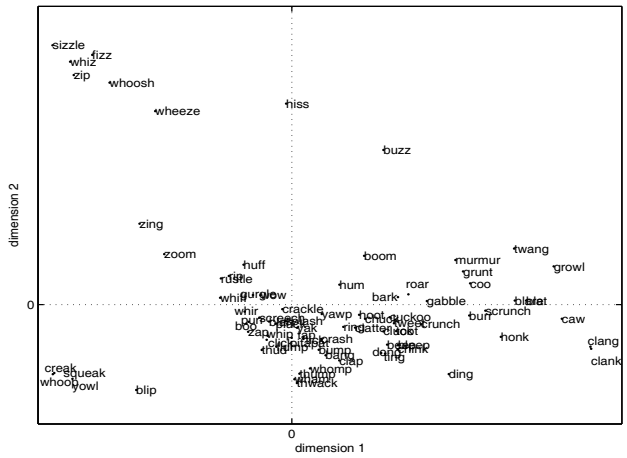

Fig. 1. Onomatopoeia words in 2-D 'meaning space' . Note: *clang*, *clank* are close to each other, but are far from *fizz*, *sizzle*

the meaning of an unfamiliar word, the sets $\{L_j\}$ and $\{L_k\}$ generated by the thesaurus have some *meaning* associated with O_j and O_k in the language. The similarity is a measure of *sameness* in meaning (number of common words). For W words, we get a symmetric $\mathcal{R}^{W\times W}$ matrix where the $(j,k)^{th}$ element is the distance between the j^{th} and k^{th} word. Also, the j^{th} row is a vector representation of the j^{th} word in terms of other words in the set. By principal component analysis (PCA) on this set of vectors, each word is represented in a reduced dimensional space $\mathcal{R}^d$ with $d < W$. The squared sum of the first eight ordered eigenvalues covered more than 95% of the total squared sum, resulting in $d{=}8$ and $W{=}87$. These points are then clustered using the k-means algorithm in this lexical, onomatopoeic 'meaning space'.

From table 1 (the list of words used here) 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 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 meaning space. Note that the words *growl* and *twang* appear nearby, due reduced 2 dimensions of the illustration.

By modelling the onomatopoeia word representation in this space as observations of a multivariate Gaussian process, the words are clustered using the Bayesian information criterion (BIC). The BIC [7] is widely used for choosing the appropriate number of clusters in unsupervised clustering [8, 9]. If each cluster in a model M_k (with k clusters) is assumed to follow a multivariate Gaussian distribution we get a closed-form expression for the BIC as given in [9]. For a set of competing models $\{M_1, M_2, \ldots, M_i\}$ we choose the model that maximizes the BIC. We use this criterion to choose k for the k-means algorithm for clustering the words. Since the number of points in the onomatopoeic meaning space is small ($W{=}87$), a bootstrapping approach is used to estimate the BIC for each k. The BIC estimates are averaged over 500 runs. The resulting variation of the BIC as a function of k is shown in Figure 2. The maximum value of $\widehat{BIC}$ was obtained for $k = 19$. Table 2 illustrates a few of the resulting clusters.

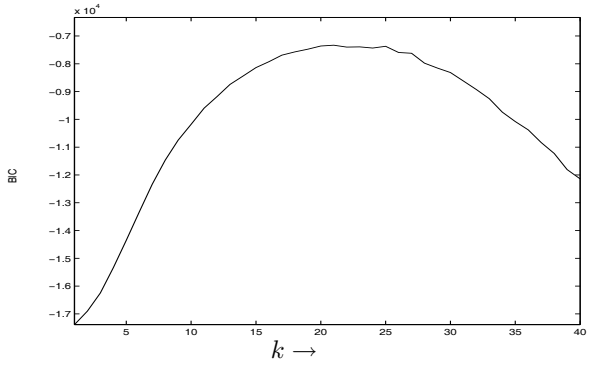

Fig. 2. Variation of $\widehat{BIC}(M_k)$ as a function of k for 500 runs

cluster 1($\mathcal{C}_1$)**:** Clang,Clank,Ding,Dong,Ting	**cluster 2** ($\mathcal{C}_2$) Beep,Bleep,Toot
cluster 3($\mathcal{C}_3$)**:** Creak,Squeak, Screech,Yawp,Yowl	**cluster 4**($\mathcal{C}_4$)**:** Cluck,Cuckoo, Hoot,Tweet
cluster 5($\mathcal{C}_5$)**:** Buzz,Fizz,Sizzle,Hiss Whiz,Wheeze,Whoosh,Zip	**cluster 6**($\mathcal{C}_6$)**:** Thump,Thwack,Wham
cluster 7($\mathcal{C}_7$)**:** Burr,Crunch,Scrunch	**cluster 8**($\mathcal{C}_8$)**:** Rip,Zing,Zoom
cluster 9($\mathcal{C}_9$)**:** Clatter,Blah,Gabble,Yak	**cluster 10**($\mathcal{C}_{10}$)**:** Meow,Moo,Yip

Table 2. Automatically derived word clusters in lexical onomatopoeic 'meaning space'.

Since word clusters can be formed it can be inferred that: (1) words within clusters have overlapping meaning, (2) words in different clusters are sufficiently distinct, and (3) the proposed metric sufficiently discerns the words by their meaning. Both general and specific perceived audio properties can be described using onomatopoeia words. Next, the procedure for tagging the clips with onomatopoeia words is described, following which the clustering of the extracted feature vectors using word-level clustering information is discussed.

2.3. Tagging the Audio Clips with onomatopoeia words

A set of 236 audio clips (belonging to categories such as: *animals, birds, footsteps, transportation, construction work, fireworks etc.*) were selected from the BBC sound effects Library [15]. Four subjects, with English as their first language, tagged this initial set of clips with onomatopoeia words. A Graphical User Interface (GUI) based software tool was designed to play each clip over a pair of headphones and have the subject click on the relevant onomatopoeia words that best described the clip. All the clips were edited to be about 10-14 seconds in duration. The clips were randomly divided into 4 sets, so that the volunteers spent only 20-25 minutes at a time, per set. Only the tags that were common to two or more volunteers were retained. In general, at least one or two tags assigned by volunteers were in agreement for a clip. Note that the resulting tags are the onomatopoeic descriptions that best represent the perceived acoustic properties. The tags from this set were then copied to other clips with similar names. For example,the clip with the name 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, 1014 clips totaling 3.34 hours of labeled/tagged data were available.

2.4. Clustering the feature vectors

As mentioned in point 2(a) of section 1, the feature vectors are first clustered using the information from clusters of onomatopoeia words of their clips. The following voting procedure was used:

1. For each clip, the number of onomatopoeia tags common with the words in each cluster $\mathcal{C}_j, j \in \{1, 2, \ldots, k\}$ was counted.
2. The features extracted from the audio clip are assigned to the cluster $\mathcal{C}_j$ with most number of common words. *Collision* (where the tags may have the same number of common words with more than one word cluster) was randomly resolved.

With this procedure, the resulting clusters of audio in the feature space are similar in terms of their onomatopoeic descriptions. In step 2, it is also possible to assign the acoustic features from a clip to more than one cluster $\mathcal{C}_j$ by having at least one word common with the clusters. But this would result in a more complex grouping of the features. Some of the clusters as a result of this procedure are listed below:

Cluster :
PERSIAN_CATS_EAT_PURR.wav,
GARGLE_BB.wav, SLURP_BB.wav
TRACTORS_WORK_IN_YARD_BB.wav
TEAPOT_BEING_FILLED_BB.wav

Cluster:
MANY_HORSES_TROTTING_BB.wav,
LRG_TACK_NAIL_HAMMERED_B2.wav,
CLIZA_OUTDOOR_MARKET_B2.wav,
BRUSH_PAINTING_B2.wav,
GAS_BLOWLAMP_LIT_FLAME_B2.wav

Cluster:
AMSTERDAM_TRAM_3_BB.wav,LONDON_SUBWAY_ARRIVES_01_B2.wav,
BUILDING_SITE_HAMMERING_B2.wav,CAN_OPENER_BB.wav,
TOOLBOX_CLOSED_B2.wav

As an example, note that AMSTERDAM_TRAM_3_BB , TOOLBOX_CLOSED_B2 are clustered together. It can be interpreted that the properties of the sound generated by the tram and the box can both be described with the words $\{$*clang, clank*$\}$. Thus the clusters resulting from this procedure have similarity in terms of their onomatopoeic descriptions.

Using word-level clustering information, features extracted from the clips were clustered into $k = 19$ clusters. Since the onomatopoeia words describe the acoustic properties, the underlying acoustic data can also be expected to have 19 clusters. As mentioned in point 2(b) of section 1 a "raw grouping" is done by clustering all the extracted features using k-means algorithm in the acoustic feature space. This was done without using information from the word-level clusters. For this, the algorithm was also initialized to have 19 clusters.

3. CLASSIFICATION EXPERIMENTS

First, by MDA the dimensionality of the problem was reduced to $(k - 1)$. Then, the data was split into train and test sets (90% and 10% respectively). Parameters of the GMAP classifier were determined using the train set. This was done for both methods of clustering. The final result for each clustering method is presented.

Classification accuracy using word-level clustering : The classification results we obtained were better for some clusters and worse for others with an overall accuracy of 54.44%. The recall and precision for those clusters are listed in Table 3. The resulting 2 nearest clusters (in terms of most confusing clusters for a given cluster) is given in Table 4.

Accuracy of "raw grouping": We obtained a classification accuracy of 85.28% for frame level tests for the raw clustering without using information from the word clusters. Indicating 19 distinct clusters indeed exist amongst the feature vectors extracted from the audio data in the acoustic space. That is, higher classification accuracy $\Rightarrow$ resulting clusters are well

	% P	% R	Cluster words
B	64.8	81.1	{*buzz,fizz,hiss,sizzle,wheeze,whiz,whoosh,zip*}
E	72.5	73.4	{*huff,hum,whiff,wow*}
S	47.8	68.3	{ *click,chink,tick*}
T	65.6	54.8	{*creak,squeak,screech,yawp,yowl*}.
W	25.6	23.0	{*cluck,cuckoo,honk,hoot,tweet*}
O	28.1	21.2	{*meow,moo,whoop,yip*}
R	39.1	18.1	{ *crackle,pluck,splash,tap*}
ST	24.1	10.7	{*clang,clank,ding,dong,ting*}

Table 3. 4 best and worst precision (%**P**) and recall (%**R**) rates for classification. Clusters are formed by using word-level grouping information

{*blare,blat,grunt,murmur*} & ,{*burr,crunch,,caw,scrunch*} FOR {*buzz,fizz,hiss,sizzle,wheeze,whiz,whoosh,zip*}
{*buzz,fizz,hiss,sizzle,wheeze,whiz,whoosh,zip*} & {*beep,bleep,toot*} FOR {*huff,hum,whiff,wow*}
{*buzz,fizz,hiss,sizzle,wheeze,whiz,whoosh,zip*} & {*creak,squeak,screech,yawp,yowl*} FOR {*click,chink,tick*}
{*burr,crunch,caw,scrunch*} & {*blah,clatter,gabble,yak*} FOR {*creak,squeak,screech,yawp,yowl*}

Table 4. The two most confusing clusters for each of the 4 clusters that have the best precision and recall rates.

separated $\Rightarrow$ acoustic features are sufficiently discriminatory.

4. CONCLUSION AND FUTURE WORK

In this work, clustering of acoustic feature vectors with and without using word-level information was analyzed. Word-level clustering was done using onomatopoeia words as means to represent perceived audio signal characteristics. These words can be used as a meta-level representation between acoustic features and language-level descriptions of audio.

Using the proposed distance metric in a 'meaning space', along with k-means algorithm, and the BIC, the words are clustered into like groups. This grouping information is used to cluster features extracted from the corresponding audio clips that were manually tagged offline with onomatopoeia words using a voting procedure. This clustering was compared with "raw" clustering: grouping feature vectors without using information from word-level grouping. The comparison was performed in terms of classification accuracy of a GMAP classifier after MDA. The results in this work indicate that certain word clusters are more separable than the others. It can be, in part, due to the fact that the acoustic features used in this work are only able to represent certain onomatopoeic descriptions. It can also be because of inconsistencies in the understanding and usage of onomatopoeia words as well. Also, some words (such as *crackle*) represent long-term temporal properties that are not well represented by frame level analysis.

Another interpretation of the results is that, the raw clustering results in partitioning of the feature vectors into regions of contiguous volume of space. However, clustering using onomatopoeia grouping information may result in fragmented partitioning, where the feature vectors of a cluster may be present in different regions of the feature space. This essentially brings out the differences in signal level measures and linguistic level descriptions. This also calls for signal measures that are representative of linguistic level descriptions.

Onomatopoeia words can be used to annotate a medium that cannot represent audio (e.g. text). Given better signal measures, this representation can be useful for computing with both in terms of language level units and signal level measures. The ability to cluster words in a quantitative meaning space implies words within the clusters have overlapping meaning and words in different clusters are sufficiently distinct. This makes them useful as they can express and represent both specific and general audio characteristics. This is a desirable trait as a meta-level representation making them suitable for automatic annotation and processing of audio.

The preliminary results obtained here mainly indicate that some of the onomatopoeia words are represented better in the acoustic feature space than the others. We would like to investigate this further to obtain a better acoustic feature set to represent all aspects of the description. Other avenues include developing a predictive model to generate appropriate onomatopoeic representations (or, less ambiguous-more intuitive language descriptions) given an audio class or audio sample. To expand the scope of this work we would like to develop a decision tree approach to explore the noun $\longrightarrow$ action(verb) $\longrightarrow$ onomatopoeic-description relationship that would be useful in a higher level representation and grouping of acoustic events.

5. REFERENCES

[1] G. Guo and S. Z. Li,"Content-Based Audio Classification and Retrieval by Support Vector Machines,"in IEEE Trans. on Neural Nets., Vol.14, No.1, January 2003.

[2] L. Liu,H.J. Zhang and H. Jiang,"Content Analysis for Audio Classification and Segmentation," IEEE Trans. on Speech and Audio Processing, Vol.10, No.7, October, 2002.

[3] T. Zhang, C. C. Kuo,"Audio Content Analysis for Online Audiovisual Data Segmentation and classification," in IEEE Trans. on Speech and Audio Processing Vol.9, No.4, May, 2001.

[4] S. Sundaram, S. Narayanan, "Vector-based representation and clustering of audio using onomatopoeia words.", In Proc. of AAAI 2006 Fall Symposia, Arlington, VA, October 2006.

[5] M. Slaney,"Semantic-Audio Retrieval," Presented at the ICASSP, Orlando, Florida, USA. May 13-17,2002.

[6] P. Cano,M. Koppenberger,S. Le Groux,J. Ricard,and P. Herrera,and N. Wack "Nearest-neighbor generic sound classification with a WordNet-based taxonomy",In Proc. 116^{th} Audio Engineering Society (AES) Convention, Berlin, Germany, 2004.

[7] G. Schwarz,"Estimating the Dimension of a Model", The Annals of Statistics, Vol.6, No.2 , pp 461-464, March 1978.

[8] B. Zhou, J. H. L. Hansen,"Unsupervised Audio Stream Segmentation and Clustering Via the Bayesian Information Criterion" In Proc. of the ICSLP2000, Beijing, China Oct 6-20, 2000.

[9] S. .S Chen and P. S. Gopalakrishnan ,"Clustering Via the Bayesian Information Criterion with applications in Speech Recognition". In Proc. of the ICASSP Vol.2 ,12-15 May 1998.

[10] D. Li, I.k. Sethi, N. Dimitrova, T. McGee,"Classification of general audio data for content-based retrieval," in Pattern Recognition Letters, Vol.22, 533-544, 2001.

[11] R. O. Duda,P. E. Hart,D.G. Stork ,"Pattern Classification," Wiley-Interscience; 2nd edition, October, 2000

[12] "Oxford English Dictionary." http://www.oed.com

[13] WordNet : http://www.cogsci.princeton.edu/∼wn

[14] G. Ward, "Moby Thesaurus http: // www.dcs.shef.ac.uk/ research/ ilash/ Moby/".

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