QUERY-BY-EXAMPLE RETRIEVAL OF SOUND EVENTS USING AN INTEGRATED SIMILARITY MEASURE OF CONTENT AND LABEL

Annamaria Mesaros[1] *Toni Heittola*[2] *Kalle Palomäki*[1]

[1] Department of Information and Computer Science, Aalto University
[2] Department of Signal Processing, Tampere University of Technology

ABSTRACT

This paper presents a method for combining audio similarity and semantic similarity into a single similarity measure for query-by-example retrieval. The integrated similarity measure is used to retrieve sound events that are similar in content to the given query and have labels containing similar words. Through the semantic component, the method is able to handle variability in labels of sound events. Through the acoustic component, the method retrieves acoustically similar examples. On a test database of over 3000 sound event examples, the proposed method obtains a better retrieval performance than audio-based retrieval, and returns results closer acoustically to the query than a label-based retrieval.

1. INTRODUCTION

Retrieval engines are an essential tool for searching and browsing databases. A straightforward way of performing retrieval is based on textual descriptors, and the results are satisfactory as long as the user knows exactly what he or she wants, and the database where the search is performed is manually indexed and structured. This is prone to errors when the label and the content do not match. In recent years, retrieval methods have taken a big step by using content-based information in order to provide better quality results, and content-based search engines are nowadays available for different types of media.

Retrieval based on content is done using objective measures to express the similarity between the content of the query and items in the database. In music information retrieval, audio similarity can reflect rhythm, timbre, chords, etc. For everyday sounds, it is harder to define what type of similarity we are looking for - perhaps the origin of the sound. For example, with a query of bird singing we would easily accept results containing different kind of birds, even if their vocalizations differ substantially, but would be disappointed to retrieve a tune whistled by a human.

A method for connecting the labels and the audio content is presented in [1, 2, 3]. Semantic features for audio files are obtained from textual descriptions containing multiple words. For each word, a binary value indicates if it appears or not in the textual description of each audio file. The final semantic model is a collection of word-level distributions modeling the distribution of audio features associated with each semantic concept. This way, the search space for similar audio examples is the semantic space.

The use of semantic analysis is not limited to retrieval. Methods for classification and automatic annotation of generic sounds and music also benefit of using semantic knowledge [4, 5]. In these papers, the experiments consisted of finding a best match for all the sounds in the database, and evaluate how well their labels match, using the WordNet taxonomy [6]. The relationship between acoustic and semantic similarity in a database of generic sounds was studied also in [7].

Words chosen to label sounds usually describe the source of the sound, which can be an object, action, or both (car horn, knocking, chair squeaking), and each person can have their own way of describing it. With an infinite amount of possible sound sources and a large variety of sounds, variability in the labels is a natural outcome. Tackling this aspect will provide the possibility of developing a general indexing and retrieval system, able to deal with the variability of the labels.

This paper proposes a new approach for combining audio similarity with semantic similarity for sound events query-by-example retrieval. A single measure of similarity that takes into account audio and semantics is created based on the audio similarity of the sound events and the semantic similarity of their labels. The search space for the similar items is the joint audio-semantic similarity space. The semantic side provides information to guide the audio-based search towards examples with similar labels. On the other side, the audio content helps eliminating the non-related audio examples that have the same label.

2. SYSTEM DESCRIPTION

We evaluate audio and semantic similarity of individual sound examples, in order to create a joint similarity matrix taking into account both aspects. Query-by-example retrieval based on both the content and the provided label of the query is evaluated in a database containing a variety of sounds.

This work was financially supported by Academy of Finland under the grants 136209 (Palomäki) and 251170 (Mesaros) Finnish Centre of Excellence Program (2012-2017) and by TEKES FuNeSoMo project (Palomäki).

2.1. Similarity of audio content and labels

Audio similarity between two examples is usually evaluated with objective measures based on distance metrics between frame-based representations of the signals or between statistical models of the signals. To calculate similarity between sound events, we start by calculating mel frequency cepstral coefficients (MFCC), with 20 ms length window and 50% overlap. For each annotated sound event, a Gaussian mixture model (GMM) is estimated based on the MFCCs. Distance between two distributions is evaluated using the symmetrical Kullback-Leibler divergence [8]. The distance from each event to every other event, collected into a matrix, can be used directly for clustering or classification.

Semantic similarity is measured using tools from natural language processing and WordNet [6]. We use the *path similarity*, that is calculated as the inverse of the shortest path in the WordNet hierarchy between the two compared concepts [9]. For example, considering the most common meanings for nouns "cat" and "dog" presented in Figure 1, the path similarity between them is 0.2 (inverse of the path containing 5 nodes). A method for calculating semantic similarity for labels containing multiple words is presented in [7]. The values of the path similarity are between 0 (not the same part of speech) and 1 (synonyms) and they are quantized (1/1, 1/2, 1/3 and so on).

Sound events similarity as a combination of acoustic and semantic similarities can be obtained by a weighted mean of the two similarity measures. The combined similarity matrix will be an $N \times N$ matrix, N being the number of sound events in the database, having form $C = (1-w) \cdot A + w \cdot S$, where A is the acoustic similarity matrix, S is the semantic similarity matrix and w is the weight of the semantic similarity. The KL divergence values represent dis-similarity, therefore they need to be transformed to represent similarity. We choose to limit values to the 98 percentile value, normalize, and subtract them from 1, to result in similarity values between 0 and 1. This brings similarity matrices A and S into the same range, with 1 for most similar items, 0 for most dis-similar. Each side can be emphasized in C by selecting w accordingly.

2.2. Retrieval methods and evaluation metrics

A sound event instance $I \in [1 : N]$ is characterized by a feature vector $\mathbf{c}_I$ representing a row in C. Each element of this feature vector represents similarity of event I to another event in the database. The most similar items to the event I can be found by directly sorting the elements of the vector $\mathbf{c}_I$. This method of retrieval represents an exhaustive search, because it involves computing similarity of the event I to every other event in the database. This will be used as a baseline for comparison.

The feature vectors $\mathbf{c}$ corresponding to all events in the database provide a mapping of the similarity space, by representing relative positions of events with regard to each other.

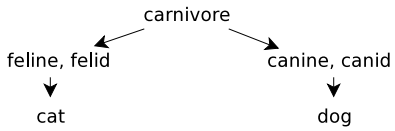

Fig. 1. The path between the most common meaning of "cat" and the most common meaning of "dog" in WordNet

Similarity between an event I and another event J is measured as euclidean distance $d(\mathbf{c}_I, \mathbf{c}_J)$ between feature vectors $\mathbf{c}_I$ and $\mathbf{c}_J$, representing rows in the similarity matrix. The most similar items to the event I are the ones with smaller distance from $\mathbf{c}_J$ to $\mathbf{c}_I$.

However, for a large database, the dimensionality of this similarity space is prohibitively large for such an approach. Dimensionality reduction can be obtained by selecting a number of k points in this space to serve as anchors, and calculating only the values with regard to these anchors [10]. Each sound event instance I will be characterized by a k-dimensional feature vector representing its similarity to the k anchors. This is equivalent with considering only a number k of columns, randomly, from the $N \times N$ similarity matrix. From this perspective, the exhaustive search has one anchor point, which is the query.

Retrieval performance is measured using the mean average precision (MAP). This provides a single measure of performance, avoiding the trade off between precision and recall, and is useful in illustrating the performance of the system in terms of ranked results. The mean average precision also generalizes for different recall levels (different number of relevant documents for each query).

The average precision for a query is calculated based on precision at each 'seen' relevant document:

$$AP = (\sum_{n=1}^{R} \frac{n}{rank_n})/R, \quad (1)$$

where R is the number of relevant documents for that query and $n/rank_n = 0$ if document n was not retrieved. The average precision is calculated for each query, to evaluate overall performance of the system as the mean average precision over all queries.

Another measure used for illustrating retrieval performance is precision at a given rank. For example precision at rank 20 (P@20) represents the percentage of relevant documents in the top 20 retrieved list. On the negative side, this measure does not average well, as it fails to account for different recall levels.

3. EXPERIMENTAL RESULTS

We present in detail three cases for the retrieval, by varying the weight of the semantic similarity: retrieval based on audio only ($w = 0$, $C = A$), based on label only ($w = 1$, $C = S$),

banging bicycle, bicycle, bicycle bell, bicycle clang, bicycle gear ticking, bicycle gears, bicycle on gravel, bicycle pedal, bicycle squeaks, bicycle stand, bicycle, ticking bicycles, bike gears ticking, braking bicycle, crate on bicycle, distant bicycle, parking bicycle, passing bicycle, slipping wheels, train wheels, unlocking bicycle, wheels

Table 1. Equivalent labels determining the number of relevant documents for label "bicycle": 22 different labels, 79 total event instances.

and combined with weight 0.9 for the semantic component ($w = 0.9$, $C = 0.1 \times A + 0.9 \times S$). When retrieval is done based on audio only, the mean average precision illustrates how many of the acoustically closest neighbors have a synonym label. When retrieval is done based on label only, the mean average precision is maximum, because of the evaluation itself being done based on labels. The acoustic similarity of the closest neighbors can however illustrate the difference between the two extreme cases. Combining the acoustic and semantic sides, we should obtain better retrieval score than the acoustic only, but without compromising in acoustic similarity.

3.1. Database description

The database used in this study is Dares database [11]. This collection contains 765 unique labels, with a variety of labels referring to the same concept (see Table 1), and a total of 3214 annotated event instances.

We select queries from labels with examples belonging to at least three different audio files. When one event instance is used as query, all other events belonging to the same audio file are excluded from the search database (to avoid audio similarity due to same recording conditions). We use only queries that have at least 20 relevant items, in order to be able to calculate P@20. The final number of selected queries is 1245, belonging to 34 classes. All results will be presented as average of these 1245 test cases.

To evaluate MAP, we consider as relevant all the event instances that have at least one synonym in common with the query in their label. This takes into account all the different labels referring to a common concept. Table 1 presents the collection of labels in the database that are considered equivalent when the query is "bicycle". This example also illustrates the errors due to polysemy ("bicycle"="wheels") for a label based retrieval.

3.2. Full list performance

We evaluate the retrieval power of the system by allowing the system to return all items in the database. This is an artificial case that serves as a baseline. This baseline illustrates the maximum performance, because by ranking the entire collection based on similarity, the system will return, sooner or later, all relevant documents. The MAP will however penalize the system which ranks relevant items at lower positions, and this allows a comparison between the exhaustive search (ranking items directly based on values in C) and the system based on anchors.

semantic weight	exhaustive search	10 anchors
w=0	0.08±0.001	0.09±0.00
w=0.9	0.05±0.001	0.44±0.07
w=1	0.05±0.001	0.57±0.09

Table 2. MAP when ranking all items in the database, i.e the system retrieves all relevant items.

Retrieval performance for exhaustive search and using ten anchors is presented in Table 2.We chose $k = 10$, as according to previous studies [10, 7], higher dimensionality does not bring considerable increase in performance. The anchors are selected randomly from the set of events in the database, to avoid the problem of carefully selecting them in different parts of the feature space [10].

The system using anchors shows a significantly higher performance, especially when the semantic side is considered. This means that the relative positions with regard to ten points provide a more accurate representation of the similarity space than the position with respect to a single point. Retrieval using anchors and only semantic similarity ($w = 1$) returns the relevant items on the highest ranks. This is practically the best possible performance for the system. However, this is purely an artificial result, since in a retrieval system one does not expect all the database to be ordered by rank.

3.3. Top 20 list performance

In a real user scenario, a retrieval system will return ranked results, out of which the user will be looking at the first few. We evaluate performance of the system using ten anchors for top 20 list, calculating the mean average precision for the top 20 returned items, and average acoustic similarity of relevant retrieved items. The maximum possible MAP of the system in this case is 0.24, when all items in top 20 are relevant. For example, we can calculate the maximum MAP for a "bicycle" query according to the numbers in Table 1: there are 79 events labeled with synonyms of "bicycle", therefore 78 relevant items. If all 20 items in top 20 are relevant, MAP is 0.13.

The highest retrieval score in Table 3 is for the situation when the retrieval is done only based on the semantic side ($w = 1$). This case brings into top 20 the highest number of relevant items, because of searching based on the label. The MAP in this case is 0.16, compared to 0.24 maximum possible value, and there are 85% relevant items in top 20.

The combined similarity provides lower retrieval performance, but the relevant items in top 20 are closer to the query from acoustic point of view. The advantage of using a combi-

semantic weight	MAP	avg. ac. similarity	P@20
w=0	0.005±0.001	0.48±0.12	0.10±0.01
w=0.9	0.09±0.01	0.47±0.11	0.58±0.08
w=1	0.16±0.02	0.41±0.08	0.85±0.08

Table 3. MAP of top 20 results, average acoustic similarity of relevant items in top 20 and P@20.

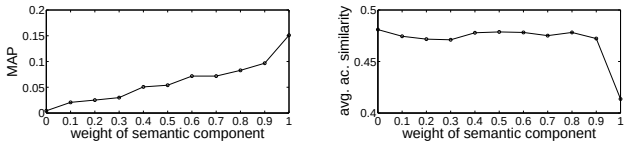

Fig. 2. MAP and average acoustic similarity of top 20 list for varying weights of the semantic component

nation of audio and semantics is the possibility of varying the parameter w for a desired result. The MAP and audio similarity values for values of w between 0 and 1 are presented in Figure 2. For the data analyzed in this work, it seems that a good value for w is close to 1, in order to achieve high MAP, but not 1, in order to obtain good acoustic match. The acoustic similarity values introduce the variation needed to distinguish between items having the same label.

We have to stress on the difficulty of evaluating the results with the available ground truth, as the number of relevant items is considered to be simply the reunion of all labels that have at least one common term. This results in overestimating the number of relevant items (maximum recall) because of words polysemy. In the example given in Table 1 , 'train wheels' should not be considered a relevant item for "bicycle", but the situation results from 'wheels' having the same meaning as "bicycle". (According to WordNet, 'wheels' = 'a wheeled vehicle that has two wheels and is moved by foot pedals'.) Such cases are also boosting the result of the presented retrieval based on semantics only, because some non-relevant items were considered correctly retrieved ('train wheels' was considered correctly retrieved when the query was "bicycle").

4. CONCLUSIONS

This paper presented a new similarity measure calculated as a weighted average of audio and semantic similarity. Combining the two sides results in a unified search space for retrieval of audio examples that are similar in content and have labels referring to the same concept. The method is able to handle variability in labels of sound events without compromising on the acoustic similarity. This approach is useful for user-contributed data where labels vary a lot. The obtained results show that the proposed method obtains better retrieval performance than an audio-based retrieval, and better acoustic similarity (objectively measured) than label-based retrieval. Future work will include a listening test for evaluating the subjective audio similarity of the retrieval results, to validate these results with human subjects.

5. REFERENCES

[1] M. Slaney, "Mixtures of probability experts for audio retrieval and indexing," in *Proc. of IEEE Int. Conf. on Multimedia and Expo ICME '02*, 2002, pp. 345–348.

[2] L. Barrington, A. Chan, D. Turnbull, and G. Lanckriet, "Audio information retrieval using semantic similarity," in *IEEE Int. Conf.on Acoustics, Speech and Signal Processing ICASSP 2007*, 2007, pp. 725–728.

[3] M. Levy and M. Sandler, "Music information retrieval using social tags and audio," *IEEE Trans. Multim.*, vol. 11, no. 3, pp. 383 –395, april 2009.

[4] P. Cano, M. Koppenberger, P. Herrera, S. Le Groux, J. Ricard, and N. Wack, "Nearest-neighbor generic sound classification with a wordnet-based taxonomy," in *Audio Engineering Society Convention 116*, 5 2004.

[5] P. Cano, M. Koppenberger, S. Le Groux, J. Ricard, N. Wack, and P. Herrera, "Nearest-neighbor automatic sound annotation with a wordnet taxonomy," *J. Intell. Inf. Syst.*, vol. 24, no. 2, pp. 99–111, May 2005.

[6] Princeton University, "About WordNet," `http://wordnet.princeton.edu`, 2010.

[7] A. Mesaros, T. Heittola, and K. Palomäki, "Analysis of acoustic-semantic relationship for diversely annotated real-world audio data," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing ICASSP 2013*, 2013.

[8] T. Virtanen and M. Helen, "Probabilistic model based similarity measures for audio query-by-example," in *Proc. of WASPAA*, 2007.

[9] T. Pedersen, S. Patwardhan, and J. Michelizzi, "WordNet::Similarity: measuring the relatedness of concepts," in *Proc. of Fifth Annual Meeting of the North American Chapter of the Assoc. for Computational Linguistics*, 2004, pp. 38–41.

[10] M. Helen, *Similarity measures for content-based audio retrieval*, Ph.D. thesis, Tampere University of Technology, 2009.

[11] M. Grootel, T. Andringa, and J. Krijnders, "DARES-G1: Database of annotated real-world everyday sounds," in *Proc. of the NAG/DAGA Meeting*, 2009.