ANALYSIS OF ACOUSTIC-SEMANTIC RELATIONSHIP FOR DIVERSELY ANNOTATED REAL-WORLD AUDIO DATA

Annamaria Mesaros[⋆] *Toni Heittola*[†] *Kalle Palomäki*[⋆]

[⋆] Department of Information and Computer Science, Aalto University
[†] Department of Signal Processing, Tampere University of Technology

ABSTRACT

A common problem of freely annotated or user contributed audio databases is the high variability of the labels, related to homonyms, synonyms, plurals, etc. Automatically re-labeling audio data based on audio similarity could offer a solution to this problem. This paper studies the relationship between audio and labels in a sound event database, by evaluating semantic similarity of labels of acoustically similar sound event instances. The assumption behind the study is that acoustically similar events are annotated with semantically similar labels. Indeed, for 43% of the tested data, there was at least one in ten acoustically nearest neighbors having a synonym as label, while the closest related term is on average one level higher or lower in the semantic hierarchy.

Index Terms— sound events, audio similarity, semantic similarity

1. INTRODUCTION

Databases are a significant aspect in any research problem: data used in training and testing of algorithms must be comprehensive, general enough to ensure that the developed algorithm generalizes well, and of satisfactory detail to the problem at hand. When collecting data for a specific application, the resulting annotation serves the purpose of the planned research but comes with a cost of large collection efforts. On the other hand, user contributed data is publicly available but annotations usually contain homonyms (words with multiple meanings), synonyms (different words with the same meaning), plurals, etc. This brings up the question how to annotate a database consistently both in detailed and meta data level so that it is useful for a variety of tasks and for many research teams, rather than oriented to a certain narrow application [1]. Evidently there is a need for methods for refining the annotations automatically, as this would solve the problem of variability or detail level of the labels, and thus transform annotations by automatically re-labeling the data to a consistent format.

This paper studies the relationship between the audio similarity and semantic similarity of annotated data. The goal is to determine whether it is possible to automatically re-label an audio database by processing audio and semantic information available in order to discover a different set of labels and associate them to groups of sounds. This would provide a method to reorganize any publicly available database to the desired level of detail. Our focus is on the area of sound event detection, and concentrates on one specific database.

Sound event detection is part of computational auditory scene analysis (CASA). This area has produced applications to detecting significant sound events in movie soundtracks [2], sports videos [3], surveillance recordings [4, 5], office [6] and other everyday environments [7, 8]. Unfortunately the databases used in these studies are not shared between teams and therefore the results are not always easy to compare. An effort to create a public database for sound event detection has resulted in the DARES database [9], containing recordings from everyday environments and a variety of labels. The authors note the difficulty of choosing the level of detail in labeling sound events: too general labels will lose the details, too detailed labels will fragment the classes and render the database unusable.

Automatic tag recommendation based on audio similarity was presented by Martinez et al. [10] to overcome some of the problems related to collaborative tagging. In [10], tags for an audio file were suggested based on a kNN classifier that selected the closest neighbors based on audio similarity. Human assessment was used to evaluate the perceived quality of the candidate tags and in 77% of the sounds used, the annotations have been successfully extended with the proposed tags. The method was proposed for enhancing the semantic annotations of scarcely tagged audio files.

In the present study we propose a study based on objective and automatic measures. The study is based on the assumption that acoustically similar sound events are annotated with semantically similar terms. These assumptions will be verified by evaluating the labels of acoustically similar events, using an objective semantic similarity measure instead of us-

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).

This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. This publication only reflects the authors' views (Palomäki).

ing judgments of human listeners as Martinez et. al. [10]. We calculate the semantic similarity between the label of the test audio example (ground truth label) and the labels of the acoustically closest neighbors, to assess the possibility of relabeling data based on the labels of neighbors.

2. SYSTEM OVERVIEW

This study consists of linking the two sides of a database: the audio similarity of sound event examples and the semantic similarity of the labels they are annotated with. Audio similarity can be measured using objective measures based on distance metrics between frame-based representations of the signals or between statistical models of the signals [11]. For calculating audio similarity between sound events, we will model the event instances using Gaussian mixture models (GMM), use Kullback Leibler divergence as distance metric between event-GMMs and finally will judge similarity using a k nearest neighbors (kNN) approach.

Semantic similarity is measured using tools from natural language processing [12] and WordNet [13]. Labels of sound events usually describe the source of the sound, be it an object, action, or both (car horn, knocking, chair squeaking). Because there is an infinite amount of sound sources, producing a variety of sounds, labeling these sounds is usually a matter of personal life experience and perception [14].

2.1. Audio similarity

In order to obtain models for each individually labeled sound event in the database, we extracted audio segments according to the annotations, between the annotated start time and end time for each labeled event. Each extracted segment represents a sound event instance, for which 20 mel frequency cepstral coefficients (MFCC) were calculated in 40 bands, with 20 ms length window and 50% overlap. Based on the static, delta and acceleration coefficients, a GMM with 5 components was estimated and the distances between each two GMMs were calculated. These steps are illustrated in Fig. 1.

The similarity between sound events is characterized using the empirical symmetric Kullback Leibler divergence [11] between event GMMs. The Kullback Leibler divergence is a measure of difference between two distributions $p(x|\lambda_n)$ and $p(x|\lambda_m)$, where λ_n and λ_m are GMMs modeling two sequences of features X_n and X_m corresponding to sound events n and m.

$$D(p(x|\lambda_n)||p(x|\lambda_m)) = \int p(x|\lambda_n) log \frac{p(x|\lambda_n)}{p(x|\lambda_m)} dx \quad (1)$$

In order to obtain a distance measure, the divergence is symmetrized by adding the term $D(p(x|\lambda_m)||p(x|\lambda_n))$. The symmetric divergence can be solved in a closed form when $p(x|\lambda_n)$ and $p(x|\lambda_m)$ are modeled using a single Gaussian distribution. For multiple Gaussians, several approximations

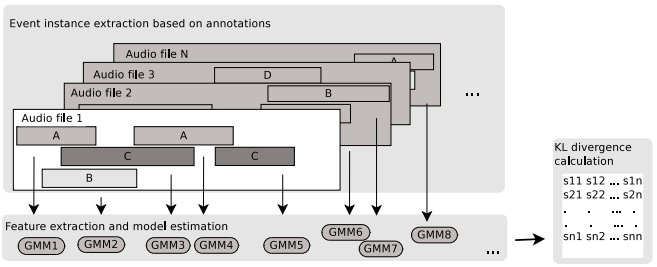

Fig. 1. Block diagram of audio processing chain

exist for the divergence [15]. A computationally efficient approximation is the *empirical Kullback-Leibler divergence*, written using the samples of the observation sequence X_n of length N that were used to train the distribution $p(x|\lambda_n)$.

$$D_{emp}(p(x|\lambda_n)||p(x|\lambda_m)) = \frac{1}{N} log \frac{p(X_n|\lambda_n)}{p(X_n|\lambda_m)}. \quad (2)$$

The empirical symmetric Kullback-Leibler Divergence is therefore calculated as:

$$E(X_n, X_m) = \frac{1}{N} log \frac{p(X_n|\lambda_n)}{p(X_n|\lambda_m)} + \frac{1}{M} log \frac{p(X_m|\lambda_m)}{p(X_m|\lambda_n)}. \quad (3)$$

The distances $E(X_n, X_m)$, collected into a similarity matrix S, can be used directly for clustering or classification. We use a dimensionality reduction method, by randomly selecting a number of q sound events from the data set that are used as anchor points indexed by a. Then we construct q dimensional feature vectors f_n for each sound event n by measuring the distance $E(X_n, X_a)$ from the sound event n to the anchor points [16]. This translates into using q random columns of the similarity matrix S, which in practice means avoiding the calculation of the full similarity matrix. In other words, each event instance will be characterized by a feature vector containing the KL divergence between its GMM and a number of q other event instance GMMs. The nearest neighbors will be calculated based on these feature vectors.

2.2. Semantic similarity

Semantic similarity calculations are based on WordNet [13]. WordNet is a lexical database for English language that groups words into sets of synonyms called synsets. The relationships between synsets are represented through hierarchies, separately for different *parts of speech* (nouns, verbs, adjectives, adverbs). For nouns, the relationships are: hypernym/hyponym (”dog” is a type of ”canine”), meronym/holonym (”finger” is part of ”hand”), coordinate terms (that share a hypernym – ”dog” and ”wolf” are both ”canine”).

These relationships can be used for example to group coordinate terms from labels into more general concepts based on their common hypernym or to link terms to each other

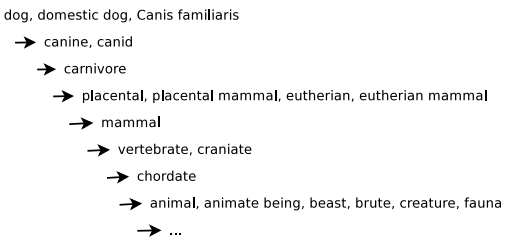

Fig. 2. WordNet hierarchy for "dog"

based on their semantic relationships. An example of WordNet hierarchy is presented in Figure 2. Words at the same level are synonyms, each lower level is a type of the upper: "dog" IS-A "canine" IS-A "carnivore" and so on. Verbs are also grouped based on IS-A relationships.

Measures of similarity use information from this IS-A hierarchy, to quantify how much a concept A is like (or is similar to) a concept B. Similarity measures can be calculated only between pairs of nouns or pairs of verbs – they do not cross the part of speech boundaries.

There are a number of similarity measures based on the path length between a pair of concepts [17]. We choose to use a measure named *path similarity*, that is calculated as the inverse of the shortest path between the two compared concepts. For example, considering the most common meanings for nouns "cat" and "dog" presented in Figure 3, the path-based similarity between them is 0.2 (inverse of the path containing 5 nodes). The value of the path similarity is bounded between 0 (not the same part of speech) and 1 (synonyms).

In this study we deal with labels that can contain multiple concepts, therefore we extend the above *path similarity* measure by considering each meaning of each concept. For consistency we will refer to it throughout this paper simply as *semantic similarity*. For example, the noun "cat" has eight meanings (a type of whip, among others), while the noun "dog" has seven meanings (hot dog, among others). The shortest path is evaluated from each meaning of "cat" to each meaning of "dog", resulting in 56 values. Out of these, the maximum value is chosen, representing the closest possible meanings of the two compared words. When the label contains more words, this process is done for each meaning of each word, and the maximum value from the entire set of results is considered as the semantic similarity between the two labels.

2.3. Labels processing

A simplification of the labels is needed for calculating semantic similarities, as the chosen similarity measure does not cross the part of speech boundaries and is not directly applicable to labels composed of multiple words. We reduce the labels to the noun(s) that it contains, converted to the basic form by stemming. In some cases the labels are collocations –

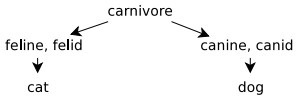

Fig. 3. The path between the most common meaning of "cat" and the most common meaning of "dog" in WordNet

multi-word expressions with a certain meaning that cannot be described by the component words separately. In such cases, when the collocation was found in WordNet, it was kept as such.

3. EVALUATION

3.1. Database

The database used in this study is DARES [9], recorded with focus on everyday sound events research. The database provides detailed annotation that describes the source that produced the sound. Each recording is accompanied by a description of the content and the location, and timed annotations of the sound sources present in the signal.

The database consists of 123 recordings of length 60 seconds. The annotations consist of a label in English, and the starting and end time. The database contains 765 unique labels, containing some duplicates (title case), spelling errors and sometimes lengthy and complex descriptions. In total there are 3214 annotated event instances. More detailed statistics about the frequency of these labels are presented in Table 1.

no.of labels	429	122	94	20	3
frequency	1	2	>5	>20	>100

Table 1. Number of annotated labels and their frequency: out of 765 unique labels, only 20 appear at least 20 times.

Simplification of the labels by extracting the nouns results in a number of 443 unique labels containing combinations of 387 unique nouns. Examples of the results of this processing are shown in Table 2. Labels containing no nouns result in empty strings; this does not change the applicability of the method, it simply reduces the number of sound events for which the similarity is calculated. From 3214 event instances, only 2881 are left to be evaluated for audio and semantic similarity, as 333 out of 3214 do not contain nouns. The frequency of the simplified labels is presented in Table 3.

3.2. Experimental results

The relationship between audio similarity and semantic similarity is evaluated for each event instance individually. For finding similar audio events for a given test event, the k nearest neighbors are sought based on q dimensional audio feature

original label	simplified label	explanation
washing machine	washing machine	collocation
putting lid on pan	lid, pan	nouns
scratching	(none)	no nouns

Table 2. Examples of labels processing outcome

no.of labels	198	68	88	29	4
frequency	1	2	>5	>20	>100

Table 3. Number of simplified labels and their frequency: out of 443 labels, 29 appear at least 20 times.

vectors f_n. Then the semantic similarity between the ground truth (simplified) label of the tested event and the (simplified) labels of these k neighbors is calculated. Different values for k and q were used in the experiments. Table 4 presents the results, average percentage of event labels among k neighbors that have a synonym as a label, and averaged semantic similarity between the label of the test event and semantically closest label among the k neighbors. Averages are calculated over all the events in the database .

A baseline value for this system is the semantic similarity between the label of the tested sound event and the label of the acoustically closest event. Using $q = 10$ randomly chosen anchors, the acoustically closest neighbor has the same label as the tested event or a synonym term in 10% of the cases. The average semantic similarity between the label of the test event and the label of the closest neighbor is 0.25, which means on average 4 nodes including the tested concepts.

We calculate the average semantic similarity of the (simplified) ground truth label to the labels of the $k = 10$ acoustically closest audio segments, in line with [10] where tags were recommended if found to characterize at least 4 of 10 nearest neighbors. The semantic similarity within a suitable number of nearest neighbors could be used as an objective measure for recommending labels for sound events, and the common label could be recommended for acoustically similar sound events. When using $k = 10$ nearest neighbors, 43% of the points have at least one out of the 10 acoustically closest examples with the same or synonym label. The average semantic similarity of the best semantic match within this neighborhood is 0.51(two nodes). This means that on average each event has at least one of ten acoustically closest neighbors having as label a direct hyponym or hypernym of its own label.

q anchors	k neighbors	synonyms	averaged semantic similarity
10	1	10 %	0.25
10	10	43%	0.51
10	20	47%	0.53
50	10	48%	0.58

Table 4. Evaluation results using q anchors and k acoustically closest neighbors: synonyms and best semantic match between neighbors, averaged over the database

4. DISCUSSION

Based on the presented evidence, we observe that in many cases the acoustically similar neighbors had semantically similar labels. It is reasonable to say that re-labeling a diversely annotated sound events according to their acoustically similar neighbors would provide acceptable outcome for databases like DARES, most probably resulting in a more compact set of new labels. In 43% of the cases there is at least one of ten nearest neighbors which has the same label or a synonym. Overall the semantic similarity was on average 0.5, which means a distance of two nodes. In WordNet, the synsets that have a distance of two nodes are direct hypo/hypernyms. This means that the terms used in the labels are very close to each other, and the re-labeling would on average result in a term which is one level higher or lower in semantic hierarchy.

Based on the results presented in Table 4, we can observe that when more neighbors are evaluated, or a higher number of anchors is chosen, the chance for finding synonym labels in the neighborhood is higher: 47% for 20 neighbors compared to 43% for only 10 neighbors when using 10 anchors, 48% for 50 anchors compared to 43% for 10 anchors, when evaluating 10 nearest neighbors. However the average semantic similarity of the closest label from this neighborhood does not change that much with the number of neighbors. For the envisioned application, of re-labeling data, we consider that it is important to use moderately small number of neighbors, in order to find a good acoustic match as well as a closely related label. Instead of extending the annotations as in [10], this system would be used in a different way, by imposing a set of labels and re-labeling the diversely annotated data into the chosen reduced set of labels.

5. CONCLUSIONS

This paper presented a study of the semantic similarity between the labels of a diversely annotated database with sound event examples. We verified the assumption that acoustically similar events are labeled with semantically similar terms, and conclude that it is a valid assumption. In most cases there was at least one in ten acoustically closest neighbors that had a synonym label, and on average all had a neighbor which is one level of detail higher or lower in the semantic hierarchy.

Planned future work includes developing strategies for selecting new labels. For each sound event a new label could be chosen based on the combined audio and semantic similarity. Such a system also needs a mechanism for discarding sound event examples, and methods for dealing with verbs and combinations of verbs and nouns.

6. REFERENCES

[1] G. Peeters and K. Fort, "Towards a (better) definition of the description of annotated M.I.R. corpora," in *ISMIR (International Symposium for Music Information Retrieval)*, 2012.

[2] R. Cai, Lie Lu, A. Hanjalic, H-J. Zhang, and L-H. Cai, "A flexible framework for key audio effects detection and auditory context inference," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 1026 – 1039, may 2006.

[3] M. Xu, C. Xu, L. Duan, Jesse S. Jin, and S. Luo, "Audio keywords generation for sports video analysis," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 4, no. 2, pp. 11:1–11:23, May 2008.

[4] Y-T. Peng, C-Y. Lin, M-T. Sun, and K-C. Tsai, "Healthcare audio event classification using hidden markov models and hierarchical hidden markov models," in *Proceedings of the 2009 IEEE international conference on Multimedia and Expo*, 2009, ICME'09, pp. 1218–1221.

[5] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *IEEE International Conference on Multimedia and Expo, ICME 2005*, july 2005, pp. 1306 –1309.

[6] R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds., *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007*, Springer-Verlag, Berlin, Heidelberg, 2008.

[7] F. Kraft, R. Malkin, T. Schaaf, and A. Waibel, "Temporal ica for classification of acoustic events in a kitchen environment," in *Proceedings of the INTERSPEECH*, 2005, pp. 2689–2692.

[8] A. Mesaros, T. Heittola, T. Virtanen, and A. Eronen, "Acoustic events detection in real life recordings," in *Proc. of the 2010 European Signal Processing Conference (EUSIPCO-2010)*, 2010, pp. 1267–1271.

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

[10] E. Martinez, O. Celma, M. Sordo, B. De Jong, and X. Serra, "Extending the folksonomies of freesound.org using contentbased audio analysis," in *Sound and Music Computing Conference*, 2009.

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

[12] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit*, O'Reilly Media, 2009.

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

[14] B. Gygi and V. Shafiro, "Environmental sound research as it stands today," *Proceedings of Meetings on Acoustics*, vol. 1, no. 1, 2007.

[15] J. R. Hershey and P. A. Olsen, "Approximating the kullback leibler divergence between gaussian mixture models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'07)*, 2007, vol. 4, pp. IV–317 – IV–320.

[16] M. Helen and T. Lahti, "Query by example in large databases using key-sample distance transformation and clustering," in *Proceedings of the Ninth IEEE International Symposium on Multimedia Workshops*, 2007, ISMW '07, pp. 303–308.

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