Audio clip classification using social tags and the effect of tag expansion

Frederic Font[1], Joan Serrà[2], and Xavier Serra[1]

[1]*Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain*

[2]*Artificial Intelligence Research Institute (IIIA-CSIC), Bellaterra, Barcelona, Spain*

Correspondence should be addressed to Frederic Font (`frederic.font@upf.edu`)

ABSTRACT

Methods for automatic sound and music classification are of great value when trying to organise the large amounts of unstructured, user-contributed audio content uploaded to online sharing platforms. Currently, most of these methods are based on the audio signal, leaving the exploitation of users' annotations or other contextual data rather unexplored. In this paper, we describe a method for the automatic classification of audio clips which is solely based on user-supplied tags. As a novelty, the method includes a tag expansion step for increasing classification accuracy when audio clips are scarcely tagged. Our results suggest that very high accuracies can be achieved in tag-based audio classification (even for poorly or badly annotated clips), and that the proposed tag expansion step can, in some cases, significantly increase classification performance. We are interested in the use of the described classification method as a first step for tailoring assistive tagging systems to the particularities of different audio categories, and as a way to improve the overall quality of online user annotations.

1. INTRODUCTION AND RELATED WORK

The internet is full of user-contributed multimedia content that is usually lacking common metadata or annotations to help in its organisation, browsing, sharing and reuse [24]. Each online platform has its own particular uploading process, featuring different functionalities and design. However, practically all such processes ask users to provide some kind of metadata to be able to easily index the uploaded content. For that purpose, it is very common the usage of collaborative tagging systems [18]. Thus, the responsibility of describing uploaded content relies on users themselves, and so depends the descriptions' accuracy and comprehensiveness. As a result, multimedia content is often sparsely annotated and with a certain degree of incoherence due to users' different annotating styles, which limits the possibilities for resource organisation and structured content browsing [5, 11].

Facing that problem, content-based techniques for the automatic classification of audio, image and video provide reasonably good results when classifying unstructured contents in reduced domains and using rather general taxonomies. In the audio domain, some research has been focused on specific problems such as distinguishing among instruments [4, 12, 17] and sound effects [6, 13, 19, 22]. In these studies, sounds are classified into non-overlapping categories such as instrument names (e.g. *trumpet*, *saxophone*), performance articulations (e.g. *pizzicato*, *staccato*) or foley and environmental sounds (e.g. *explosions*, *doors*, *automobiles*, *animals*). Other approaches classify audio clips in broader categories such as *environmental sounds*, *music* or *speech* [6, 19]. All these approaches follow a very similar structure. Typically, a set of low-level audio features is extracted from sound samples in a given collection, yielding a feature vector representation of every sound. Also, sound samples are manually annotated using the concepts of a taxonomy representing the particular classification domain (e.g. a taxonomy of musical instruments or sound effects). These taxonomies tend to be rather small (between 2 and 20 concepts). Then, supervised learning is performed using SVM, HMM, k-NN or Bayesian classifiers trained with the feature vectors corresponding to annotated sound samples.

Further research has been focused in music genre classification [7, 23] and mood classification [2, 14]. Again, the followed approach is very similar to the one previously described. However, some of these studies also make use of use of textual data such as lyrics and social

tags. Chen et al. [7] take advantage of social tags to build a song similarity graph (on the basis of shared tags) that is used to propagate the output of a content-based music genre classifier among neighbouring songs in the graph. In that case, no classifier is directly trained with tagging data. Laurier et al. [14] use latent semantic analysis on songs' lyrics to train a k-NN classifier which, in combination with an SVM trained with low-level audio feature vectors, is used for distinguishing between moods. In a similar way, Bischoff et al. [2] also present an hybrid strategy for mood classification which linearly combines an SVM classifier trained with low-level audio feature vectors and a Naive Bayes classifier trained with social tags extracted from Last.fm. A bag-of-words approach is followed to represent songs as high-dimensional vectors of tag occurrences, but no dimensionality reduction techniques are applied to these vectors and therefore the Naive Bayes classifier is trained with very high dimensional data. To the best of our knowledge, no audio classification systems have been researched that only use information coming from social tags.

Generally speaking, the classification of image [16] and video [3] content also follows very similar strategies to the ones outlined for sound and music. As far as we know, no methods for image or video classification have been designed that use only tag information. Given the high availability of user annotated multimedia content in the internet we believe that despite annotations' noisiness, classification methods based on information such as social tags deserve more attention.

In this paper we propose a method for audio clip classification that makes use of well-known supervised learning techniques to automatically categorize audio clips. The general structure of the proposed method is very similar to what can be found in the literature, except for two main differences: (i) we use user annotations, i.e., tags, to represent audio clips instead of low-level audio features, and (ii) we propose a novel step consisting in expanding audio clips' annotations by automatically adding new related tags prior to classification. We evaluate our method using two different classifiers and a dataset of publicly-available annotated audio clips extracted from Freesound[1]. We focus our study on the classification accuracy of scarcely annotated audio clips.

[1]Freesound (www.freesound.org) data, including audio clips and annotations, can be gathered using the pubic Freesound API (www.freesound.org/help/developers/).

Automatic audio classification is fundamental for improving search, browsing and reuse of online content. We are however very interested in the use of audio classification as a first step for further applying different treatments and specialized processing to audio samples depending on the classification results. We are particularly interested in the use of audio classification in assistive tagging systems [24]. Such a system could predict the category of a given audio sample through a classification process and then ask users to annotate with tags regarding meaningful audio properties relevant for the particular category. The classification system we describe here is designed to be a part of a bigger assistive tagging system for the annotation of audio samples.

The rest of the paper is organized as follows. In Sec. 2 we describe our dataset, the classification methodology and the evaluation strategy. Sec. 3 reports the results of the evaluation and Sec. 4 summarizes some conclusions and outlines future work.

2. METHODOLOGY

2.1. Data set

Freesound is a popular site for audio sharing that contains more than 170,000 audio clips and has more than three million registered users. In Freesound, users contribute with self-created audio clips of very different nature, and annotate them using tags and textual descriptions (in this article we only focus on the tag annotations). By audio clips we understand any kind of sounds including effects, field recordings, environmental sounds, melodies played by instruments, rhythmic loops, etc., but not including musical recordings in the most traditional sense of "songs". The Freesound tagging system is not very sophisticated, it does not provide any kind of guidance for users such as tag recommendation nor does it restrict in any way the vocabulary of tags that can be used. Previous research has shown that audio clip descriptions tend to be quite noisy and that there is not much agreement among users regarding tagging patterns and styles, yielding a noisy folksonomy [10].

Given such heterogeneity, we have defined the categories we want to infer from tag information in a way that they can include the whole range of sounds that can be found in the original collection. We have also defined such categories so that they can, in a near future, allow us to apply meaningful different treatments tailored to different types of audio clips (see the discussion in Sec. 4). The

resulting categories are quite general and are inline with other categorisations of audio clips reported in the literature [6, 19]:

1. SOUNDFX: here we include all kinds of what is generally known as sound effects, including *foley*, footsteps, opening and closing doors, alarm sounds, cars passing by, animals, and all kinds of noises or artificially created glitches. In general these tend to be short clips.

2. SOUNDSCAPE: this category includes generally longer recordings resulting of the addition of multiple sounds that, in isolation, would be classified under SOUNDFX. Examples would be environmental recordings, street ambiances or artificially constructed complex soundscapes.

3. SAMPLES: this category represents all sorts of instrument samples, including single notes, chords and percussive hits. Typical examples of this category include single notes of a piano recorded one by one and uploaded as different audio clips, or samples from a complete drum set.

4. MUSIC: here we include more complex musical fragments such as melodies, chord progressions, and drum loops. In the same way as SOUNDSCAPE sounds can be understood as the addition of multiple SOUNDFX, audio clips under MUSIC category can be conceived as combinations of SAMPLES.

5. SPEECH: the last category includes all sorts of speech-related audio clips such as text reading, single words or recordings of text-to-speech processors.

We have not investigated the use of more precise categories as our current goal is the classification of audio clips in Freesound in broad categories that allow further tailored treatment, and not the classification of these audio clips into a more specific taxonomy that could be used as an interface for browsing Freesound content.

In order to create a data set for the supervised learning process we manually assigned one of the above categories to a number of audio clips from Freesound. To do that we have been iteratively presented with randomly chosen audio clips and assigned them to one category. As it can be imagined, these categories are not completely orthogonal and there are some clips for which the decision has not been straightforward just by listening to the audio. In these cases, we also relied on provided textual descriptions. The crafted data set includes a minimum of 2,088 sounds per category (corresponding to the case of SPEECH) and a maximum of 6,341 (for the case of SAMPLES). Comparing the totality of Freesound audio clips and the manually annotated subset, we observe qualitatively similar relative distributions of tag occurrences and number of tags per audio clip. Fig. 1 shows typical examples of tags that Freesound users assigned to audio clips for the five defined audio categories.

2.2. Classification method

To classify audio clips we follow a bag-of-words approach where each clip is represented as a vector whose elements indicate the presence or absence of a particular tag. Feature vectors contain all possible tags in the collection, thus their dimensionality is very high. Here we do not carry on any dimensionality reduction step to lower the size of the feature vectors. Instead, in order to keep them in manageable sizes, we remove all tags that, considering tag assignments for all Freesound audio clips, are used less than 10 times. This leaves us with a total of 7,712 tags, yielding binary vectors of 7,712 dimensions. Notice moreover that these vectors are very sparse, as audio clips are usually tagged with only a few tags (actually, they are annotated with an average of 6.79 tags per clip [10]). These particularities make the problem close to what is normally found in text classification, where high dimensionality and sparseness are commonplace [20].

We use the aforementioned feature vectors to fit a classifier, using the same number of examples for each class. We test our method using both a support vector machine (SVM) and a naive Bayes (NB) classifier[2]. The choice of these specific classifiers is motivated by their popularity in multimedia classification tasks, and because they have been shown to be well suited for high dimensional and sparse classification tasks such as the one we are facing here [1, 20] (details about the exact number of examples per class used to fit the classifier and the consideration of training and test sets are given in Sec. 2.3). For the sake of simplicity, our approach consists in training one sin-

[2]We implement the classifiers using the "scikit-learn" Python package (http://scikit-learn.org/). We use the classes `LinearSVC` and `BernoulliNB` for SVM and NB, respectively, with default parameters. `LinearSVC` follows the "one versus all" approach for multiclass classification.

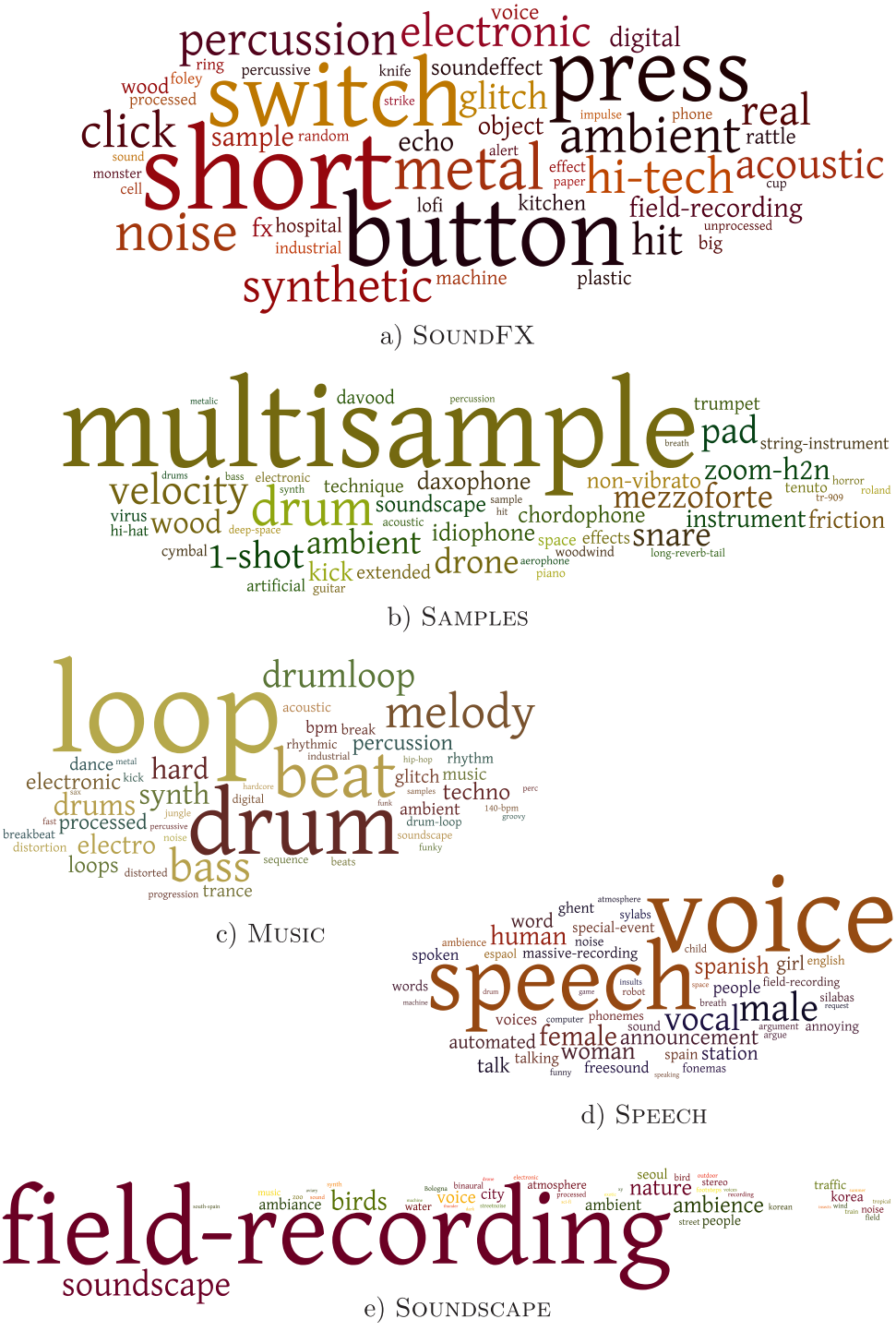

Fig. 1: Tagclouds of the 50 most used tags in the five defined audio categories. The size of the tags is proportional to the frequency of occurrence among all the clips annotated under each category. For building these tagclouds we only considered the set of clips manually annotated as ground truth. Tagclouds were generated with an online tool available at www.wordle.net.

gle classifier for distinguishing among the five categories described above.

After fitting the classifier, but prior to assigning a category to our audio clips, we introduce an extra step which has the goal of improving classification accuracy, specially for those audio clips that are poorly labeled (e.g., with only one or two tags). Such step consists in automatically expanding the annotations of audio clips by adding other related tags. The idea is that, this way, we give more information to the classifier for predicting the category of otherwise scarcely-labeled audio clips. To perform the tag expansion step we use the tag recommendation system described in [9], which is solely based on tag assignments and basically computes a tag-tag similarity matrix on the basis of tag co-occurrence in audio clips. This similarity matrix is used to select some candidates given a set of input tags, and then a number of heuristics are applied to sort these candidates and to automatically determine how many of them should be recommended. Therefore, given a list of some tags we can use the tag recommendation system to expand the list with some other presumably relevant tags. The tag recommendation system is configured with the combination of parameters that reports better average precision and recall according to [9].

2.3. Evaluation strategy

We follow a random sub-sampling cross-validation strategy where we split our data set into training and testing sets. We then compute the out-of-sample accuracy as a percentage of well-classified instances from the testing set when using the fit from the training set. This process is repeated 100 times for each classifier and parameter configuration that we test (see below), and overall accuracy is obtained by averaging over the results of all repetitions. In each repetition, our data set is composed of a random selection of 1,000 audio clips from every category, adding up to a total of 5,000. This way we maintain a balance in the number of audio clips per category. We additionally impose the limit of not getting more than 50 clips of the same category uploaded by the same Freesound user. We do that to avoid what could potentially be an equivalent of the *album effect* that is known to happen in automatic music artist recognition [26]. In each repetition, the testing set is selected as a random subset representing 10% of the data, and being equally-distributed among categories (i.e., 100 audio clips per category).

As mentioned in Sec. 2.2, we test our method using SVM and NB classifiers. We also added a random classifier to serve as a baseline. To understand the effect of the tag expansion step, we also test the method for two separate configurations where this step is turned on and off. In addition, we are specially interested in evaluating the accuracy of the classification system in those cases where only a few input tags are available. Hence, we introduce a limitation to the testing set consisting in randomly removing tags from audio clips prior to classification, only leaving a particular number of N input tags per audio clip. We consider values of N ranging from 1 to 5. This obviously adds another constraint to the selection of the testing set, which is to make sure that selected audio clips have at least N tags. The whole evaluation process is performed for all the different values of N, for both SVM and NB classifiers, and for the configurations with tag expansion turned on and off, yielding a total of 20 evaluated experiment combinations.

3. RESULTS

Fig. 2 shows the accuracy results of our classification method for all the experiment combinations described in Sec. 2.3. Note that all experiment combinations are far above the random classifier accuracy. The NB classifier reports overall a higher accuracy than the SVM, with a statistically significant[3] average accuracy increase of 10% ($p < 10^{-12}$). The tag expansion step is shown to be very useful for the SVM case, reporting a statistically significant average accuracy increase of 9% ($p < 10^{-9}$). Also, tag expansion tends to add more tags to scarcely labeled samples. This means that the lower the number of input tags, the bigger the number of added tags (average of 7.18 added tags per evaluated sample). Importantly, for the SVM case, we see that the smaller the number of input tags, the larger the increase in accuracy when using tag expansion compared to switching it off. However, the tag expansion step does not exhibit similar results for the NB case, and it even shows an statistically significant small decrease in accuracy (average of -2%, $p < 0.05$). Overall, the classification system is able to successfully classify audio clips among five generic categories inside the audio domain (Sec. 2.1), with accuracies ranging from 70% to 90%, and depending on the number of input tags available for classification.

We have performed additional experiments with different

[3] Statistical significance is assessed by considering the maximum p-value across pairwise comparisons between experiment combinations and using the well-known Wilcoxon signed-rank test with Bonferroni adjustment [8].

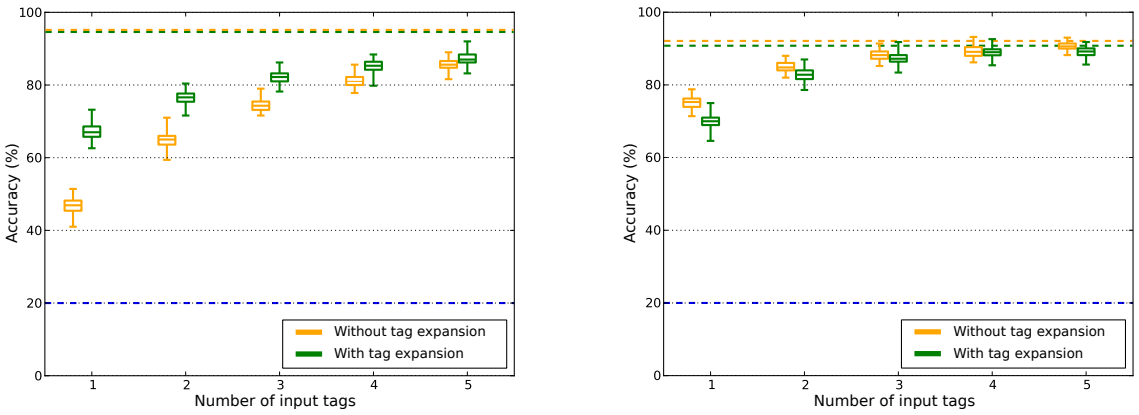

Fig. 2: Classification accuracy using SVM (left) and NB (right) classifiers. The dashed line at 20% accuracy corresponds to the random baseline (see Sec. 2.3). The dashed lines around 95% (SVM) and 90% (NB) correspond to the accuracy achieved when no restriction on the number of tags for the testing set is performed.

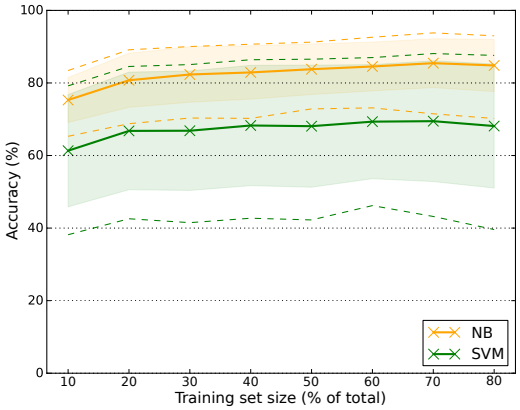

Fig. 3: Average classification accuracy for SVM and NB classifiers when using different training set sizes. We aggregated the accuracy results of the experiments with different numbers of input tags ($N = 1,...,5$) for every classifier. Dashed lines indicate minimum and maximum accuracies while the shaded zones indicate standard deviation. These experiments are computed without the tag expansion step.

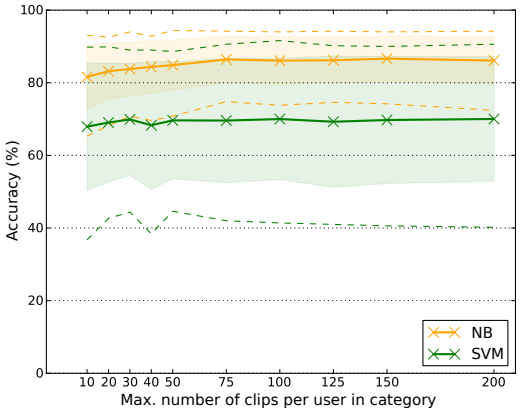

Fig. 4: Average classification accuracy for SVM and NB classifiers with different maximum number of uploaded clips of the same Freesound user in the same audio category. We aggregated the accuracy results of the experiments with different numbers of input tags ($N = 1,...,5$) for every classifier. Dashed lines indicate minimum and maximum accuracies while the shaded zones indicate standard deviation. These experiments are computed without the tag expansion step.

training set sizes (i.e., using less than than 90% of audio clips for training). The results we obtained are consistent with those reported above with very few variation on accuracy for training set percentages higher than 50% (Fig. 3). This reinforces the validity of the classification results as the use of smaller training sets does not heavily affect classification accuracy. Furthermore, we also tested different values for the imposed maximum of 50 clips uploaded by the same user in the same category (Sec. 2.3). Our results show that the accuracy does not seem to be very influenced by such limit (Fig. 4), thus

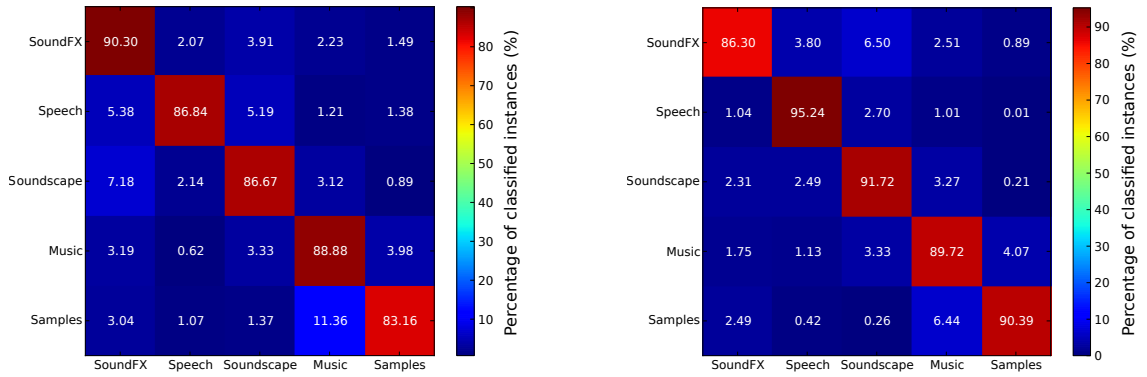

Fig. 5: Confusion matrix for the best scoring experiment combinations of SVM (left, $N = 5$, using tag expansion) and NB (right, $N = 5$, not using tag expansion) classifiers. We only reproduce these two matrices as are representative of the resulting matrices of other experiment combinations.

partially questioning the existence of the aforementioned *user effect*.

Considering the confusion matrix of different experiment combinations (Fig. 5) it can be observed that, although there are not very strong patterns regarding the confusion between category pairs, it happens for most experiment combinations that the pairs SAMPLES-MUSIC, SPEECH-SOUNDSCAPE and SOUNDSCAPE-SOUNDFX tend to be more confused than the others. It is intuitively plausible that MUSIC and SAMPLES are confused given that tags like instrument names could be used to annotate clips in both categories. Moreover, tags used for SOUNDFX are also typically used in SOUNDSCAPE to designate particular sound sources that appear in the recording. Confusions between SOUNDSCAPE and SPEECH can be intuitively explained because a lot of ambient recordings contain background voices and are therefore annotated with tags such as `voice`, `talking` and language names.

4. DISCUSSION AND FUTURE WORK

In this paper we have proposed a method for the automatic classification of audio clips solely based on user-provided tags. The method is able to successfully classify audio clips into five broad categories with accuracies ranging from 70% to 90%. We described a novel tag expansion step intended to improve accuracy when classifying poorly annotated audio clips. Although results show that this step does not seem to be useful when using a naive Bayes classifier (it does not contribute to an increase of accuracy), it certainly improves the results when using a support vector machine classifier. In that case, the tag expansion step contributes a significant accuracy increase of 9%, particularly prominent when classifying scarcely annotated clips (Fig. 2, left). We believe that the tag expansion step could potentially contribute in increasing accuracy when using other kinds of classifiers too.

A good aspect of the method we propose is that, as it is solely based on social tags, it could presumably be directly generalised to other multimedia domains. As a further improvement, we shall include a pre-processing step for reducing the tag noisiness prior to the training and classification steps. Such noisiness reduction could include natural language processing techniques like stemming or keyword extraction from user-provided textual sound descriptions, which are not considered in this paper.

We believe that a better organisation of user-contributed audio (and multimedia content in general) should not only be approached with the modelling of more sophisticated and accurate classifiers. In particular, tagging systems should be able to promote the generation of more reliable annotations from the users' side. Noticeably, there exist some efforts in that direction which describe systems that assist users in the tagging process [24]. We think that such systems could take advantage of automatic classification methods such as the one proposed here. One idea that we want to explore is the use of automatic classification as a first step for a tag recommendation system that would adapt its output to the detected category. Assistive tagging systems could even take ad-

vantage of automatic classification to require category-specific types of relevant information during a tagging process (e.g., asking for tags describing properties such as "pitch" or "instrument name" when annotating an audio clip belonging to SAMPLES). We believe that better annotations would result in less noisy folkonomies and would not only benefit the search and retrieval of online content but would also leverage the value of folksonomies as sources for knowledge and ontology mining.

5. ACKNOWLEDGEMENTS

This work has been supported by BES-2010-037309 FPI from the Spanish Ministry of Science and Innovation (TIN2009-14247-C02-01; F.F.), 2009-SGR-1434 from Generalitat de Catalunya (J.S.), JAEDOC069/2010 from Consejo Superior de Investigaciones Científicas (J.S.), ICT-2011-8-318770 from the European Commission (J.S.), and FP7-2007-2013 / ERC grant agreement 267583 (CompMusic).

6. REFERENCES

[1] K. P. Bennett and C. Campbell. Support vector machines: Hype or hallelujah? *ACM SIGKDD Explorations Newsletter*, 2(2):1–13, 2000.

[2] K. Bischoff, C. S. Firan, R. Paiu, W. Nejdl, C. Laurier, and M. Sordo. Music mood and theme classification-a hybrid approach. In *Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)*, pages 657–662, 2009.

[3] D. Brezeale and D. Cook. Automatic video classification: A survey of the literature. *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews*, 38(3):416–430, 2008.

[4] P. Cano, M. Koppenberger, S. Le Groux, J. Ricard, N. Wack, and P. Herrera. Nearest-neighbor generic sound classification with a WordNet-based taxonomy. In *Proc. of the Audio Engineering Society Convention*, 2004.

[5] I. Cantador and I. Konstas. Categorising social tags to improve folksonomy-based recommendations. *Web Semantics: Science Services and Agents on the World Wide Web*, 9(1):1–15, Mar. 2011.

[6] M. Casey. General sound classification and similarity in MPEG-7. *Organised Sound*, 6(2):153–164, 2002.

[7] L. Chen, P. Wright, and W. Nejdl. Improving music genre classification using collaborative tagging data. In *Proc. of the ACM Int. Conf. on Web Search and Data Mining (WSDM)*, pages 84–93, 2009.

[8] Dunn, O. J. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

[9] F. Font, J. Serrà, and X. Serra. Folksonomy-based tag recommendation for collaborative tagging systems. *International Journal of Web Semantics and Information Systems*, 9(2):1–30, 2013.

[10] F. Font and X. Serra. Analysis of the folksonomy of freesound. In *Proc. of the CompMusic Workshop*, pages 48–54, 2012.

[11] H. Halpin and V. Robu. The dynamics and semantics of collaborative tagging. In *Proc. of the Semantic Authoring and Annotation Workshop*, pages 1–21, 2006.

[12] P. Herrera, G. Peeters, and S. Dubnov. Automatic classification of musical instrument sounds. *Journal of New Music Research*, 32(1):3–21, 2003.

[13] C.-C. J. Kuo and T. Zhang. Classification and retrieval of sound effects in audiovisual data management. In *Proc. of the 33rd Asilomar Conf. on Signals, Systems, and Computers*, volume 1, pages 730–734. Ieee, 1999.

[14] C. Laurier, J. Grivolla, and P. Herrera. Multimodal Music Mood Classification Using Audio and Lyrics. In *Proc. of the Int. Conf. on Machine Learning and Applications (ICMLA)*, pages 688–693, 2008.

[15] M. Levy and M. Sandler. Music Information Retrieval Using Social Tags and Audio. *IEEE Transactions on Multimedia*, 11(3):383–395, 2009.

[16] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, 2007.

[17] A. Livshin, G. Peeters, and X. Rodet. Studies and improvements in automatic classification of musical sound samples. In *Proc. of the Int. Computer Music Conf. (ICMC 2003)*, pages 171–178, 2003.

[18] C. Marlow, M. Naaman, M. Davis, and S. Hall. HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, To Read. In *Proc. of the Conf. on Hypertext and Hypermedia*, pages 31–39, 2006.

[19] G. Roma, J. Janer, S. Kersten, M. Schirosa, P. Herrera, and X. Serra. Ecological Acoustics Perspective for Content-Based Retrieval of Environmental Sounds. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010:7, 2010.

[20] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[21] N. Scaringella, G. Zoia, and D. Mlynek. Automatic genre classification of music content: a survey. *IEEE Signal Processing Magazine*, 23(2):133–141, 2006.

[22] S. Sundaram and S. Narayanan. Classification of sound clips by two schemes: Using onomatopoeia and semantic labels. In *Proc. of the IEEE International Conference on Multimedia and Expo (ICME 2008)*, pages 1341–1344, June 2008.

[23] R. Tao, Z. Li, and Y. Ji. Music genre classification using temporal information and support vector machine. In *Proc. of the 16th Advanced School for Computing and Imaging Conf. (ASCI 2010)*, 2010.

[24] M. Wang, B. Ni, X.-S. Hua, and T.-S. Chua. Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Computing Surveys*, 44(4):1–24, 2012.

[25] J. Weston, S. Bengio and P. Hamel. Multi-Tasking with Joint Semantic Spaces for Large-Scale Music Annotation and Retrieval. *Journal of New Music Research*, 40(4):337–348, 2011.

[26] B. Whitman, G. Flake, and S. Lawrence. Artist detection in music with Minnowmatch. In *Proc. of the IEEE-SPS Workshop on Neural Networks for Signal Processing*, pages 559–568, 2001.