TOWARDS TIME-VARYING MUSIC AUTO-TAGGING BASED ON CAL500 EXPANSION

Shuo-Yang Wang[1], *Ju-Chiang Wang*[1,2], *Yi-Hsuan Yang*[1], *and Hsin-Min Wang*[1]

[1]Academia Sinica, Taipei, Taiwan
[2]University of California, San Diego, CA, USA
Emails: {raywang, asriver, yang, whm}@iis.sinica.edu.tw

ABSTRACT

Music auto-tagging refers to automatically assigning semantic labels (tags) such as genre, mood and instrument to music so as to facilitate text-based music retrieval. Although significant progress has been made in recent years, relatively little research has focused on semantic labels that are time-varying within a track. Existing approaches and datasets usually assume that different fragments of a track share the same tag labels, disregarding the tags that are time-varying (e.g., mood) or local in time (e.g., instrument solo). In this paper, we present a new dataset dedicated to *time-varying* music auto-tagging. The dataset, called *CAL500exp*, is an enriched version of the well-known CAL500 dataset used for conventional track-level tagging. Given the tag set of CAL500, eleven subjects with strong music background were recruited to annotate the time-varying tag labels. A new user interface for annotation is developed to reduce the subject's annotation effort yet increase the quality of labels. Moreover, we present an empirical evaluation that demonstrates the performance improvement CAL500exp brings about for time-varying music auto-tagging. By providing more accurate and consistent descriptions of music content in a finer granularity, CAL500exp may open new opportunities to understand and to model the temporal context of musical semantics.

Index Terms— Music auto-tagging, temporal context, time-varying, annotation interface, dataset construction

1. INTRODUCTION

Fueled by the tremendous growth of digital music libraries, a large number of example-based and text-based music information retrieval (MIR) methods have been proposed in the literature. The former retrieval scenario allows users to query music with audio examples, such as a hummed melody or a fragment of a desired song [1, 2], whereas the latter helps users to search music through a few keywords related to high-level music semantics or metadata such as artist name, song title, genre, style, mood, and instrument [3–5]. The task of automatically tagging musical items (e.g., artists, albums, or tracks) with such high-level musical semantics is usually referred to as music *auto-tagging* in the MIR literature [6–23].

In many previous works, music auto-tagging has been devoted to labeling music in the *track-level*, assuming that the overall content of a track can be summarized by a set of tags [8, 9, 13, 18]. That is, they usually collect the ground-truth associations between tag and music in the track level [24], develop a set of track-level auto-taggers, and then evaluate the accuracy by comparing the predicted labels against the ground-truth ones. This approach is straightforward since it is natural for people to talk about music in the track-level. However, it might not be adequate for tracking the tags that vary with time as different fragments of a track might be semantically non-homogenous. For example, it is well-known that the music emotion aspect is better modeled as *time-varying* [25, 26]. For local musical events such as instrument solo, it is also preferable to consider the corresponding audio content in a finer granularity (i.e., smaller temporal scale) [22].

The prevalence of the track-level approach might be partly due to the difficulty of collecting tag labels in smaller temporal scale. It requires people to listen to a track and make the moment-by-moment annotations consecutively. An annotator would have to listen to the same track several times to ensure that the annotation is accurate and complete, which is enormously labor-intensive and time consuming. Therefore, existing datasets for auto-tagging usually employ track-level tags [14, 27], without specifying the exact temporal positions in a track with which a given tag is associated.

Mandel *et al.* presented an early attempt to address this issue [7, 15]. For each track, they sampled five fixed-length (10-second) segments evenly spaced throughout the track. Then, the crowdsourcing platform Mechanical Turk [29] was adopted to collect the tags for each segment. They found that different parts of the same track tend to be described differently by the human listeners. However, obtaining a segment for annotation without concerning its possible acoustic homogeneity and the corresponding duration variability may result in degrading the tag label quality, as the annotators might not easily catch the local musical event. By describing tags in a shorter and variable temporal scale that is acoustically homogeneous, the connection between natural language

This work was supported by the Ministry of Science and Technology of Taiwan under Grant NSC 101-2221-E-001-019-MY3 and the Academia Sinica–UCSD Fellowship to Ju-Chiang Wang.

Table 1. Existing datasets for music auto-tagging

dataset	stimuli	annotation method	taxonomy	label	# tags	public
CAL500 [27][1]	500 tracks	university students	expert	strong	174	yes
CAL10k [14][1]	10,870 tracks	professional editors	expert	weak	1,053	yes
MSD [28][3]	1,000,000 tracks	social tags	folksonomy	weak	7,643	yes
MajorMiner [6][4]	2,600 segments (10 sec)	game with a purpose	folksonomy	weak	6,700	no
Magnatagatune [10][5]	25,860 segments (30 sec)	game with a purpose	folksonomy	weak	188	yes
Mech. Turk [15]	925 segments (10 sec)	crowdsourcing	folksonomy	weak	2,100	no
CAL500exp[2] (this work)	3,223 segments (3–16 sec) from 500 tracks	experts	expert	strong	67	yes

(i.e., tags) and music would be better defined, leading to new opportunities to bridge the so-called semantic gap [4].

To this end, our goal of *time-varying music auto-tagging* is to train the auto-taggers based on length-variable homogeneous segment tag labels so as to make more accurate tag predictions for contiguous, overlapping short-time segments (with variable length) of a track. The concept of time-varying music auto-tagging lends itself to applications such as audio summarization, *playing-with-tagging (PWT)* [22] (i.e., visualizing music signals by tracking the tag distribution during playback), *automatic music video generation* [30, 31] (i.e., matching between the music and video signals in a more fine-grained temporal scale), and *audio remixing* [32] (i.e., jumping from a fragment of a track to a fragment of another track.

Following this research line, in this paper we present a novel dataset to foster time-varying music auto-tagging. The dataset, which is called *CAL500 Expansion* (CAL500exp), is an enriched version of the well-known CAL500 dataset [9].[1] Below we highlight three main contributions of this work.

- We present a novel protocol with three new elements tailored for constructing a time-varying music auto-tagging dataset. First, instead of using segments of fixed duration, we perform audio-based segmentation to extract acoustically homogenous segments with variable length and inter-segment clustering to select the representative segments for annotation (cf. Section 3.1). Second, instead of annotating each segment from scratch, we initialize the annotation of each segment based on the track-level labels of CAL500 and ask subjects to check and refine the labels to save annotation burden (cf. Sections 3.2–3.3). Third, instead of resorting to crowdsourcing, we recruit subjects with strong music background and devise a new user-interface for better annotation quality (cf. Section 3.4).
- We present a comparative study that validates the performance gain brought about by CAL500exp for time-varying music auto-tagging (cf. Section 4).
- We have made CAL500exp available upon request to the research community.[2]

[1] http://cosmal.ucsd.edu/cal/projects/AnnRet/

[2] http://slam.iis.sinica.edu.tw/demo/CAL500exp/

2. RELATED WORK

Music auto-tagging has been studied for years [13]. Many sophisticated machine learning algorithms have been proposed to improve the accuracy of auto-tagging, including the consideration of tag correlation [11], cost-sensitive ensemble learning [19], time series models [20] and deep neural network [21]. In this paper, we attempt to improve the performance of auto-tagging via constructing a new dataset whose labels are more accurate, consistent and complete, with a specific focus on handling music semantics that are local or time-varying.

Tagged music database can be obtained from different sources [24], including conducting human surveys, deploying games with a purpose, collecting web documents or harvesting social tags. One can have an overview with Table 1 that, existing datasets usually differ in the granularity of annotation (track- or segment-level), number of musical pieces and tags, annotation methods, level of expertise of the annotators (e.g. crowd or experts), taxonomy definition (expert or folksonomy [8]), and the label type (strong or weak).[6]

We note that the CAL500 dataset, which consists of 500 Western Pop songs, is a widely-used track-level dataset [9, 11, 20, 21]. It employs 174 expert-defined tags covering 8 semantic categories including emotion, genre, best-genre, instrument, instrument solo, vocal style, song characteristic and usage. The decision of each tag label is made by "majority voting" over at least three paid university students. We build the new dataset (CAL500exp) based on CAL500, because of its complete and balanced taxonomy and relatively high label quality (cf. Table 1).

CAL500exp, which is introduced in this paper, stands out as the only segment-level dataset using variable-length (3–16 second) segments. On average, the length of a segment is 6.58 ± 2.28 seconds. In contrast, other segment-level datasets use fixed-length segments and usually do not con-

[3] http://labrosa.ee.columbia.edu/millionsong/

[4] http://majorminer.org/

[5] http://tagatune.org/Magnatagatune.html

[6] Tag labels elicited from social websites or game with a purpose, called "weak labels," could be fairly noisy and sparse and in particular have enormous false negative labels [33]. In contrast, "strong labels" indicate that each tag is carefully verified for each song.

sider whether the segments are acoustically homogeneous or representative of the corresponding track. Moreover, CAL500exp is characterized by its "backward compatibility" with CAL500 and therefore inherits the expert-defined taxonomy. Accordingly, researchers can use the original audio sources of CAL500 and the label information of CAL500exp in their study. Although the quantity of CAL500exp is relatively smaller than datasets such as Magnatagatune [10], CAL10k [14] and the million song dataset (MSD) [28], it offers unique opportunities to study music auto-tagging in shorter temporal scale.

We also note that the PWT system [22], which is a direct application of time-varying music auto-tagging, requires a real-time auto-tagger that makes the short-time tag prediction with a "sliding chunk" (in the segment-level) and displays the predicted results in sync with music playback. One can expect better performance by training a PWT system on the segment-level tag labels of CAL500exp.

3. CAL500 EXPANSION

3.1. Data Preprocessing

Some minor problems of CAL500 have been identified and addressed by Sturm [34]. We follow his guidelines and assume that the song order of annotations in the annotation text files complies with what indicated in the text file of the song names. Then, we select 500 out of 502 songs that both sound files and tag annotations are available.[7] Finally, we replace the sound file "jade_leary-going_in.mp3", which was originally overly short (313 bytes), with the one obtained from [34]. Before content analysis, we downsample each sound file to 22,050 Hz and merge stereo to mono, a common practice in MIR [4].

To obtain acoustically homogenous segments, we adopt Foote & Cooper's segmentation algorithm [35] implemented by the MIRToolbox [36] to process every track in CAL500. The idea is to first detect the changes in spectrum on the self-similarity matrix of a track and then find local peaks from the resultant novelty curve as the segment boundaries. After segmentation, there are in total 18,664 segments, with each track being partitioned to 37.3 segments on average.

Because many segments of a song could be similar, it is time-consuming and perhaps redundant to annotate every segment. Therefore, we perform k-medoids clustering [37] on the segments of each song. A 140-dimensional acoustic feature vector (cf. Section 4.1) is used to represent each segment. The medoid of each cluster is selected as a representative segment to annotate. The cluster number k (ranging from 1 to 8) is set in proportion to the number of segments of a track. To ensure the quality and diversity of the k-medoids result, we repeat the algorithm 20 times (with random initialization) and select the result with the smallest cumulative distance between a segment and its medoid. Eventually, we obtain on average 6.4 representative segments per track.

During playback, we hope that subjects can annotate tag labels according to the middle part of a segment. Thus, we emphasize the middle part by integrating a volume weight vector $\mathbf{v}$ (with length t) based on a Hamming window $\mathbf{w}$ (with length $t/2$) to fade-in and fade-out the segment, where $\mathbf{v} = [\text{left part of } \mathbf{w},\ \mathbb{1}(t/2),\ \text{right part of } \mathbf{w}]$, where $\mathbb{1}(n)$ is the n-dimensional vector with all ones.

3.2. Taxonomy for Time-varying Music Tags

To determine the tag set of CAL500exp, we remove some contrary tags that begin with 'NOT' in CAL500, because each 'NOT' tag has its positive counterpart. For example, we discard 'NOT-Emotion-Angry/Agressive' as it can be represented by a negative label of 'Emotion-Angry/Agressive.' This reduces the total number of unique tags to 144.

To prevent chaos, we show one category of tags to subjects at a time. Moreover, we observe in our pilot study that the tag labels of some categories are not time-varying and almost identically annotated among all the segments of a track. In consequence, we define two types of tag categories, namely *time-varying tags* (i.e., Instrument, Instrument-Solo, Vocal, and Emotion) for the segment-level annotations and *time-invariant tags* (i.e., Genre, Genre-Best, Song, and Usage) in the track-level scale. In this paper, we focus on the 67 time-varying tags to be annotated in the segment level.

3.3. Tag Label Initialization

To alleviate the annotation labor, we provide initialized tag labels as default for each segment and ask the subjects to modify the default labels by insertion (adding tags) and deletion (removing tags). From the pilot study, we also find that removing tags is easier than adding tags. Therefore, the following two strategies are considered to generate the default tag labels. First, we generate the tag labels for each segment of a track as long as an annotator of CAL500 has applied the tag to that track, instead of using the "hard" label obtained by majority voting [9]. Second, we "re-tag" each segment by using audio-based auto-taggers trained on the track-level tag labels of CAL500. Specifically, we train auto-taggers using all the segments with the tag labels of their originated tracks and individually re-tag each segment with binary outputs. Finally, the default tag labels are derived by unifying the results obtained from the two strategies. Obviously, our strategies lead to many false positive labels (especially for instrumentation and vocal tags) comparing to the possible ground-truth that subjects are going to give. For example, 'Electric Guitar Solo' may be initially assigned to every segment according to CAL500 but may not appear in all segments in reality. We expect that the subjects recruited for annotating CAL500exp can identify and remove such false positive labels in most cases.

[7] The 500 selected songs can be found in the website of CAL500exp.

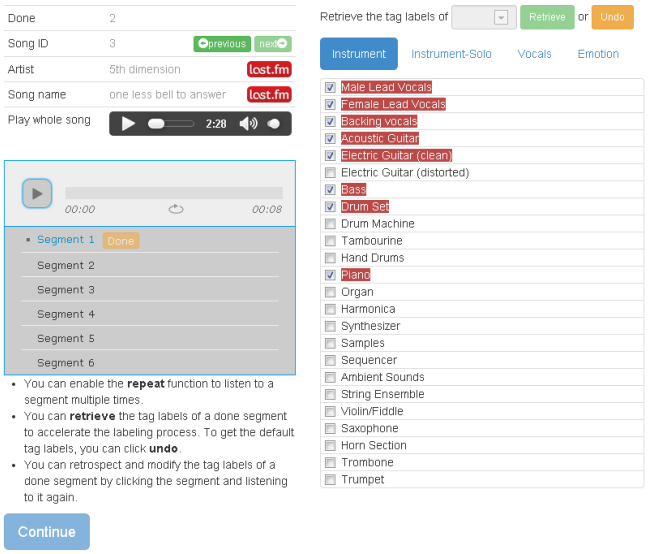

Fig. 1. A snapshot of the user interface we develop for segment-level tag annotation. The annotators are requested to annotate the tags category-by-category, by refining the default set of tags generated by tag label initialization.

3.4. User Interface

Figure 1 shows the designed user interface. The left hand side of the interface shows, from top to bottom, the information of the track,[8] the whole track preview player, segment-level music player, the list of segments of the track, and annotation instructions. On the other hand, the right hand side shows the candidate tags grouped by categories (organized by using tabs) where the initialized tag labels (cf. Section 3.3) are checked and highlighted initially.

The interface employs a two-stage process for annotating each track. A subject has to first listen to and annotate all the segments of a track with time-varying tags in the segment-level before proceeding to annotate the time-invariant tags of the track (in the track-level). Accordingly, with the time-varying tags in mind, it may be easier for the subject to annotate the time-invariant tags.

According to the pilot study, a subject usually has to listen to a segment several times when verifying its time-varying tag labels. Hence, we provide a 'repeat' function (shown in the middle of the left hand side, under the play bar) in the music player. In addition, we have also found that the tag labels of some representative segments of a track might be still similar. We then include a 'copy' function (shown in the upper-right corner of the interface), so that for a new segment, subjects can copy the tag labels of a previously done segment and make modification upon them.

Once subjects have done a segment, the segment block will become green. Then, they can retrospect and modify the tag labels by clicking on segments with green block. They can also modify the tag labels of a previously done track with the 'previous' and 'next' buttons beside 'Song ID.'

The interface is web-based and built by WampServer, which allows web applications created with Apache2, PHP, and MySQL database under Microsoft Windows environment. On the client side, we utilize jPlayer to play the audio contents, and Bootstrap 3 as the front-end framework.

[8] We also provide Last.fm links for more detailed information of the artist and track, such as social tags, high quality audio sources, and user comments.

Table 2. The statistics of average tag insertions (ins), deletion (del), and operation (opr), and the numbers (num) of annotated segments among different subjects (sbj).

sbj	ins	del	opr	num
1	1.15	2.75	3.90	316
2	2.97	5.66	8.63	652
3	2.32	5.99	8.31	1278
4	2.02	7.30	9.32	656
5	2.22	6.21	8.43	2612
6	2.61	8.55	11.16	1615
7	1.91	7.92	9.83	979
8	2.93	6.40	9.33	626
9	3.05	8.54	11.6	982
10	1.84	5.69	7.52	647
11	3.24	5.63	8.86	642
avg	2.42	6.77	9.18	—

3.5. Analysis of Subjects' Annotating Behaviors

Table 2 reports some information of the subjects' annotating behaviors. We recruited and paid eleven subjects with strong musical background, including professional musicians (user IDs: 1, 2, 4, 5, 6, 9, and 10), studio engineers (IDs 1, 5, 9, and 10), MIR researchers (IDs 3 and 7), amateur musicians (IDs 3, 2, 8, and 11) and students graduated from music degree programs (IDs 6 and 8). All subjects can determine the number of tracks they like to label. Each subject was rewarded 1.2 USD per track and not allowed to label a certain track twice.

The annotation process lasted about three weeks. Each segment and track have been completely annotated by at least three subjects. Following the method of CAL500, we perform majority voting to determine the binary ground-truth labels for both time-varying and time-invariant tags.

In Table 2, one can see the average numbers of insertion, deletion, and operation (the sum of insertion and deletion) made by the eleven subjects for the time-varying tags. Two observations can be made. First, the average operation rate is not small (9.2/67=13.7%), suggesting that the subjects might have taken this annotation job seriously, rather than just using the default tag labels. Second, the number of deletion is generally much larger than that of insertion. This is expected, as the tag label initialization methods (cf. Section 3.3) would generate many false positive labels in the default set.

4. EXPERIMENT

This section presents empirical evaluations on time-varying music auto-tagging. The purpose of this study is to verify whether the subjects' operations lead to better consistency in response to the audio content, and to demonstrate the performance improvement brought about by CAL500exp.

4.1. Experiment Setup

For a frame-based feature vector, a hybrid set of frame-level energy, timbre and harmonic descriptors were computed by using the MIRToolbox [36], with a frame size of 50 ms and half overlap. The features include root-mean-square energy, zero-crossing rate, spectral flux, spectral moments, MFCCs, chroma vector, key clarity, musical mode, and harmonic detection. The segment-level feature vector is represented by concatenating the weighted mean and standard deviation (STD) of the frame-based feature vectors using $\mathbf{v}$ as the weights, forming a 140-dimensional vector. Finally, each feature dimension is normalized to zero mean and unit standard deviation throughout all the segments of the dataset.

For classification, we adopt the standard binary relevance multi-label classification scheme [9] and train each tag classifier with the linear-kernel SVM implemented by LIBLINEAR [38]. While predicting the tags of a segment, each tag classifier outputs a probability for a tag. As for binary output, we annotate the tags of a segment as positive if their probabilities are greater than the threshold determined by an inner (training set) cross-validation (denoted as 'CV'). The fold splitting is performed in the track level.

We conduct both *intra-dataset* and *inter-dataset* evaluations using CAL500 and CAL500exp. The intra-dataset case, denoted by '$\mathcal{D}$`(CV)`,' uses standard five-fold CV on one of the datasets (i.e., $\mathcal{D}$ can be CAL500 or CAL500exp). For the inter-dataset evaluation, denoted by '$\mathcal{D}_1 \rightarrow \mathcal{D}_2$,' we note that the two datasets share the same audio sources and features, and thus we perform the training and tag prediction in the scenario of five-fold CV using $\mathcal{D}_1$, but then evaluate the test accuracy using the ground-truth labels of the corresponding fold from $\mathcal{D}_2$. For instance, `CAL500`$\rightarrow$`CAL500exp` stands for training on CAL500 and then evaluating based on the labels of CAL500exp. Note that, for CAL500 the ground-truth label of a segment is obtained from that of its originated track.

To evaluate the performance of time-varying music auto-tagging (e.g., in the scenario of automatic music tag tracking applications [22]), we can treat the segments in the test fold as the representative segments sampled by a sliding chunk from the test tracks. The performance of the binary outputs is measured in terms of *per-tag* precision, recall, and F-score (the harmonic mean of precision and recall) [9]. As for the performance of the probabilistic outputs, we report the *per-segment* AUC (the area under the ROC curve) to outline how accurate the predicted tag distribution is.

Table 3. (a) presents the results of Instrument, Instrument-Solo and Vocal tags, and (b) shows the result of Emotion tags. We use P, R, and F to denote per-tag precision, recall, and F-score, respectively.

(a) instrument & vocal	P	R	F	AUC
`CAL500(CV)`	0.157	0.371	0.213	0.833
`CAL500exp(CV)`	0.257	0.456	0.317	0.884
`CAL500`$\rightarrow$`CAL500exp`	0.226	0.384	0.267	0.842
`CAL500exp`$\rightarrow$`CAL500`	0.172	0.445	0.225	0.844
(b) emotion	P	R	F	AUC
`CAL500(CV)`	0.301	0.701	0.417	0.735
`CAL500exp(CV)`	0.455	0.759	0.561	0.842
`CAL500`$\rightarrow$`CAL500exp`	0.443	0.751	0.543	0.802
`CAL500exp`$\rightarrow$`CAL500`	0.303	0.747	0.422	0.744

4.2. Result and Discussion

The result is shown in Table 3, which divides the time-varying tag set into two groups: (a) Instrument, Instrument-Solo, and Vocal tags, and (b) Emotion tags. We make the following observations. First, by comparing the results of `CAL500(CV)` and `CAL500exp(CV)`, we see that CAL500exp leads to better performance for all performance measures and tag groups, showing that the connection between audio and tag for CAL500exp is relatively easier to model. This may also suggest better tag label consistency among different segments of CAL500exp. Second, considering the case of fixing the test set to CAL500 and using either CAL500 or CAL500exp for training, we see that `CAL500exp`$\rightarrow$`CAL500` consistently outperforms `CAL500(CV)` in most cases (e.g., see the first and fourth rows of Table 3). This implies that the tag labels of CAL500exp are more accurate, so thay can even achieve better performance when using lower-quality labels of CAL500 for testing. Third, we find that `CAL500exp(CV)` yields better performance than `CAL500`$\rightarrow$`CAL500exp` (second and third rows). The differences in F-score and AUC are significant, showing that we can get more accurate auto-taggers for time-varying auto-tagging by using CAL500exp instead of its predecessor CAL500 for training. Such result also validates the motivation of this paper. Finally, the performance difference between `CAL500exp(CV)` and `CAL500`$\rightarrow$`CAL500exp` is larger for the instrument & vocal tags than for the emotion tags. This is reasonable due to the factor that instrument & vocal tags are less subjective so the improvement of CAL500exp can be easily reflected.

5. CONCLUSION

In this paper, we have presented a new publicly available dataset, called CAL500exp, to facilitate music auto-tagging in a smaller temporal scale, which holds the promise of enabling applications such as play-with-tagging. The dataset has been constructed by taking many issues into considera-

tion so as to improve its usefulness for the research community. For instance, music segmentation is used to make the connection between tags and music better-defined; a new annotation user interface, representative segment selection, and music re-tagging are performed to reduce user burden and improve annotation quality. We have also presented a comprehensive performance study that demonstrates the advantage of the new dataset for time-varying auto-tagging. We hope that the dataset can call for more research towards understanding the temporal context of musical semantics.

6. REFERENCES

[1] A. L.-C. Wang, "An industrial-strength audio search algorithm," in *ISMIR*, 2003.

[2] C. Bandera et al., "Humming method for content-based music information retrieval," in *ISMIR*, 2011.

[3] B. Whitman and R. Rifkin, "Musical query-by-description as a multiclass learning problem," in *IEEE MMSP*, 2002.

[4] M. A. Casey et al., "Content-based music information retrieval: Current directions and future challenges," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 668–696, 2008.

[5] Y.-H. Yang and H.-H. Chen, "Machine recognition of music emotion: A review," *ACM Trans. Intelligent Systems and Technology*, vol. 3, no. 4, 2012.

[6] M. I. Mandel and D. P. W. Ellis, "A web-based game for collecting music metadata," *JNMR*, vol. 37, pp. 151–165.

[7] M. I. Mandel and D. P. W. Ellis, "Multiple-instance learning for music information retrieval," in *ISMIR*, 2008, pp. 577–582.

[8] P. Lamere, "Social tagging and music information retrieval," *JNMR*, vol. 37, no. 2, pp. 101–114, 2008.

[9] D. Turnbull et al., "Semantic annotation and retrieval of music and sound effects," *TASLP*, vol. 16, no. 2, pp. 467–476, 2008.

[10] E. Law and L. von Ahn, "Input-agreement: A new mechanism for collecting data using human computation games," in *Proc. ACM CHI*, 2009, pp. 1197–1206.

[11] S.R. Ness, A. Theocharis, G. Tzanetakis, and L.G. Martins, "Improving automatic music tag annotation using stacked generalization of probabilistic svm outputs," in *ACM MM*, 2009.

[12] Y.-H. Yang, Y.-C. Lin, A. Lee, and H.-H. Chen, "Improving musical concept detection by ordinal regression and context fusion," in *ISMIR*, 2009, pp. 147–152.

[13] T. Bertin-Mahieux et al., "Automatic tagging of audio: The state-of-the-art," in *Machine Audition: Principles, Algorithms and Systems*, Wenwu Wang, Ed. IGI Global, 2010.

[14] D. Tingle, Y. E. Kim, and D. Turnbull, "Exploring automatic music annotation with acoustically objective tags," in *Proc. ACM MIR*, 2010, pp. 55–62.

[15] M. I. Mandel et al., "Contextual tag inference," *ACM Trans. Multimedia Computing, Communications & Applications*, vol. 7S, no. 1, pp. 1547–1556, 2011.

[16] J.-C. Wang et al., "Query by multi-tags with multi-level preferences for content-based music retrieval," in *IEEE ICME*, 2011.

[17] J.-C. Wang et al., "Colorizing tags in tag cloud: A novel query-by-tag music search system," in *ACM MM*, 2011, pp. 293–302.

[18] G. Marques et al., "Three current issues in music autotagging," in *ISMIR*, 2011, pp. 795–800.

[19] H.-Y. Lo et al., "Cost-sensitive multi-label learning for audio tag annotation and retrieval," *IEEE TMM*, vol. 13, no. 3, pp. 518–529, 2011.

[20] E. Coviello, A. B. Chan, and G. R. G. Lanckriet, "Time series models for semantic music annotation," *IEEE TASLP*, vol. 19, no. 5, pp. 1343–1359, 2011.

[21] J. Nam, J. Herrera, M. Slaney, and J. O. Smith, "Learning sparse feature representations for music annotation and retrieval," in *ISMIR*, 2012, pp. 565–560.

[22] J.-C. Wang, H.-M. Wang, and S.-K. Jeng, "Playing with tagging: A real-time tagging music player," in *ICASSP*, 2012.

[23] C.-C. M. Yeh, J.-C. Wang, Y.-H. Yang, and H.-M. Wang, "Improving music auto-tagging by intra-song instance bagging," in *ICASSP*, 2014.

[24] D. Turnbull et al., "Five approaches to collecting tags for music," in *ISMIR*, 2008, pp. 15–20.

[25] E. Schubert, "Modeling perceived emotion with continuous musical features," *Music Perception*, vol. 21, no. 4, pp. 561–585, 2004.

[26] E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions from audio," in *ISMIR*, 2010.

[27] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Towards musical query-by-semantic-description using the CAL500 data set," in *ACM SIGIR*, 2007, pp. 439–446.

[28] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *ISMIR*, 2011.

[29] W. Mason and S. Suri, "Conducting behavioral research on Amazon's Mechanical Turk," *Behavior Research Methods*, vol. 44, no. 1, pp. 1–23, 2012.

[30] J.-C. Wang et al., "The acousticvisual emotion Gaussians model for automatic generation of music video," in *Proc. ACM MM*, 2012, pp. 1379–1380.

[31] C. Liem, A. Bazzica, and A. Hanjalic, "MuseSync: standing on the shoulders of Hollywood," in *ACM MM*, 2012.

[32] M. E. P. Davies et al., "AutoMashUpper: An automatic multi-song mashup system," in *ISMIR*, 2013.

[33] Y.-Y. Sun, Y. Zhang, and Z.-H. Zhou, "Multi-label learning with weak label," in *Prof. AAAI*, 2010.

[34] Bob L. Sturm, "Using the CAL500 dataset?," [Online] `http://media.aau.dk/null_space_pursuits/2013/03/using-the-cal500-dataset.html`.

[35] J. T. Foote and M. L. Cooper, "Media segmentation using self-similarity decomposition," in *Proc. SPIE*, 2003, pp. 167–175.

[36] O. Lartillot and P. Toiviainen, "A Matlab toolbox for musical feature extraction from audio," in *Prof. DAFx*, 2007.

[37] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3336–3341, 2009.

[38] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.