INTRODUCING THE CAL500EXP DATASET: TOWARDS TIME-VARYING MUSIC AUTO-TAGGING

Ju-Chiang Wang[1,2], **Shuo-Yang Wang**[1], **Yi-Hsuan Yang**[1], **and Hsin-Min Wang**[1]
[1] Academia Sinica, Taipei, Taiwan
[2] University of California, San Diego, La Jolla, CA, USA
asriver.wang@gmail.com, {raywang, yang, whm}@iis.sinica.edu.tw

1. INTRODUCTION

With tremendous growth of digital music libraries online, a large number of text-based music information retrieval (MIR) methods have been proposed in the literature [1, 2, 4, 6, 8, 11, 13, 17–19, 22]. These methods hold the promise of helping users search for music in a content-based way through a few keywords related to high-level music semantics or metadata such as artist name, song title, genre, style, mood, and instrument [5]. The task of automatically annotating musical items (e.g., artists, albums, or tracks) with high-level musical semantics is usually referred to as music *auto-tagging*.

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 [1,6,17]. That is, they usually collect the ground-truth associations between tag and music in the track level [15], 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* [12]. 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) [19].

The prevalence of the track-level approach might be partly due to the difficulty of collecting tag labels at smaller temporal scales. 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, 16], 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 [9, 10]. For each track, they sampled five fixed-length (10-second) segments evenly spaced throughout the track. Then, an online crowdsourcing platform, Amazon Mechanical Turk, [1] was adopted to collect the tags for each segment. It is found that different parts of the same track tend to be described differently by the human listeners. However, obtaining a short music 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 (i.e., tags) and music would be better defined, leading to new opportunities to bridge the so-called semantic gap.

To this end, the 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)* [19] (i.e., visualizing music signals by tracking the tag distribution during playback), *automatic music video generation* [7, 20] (i.e., matching between the music and video signals in a more fine-grained temporal scale), and *audio remixing* [3] (i.e., jumping from a fragment of a track to a fragment of another track).

In light of above discussion, 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 [17]. [2] To provide more accurate and consistent labels of music content in a finer granularity, a novel protocol with three new elements tailored for constructing a time-varying music auto-tagging dataset is proposed.

- 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

[1] https://www.mturk.com/
[2] http://cosmal.ucsd.edu/cal/projects/AnnRet/

segments for annotation.

- 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.

- Instead of using crowdsourcing, we recruit subjects with strong music background and devise a new user-interface for better annotation quality.

Furthermore, we have also presented a comparative study that validates the performance gain brought about by the CAL500exp dataset over its predecessor CAL500 for time-varying music auto-tagging. For more details, we refer readers to [21].

To call for more attention to time-varying auto-tagging, we have made CAL500exp available upon request to the research community. [3] We believe that CAL500exp may open new opportunities to understand and to model the temporal context of musical semantics.

2. ACKNOWLEDGEMENTS

This work was supported by Academia Sinica–UCSD Postdoctoral Fellowship to Ju-Chiang Wang, and the Ministry of Science and Technology of Taiwan under Grants NSC 101-2221-E-001-019-MY3 and 102-2221-E-001-004-MY3.

3. REFERENCES

[1] T. Bertin-Mahieux, D. Eck, and M. Mandel. Automatic tagging of audio: The state-of-the-art. In Wenwu Wang, editor, *Machine Audition: Principles, Algorithms and Systems*. IGI Global, 2010.

[2] E. Coviello, A. B. Chan, and G. R. G. Lanckriet. Time series models for semantic music annotation. *IEEE TASLP*, 19(5):1343–1359, 2011.

[3] M. E. P. Davies, P. Hamel, K. Yoshii, and M. Goto. AutoMashUpper: An automatic multi-song mashup system. In *Proc. ISMIR*, 2013.

[4] D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green. Automatic generation of social tags for music recommendation. In *Advances in neural information processing systems*, pages 385–392, 2008.

[5] P. Grosche, M. Müller, and J. Serrà. Audio content-based music retrieval. *Multimodal Music Processing*, 3:157–174, 2012.

[6] P. Lamere. Social tagging and music information retrieval. *JNMR*, 37(2):101–114, 2008.

[7] C. Liem, A. Bazzica, and A. Hanjalic. MuseSync: standing on the shoulders of Hollywood. In *Proc. ACM MM*, 2012.

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

[8] H.-Y. Lo, J.-C. Wang, H.-M. Wang, and S.-D. Lin. Cost-sensitive multi-label learning for audio tag annotation and retrieval. *IEEE TMM*, 13(3):518–529, 2011.

[9] M. I. Mandel and D. P. W. Ellis. Multiple-instance learning for music information retrieval. In *Proc. ISMIR*, pages 577–582, 2008.

[10] M. I. Mandel, R. Pascanu, D. Eck, Y. Bengio, L. M. Aiello, R. Schifanella, and F. Menczer. Contextual tag inference. *ACM Trans. Multimedia Computing, Communications & Applications*, 7S(1):1547–1556, 2011.

[11] G. Marques, M. A. Domingues, T. Langlois, and F. Gouyon. Three current issues in music autotagging. In *Proc. ISMIR*, pages 795–800, 2011.

[12] E. Schubert. Modeling perceived emotion with continuous musical features. *Music Perception*, 21(4):561–585, 2004.

[13] M. Sordo, C. Laurier, and O. Celma. Annotating music collections: how content-based similarity helps to propagate labels. In *Proc. ISMIR*, 2007.

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

[15] D. Turnbull, L. Barrington, and G. Lanckriet. Five approaches to collecting tags for music. In *Proc. ISMIR*, pages 15–20, 2008.

[16] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Towards musical query-by-semantic-description using the CAL500 data set. In *Proc. ACM SIGIR*, pages 439–446, 2007.

[17] D. Turnbull, L. Barrington, D. Torres, and G. R. G. Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE TASLP*, 16(2):467–476, 2008.

[18] J.-C. Wang, Y.-C. Shih, M.-S. Wu, H.-M. Wang, and S.-K. Jeng. Colorizing tags in tag cloud: A novel query-by-tag music search system. In *Proc. ACM MM*, pages 293–302, 2011.

[19] J.-C. Wang, H.-M. Wang, and S.-K. Jeng. Playing with tagging: A real-time tagging music player. In *Proc. IEEE ICASSP*, pages 77–80, 2012.

[20] J.-C. Wang, Y.-H. Yang, I.-H. Jhuo, Y.-Y. Lin, and H.-M. Wang. The acousticvisual emotion Gaussians model for automatic generation of music video. In *Proc. ACM MM*, pages 1379–1380, 2012.

[21] S.-Y. Wang, J.-C. Wang, Y.-H. Yang, and H.-M. Wang. Towards time-varying music auto-tagging based on CAL500 expansion. In *Proc. IEEE ICME*, 2014.

[22] C.-C. M. Yeh, J.-C. Wang, Y.-H. Yang, and H.-M. Wang. Improving music auto-tagging by intra-song instance bagging. In *Proc. IEEE ICASSP*, pages 2139 – 2143, 2014.