BUILDING K-POP SINGING VOICE TAG DATASET: A PROGRESS REPORT

KeunHyoung Luke Kim, Sangeun Kum, Chae Lin Park, Jongpil Lee, Jiyoung Park, Juhan Nam
Graduate School of Culture Technology, KAIST
{doiluvu,lynn08,keums,richter,jypark527,juhannam}@kaist.ac.kr

ABSTRACT

We present a progress report on building a new music auto-tagging dataset that focuses on singing voice tags and K-pop music. The dataset uses 70 tag words that describe the quality of singing voice in detail. The tags are annotated to a set of K-pop music in both song-level and segment-level. In this report, we show a brief analysis on song-level tagging results.

1. INTRODUCTION

In popular music, singing voice is the central sound source that determines the song quality, as it conveys melody, lyrics, emotion and humanity with its high expressivity. Therefore, it will be useful for music search or recommendation if we have detailed annotation of singing voice descriptions to songs. In this context, the majority of datasets so far have handled general song quality such as genre, mood and instruments [2, 3, 5]. In addition, the songs with the labels are mostly western pop music. In this report, we focus on collecting tag words that describe timbre and expressions of singing voice in detail. Also, we target to annotate contemporary Korean pop music, often called *K-pop*, using the tags.

2. METHOD

We used the K-pop Vocal Analysis website [1] as a main source to collect singer and tag data. It contains expert-level analysis of K-pop singers focusing on their timbre and singing styles. From the website, we mined about 300 tag words that describe voice quality. After filtering out inappropriate words, we collected 70 voice quality description words and also a list of 114 singers. We obtained five audio files per singer and filtered out songs with duet, chorus singers or rap. As a result, we collected 469 songs. Five human annotators participated in song-level tagging with regard to the 70 singing voice tag words. Each song was tagged by three different annotators.

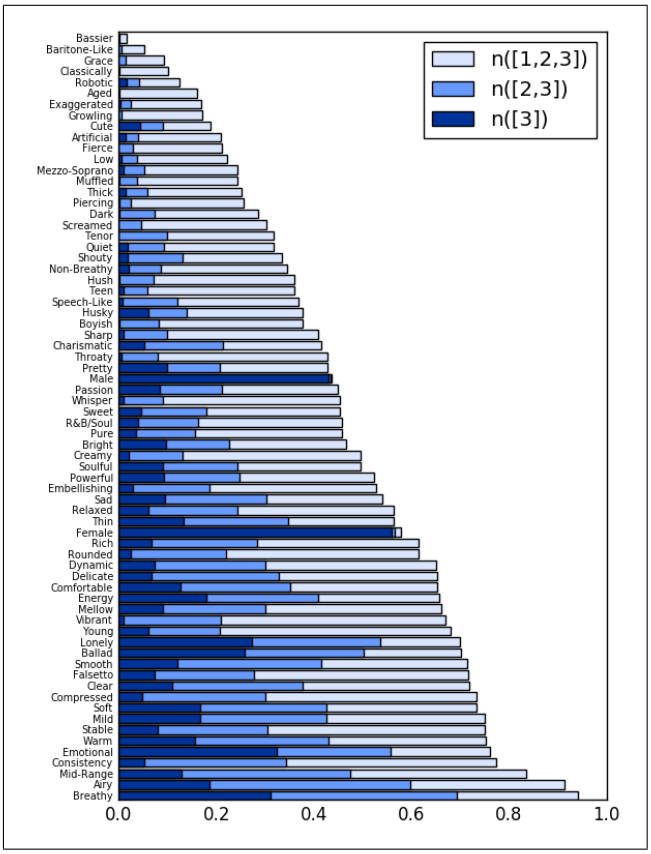

Figure 1: Frequency and coincidence of singing voice tag annotation. n([i,j,k]) is the number of songs that was tagged i, j or k times for each tag (Note that each song is tagged by three different annotators for all 70 tags).

Considering that the voice quality changes over different sections of a music piece (e.g. verse, chorus) [6], we also annotate songs in a segment-level. Using a singing voice detector, we trimmed audio files into 10-sec long segments with voice, thereby obtaining 6787 examples. Currently, we are collecting annotation data for the segments using a web-based interface. For this segment-level annotation, we are using 42 tags after analyzing redundancy from the song-level tagging data.

3. ANALYSIS

For the song-level tagging data, we calculated the frequency and coincidence of singing voice tags. We counted the number of songs annotated with each tag, distinguishing three possible coincidence. The frequency and coin-

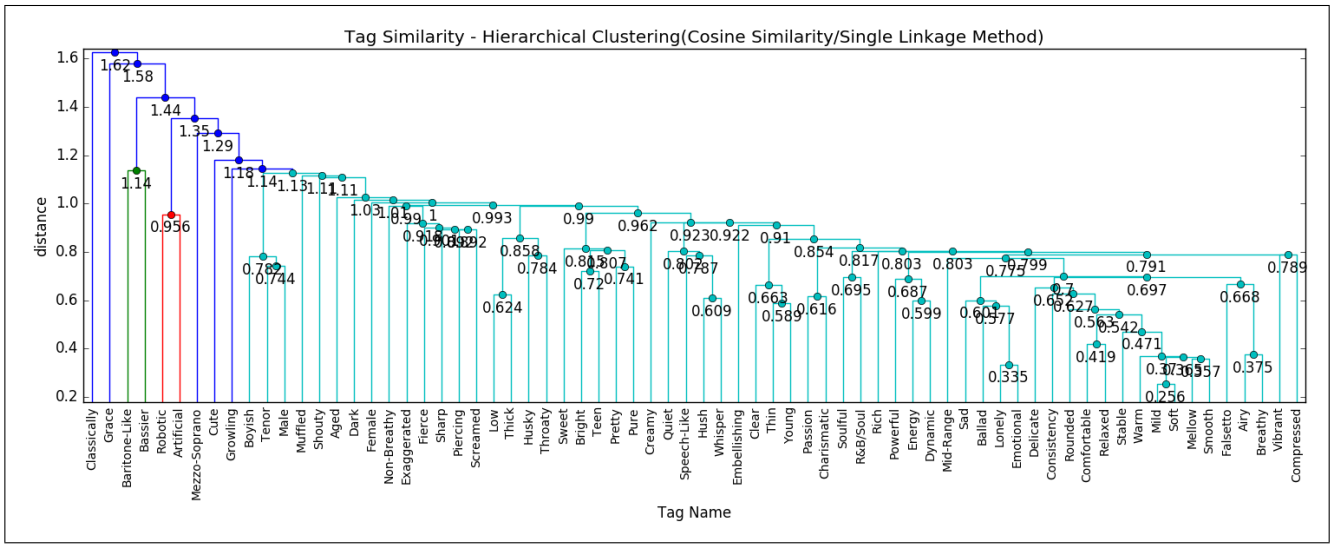

Figure 2: Hierarchical Clustering by Tag Similarity

cidence are displayed in Figure 1. It shows that tags such as *Bassier*, *Baritone-like* and *Grace* are annotated once in less than 10% of the songs. Tags such as *Male* and *Female* have almost 100% coincidence as they are easy to recognize. On the other hand, *Throaty* and *Vibrant* have very small coincidence as they might be difficult or ambiguous to identify.

We also display a hierarchical clustering result using similarity between tags. The similarity matrix was computed by multiplying from the tag by song matrix and its transpose matrix. The tag-to-tag distance was computed using cosine distance. The result is plotted in Figure 2. It shows that tag pairs such as *Mild-Soft* and *Airy-Breathy* are very close to each other. These similar or redundant words were merged into a single word for segment-level annotation.

4. CONCLUSIONS

The K-pop singing voice tag dataset provides both song-level and segment-level tags for detailed description of singing voice. We reported our ongoing work and a brief analysis of song-level tagging data. After collecting segment-level tag annnotation data, we plan to release the dataset on public domain. In addition, we plan to train a classifier so that it can automatically predict the detailed singing voice tags. For example, we can apply our recently proposed music auto-tagging model to the K-pop singing voice tag dataset [4]. Finally, we will utilize the auto tagging results for content-based music search and recommendation.

5. REFERENCES

[1] "K-pop vocalists' vocal analyses", web resource, available:. `http://kpopvocalanalysis.net`. Accessed: 2017-10.

[2] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2011.

[3] Edith Law, Kris West, Michael Mandel, Mert Bay, and J. Stephen Downie. Evaluation of algorithms using games : The case of music tagging. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2009.

[4] Jongpil Lee and Juhan Nam. Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging. *IEEE Signal Processing Letters*, 24(8):1208–1212, 2017.

[5] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. Towards musical query-by-semantic-description using the cal500 data set. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 439–446. ACM, 2007.

[6] Shuo-Yang Wang, Ju-Chiang Wang, Yi-Hsuan Yang, and Hsin-Min Wang. Towards time-varying music auto-tagging based on cal500 expansion. In *2013 IEEE International Conference on Multimedia and Expo (ICME)*, 2013.