Semantic Tagging of Singing Voices in Popular Music Recordings

Keunhyoung Luke Kim, Jongpil Lee, Sangeun Kum, Chae Lin Park, Juhan Nam, *Member, IEEE*

***Abstract*—Singing voice is a key sound source in popular music. As recent music streaming and entertainment services call for more intelligent solutions to retrieve songs or evaluate musical characteristics, automatic analysis of popular music targeted to singing voice has been a significant research subject. The majority of studies have focused on quantitative or objective information of singing voice such as pitch, lyrics or singer identity. However, singing voice has a wide variety of dimensions that are somewhat difficult to quantify and therefore we often describe by words. In this paper, we address the qualitative analysis of singing voice as a music auto-tagging task that annotates songs with a set of tag words. To this end, we build a music tag dataset dedicated to singing voice. Specifically, we define a vocabulary that describes timbre and singing styles of K-pop vocalists and collect human annotations for individual tracks. We then conduct statistical analysis to understand the global and temporal characteristics of the tag words. Using the dataset, we train a deep neural network model to automatically predict the voice-specific tags from popular music recordings and evaluate the model in different conditions. We discuss the results by comparing them to the statistical analysis of tag words. Finally, we show potential applications of the vocal tagging system in music retrieval, music thumbnailing and singing evaluation.**

***Index Terms*—singing voice, vocal, semantic analysis, music tagging, convolutional neural networks, timbre, K-Pop**

I. Introduction

Singing voice is one of the most essential sound sources in music. It can deliver not only melody but also lyrics with great emotional expressions. In popular music, the role of singing voice is more important as the vocal quality of singers is critical in attracting people and gaining popularity. Therefore, a song is written often considering the vocal characteristics of singers. In music production, vocal tracks are usually the main focus in mixing them with accompanying instrumental tracks. This importance has led to active research on computational analysis of singing voice [1]. In particular, singing voice analysis in popular music recordings has drawn much attention as music streaming services have grown and online karaoke services (e.g., mobile apps) have become widespread [2].

Singing voice analysis has been carried out to acquire diverse types of information in music and utilize it for various applications. Since the main vocal in popular music usually accounts for melody and lyrics, extracting the pitch and formant features of singing voice from the mixture track can allow for obtaining the melody contours [3]–[6] or transcribing the lyrics of the song [7], [8]. Furthermore, the melodic or phonetic features can be used for other related tasks such as query-by-humming [9], singing skill evaluation [10], and audio-to-lyrics alignment [11]. Since singing voice contains information about the singer as well, the acoustic features can be used for singer-oriented applications such as singer identification [12] or music retrieval based on vocal timbre similarity [13], [14].

While extracting the quantitative or objective information such as melody, lyrics, and singer identity has been actively studied so far, qualitative elements of singing voice such as vocal timbre or singing styles have been paid less attention. This might attribute to the difficulty of defining appropriate measures in the qualitative analysis. One approach to tackle the problem is using a set of words that describe the impression of singing voice and rate the matching between the individual words and a singing voice [15].

As for popular music, this word-based music analysis has been handled in the context of music auto-tagging, which has been a widely explored topic in the area of music information retrieval (MIR) [16]–[18]. However, the majority of tag vocabulary in existing music tag datasets account for general song characteristics such as genre or mood, not sufficiently covering attributes on singing voice. This poses a new setup for qualitative analysis of singing voice in popular music recordings, that is, music auto-tagging that focuses on vocal-specific tags.

In this regard, this paper presents a comprehensive study on semantic tagging of singing voice. We first introduce a new music tag dataset dedicated to singing voice that contains up to 70 vocal tags and human annotations for 466 K-pop music tracks. The human annotations include not only track-level but also segment-level, considering that timbre and expressions of singing voice can dynamically change over different sections within a track (e.g., verse or chorus) [19], [20]. For the segment-level annotation, we searched vocal segments in the music tracks using a voice detection algorithm and obtained 6,787 10-second-long vocal segments. In addition, we assigned three human annotators to each segment, considering the subjective nature of tag words. We term this as the K-pop Vocal Tag (KVT) dataset. Using the KVT dataset, we conducted two primary studies. One is statistical analysis of the human annotations in terms of frequency, agreement, temporal dynamics and correlation of the vocal tags. This will provide an understanding of characteristics of the vocal tag words. The other is auto-tagging of singing voice by training a deep neural network in a supervised setting using the KVT dataset. We train and evaluate the model with both track-level and segment-level annotations, and also cross-test it with the two different levels of annotations to investigate the effect

The authors are with the Graduate School of Culture Technology, Korea Advanced Institute of Science and Technology, Daejeon, 34141, South Korea (e-mail:{dilu,richter,keums,lynn08,juhannam}@kaist.ac.kr).

of segment-level annotations. We also attempt to use isolated vocal tracks as input to compare that against the original mixed tracks as input. Furthermore, we compare the prediction results to the statistical analysis of human annotations to interpret the prediction results. Finally, we discuss potential applications of the qualitative singing analysis including vocal-oriented music retrieval, highlight detection and singing evaluation. To the best of our knowledge, this is the first research on music auto-tagging focusing on singing voice in popular music recordings.

This paper is organized as follows. Section II reviews previous work related to characterizing singing voice using words. Section III describes the detail of constructing the vocal tag dataset. Section IV analyzes the human annotations using various statistical measures. Section V handles automatic music tagging using a convolutional neural network and the dataset. Section VI suggests potential applications of this research. The conclusion section wraps up the discussion and outlines future work.

II. Related Work

Singing voice can be explained in terms of pitch, loudness and timbre as a sound source [2]. While pitch and loudness are measurable features as a single continuum, timbre is a multi-dimensional concept that has different levels of abstractions [22]. Thus, instead of quantitative measures, researchers attempted to explain it using verbal attributes to capture the perceptual nuance. For example, Lichte [23] used pairs of words with opposite meanings such as "bright/dull" or "thin/full" as a scale of timbre space. Von Bismarck [24] investigated orthogonal sets of word pairs based on a factor analysis of human ratings.

As for singing voice, this qualitative analysis has mainly focused on performance expressions [15], [25], [26]. For example, Scherer et al. [26] studied the acoustic correlates of emotional expressions in opera singing using "anger", "fear", "tender", "joy", "sad" and "pride" as representative words of different emotions. They conducted a multivariate analysis of variance to examine the effects of the emotion targets on various acoustic features and also a multiple discriminant analysis to confirm the reliability of the acoustic patterns. Kanato et al. [15] investigated more generalized descriptions of singing styles using words extracted from various documents. They collected human ratings of the words on a 7-point scale for recordings from singers and conducted factor analysis on the scale data. They showed that three words, "powerful", "cautious" and "cheerful" can be used as basis factors that can explain other words such as "joyful", "lightly", "weak" and "clear". Also, they conducted multiple regression to find how the factor words are related to acoustic features. While these studies provide meaningful methods to define words and associate them with the acoustic features of singing voice, the audio data were recorded in a highly controlled setting to focus on the performance aspect of singing voice. Specifically, the audio data were recorded as monaural solo vocal without any instrumental background. They even marginalized the emotional effect from song itself by using an original composition unknown to singers or non-melodic notes (e.g., music scales). While these controlled settings can provide scientifically rigorous analysis on the effect of singing voice, they are not applicable to commercial popular tracks. Also, they did not explicitly deal with timbral traits of individual singers.

In this study, we aim to analyze singing voice in popular music recordings where professional vocals are accompanied by background instrumental tracks. In order to collect the vocal description data, we obtained highly vocal-specific tags and guided the annotators to focus on singing voice. We also asked them to make a binary decision for each tag to mitigate the annotation efforts. This is in fact the same setup in music auto-tagging tasks. While general music auto-tagging datasets include a highly broad scope of words including genre, mood, instruments or other attributes, our dataset is specialized to singing voice. Some existing music tag datasets actually contain some vocal tags but they are not sufficient to cover the diverse aspects of singing voice [16], [17], [21]. Table I compares vocal tags from popularly used music tagging datasets to those in the proposed KVT dataset. The table evidently shows that our proposed dataset provides much richer tags that describe singing voice (note that the KVT dataset focuses on solo singers and so it is empty in the ensemble type category). We will explain the detail of collecting the data in the next section.

III. Data Collection

This section presents the design process of collecting the vocal tag vocabulary and human annotations of popular music recordings.

A. Tag Vocabulary and Audio Data

The most essential step in creating the tag dataset is defining a tag vocabulary that precisely describes diverse attributes of singing voice. In general, a tag vocabulary can be collected in several different ways including expert curation [16], [27], web document mining [28] or tag collection from social communities [21]. Our approach is based on mining web documents, specifically, taking advantage of a website that provides expert-level reviews of vocalists by professional vocal trainers.[1] The website contains highly detailed analysis of K-pop vocal singers in terms of timbre, vocal techniques and pitch range and it also allows the page viewers to freely discuss the analysis (some popular posts have more than 1,000 responses).

We first collected a large set of words from the website via web-crawling and extracted about 300 tag words associated with singing voice. After filtering out inappropriate words manually, we downsized them to 70 tags. They are listed in Table I. Note that, since we focus on vocal characteristics for solo vocals, we did not include any tag of ensemble voices. In order to collect songs, we first extracted a list of 114 K-pop singers from the website and selected 4 or 5 audio tracks per singer based on song popularity. We filtered out songs with duet, chorus singers or rap. As a result, we

[1] http://kpopvocalanalysis.net

TABLE I
A COMPARISON OF VOCAL TAG VOCABULARY IN VARIOUS DATASETS. IN KVT, TAGS USED IN SEGMENT-LEVEL ANNOTATION ARE IN BOLD. OTHER TAGS ARE USED IN TRACK-LEVEL ANNOTATIONS ONLY. IN LAST.FM, TAGS WITH MORE THAN 200 ANNOTATIONS ARE LISTED.

Category	KVT	Magnatagatune [17]	CAL500 [16]	Last.fm [21]
Range	**Low-range, Mid-range, High-range** Baritone-like, Bassier, Mezzo-Soprano Tenor, Low		High-pitched Low-pitched	
Timbre (low-level)	**Husky/Throaty, Thick, Thin** **Warm, Bright, Clear, Relaxed** **Dark, Energetic, Mild/Soft** **Sharp, Rich, Rounded, Stable, Breathy** Growling, Muffled, Fierce, Piercing, Screamed, Husky, Throaty Quiet, Hush, Smooth, Airy		AlteredwithEffects Gravelly Screaming Strong	Cry, Scream Chilled vocal
Timbre (high-level)	**Lonely, Sad, Passion, Charismatic** **Pretty, Cute, Delicate, Emotional** **Pure, Robotic/Artificial, Embellishing** **Sweet, Young, Compressed, Dynamic** Classically, Grace, Boyish, Aged, Exaggerated, Teen, Creamy, Soulful, Powerful, Delicate Consistensy, Comfortable		Aggressive Emotional	Beautiful voice Sexy female vocals A dynamic male vocalist An emotional male lead vocal performance
Gender	**Male, Female**	Male, Female	Female Lead Vocals Male Lead Vocals	Female, Male
Genre	**Soulful/R&B, Ballad**	Male Opera Female Opera Chant		Jazz Vocal
Technique	**Whisper/Quiet, Shouty, Vibrato** **Falsetto, Speech-like** Non-breathy	Talking	Breathy, Falsetto Monotone Rapping, Spoken	
Ensemble type		Choir, Duet	Call&Response, Duet Vocal Harmonies Backing vocals	Vocal Harmony, Singalong Duet, Choir

acquired 466 songs. The audio tracks include not only vocals but also instrumental sounds. Although we could separate out instrumental sounds from the mixed audio for annotators to focus on singing voice, we decided to use the original mixed tracks and so let the human annotators separate out the vocal audio via their hearing system for two reasons. First, isolated vocals from the mixed audio sound quite unnatural. Singers perform the song while listening to the instrumental sound in the background and thus vocals are affected by the instrumental sound. In addition, mixing engineers process the vocal and instrumental tracks to be cohesively mixed using various audio effects. This is, the mixed audio track is not simply the sum of vocal and background music. This makes the isolated vocal sounds unfamiliar and even awkward. Second, although current singing vocal separation algorithms have made significant advances [29], [30], they still have audible artifacts or contain background vocals or chorus as they separate the sources in stem-level. Thus, we decided that using the separated vocals are not appropriate for human annotation.

B. Human Annotation

Human annotation data for music tracks can be also collected from various resources or strategies [31]. For example, they include conducting a survey [16], [27], harvesting social tags [21], playing annotation games [17], [32] and mining web documents [28]. Since we extracted the tag vocabulary from the vocal analysis website, we could exploit it again to collect the tag annotation for songs from the same web documents. However, the reviewers do not necessarily check the entire tag vocabulary that we defined and thus the annotation can include a significant amount of false negatives. Also, they often use the words to describe general characteristics of singers without being conditioned on a specific song. Therefore, we collected human annotations through a separate survey, using the tag vocabulary and audio data.

We conducted the survey in two steps. The first step is track-level annotation for every tag based on the overall impression after listening to a single piece of music tracks. This track-level strong labeling is the approach in the CAL500 dataset [16] and the Music Genome Project [27]. While this track-level annotation provides a succinct summary of tag annotations per song, it ignores the time-varying nature of vocal tags according to music structure in a song. For example, "High-range" and "Shouty" voices usually appear in the chorus part or the climax of a song. Annotators may become confused with whether these locally active tags should be positive or not in the track-level annotation. To address the temporal inconsistency of vocal tags within a song, in the second step, we conducted segment-level annotation on short vocal segments. This segment-level annotation was the approach in the CAL500exp dataset [20]. The following subsections will explain the two-step human annotations in detail.

C. Track-level Annotation

We collected track-level annotations from a small group of semi-experts (5 participants). They are graduate students who have educational backgrounds in music technology. They are also either amateur or professional musicians. They annotated each song with either '1' (positive) or '0' (negative), listening

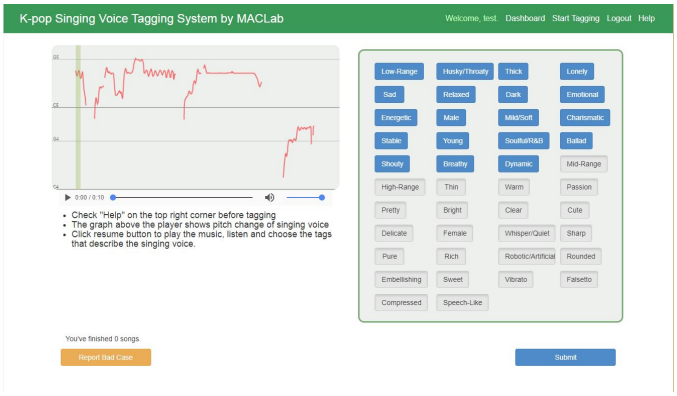


Fig. 1. A screenshot of the crowd-sourcing system for segment-level annotation of vocal tags. The left panel shows the melody contour of singing voice and the right side lists on/off buttons for each tag. The initial states of tags are determined by the track-level annotation that the segment belongs to.

to individual music tracks as many times as they want. They were guided to annotate the song with positive when they are confident with the matching between the tag and the song. Otherwise, they annotated it with negative. Each song was annotated by 3 different participants.

After collecting the track-level annotation data, we investigated the frequency, ambiguity and redundancy of tags as presented in our preliminary study [33]. We measured the frequency by counting the number of annotations per tag. The results showed that some tags such as "Bassier", "Baritone-like" and "Grace" are rarely used, indicating that there are few vocals with the characteristics or annotators are uncertain about them. We also measured tag ambiguity by examining the agreement from different annotators. While "Male" and "Female" have high agreement in general, "Throaty" and "Vibrant" have very low one. Lastly, we measured the redundancy by carrying out hierarchical clustering on the tag-to-tag similarity matrix. Specifically, we computed tag-to-tag cosine distance where each tag vector is binary annotations of a song and used a single linkage method for hierarchical clustering. A plot of the clustering result is shown in our previous work [33]. The results showed that tag pairs such as "Mild-Soft" or "Airy-Breathy" are highly close to each other. We refined the tag words by merging such redundant tags or removing tags with too low activation. As a result, we downsized the tag vocabulary to 42 tags. They are shown in bold in Table I.

D. Segment-level Annotation

Segment-level annotation is labeling short vocal segments within a track. This involves detecting singing voice in the mixed tracks and cutting them off with an appropriate length. To this end, we used a singing voice detector based on CNN [34]. We trained it using datasets with vocal detection labels (RWC [35] and Jamendo [36]) and ensured that the voice detector generally works well for the K-pop songs by inspecting the detected results on a randomly selected subset. Using the pre-trained CNN, we computed the confidence levels of the detector over a segment window. Whenever the average likelihood level within the window is above 80%, we took the segment window and slided it such that the next segment has no overlap with the current one. In our preliminary test, we tried different lengths of segments. When the segment was shorter, the tags were likely to be more consistent within the segment but it became more difficult to confirm the matching between tags and the segment. When the segment was longer, on the other hand, the result was opposite. We found that 10 second is appropriate in the test. This resulted in 6,787 vocal segments out of the 466 music tracks.

Using the vocal segments and the downsized tag vocabulary, we conducted segment-level annotation. Considering that the number of audio examples is much larger in segment-level, we crowd-sourced the annotation survey. To this end, we built a web-based system as shown in Figure 1. The user interface includes two panels. The left one shows the pitch contour of singing voice extracted using a melody extraction algorithm based on deep neural networks [6]. This pitch contour visually helped annotators track and focus on vocal sounds. The right one lists the tag vocabulary and their binary status. Also, Korean translations and tagging instructions are provided on the page. Following the method in [20], the initial decisions of tags were set by the track-level annotation where the segment belongs to. The participants were guided to change the decisions only when they do not agree with the initial decisions. This significantly mitigated the annotation efforts, compared to annotating them from scratch. Each participant annotated from 50 to 300 randomly assigned segments and each segment was annotated by 3 different participants. The total number of participants was 82 and most of them are undergraduate and graduate students of KAIST who have high-level English proficiency.

The system tracked the activity of annotators including the elapsed time after audio loaded for every annotation session, time-stamps of detailed events such as audio loading and tag annotation. This information was used to verify the integrity of annotators. Annotators are marked invalid if: a) the user has less than 50 annotation sessions, b) median elapsed time is less than 15 seconds, c) number of annotation sessions with less than 10 seconds taken is too large (about 10 percents). All annotations from invalid annotators are not included in the dataset as well as the statistic analysis. The median of elapsed time was 35.22 seconds and the mean of tag insertion and deletion actions from the initial states was 7.5. The ratio of operations given 42 tags ($0.178 = 7.5/42$) is slightly greater than that in the CAL500exp dataset which reported the average operation of 9.18 given 67 tags ($0.137 = 9.18/67$) [20].

IV. Statistical Analysis

This section reports statistical analysis of the tag vocabulary focusing on the segment-level annotation data. We first define symbolic expressions for the annotation data, and then investigate frequency and agreement as global characteristics, temporal activations and intra/inter-song frequency as within-song characteristics, and similarity between tags.

A. Annotation Data Definitions

For annotator a, tag t and segment s, we define the annotation record as $r_{a,t,s}$, the number of positive annotations

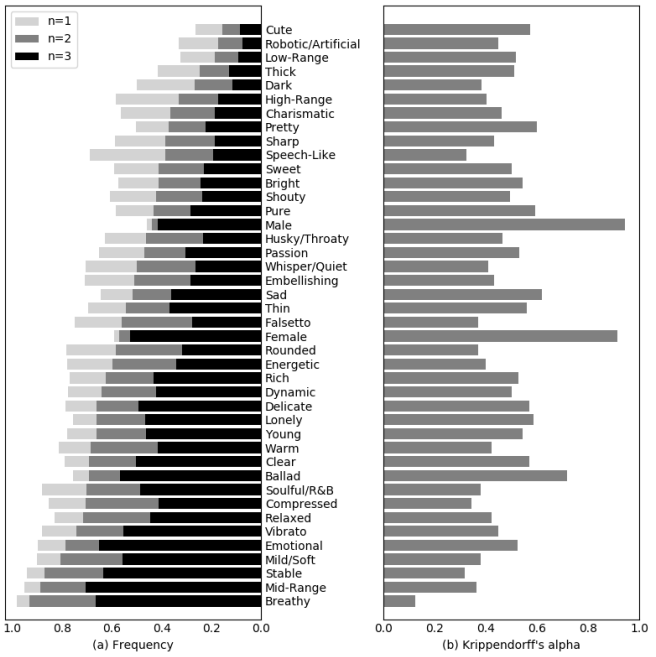


Fig. 2. Frequency distribution and agreement of tags. n-frequency is the ratio of segments that a tag is annotated as positive at least n times (n = 1, 2 or 3) out of three annotators (a). Krippendorff's alpha is used to calculate the agreement among annotators (b).

as $p_{t,s}$, and the binary activation as $b_{t,s}$ as below:

$$r_{a,t,s} = \begin{cases} 1 \text{ if positive} \\ 0 \text{ if negative} \end{cases} \tag{1a}$$

$$p_{t,s} = \sum_{a=1,2,3} r_{a,t,s} \tag{1b}$$

$$b_{t,s} = \begin{cases} 1 \text{ if } p_{t,s} \geq 2 \\ 0 \text{ if } p_{t,s} < 2 \end{cases} \tag{1c}$$

Note that, since we assigned three different annotators to each audio segment, $p_{t,s}$ ranges from 0 to 3.

B. Global Characteristics

1) n-frequency: The frequency distribution of tag annotations across the entire audio segments is straightforward but an important measure that can deliver insight to understand general characteristics of the tags and the dataset. In our preliminary work, we called the relative distribution of positive annotation simply "frequency" [33]. Here we change the name to *n-frequency* and formally define it as below:

$$F_{t,n} = \frac{|\{p_{t,s} \mid p_{t,s} \geq n\}|}{N}, \tag{2}$$

where N is the total number of audio segments and the numerator is the total number of audio segments whose positive annotations are greater than or equal to $n \in \{1, 2, 3\}$.

Figure 2 (a) shows the distributions of n-frequency using the segment-level human annotation data. We sorted the tag words in ascending order of $F_{t,2}$. Note that $F_{t,2}$ is equivalent to the average of $b_{t,s}$ as a binary random variable, as it represents how likely the tag is positive given the audio segment. Compared to other strongly-labelled music tag datasets such as CAL500 [16] or CAL500exp [20], the median of $F_{t,2}$ is somewhat high, which is about 0.5. This is probably because the vocal-specific tags are less exclusive to each other than those in the general music tag datasets where genre and mood categories often entail one out of many choices. Also, compared to the track-level tag annotations in our preliminary work [33], the segment-level tag annotations have more positive labels, that is, higher $F_{t,2}$ values. This means that the human annotators tend to change the annotations more from negative to positive than from positive to negative in the crowd-sourcing annotation system (note that we initialized the tags with track-level annotations as explained in Section III-D), as the segment-level audio makes them focus on the local characteristics.

This distribution of n-frequency also reflects general characteristics of the vocals in the dataset. For example, the top tags in F_2 including "Breathy", "Mid-range", "Stable" and "Emotional" can be considered as common vocal characteristics of the K-pop singers. Another notable result is that there is quite significant disagreement among annotators as indicated by the light and dark gray bar in Figure 2 (a). This is expected due to the subjective nature of the vocal tags, many of which are associated with high-level impression in timbre. To better understand the subjectivity of vocal tags, we examine the human annotation data further in the following subsection.

2) Agreement: When multiple annotators participate in annotation, agreement is an important measure to validate the dataset. Annotators may be more subjective with the KVT dataset since interpretation of semantic tags and effect of background music can vary for each individual. Agreement among human annotators has been studied in many MIR tasks such as music genre classification [37], [38], music emotion recognition [39], [40], audio music similarity [41] and chord recognition [42]. They quantified the degree of coincidence among human annotators, which is often referred to as *inter-annotator agreement (or inter-rater reliability)*, using different types of statistical measures. Among others, Krippendorff's alpha has been a frequent choice as it is applicable to any number of annotators, partial annotations (missing or unequal samples per annotator) and various types of data [39], [40], [42], [43]. This versatility fits well on the crowd-sourced annotations in the KVT dataset where raters and ratings are fragmented and irregular. If a tag has a high probability (F_2-frequency), the expected agreement of tags will also be high. Thus, Krippendorff's alpha calculates the amount of agreement against chances with the same distribution, that is, by dividing the observed agreement by the likelihood [44]. Therefore, if a tag has a high value of F_2 frequency, it is expected to have an even higher agreement to maintain the same value of Krippendorff's alpha. The range is on a scale from 0 (no agreement) to 1 (full agreement). Values between 0.4 and 0.75 are considered as a fair agreement beyond chances [42].

Figure 2(b) shows the calculated Krippendorff's alpha on each vocal tag in the KVT dataset. As expected, F_2-frequency and Krippendorff's alpha are not strongly correlated to each other. "Male" and "Female" have the highest values, which

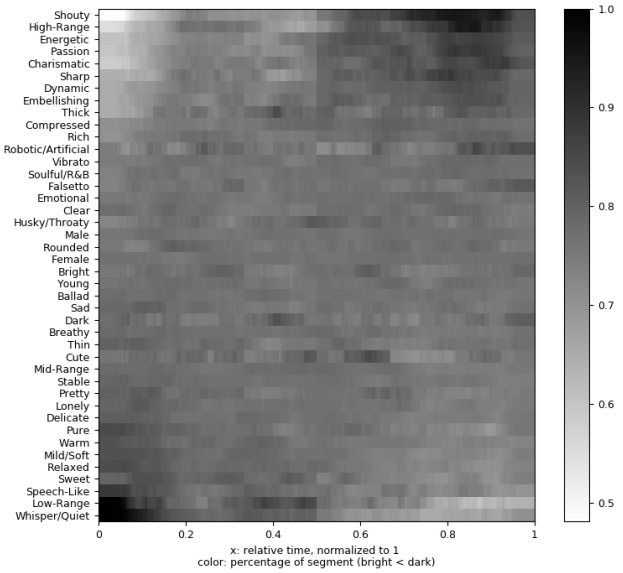


Fig. 3. Temporal probability distribution of tags.

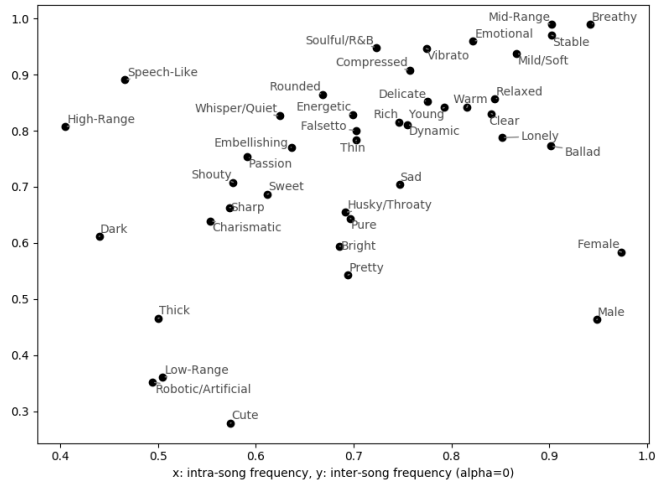


Fig. 4. Intra-song frequency vs. inter-song frequency.

are above 0.9. This is expected as they are easy to recognize. All other tags have much lower values but the majority are in the fair agreement range (between 0.4 and 0.75). About 7 tags are below the fair agreement. We can also observe that tags with high n-frequency such as "Breathy", "Mid-range", "Stable" and "Mild/Soft" has relatively low Krippendorff's alpha values. This is because Krippendorff's alpha, by definition, cancels the effect of the tag frequency. "Compressed", "Rounded" and "Speech-Like" might be because the meanings are more ambiguous than the others. However, the majority of them are marginally less than 0.4. Overall, the Krippendorff's alpha values are similar to or greater than results reported on emotion recognition (0.54 and 0.55 in [39] and 0.360 and 0.222 in [40]).

C. Within-Song Characteristics

1) Temporal Activations: Vocal timbre and expressions can vary from one section to another section within a song. The segment-level annotation allows for analyzing the time-varying characteristics of tags. A simple way of observing the temporal characteristics of vocal tags is calculating the distribution of tag activations over time within a song. Figure 3 shows the temporal activation of tags. Since every song has a different length, we normalized the duration of song by resampling the binary sequence of tag activations, $b_{s,t}$, to have the same number of elements and averaging the activations across all songs. Also, we sorted the tags in the descending order of temporal centroid of average activations, which is defined as $\sum_n n \cdot a_{t,n} / \sum_n a_{t,n}$ where n is the resampled time index and $a_{t,n}$ is the average of the binary activations across songs for tag t. This allows to easily observe the trend of tag appearance by the temporal location within a song. For example, tags in the top rows such as "Shouty", "High-range" and "Energetic" have strong activations at the end of songs, confirming the common song arrangement that the last part is the climax that contains powerful and highly tense voices. On the other hand, tags in the bottom rows such as "Whisper/Quite", "Low-Range" and "Speech-Like" have strong activations at the beginning of songs. This also explains general characteristics of verse part in popular songs. The rest of tags in the middle have some degree of fluctuations but they tend to be consistent over time.

2) Intra-Song Frequency and Inter-Song Frequency: Although the temporal activations show general trends of tag appearance over time, the averaging across songs dilutes how frequently a tag is activated within each song. In order to measure the degree of tag activation within a song, we define *intra-song frequency* as below:

$$D_{t,i} = \sum_j b_{t,i,s} / N_i \tag{3a}$$

$$D_t = \frac{\sum_i D_{t,i}}{|\{i \mid D_{t,i} > 0\}|} \tag{3b}$$

where $b_{t,i,s}$ is the tag binary activation for tag t, song i and segment s in the song i, and N_i is the number of segments in the song. This calculates, when a tag is activated at least once in a song, how much the tag is activated over all segments in the song. This measures the locality of the tag activations within a song. In order to observe intra-song frequency in a contrasting view, we also define *Inter-song frequency*:

$$I_t = \frac{|\{i \mid D_{t,i} > 0\}|}{N} \tag{4}$$

This measures the locality of the tag activation across songs, that is, if the tag appears only in a few songs or more commonly over many songs.

We calculated the intra-song frequency and inter-song frequency values for each tag and plotted them in Figure 4. The frequency measures are generally correlated to n-frequency. However, their relative ratio explain locality characteristics of tags. For example, "High-range" and "Speech-Like", "Whisper/Quiet" and "Shouty" have low intra-song frequency and high inter-song frequency. This makes sense because the tag characteristics are typical in pop music but they usually

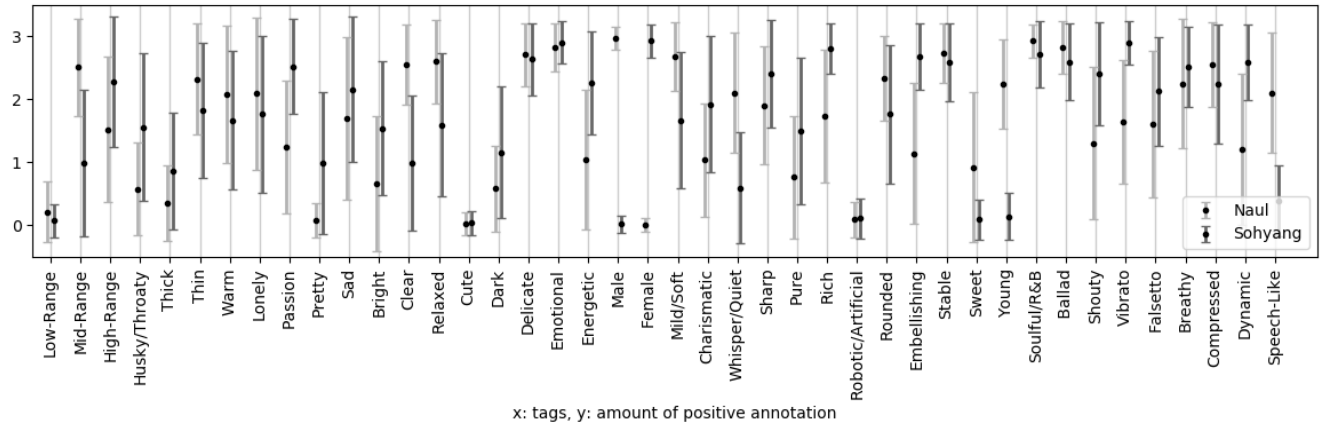


Fig. 5. Case study examples of tag activation patterns that compare two artists.

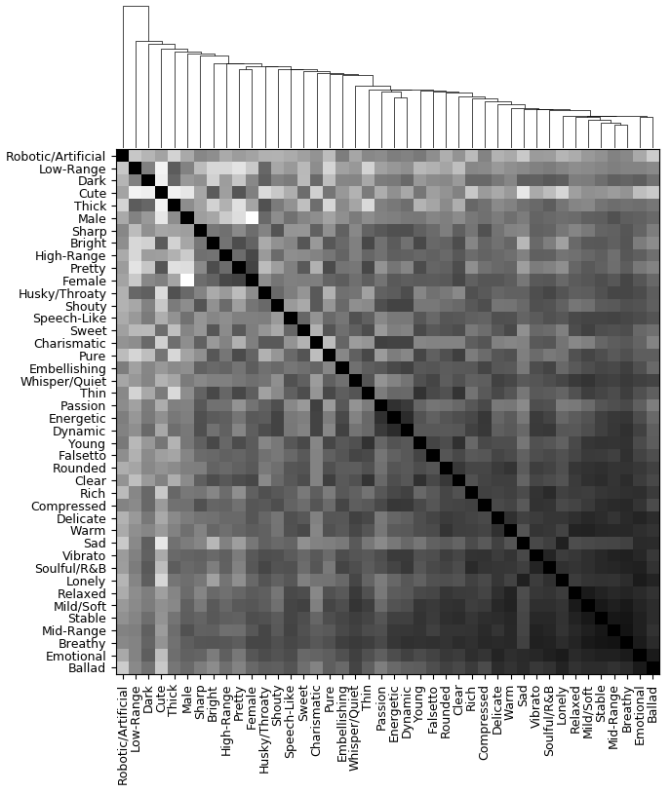


Fig. 6. Tag-to-tag distance. Dark cells mean close distance.

appear in a specific part in the song. That is, they tend to explain singing techniques related to song arrangement rather than characteristics of individual singer. On the other hand, "Lonely", "Ballad", "Warm" and "Sad" have relatively high intra-song frequency compared to inter-song frequency. They are mainly associated with mood of vocal sound, which are usually consistent within a song. That is, they tend to explain characteristics of individual singer for a song. "Male" and "Female" are extreme cases. That is, they are obviously consistent within a song and thus they have very high intra-song frequency but about half inter-song frequency.

D. Tag Similarity

Co-occurrence between tags is common in music tagging because they intrinsically share some (hidden) factors [15], [45], [46]. We investigate the similarity of vocal tags using the segment-level annotation data. We first formed a segment-by-tag matrix with the number of positive annotations $p_{t,s}$ and obtained the tag-by-tag correlation matrix by computing cosine similarity distances between each pair of tag-wise vector in the segment-by-tag matrix. Figure 6 shows the distances between all tag pairs as a matrix. We applied a hierarchical clustering algorithm [47] to effectively visualize cluster groups on the similarity matrix. It shows a couple of distinctive clusters. For examples, "Sharp", "Bright", "High-Range", "Pretty" and "Female" forms a cluster in the upper part. They are seen to explain common characteristics of female K-pop singers. "Passion", "Energetic", and "Dynamic" in the middle seem to be closer together because they are similar words. Some pairs such as "Young" and "Clear", "Sad" and "Lonely", "Pretty" and "Cute" also appear quite commonly. While many vocal tags co-occur and contain some degree of redundancy, they have subtle differences in meaning and some of them elaborate other tags. Thus, the use of multiple tags can strengthen the specificity of vocal description.

E. Case Study of Tag Annotations in Artist Level

Previous sections focused on characteristics of tag itself. Here we shift the perspective to singers who are the target of the tags. We can represent the vocal characteristics of a singer by summarizing the tag annotations of all songs that belong to the singer. Since we have more than 100 singers in the dataset, we selected two representatives, *Naul* and *Sohyang*, as examples of the singer-level tag analysis. *Naul* is a male R&B singer and *Sohyang* is a female Gospel singer. They are well-known for love songs, proficiency in singing skill and emotional expressions. Figure 5 shows the mean and standard deviation of tag annotations of the two singers. Both of them commonly have high positive annotations of "Emotional", "Soulful/R&B", "Delicate", "Stable" and "Ballad". These are expected results considering the music genre and their singing

style and skill. Differences are found in many tags including "Clear", "Whisper/Quiet", "Embellishing", "Young" and "Speech-like". We plotted the standard deviation of tag annotations as a complementary measure. The standard deviation indicates consistency of the tag for a singer. For example, *Sohyang* has very high average but low standard deviation in "Rich", "Embellishing" and "Vibrato", indicating that she tends to maintain the singing style in her songs. This is contrasted to the activations with high variations on the same tags in *Naul*, indicating that he tends to dynamically change the singing style in different segments of his songs.

V. Auto-Tagging

This section presents music auto-tagging experiments using the KVT dataset and a CNN-based classifier. We use the binary activation $b_{s,t}$ as ground truth for the output label. Due to the disagreement among human annotators, the labels may be noisy and this can hinder the training. However, this noisiness in label is commonly found in music tag datasets, particularly when the tags are based on "folksonomy" and weakly labelled [46]. Nonetheless, they can be used to effectively train a neural network or learn meaningful audio embedding via the deep representation learning [46], [48].

We conduct three different experiments to identify the characteristics of the vocal tags. In the first experiment, we evaluate the performance of the auto-tagging model when track-level annotations or segment-level annotations are used both in training and testing sets or the different levels of annotations are used between training and testing sets. By comparing the results with label settings, we will show that the segment-level annotations are essential for vocal tagging. In the second experiment, we examine correlations between the performance and the agreement among human annotators. This will show how auto-tagging accuracy is related to the annotation agreement for each tag. In the third experiment, we separate out the vocal sound from mixed songs and perform auto-tagging on the isolated vocal sound. This will show how the voice-specific model is affected when the background music is suppressed.

A. Model

Recently, deep neural networks have been used as a standard model for music classification and auto-tagging with audio data. They can be roughly categorized into three models: 1-D CNN, 2-D CNN, and SampleCNN, depending on model flexibility [18]. 1-D and 2-D CNN models use spectrogram-based representation for their input, hence the models learn features from the time-frequency representation of audio data. On the other hand, SampleCNN uses raw wave as input to provide more flexibility and learn more fine-grained filter banks while it requires a substantial amount of data for training. We adopted a spectrogram-based model in this work. Our preliminary experiment showed that 2-D CNN models generally outperform 1-D CNN models. Thus we focused on 2-D CNN models and chose a so-called VGG-like model with a filter size of 3×3 [49]. The model configuration is shown in Table II.

TABLE II
Model Detail. Upper row of each cell shows type, kernel size, stride size, and padding(optional) of the layer. Lower row is dimension of output of the layer.

Input(128, 107, 1)
Conv., (3, 3, 32), 1 (126, 105, 32)
Max pool, (3, 3, 1), 3 (42, 35, 32)
Conv., (3, 3, 64), 1, padded (42, 35, 64)
Max pool, (3, 3, 1), 3 (14, 12, 64)
Conv., (3, 3, 128), 1, padded (14, 12, 128)
Max pool, (3, 3, 1), 3 (5, 4, 128)
Fully connected, (42)

B. Experimental Settings

The detailed experimental settings are as follows. We divided the KST data into 298, 70, and 100 songs for training, validation and test splits. They correspond to 4399, 1013 and 1375 segments, respectively. We split the songs for each split to have the same frequency distribution of tags approximately. Each song segment is 10 second long and sampled at 22,050 Hz. We used log mel-spectrogram as input of the CNN model. We first computed Short-time Fourier Transform with 1024 samples of Hanning window, 512 samples of hop size. We then converted the frequency scale to 128 mel bins, and compressed the magnitude compression with a nonlinear curve, $\log(1+C|A|)$ where A is the magnitude and C is set to 10. This returns 431 frames of mel-spectrogram per segment.

This was divided further into four sub-segments to use each of them (107 frames ≈ 2.48 sec) as the input of the CNN model (the labels are copied from the segment annotation). Once the model was trained, we made a tag prediction by averaging the model outputs over the 10-second-long segment. This segmentation strategy is widely adopted for the music auto-tagging task [18]. We used the Keras library to implement the model and train it with an SGD optimizer with learning rate 0.01 and batch size 25. To calculate F-score, we tested thresholds between 0 and 1 with a resolution of 0.01 and chose the one that achieves the maximum F-score for each tag.

For the experiment with separated vocals, we used a pre-trained Wave-U-net model [29], a state-of-the-art algorithm that separates mixed tracks into foreground (vocal) and background (other instruments) audio in the waveform domain. For this experiment, we trained and tested the model with segment-level annotation only. Other components of the experiment remain unchanged for comparison.

C. Results and Discussion

1) Track-Level and Segment-Level Annotations: Table III shows performance scores of the models. They are average values of AUC and F-score across all vocal tags. Both of

TABLE III
ANNOTATION AND RETRIEVAL RESULTS WITH SEGMENT-LEVEL TAG ANNOTATIONS AND TRACK-LEVEL TAG ANNOTATIONS. ALL REPORTED METRICS ARE AVERAGED ACROSS 10 TRAINING RUNS OF THE CNN MODELS WITH DIFFERENT INITIALIZATIONS.

Training Label	Test Label	AUC	F-score	Precision	Recall
Segment	Segment	**0.7352**	**0.7385**	0.6495	0.8144
Track	Segment	0.7099	0.7237	0.6281	0.8836
Segment	Track	0.7282	0.5256	0.4260	0.6633
Track	Track	0.7192	0.5207	0.4312	0.7409

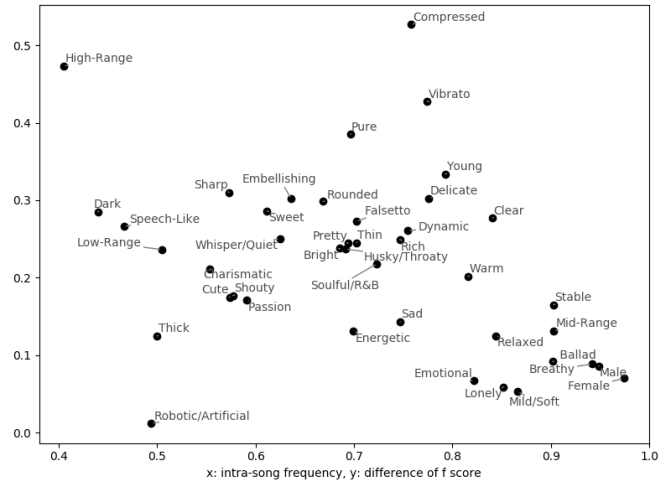


Fig. 7. A 2D plot of F-score drop and intra-song frequency when the labels changes from segments-level to track-level in both training and test sets. Except a few outliers, they are loosely inversely proportional to each other; The Pearson correlation coefficient value is -0.35 with a p value of 0.02.

them are higher when segment-level labels are used for training. This indicates the effectiveness of segment-level tag annotations. It is notable that, when track-level label is used for testing, the F-score values significantly drop (about 0.2) whereas the AUC values remain almost unchanged. This is mainly because F-score is attenuated by the data imbalance whereas ROC is not much affected by that [50]. As stated in Section IV-B1, the track-level labels have higher skew values (the negative-to-positive ratio) for each tag than segment-level labels. Since the difference of data imbalance is related to the locality of tags within a song, we investigate the F-score further by plotting the performance drop against intra-song frequency for each tag in Figure 7. It shows that vocal tags with low intra-song frequency (more local) such as "High-Range", "Dark", "Speech-like" and "Low-Range" have high drops of F-score, whereas vocal tags with high intra-song frequency (more insistent) such as "Male", "Female", "Ballad" have small changes. "Compressed" and "Vibrato" have higher differences of F-scores than other tags with similar intra-song frequency. These tags can be said to be more sensitive to locality that become less predictable with track-level labels. "Thick" and "Robotic/Artificial" are the opposite cases. In summary, when segment-level annotations are used in both training and test sets, we achieve the best results. This supports the time-varying nature of vocal tags.

2) Agreement Analysis: While the segment-level annotations make improvement, the absolute level of AUC score,

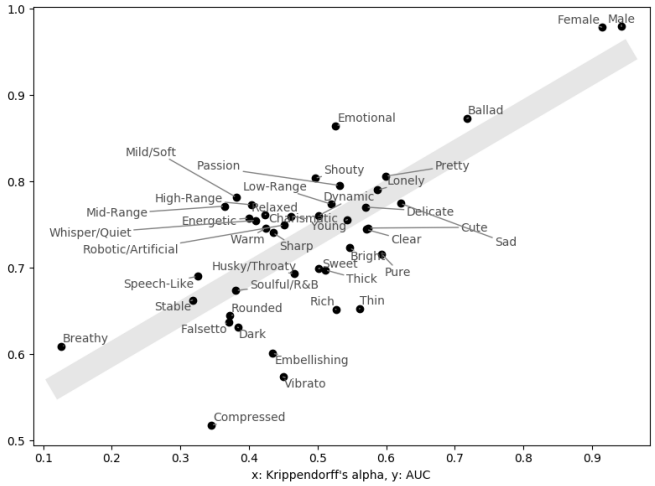


Fig. 8. A 2D plot of AUC and Krippendorff's alpha. The Pearson correlation coefficient value is 0.7278 with a p value of 4.7e-8.

which is about 0.735, is significantly lower than those reported in other music auto-tagging studies [18]. Although there will be many possible reasons for this, we assume that the ambiguity of vocal tags is an important factor and thus we investigate it, associating the performance with tag annotation agreement. In Section IV-B2, we measured the annotation agreement among different human annotators using Krippendorff's alpha. Figure 8 plots the AUC scores and the alpha values for each tag. The plot shows a strong positive relationship between them, as indicated by the Pearson correlation coefficient value of 0.7278. This is expected because the agreement measure contains potential noisiness and it is reflected on the trained model. A notable result is that some tags such as "Breathy", "Compressed" and "Stable" appear frequently (high values of n-frequency) but they have low performance (low values of AUC). This shows strong influence of the low agreement.

3) Isolated Vocal Input: Table IV compares the performance scores when the input audio is mixed (without any modification), vocal only and background music only in the auto-tagging system. The results show that the model trained with the separated vocal has better performance than the one with mixed audio, while the model trained with background music becomes relatively worse. This ensures that the tags are targeted to vocal sources in the song, in other words, human annotators pay more attention to the vocal in the mix. We examine the difference further by looking into the performance deviation for each tag. Figure 9 shows the AUC score ratio between the model with vocal and the model with background music for each tag. Tags with higher ratios are associated with singing technique ("Falsetto", "Breathy" and "Vibrato") or voice tone ("Clear", "Pretty", and "Young"). These tags are usually used to describe voice, not appropriate for background music. Therefore, they benefited more when the isolated vocal is used as input. Moreover, "Falsetto", "Breathy" and "Vibrato" have lower AUC scores with the mixed audio as seen in Figure 7. This implies that vocal features relevant to these tags are more likely to be shadowed by background music. On the other hand, we observe that

TABLE IV
PERFORMANCE COMPARISON WITH ISOLATED VOCAL AND BACKGROUND MUSIC. SEGMENT-LEVEL ANNOTATIONS ARE USED IN BOTH TRAINING AND TEST SETS. ALL REPORTED METRICS ARE AVERAGED ACROSS 10 TRAINING RUNS OF THE CNN MODEL WITH A DIFFERENT RANDOM INITIALIZATION.

Audio	AUC	F score	Precision	Recall
Mix	0.7352	0.7385	0.6495	0.8144
Vocal	**0.7491**	**0.7422**	0.6606	0.8074
Background	0.7119	0.7259	0.6268	0.8229

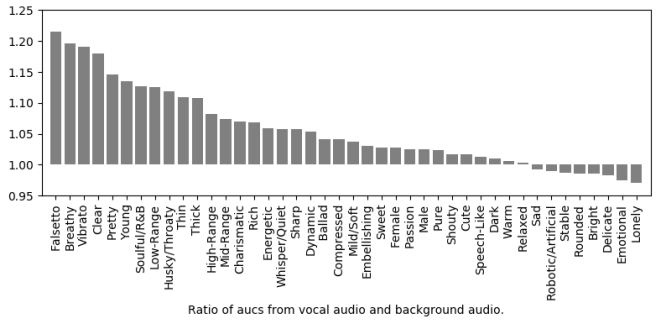


Fig. 9. Each tag's (relative) difference of AUC scores from vocal audio and background audio. Tags with the larger difference are the more independent to background music. Values below one mean the tags scored better in background audio.

tags associated with mood or overall impression ("Lonely", "Emotional", "Bright" and "Sad") have lower score ratios. This is because the words can explain not only voice but also background music.

This experiment shows that using the isolated vocal as input generally improves the model performance, and the extent depends on how exclusively the tag explains voice. However, the performance discrepancy is not sufficiently large. We conjecture there are several reasons for this result. First, the source separation algorithm that we used cannot perfectly divide the vocal and background music. When we listened to the separated background music, we were able to perceive some residuals of vocals. Second, vocals and accompaniment sound actually have some degree of correlation in general. For example, R&B music has a typical set of arrangements (e.g., medium or slow tempo, drum beat, electronic keyboard, and so on) and R&B vocalists also share common characteristics (e.g. high vibrato, pitch embellishment, and emotional). Third, annotators could be affected by the background instrumental sounds when they annotate the songs as you pointed out. Lastly, we used F-score and AUC to report the difference. The accuracy-based metrics may not be the best to differentiate the models with different inputs.

We can compensate for the accuracy-based metric with representational dissimilarity analysis, which is used to compare response patterns elicited in a brain region or model representation of neural networks [51]. This shows overall representational similarity (instead of accuracy-based performance similarity) and measures how similarly the models respond to input. The representational similarity can be computed by correlation between two different responses from the compared models. In our setting, we made tag prediction vectors for all

TABLE V
SPEARMAN'S RANK CORRELATION COEFFICIENTS BETWEEN PAIRS OF REPRESENTATIONAL SIMILARITY MATRICES FROM MODELS TRAINED WITH DIFFERENT AUDIO SAMPLES.

Representation pair	Spearman's r
Mixed - Vocal	0.9158
Mixed - Background	0.8504
Vocal - Background	0.7783

training examples from two models with different inputs and form a example-by-tag matrix for each. By multiplying the two matrices, we obtained an example-by-example matrix, which shows the representational similarity. This is often summarized using the Spearman rank correlation. We show the results for different pairs of models in Table V. The mixed audio input and the isolated vocal input have a strong correlation given the tag annotations. Thus, the two models respond to their inputs (or make the tag predictions) in a more similar manner. On the other hand, the isolated vocal and the background music have a significantly attenuated correlation. This result ensures that the vocal tags are more strongly tied to the vocal sounds in the mixed audio.

VI. APPLICATIONS

Music auto-tagging systems can be applied to various music services such as query-based music retrieval and music recommendation (or playlist generation) [52]. The voice-specific auto-tagging system can provide more unique and artist-oriented services. This section proposes potential applications.

A. Querying and Retrieval

A straightforward application is query-based music retrieval based on the vocal tags. Compared to other music auto-tagging systems, this system can be useful for searching songs with a certain vocal characteristic. For example, users can find songs with "Clear" voice using the corresponding tag. The query can be not only one of the vocal tags but also an audio track. For example, if a user likes a song because of the vocal, the system can predict vocal tag activations of the song from the audio track, and search other songs with similar vocal tag activations.

The query can be also artist-level. Given a set of songs that belong to an artist, we can compute average activations of the vocal tags to represent the artist. The resulting tag activation vector will be similar to those in Figure 5. Note that this vector itself explains the artist voice in a human-friendly way with the tag words. Using similarity between the two tag activation vectors, we can retrieve artists with similar vocal characteristics. Table VI shows several examples of retrieval results. For example, "Lee Hi" has a tag activation vector where top 5 elements correspond to "Female", "Soulful/R&B", "Rich", "Vibratio", and "Mid-Range". Using the tag activation vector as a query, the most similar artists were searched and listed in the Table. They are all female singers in common and are well-known for very skillful and expressive singing styles with the R&B feel.

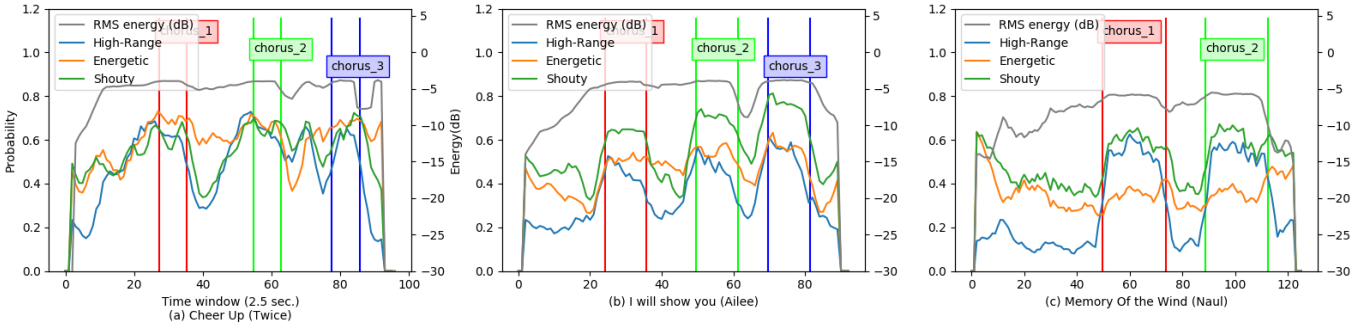


Fig. 10. Temporal probabilities of three tag prediction relevant to chorus sections and RMS energy in dB in three selected songs. We smoothed the temporal plots with a 5-point moving average filter.

TABLE VI
RETRIEVAL RESULTS OF ARTIST-LEVEL QUERIES. THE TOP 5 TAGS EXPLAIN THE QUERY ARTIST. SIMILAR ARTISTS ARE SHOWN WITH THE CORRESPONDING COSINE DISTANCES.

Query (artist)	Top 5 Tags	Similar Artists
Lee Hi	Female, Soulful/R&B, Rich, Vibrato, Mid-Range	Ann (0.0473) So Jung (0.0538) Son Seung-yeon (0.0606) ALi (0.0733) Ailee (0.8500)
IU	Female, Pretty, Sweet, Breathy, Young	Kei (0.0354) Raina (0.0448) Narsha (0.0495) So Jin (0.0523) Minah (0.0560)
John Park	Male, Stable, Vibrato, Soulful/R&B, Husky/Throaty	Taeyang (0.0510) Hwanhee (0.0606) Park Hyo Shin (0.0611) Zio (0.0630) Chang Min (0.0637)

B. Music Thumbnailing

Music thumbnailing is a task of finding the most representative or highlight section of a song [53]. This is useful for fast music browsing or preview in music service. Previous work tackled the problem by exploiting structural similarity within a song (e.g., detecting repetitive or chorus sections) [53] or semantic attention mechanism [54]. Since the KVT dataset provides tags that are explicitly relevant to highlight sections such as "High-Range", "Energetic" or "Shouty", we can use the tag prediction levels as an indicator of highlight sections. Figure 10 shows three examples of temporal predictions of the tags. To easily interpret the results, we manually annotated them with a structural label, "chorus", which is typically used as a music thumbnail [53]. We also compare the tag prediction levels to root-mean-squares (RMS) energy which is also a strong indicator of highlight section [55]. The song in Figure 10 (a) is an exciting dance music sung by a girl group. While the RMS energy does not change much, the rise and fall of the three tag predictions effectively represent the chorus section. The other two songs also show the similar correspondence. An exception is that the "Energetic" tag is not strong in Figure 10 (c). This might be because the track is an R&B song with soft mood. In general, these results show a potential use of the specific vocal tags in music thumbnailing.

C. Qualitative Singing Evaluation

Singing evaluation is an essential module of karaoke systems. Traditional approaches focused on the preciseness of pitch and rhythm in singing, which can be rated by comparing singer's pitch and energy contours to a given melodic score [10]. Beyond the standard singing evaluation, recent approaches have attempted to measure singing techniques such as vibrato [56] or even "singing enthusiasm" [57]. Our proposed vocal auto-tagging system can be also used to evaluate singing but focusing on voice timbre or singing styles. This qualitative analysis can be used not only as a complement to scoring the preciseness of pitch and rhythm but also for characterizing users' voice. Recent online karaoke services allow for recording singing along with background music and sharing the audio tracks in their websites. As aforementioned in Section VI-A, we can represent one's voice with a vocal tag vector and explain general characteristics using the auto-tagging system. Furthermore, the tag vector can be used to find other users with similar vocal characteristics and so promote social activity among users in the karaoke services. In addition, the voice-based retrieval can be used to recommend songs from artists with similar voices.

VII. CONCLUSIONS

We presented a study on qualitative analysis of singing voice in popular music. We described the details of data collection process and conducted statistical analysis of vocal tag annotations in various perspectives including frequency, agreement, temporal dynamics and similarity. Through the auto-tagging experiments, we showed that segment-level tag annotation is crucial to handle the dynamics nature of singing voice, agreement of tag annotation is proportional to auto-tagging performance, and isolated vocal as input boosts the performance further when vocal tags are more specific to voice. We also showed potential applications which can be useful in music streaming or music entertainment services. However, our study has several limitations. First, the collected dataset is not sufficiently large and so we had to use a simple

CNN for the auto-tagging system as a baseline. To overcome the limitation of small data, we could use transfer learning. For example, we can train a deep neural network model using a large-scale of vocal audio segments and artist labels [58], [59]. This provides vocal embedding space from which we can extract the vocal features and apply them to the vocal tagging. Second, some of vocal tags are still ambiguous, which is a obstacle to train the reliable auto-tagging system. We need to have a systemic approach to control it. For example, we can measure the agreement such as Krippendorff's alpha while collecting the human annotation data, and if the agreement is sufficiently low, we can discard the tag. Third, our dataset covers K-pop music only. Although the K-pop music is not exotic but rather similar to Western pop in terms of musical melody and arrangement, it definitely retains some unique characteristics due to the different cultural background. For example, lyrics is Korean and many of songs are highly emotional. Nonetheless, our study takes a first step toward qualitative analysis of singing voice in popular music and provides directions for further research. We share the KVT dataset and demo examples. The audio data is not available because of copyright issues but we provide links to access to the audio data. We described the details in the project website [2].

VIII. Acknowledgement

We appreciate Matheus Dias for permission to using the web documents on K-pop vocal analysis. This research was supported by Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Science, ICT & Future Planning (2015R1C1A1A02036962).

References

[1] M. Goto, "Singing information processing," in *Proceedings of the 12th IEEE International Conference on Signal Processing (ICSP)*, vol. 10, 2014, pp. 2431–2438.

[2] E. J. Humphrey, S. Reddy, P. Seetharaman, A. Kumar, R. M. Bittner, A. Demetriou, S. Gulati, A. Jansson, T. Jehan, B. Lehner, A. Kruspe, and L. Yang, "An introduction to signal processing for singing-voice analysis: High notes in the effort to automate the understanding of vocals in music," in *IEEE Signal Processing Magazine*, 2019, pp. 82–94.

[3] J.-L. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 169–172.

[4] V. Rao and P. Rao, "Vocal melody extraction in the presence of pitched accompaniment in polyphonic music," in *IEEE Trans. Audio, Speech, Language Process.*, 2010, pp. 2145–2154.

[5] J. Salamon, E. Gómez, D. P. Ellis, and G. Richard, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," *IEEE Signal Processing Magazine*, vol. 31, no. 2, pp. 118–134, 2014.

[6] S. Kum, C. Oh, and J. Nam, "Melody extraction on vocal segments using multi-column deep neural networks," in *Proc. of the International Society for Music Information Retrieval (ISMIR)*, 2016, pp. 819–825.

[7] T. Hosoya, M. Suzuki, A. Ito, and S. Makino, "Lyrics recognition from a singing voice based on finite state automaton for music information retrieval," in *Proc. 6th Conf. Int. Society for Music Information Retrieval (ISMIR)*, 2005, pp. 532–535.

[8] M. McVicar, D. P. Ellis, and M. Goto, "Leveraging repetition for improved automatic lyric transcription in popular music," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 3117–3121.

[2]https://khlukekim.github.io/kvtdataset

[9] J. Song, S. Y. Bae, and K. Yoon, "Mid-level music melody representation of polyphonic audio for query-by-humming system," in *Proc. 3rd conf. Int. Society for Music Information Retrieval (ISMIR)*, 2002.

[10] W.-H. Tsai and H.-C. Lee, "Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1233–1243, 2012.

[11] M.-Y. Kan, Y. Wang, D. Iskandar, T. L. Nwe, and A. Shenoy, "Lyrically: Automatic synchronization of textual lyrics to acoustic music signals," in *IEEE Trans. Audio, Speech, Language Process.*, 2008, pp. 338–349.

[12] Y. Kim and B. Whitman, "Singer identification in popular music recordings using voice coding features," in *Proc. Int. Society for Music Information Retrieval Conf.*, 2002.

[13] H. Fujihara, M. Goto, T. Kitahara, and H. G. Okuno, "A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 638–648, 2010.

[14] T. Nakano, K. Yoshii, and M. Goto, "Vocal timbre analysis using latent dirichlet allocation and cross-gender vocal timbre similarity," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 5202–5206.

[15] A. Kanato, T. Nakano, M. Goto, and H. Kikuchi, "An automatic singing impression estimation method using factor analysis and multiple regression," in *Proc. of ICMC SMC*, 2014, pp. 1244–1251.

[16] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 467–476, 2018.

[17] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games: The case of music tagging." in *Proc. 10th Conf. Int. Society for Music Information Retrieval (ISMIR)*, 2009, pp. 387–392.

[18] J. Nam, K. Choi, J. Lee, S.-Y. Chou, and Y.-H. Yang, "Deep learning for audio-based music classification and tagging: Teaching computers to distinguish rock from bach," *IEEE Signal Processing Magazine*, pp. 41–51, 2019.

[19] E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions from audio," in *Proc. 11th Conf. Int. Society for Music Information Retrieval (ISMIR)*, 2010, pp. 465–470.

[20] S.-Y. Wang, J.-C. Wang, Y.-H. Yang, and H.-M. Wang, "Towards time-varying music auto-tagging based on cal500 expansion," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2014, pp. 1–6.

[21] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proc. Int. Society for Music Information Retrieval Conf.*, 2011.

[22] S. Donnadieu, *Mental Representation of the Timbre of Complex Sounds*. New York, NY: Springer New York, 2007, pp. 272–319.

[23] W. H. Lichte, "Attributes of complex tones," *Journal of Experimental Psychology*, vol. 28, no. 6, pp. 455–480, 1941.

[24] G. von Bismarck, "Sharpness as an attribute of the timbre of steady sounds," *Acustica*, vol. 30, no. 3, pp. 159–172, 1974.

[25] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 46, no. 12, pp. 227–256, 2003.

[26] K. R. Scherer, J. Sundberg, B. Fantini, S. Trznadel, and F. Eyben, "The expression of emotion in the singing voice: Acoustic patterns in vocal performance," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 1805–1815, 2017.

[27] M. Prockup, A. F. Ehmann, F. Gouyon, E. M. Schmidt, Ò. Celma, and Y. E. Kim, "Modeling genre with the Music Genome Project: Comparing human-labeled attributes and audio features," in *Proc. Int. Society for Music Information Retrieval Conf.*, 2015, pp. 31–37.

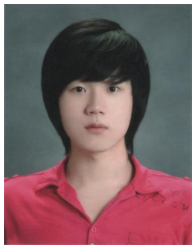
[28] B. Whitman and D. Ellis, "Automatic record reviews," in *Proc. Int. Society for Music Information Retrieval Conf.*, 2004.

[29] D. Stoller, S. Ewert, and S. Dixon, "WAVE-U-NET: A multi-scale neural network for end-to-end audio source separation," in *Proc. 19th Conf. Int. Society for Music Information Retrieval (ISMIR)*, 2018, pp. 334–340.

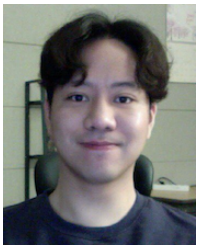
[30] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-unmix-a reference implementation for music source separation," *Journal of Open Source Software*, vol. 4, no. 41, p. 1667, 2019.

[31] D. Turnbull, L. Barrington, and G. Lanckriet, "Five approaches to collecting tags for music," in *Proc. Int. Society for Music Information Retrieval Conf.*, 2008.

[32] M. I. Mandel and D. P. Ellis, "A web-based game for collecting music metadata," *Journal of New Music Research*, vol. 37, no. 2, pp. 151–165, 2008.
[33] K. L. Kim, S. Kum, C. L. Park, J. Lee, J. Park, and J. Nam, "Building k-pop singing voice tag dataset: A progress report," in *Late-Breaking/Demo in the 18th International Society for Musical Information Retrieval Conference*, 2017.
[34] J. Schlüter and T. Grill, "Exploring data augmentation for improved singing voice detection with neural networks," in *Proc. Int. Society for Music Information Retrieval Conf.*, 2015.
[35] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: popular, classical and jazz music databases." in *In Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2002, pp. 287–288.
[36] M. Ramona, G. Richard, and B. David, "Vocal detection in music with support vector machines," in *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 1885–1888.
[37] S. Lippens, J. Martens, and T. D. Mulder, "A comparison of human and automatic musical genre classification," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2004, pp. 233–236.
[38] K. Seyerlehner, G. Widmer, and P. Knees, "A comparison of human, automatic and collaborative music genre classification and user centric evaluation of genre classification systems," *Adaptive Multimedia Retrieval. Context, Exploration, and Fusion*, pp. 118–131, 2011.
[39] M. Soleymani, A. Aljanaki, and Y.-H. Yang, "Emotional analysis of music: A comparison of methods," in *Proc. ACM Conference on Multimedia (MM)*, 2017, pp. 1161–1164.
[40] J. Fan, K. Tatar, M. Thorogood, and P. Pasquier, "ranking-based emotion recognition for experimental music," in *Proc. 18th Conf. Int. Society for Music Information Retrieval (ISMIR)*, 2017, pp. 368–375.
[41] A. Flexer and T. Grill, "The problem of limited inter-rater agreement in modelling music similarity," *Journal of New Music Research*, vol. 45, no. 3, pp. 239–251, 2016.
[42] H. V. Koops, W. B. de Haas, J. A. Burgoyne, J. Bransen, A. Kent-Muller, and A. Volk, "Annotator subjectivity in harmony annotations of popular music," *Journal of New Music Research*, vol. 48, no. 3, pp. 232–252, 2019.
[43] M. Schedl, H. Eghbal-Zadeh, E. Gomez, and M. Tkalci, "An analysis of agreement in classical music perception and its relationship to listener characteristics," in *Proc. of 17th Conf. Int. Society for Music Information Retrieval (ISMIR)*, 2016, pp. 578–583.
[44] K. Krippendorff, "Estimating the reliability, systematic error and random error of interval data," *Educational and Psychological Measurement*, vol. 30, no. 1, pp. 61–70, 1970.
[45] R. Miotto, L. Barrington, and G. Lanckriet, "Improving auto-tagging by modeling semantic co-occurrences," in *Proc. of 11th Conf. Int. Society for Music Information Retrieval (ISMIR)*, 2010, pp. 297–302.
[46] K. Choi, G. Fazekas, K. Cho, and M. Sandler, "The effects of noisy labels on deep convolutional neural networks for music tagging," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 139–149, 2018.
[47] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," *arXiv:1109.2378*, 2011.
[48] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Transfer learning for music classification and regression tasks," in *Proc. of 18th Conf. Int. Society for Music Information Retrieval (ISMIR)*, 2017, pp. 141–149.
[49] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," in *Proc. Int. Society for Music Information Retrieval Conf.*, 2016.
[50] L. A. Jeni, J. F. Cohn, and F. D. la Torre, "Facing imbalanced data–recommendations for the use of performance metrics," *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 245–251, 2013.
[51] N. Kriegeskorte, M. Mur, and P. A. Bandettini, "Representational similarity analysis-connecting the branches of systems neuroscience," *Frontiers in systems neuroscience*, vol. 2, p. 4, 2008.
[52] M. Schedl, E. Gómez, and J. Urbano, "Music information retrieval: Recent developments and applications," *Foundations and Trends in Information Retrieval*, vol. 8, no. 2-3, pp. 127–261, 2014.
[53] M. A. Bartsch and G. H. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 96–104, 2005.
[54] Y.-S. Huang, S.-Y. Chou, and Y.-H. Yang, "Pop music highlighter: Marking the emotion keypoints," *Transactions of the International Society for Music Information Retrieval*, vol. 1, no. 1, pp. 68–78, 2018.
[55] J.-W. Ha, A. Kim, C. Kim, J. Park, and S. Kim, "Automatic music highlight extraction using convolutional recurrent attention networks," *arXiv preprint arXiv:1712.05901*, 2017.
[56] T. Nakano, M. Goto, and Y. Hiraga, "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," in *Proc. Int. Conference on Speech Communication and Technology(Interspeech)*, 2006, pp. 1706–1709.
[57] R. Daido, M. Ito, S. Makino, and A. Itoa, "Automatic evaluation of singing enthusiasm for karaoke," *Computer Speech & Language*, vol. 28, no. 2, pp. 501–517, 2014.
[58] J. Park, J. Lee, J. Park, J.-W. Ha, and J. Nam, "Representation learning of music using artist labels," in *Proc. 18th Conf. Int. Society for Music Information Retrieval (ISMIR)*, 2018, pp. 717–724.
[59] J. Park, D. Kim, J. Lee, S. Kum, and J. Nam, "A hybrid of deep audio feature and i-vector for artist recognition," in *Proc. of the 35 th International Conference on Machine Learning, The 2018 Joint Workshop on Machine Learning for Music*, 2018.



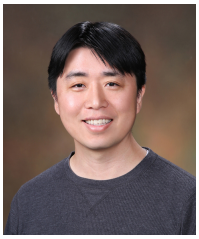
Keunhyoung Luke Kim is currently a Ph. D. candidate in the Graduate School of Culture Technology at the Korea Advanced Institute of Science and Techonology(KAIST), South Korea. He received his M.S. in culture technology from KAIST, in 2010. His research interests include musical timbre, deep learning and practical applications of them.



Jongpil Lee received the B.S. degree in electrical engineering from Hanyang University, Seoul, South Korea, in 2015, the M.S. degree, in 2017, from the Graduate School of Culture Technology, Korea Advanced Institute of Science and Technology, Daejeon, South Korea, where he is currently working toward the Ph.D. degree. He interned at Naver Clova Artificial Intelligence Research in the summer of 2017 and at Adobe Audio Research Group in the summer of 2019. His current research interests include machine learning and signal processing applied to audio and music applications.

Sangeun Kum received the B.S. degree in Electronics Engineering from Kyungpook National University, Korea, in 2014, the M.S. degree, in 2016, from the Graduate School of Culture Technology, Korea Advanced Institute of Science and Technology, Daejeon, South Korea, where he is currently working toward the Ph.D. degree. He interned at Naver AiTEMS in the summer of 2018. He is interested in machine learning, signal processing, and singing voice, which can be applied to music applications.

Chae Lin Park received the B.S. degree in Art& Techonology from Sogang University, Seoul, South Korea, in 2017, the M.S. degree, in 2019, from the Graduate School of Culture Technology, Korea Advanced Institute of Science and Technology, Daejeon, South Korea. She interned at Samsung Electronics VIP Center in the summer of 2014 and at Naver AiTEMS in the summer of 2018. Now, she is a UX designer at LOTTE e-commerce search service planning from 2019.



Juhan Nam is an Associate Professor of the Graduate School of Culture Technology at the Korea Advanced Institute of Science and Technology (KAIST), South Korea. Before joining KAIST, he was a staff research engineer at Qualcomm from 2012 to 2014. He was also a software/DSP engineer at Young Chang (Kurzweil) from 2001 to 2006. He received his Ph.D. degree in music from Stanford University, studying at the Center for Computer Research in Music and Acoustics (CCRMA). He is interested in various topics at the intersection of music, signal processing, and machine learning. He is a member of the IEEE.