Non-Reference Audio Quality Assessment for Online Live Music Recordings

Zhonghua Li[1], Ju-Chiang Wang[1,2], Jingli Cai[1], Zhiyan Duan[1], Hsin-Min Wang[2], Ye Wang[1]
[1]School of Computing, National University of Singapore, Singapore
[2]Institute of Information Science, Academia Sinica, Taiwan
{lzhlynn,asriver.wang}@gmail.com; {jingli,zhiyan}@comp.nus.edu.sg;
whm@iis.sinica.edu.tw; wangye@comp.nus.edu.sg

ABSTRACT

Immensely popular video sharing websites such as YouTube have become the most important sources of music information for Internet users and the most prominent platform for sharing live music. The audio quality of this huge amount of live music recordings, however, varies significantly due to factors such as environmental noise, location, and recording device. However, most video search engines do not take audio quality into consideration when retrieving and ranking results. Given the fact that most users prefer live music videos with better audio quality, we propose the first automatic, non-reference audio quality assessment framework for live music video search online. We first construct two annotated datasets of live music recordings. The first dataset contains 500 human-annotated pieces, and the second contains 2,400 synthetic pieces systematically generated by adding noise effects to clean recordings. Then, we formulate the assessment task as a ranking problem and try to solve it using a learning-based scheme. To validate the effectiveness of our framework, we perform both objective and subjective evaluations. Results show that our framework significantly improves the ranking performance of live music recording retrieval and can prove useful for various real-world music applications.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering, Retrieval models; H.5.5 [**Sound and Music Computing**]: Systems

General Terms

Algorithms, Design, Experimentation

Keywords

Live music videos; audio quality assessment; learning-to-rank; and music information retrieval

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2502106.

Table 1: Different live video recordings and the corresponding audio quality ratings of the song "Hold it against me" by Britney Spears.

No	URL	AQ	Ord
1	http://www.youtube.com/watch?v=IkAI6MHtJCc	Good	1
2	http://www.youtube.com/watch?v=HtCkp40b5vA	Poor	4
3	http://www.youtube.com/watch?v=7P7JwaLRToM	Poor	5
4	http://www.youtube.com/watch?v=1CvxILczzgs	Good	3
5	http://www.youtube.com/watch?v=vOBWvXOxck8	Good	2
6	http://www.youtube.com/watch?v=kCI3It5YI_0	Poor	6

1. INTRODUCTION

The last decade has witnessed an explosion of musical data on video sharing websites. YouTube[1], Youku[2], and Nico Nico Douga[3] are now counted among the largest and most important sources of music information serving users, who can not only enjoy the official audio and video releases (both live and studio versions) of their favorite artists but also share their own recordings of live concerts with others who do not have the experience. The audio quality of the huge amount of live music recordings, however, varies significantly due to different recording factors, such as environmental noise, location, and recording device. In view of this, we try to improve live music video search by incorporating audio quality assessment.

When given a query, existing video search engines match it with the metadata of each video in the database and rank results according to their relevance. User feedback, such as view count and rating, may also be considered in the ranking or re-ranking process. To the best of our knowledge, however, most popular video search engines do not take the audio quality of the recordings into consideration. For example, the song "Hold it against me" by Britney Spears has several live recordings available on YouTube. We invited ten subjects to rank the order (Ord) of the top six videos retrieved using the query "Britney Spears Hold it against me live" after evaluating their overall audio quality (AQ). The result, summarized in Table 1, indicates that audio quality varies among different recordings. Not only could officially released versions (i.e., video No.2) have poor audio quality, videos with better audio quality could often be outranked by those with poorer quality. Because most users prefer live music videos with better audio quality, it is important to account for it when searching for live music. Although user

[1]http://www.youtube.com/
[2]http://www.youku.com/
[3]http://www.nicovideo.jp/

feedback may indirectly reflect audio quality (videos with better quality are preferred by users, leading to higher view counts and ratings), videos newly added to the database have little or no user feedback information and are thus given low rankings despite their superior audio quality. Therefore, a content-based audio quality assessment strategy is in great demand for long-tail retrieval of live music videos.

Previous research on perceptual assessment of sound quality started in the early 1990s. Most published regulations can be found in the ITU (International Telecommunications Union) Recommendations, which cover various assessment methods for both audio quality (e.g., perceptual evaluation of audio quality (PEAQ) [16, 31–33]) and speech quality (e.g., perceptual evaluation of speech quality (PESQ) [18]). These approaches generally compare the quality of the sound signals processed/affected by a test system (e.g., a multimedia device, codec, and telecommunication network) with that of a reference signal in order to evaluate or improve the performance of the system. However, this so-called reference-based audio quality assessment may be inadequate for live music videos. Because recordings of the same performance can be taken under vastly different circumstances, it is seldom possible to find an appropriate reference video.

In this paper, we propose an automatic, non-reference audio quality assessment method for live music video search online. We define audio quality as a subjective metric that describes music audio content. As music is performed, recorded, and then perceived by users in a sequential manner, audio quality can also be assessed from different perspectives along this process. We thus summarize six aspects to assess audio quality, including "instrumental" and "vocal" aspects for the live performance, "environmental" and "equipment" aspects for the recording conditions during the concert, and "pleasantness" and "overall quality" aspects for the end-user perception of the recording.

Because the assessment procedure can be very subjective, we formulate the assessment task as a ranking problem and try to solve it using a learning-based scheme. There are two advantages for assessing audio quality in this way. First, a ranking mechanism (i.e., by saying that song A has better quality than song B) better controls the subjectivity than mere classification (i.e., by saying that song A has good quality while song B has poor quality). Second, the ranking-based assessment method can be directly applied to re-rank the "relevant" live music recordings for a query retrieved by video search engines.

To implement our framework, we first obtain human annotations on all the said six aspects for a set of live recordings downloaded from YouTube. Then, several learning-to-rank (LTR) models are trained using the annotated data to achieve effective online re-ranking of music video search results. The effectiveness of the proposed framework is demonstrated through objective cross-validation. We also create a large database of synthetic recordings to expand the scope of our experiments.

There are four main contributions in this paper:

- To the best of our knowledge, this work presents the first attempt to develop automatic audio quality assessment for online live music recording retrieval.
- We establish a collection of 500 live music recordings with human annotations obtained via a web-based interface and a database of 2,400 synthetically altered live music recordings in 8 different quality conditions.
- We formulate the audio quality assessment task as a ranking problem and tackle it using LTR approaches with various audio feature sets.
- We conduct both objective and subjective evaluations to demonstrate the effectiveness and the usability of the proposed framework.

Beyond the field of multimedia information retrieval, our work can make meaningful contribution to the following topics. First, audio quality assessment can function as an additional clip selection criterion for creating better live concert video mashups [28]. Second, as an auditory characteristic of a music piece, audio quality can be readily integrated into music similarity measures [35]. Consequently, our work can also facilitate the re-ranking stage in general music retrieval and recommendation services.

The rest of the paper is structured as follows. Section 2 reviews related literature on audio quality assessment. Section 3 outlines our proposed framework. Section 4 details our method from data collection, audio feature extraction, to model learning and testing. Objective and subjective evaluations are presented in Sections 5 and 6, respectively. Section 7 concludes our work and explores future directions.

2. RELATED WORK

Initially, the assessment of sound quality was carried out through subjective tests only [14,17], in which subjects rated the overall quality of a test sound (distorted signal) against the reference sound (original signal) using a five-point score based on the ITU-RBS.1284 standard [15]. For example, sounds with the quality ranging from very annoying to imperceptible are scored from 1 to 5. Although subjective test yielded reliable results, it was expensive and time-consuming. Therefore, methods were proposed to assess sound quality objectively and automatically.

At first, the objective methods compared the reference sound and the test sound using traditional measures developed purely from engineering principles (e.g., signal-to-noise ratio and total harmonic distortion). However, their performance was no match against that of methods incorporating the psychoacoustic characteristics of human auditory system. Moreover, as more non-linear and non-stationary distortions appear, the shortcomings of these algorithms become more evident.

To emulate the subjective assessment process, researchers constructed perceptual models by taking into account multiple psychoacoustic phenomena (e.g., absolute hearing threshold and masking) of human auditory system. For example, Karjalainen [22] was one of the first to use the auditory model, such as the noise loudness, for sound quality assessment. Brandenburg explored the level difference between the noise signal and the mask threshold and then proposed a noise-to-mask ratio for audio quality assessment [2,3]. Brandenburg's method was later extended to include the mean opinion scores [29, 30]. These efforts eventually led to the standardization of perceptual evaluation of audio quality (PEAQ) [16, 31–33] and of speech quality (PESQ) [18].

PEAQ performs quite well on most of the test signals [32, 33]. However, it mainly focuses on low-bit-rate coded signals with small impairments. Therefore, recent research

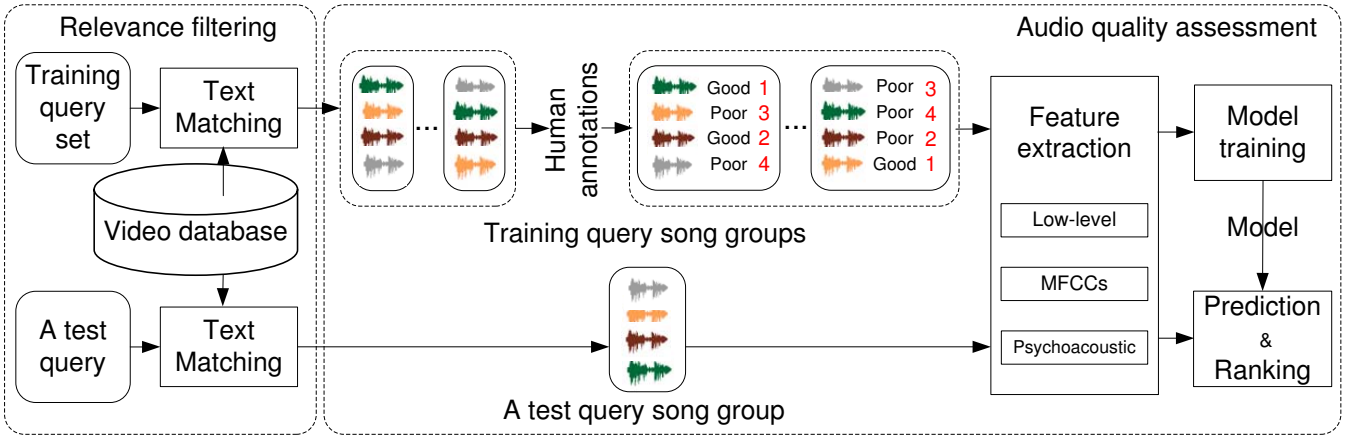

Figure 1: The system framework.

has honed PEAQ in several aspects. Barbedo and Lopes [1] developed a new cognitive model to map the output perceptual models to subjective ratings. Huber and Kollmeier [13] proposed a novel audio quality assessment method and extended the range of distortions on speech and music signals. More work is summarized in [4, 7, 32].

However, the ultimate goal of all the work mentioned above was to develop, test, or compare multimedia devices, codecs and networks for high-end audio or video services (e.g., VoIP and telepresence services) [4, 7, 32]. Moreover, both the reference signal and the distorted signal processed by the test system were available and generally well aligned. Doets and Lagendijk [8] studied the relationship between the parameters extracted from audio fingerprintings and the perceptual quality of compressed audio. However, the reference signals and the specific types of noises that may degrade the quality are required.

Recently, non-reference approaches have also been developed to assess the perceptual quality non-intrusively. Most work is related to quality assessment of speech, image or video signals. However, there is very little published work on non-reference audio quality assessment. For example, Malfait *et al.* [26] studied a non-reference quality assessment algorithm for speech, but it in fact relied on a semi-reference obtained from the reconstructed speech signal. Rix *et al.* [27] reviewed more work on non-reference quality assessment of speech. Hemami and Reibman [12] reviewed the background and related work in designing effective non-reference quality estimator for images and videos. Kennedy and Naaman [23] proposed a system that can manage and create a high quality concert video mashup by employing the audio fingerprintings of different video clips of the same event. They found that the most selected video clips tend to have higher audio quality. However, no systematic study was conducted for audio quality assessment in their work. Saini *et al.* [28] evaluated the quality of the visual information in live performance videos for creating a good quality mashup.

In the case of live music recordings created by common users, existing methods face two major difficulties. First, the audio quality of different recordings for the same live performance can be significantly different owing to various factors, such as the recording location/environment, starting point, duration, and recording device. Among the live music video recordings for a specific live concert, it is difficult to set up a reference due to the complex characteristics caused by the artist's performance, the environment of the concert venue, and the subjectivity of user perception. Second, unlike speech and studio versions of music, live music signals contain numerous and complex layers of information that may influence the assessment of audio quality (e.g., singing voice, various instrumental sounds, and unpredictable background noises). We hope that, by learning from annotated data, the learning-to-rank (LTR) algorithms can automatically assess the audio quality and predict the ranking of any live music recordings.

3. FRAMEWORK OVERVIEW

Our framework is based on how people assess the audio quality across multiple "relevant" recordings of the same song. We implement our framework in both the training and testing phases with a two-stage process (Figure 1), namely *relevance filtering* and *audio quality assessment*. Relevance filtering aims to identify the relevant songs, which should be live versions of the query song, regardless of audio quality. The second stage assesses the audio quality of each relevant song and produces a quality-based ranking.

3.1 Relevance Filtering

Searching music videos by metadata, such as artist name, song title and version type (e.g., official or live), is the most common and established method in general video retrieval systems, since most users tend to mark correct and sufficient metadata when they upload the music videos. In turn, the descriptive metadata can be used to accurately identify the live music videos relevant to a text-based query.

In this work, we use a text query with the format "artist_name song_name live" to search for live versions of a specific song on YouTube, assuming that this format can filter out most of the irrelevant as well as non-live videos. Given such a song query, the group of live versions returned is termed a *query song group*. At present, some manual efforts are also made to further check the returned recordings to ensure that they are relevant live versions to the given query. Our future work will make this filtering process fully automatic.

3.2 Audio Quality Assessment

In the training phase, we apply the learning-to-rank (LTR) algorithm to learn the model for audio quality assessment.

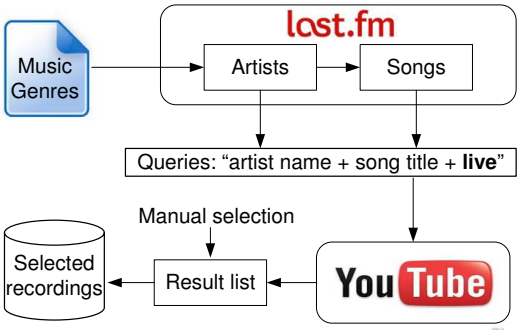

Figure 2: The diagram of collecting the live music videos.

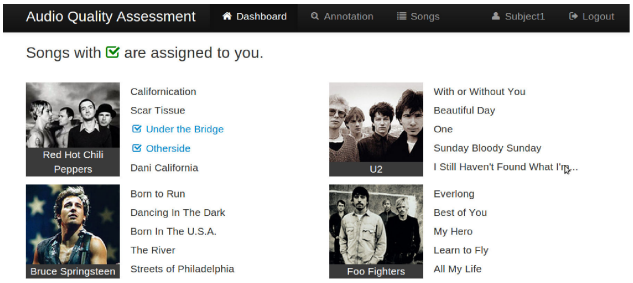

Figure 3: A snapshot of the song assignment for human annotation.

We first create a set of training song queries and then generate the query song groups via relevance filtering. Next, subjects are invited to annotate each query song group (see Section 4.2) by rating the audio quality of each component version based on the six aspects of audio quality as well as ranking the order among all versions in a query song group. With the annotated data, the LTR algorithm learns a global model that considers the ranking relationship among the component versions of each training query song group. We will explain the rationale and realization of this learning scheme in Section 4.4.

In the testing phase (i.e., the proposed audio quality assessment system), relevant filtering first generates the query song group for a test query. Then, the learned LTR model predicts the ranking score for each component version in the query song group (see Section 4.4 for details). The system then outputs a ranked list based on these scores, with the versions in descending order in terms of audio quality. This way, the assessment stage can be seen as a re-ranking over the relevance filtering results.

Both training and testing data undergo the feature extraction procedure (see Section 4.3). Various audio features sets, such as low-level acoustic features, MFCCs, and psychoacoustic features, are extracted from all recordings. For efficient audio quality assessment online, feature extraction can be performed offline beforehand.

4. THE PROPOSED APPROACH

As there is no benchmark dataset available for the proposed application, our work starts with collecting the live music database and the associated annotations. Then, we discuss several audio feature sets and present the training and testing procedures of the employed LTR algorithms.

4.1 Data Collection

For objective performance study, we have constructed two annotated datasets, one human-annotated (ADB-H) and the other synthetically generated (ADB-S). For subjective evaluation, we have also created a non-annotated (NDB) dataset.

Figure 2 outlines the live music data collection process and also illustrates the relevance filtering stage in more detail. To create a diverse dataset, we choose four genres of music (i.e., rock, pop, electronic, and country) that tend to have more live recordings of concerts. For each genre, we select a number of popular artists with their signature songs by finding each artist's top tracks on Last.fm API.[4] We then search for live music recordings on YouTube using the query format "artist_name song_name live" and manually select the most relevant results.

For ADB-H, we generate 100 song queries (4 genres, 5 artists per genre, and 5 songs per artist) on YouTube and select 5 relevant versions of each song query from its top 20 results.[5] Efforts are also made to ensure that the 5 versions vary in their audio quality. ADB-H thus contains $100 \times 5 = 500$ live music recordings.

For ADB-S, we generate 200 song queries (4 genres, 5 artists per genre, and 10 songs per artist) in addition to the 100 queries used in ADB-H. The artists for these 200 song queries are different from those in ADB-H. Moreover, unlike ADB-H, we only select the relevant version with the best audio quality (denoted as the "clean version"). As will be detailed later, each of these $100 + 200 = 300$ clean versions are subsequently used to generate 7 additional noisy versions. ADB-S thus contains $300 \times (7 + 1) = 2,400$ items.

For NDB, we generate 100 queries (4 genres, 5 artists per genre, and 5 songs per artist) from a different group of artists. Then, ten versions among the top 20 search results are randomly selected for each query song, regardless of their relevance and live-ness. NDB thus contains 1,000 non-annotated music pieces.

4.2 Data Annotation

4.2.1 ADB-H

To obtain the ground truth for ADB-H, we build an audio quality assessment interface online (Figures 3, 4, and 5) and recruit 60 subjects with normal hearing for annotation. Each subject is assigned eight query song groups (see Figure 3) from two randomly chosen genres and performs two types of annotations within each group. For the first type, the visual channel of the recording is masked, and subjects rate the audio quality in the six perception aspects (Figure 4) – instrumental, vocal, environmental, equipment-related, pleasantness, and overall quality. The first five aspects are based on a scale of 1 to 5, and the overall rating is a binary choice. Only when the audio ratings are completed can subjects enable the visual channel (by clicking the "Toggle Visual" button) to rate the seventh aspect, visual quality (see Figure 4). Currently we only consider the audio-related quality aspects; the visual quality rating is intended for fu-

[4] `http://www.last.fm/api`

[5] It may happen that YouTube returns some non-live versions in the top 20 results.

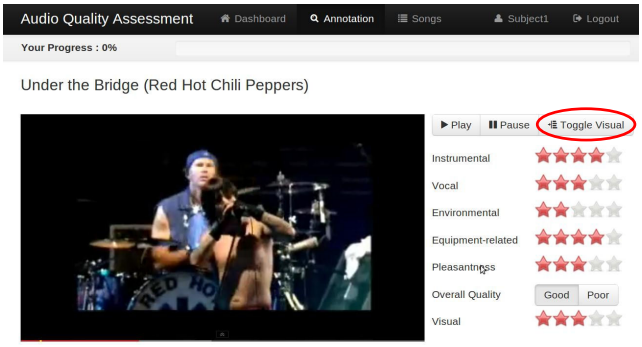

Figure 4: A snapshot of the interface for annotating the audio and visual quality.

Figure 5: The designed interface for labeling the ranking. The index of each live recording is dragged and dropped into a desired ranking position.

ture research. The second type of annotation takes place after all songs in the query song group have been rated. Subjects rank the overall audio quality of the five versions by a drag-and-drop mechanism (see Figure 5).

Since all live recordings used in our datasets are songs performed by their original artists, we assume that the subjects would not debate the merit of the performance itself. Therefore, subjects are reminded by the interface to neglect the intrinsic quality of the performance, e.g., whether the vocal is out-of-tune, cracked, off-beat, etc. In addition, we further clarify that, by "instrumental" and "vocal," we mean whether they sound clear, with the volume well-balanced, in the recording.

4.2.2 ADB-S

The significant time and manpower costs make it difficult to scale human-annotated datasets. We thus have created another annotated dataset by automatically generating various types of poor-quality recordings.

We applied the following noise effects, which are widely used in speech, TV, and movie productions [5], to the 300 clean versions.

- *Amplitude compression and then amplification*, where the compression uses the ratios of 4.47:1 for $|A| \geq -28.6$ dB, 0.86:1 for -46.4 dB $< |A| < -28.6$ dB, and 1:1.61 for $|A| \leq -46.4$ dB ($|A|$ is the absolute amplitude value), and the amplification gains the volume of the entire song with a ratio of 1:4. This simulates clipped live music sound produced by non-professional recordings devices whose dynamic range may not be able to adapt to the extremely loud musical sounds in the concert.

- *Band-pass filtering* using a second order Butterworth filter with cut-off frequencies at 100Hz and 6000Hz. This can simulate a muffling effect.

Table 2: Applied noise effects and audio quality rankings (averaged across ten subjects) of the 8 synthetic versions in a query song group in ADB-S (AQ = 1: Good, 0: Poor; Ord = quality rank).

Version	1	2	3	4	5	6	7	8
Comp & Amp		√				√		
Band-pass			√				√	
White noise				√				√
Crowd noise					√	√	√	√
AQ	1	0	0	0	0	0	0	0
Ord	1	4	2	3	5	8	6	7

- *White noise addition* with a maximum magnitude of 512 quantization steps. White noise is a signal with an approximately uniform (constant) power spectral density. This effect simulates the noise generated by the recording devices.

- *Crowd noise addition* with real-life noises from the concert audience, such as clapping, cheering, singing, and screaming. We collect 30 crowd noise samples from Freesound,[6] a collaborative sound database online, using queries "concert noise," "cheering," and "screaming." A set of 30 audio clips of crowd noises are uniformly added into each clean version with the signal-to-noise ratio of 1:1.

Since live music recordings generally contain noise from multiple sources, we also experiment with certain combination of noise effects to simulate the poor-quality examples. Seven poor-quality versions (labeled "Poor") are generated for each clean version (labeled "Good"). Each query song group in ADB-S thus contains 8 component versions. To assess which noise effect degrades audio quality more, ten subjects are invited to rank the order of the 7 poor-quality versions for 20 randomly chosen query songs (see Table 2).

4.3 Audio Feature Sets

We consider three types of audio feature sets in this work. The feature extractor is implemented based on MIRToolbox [24] and the methods proposed in [9].

- **Low-level features** (13 dim): The feature set includes root-mean-square, brightness, zero-crossing rate, spectral flux, rolloff at 85%, rolloff at 95%, spectral centroid, spread, skewness, kurtosis, entropy, flatness, and irregularity.

- **Mel-frequency cepstral coefficients** (39 dim): This feature set contains static MFCCs, delta MFCCs, and delta-delta MFCCs.

- **Psychoacoustic features** (20 dim): This feature set covers loudness, sharpness, roughness, and tonality features (key clarity, mode, harmonic change, and the normalized chroma weights).

For each audio feature set, we extract all the features with the same frame decomposition of 50ms and 50% hop sizes to ensure easy alignment, thereby obtaining a set of frame-level feature vectors for a live recording. We summarize the song-level audio feature representation for a recording by taking the mean and standard deviation of its frame-level vectors.

[6] `http://www.freesound.org/`

4.4 LTR Model Training and Testing

4.4.1 Notations and Rationale

Suppose that the annotated dataset $\mathcal{S}$ contains M query song groups, $\mathcal{S} = \{\mathbf{s}^{(i)}\}_{i=1}^{M}$, where each group has N_i component versions $\mathbf{s}^{(i)} = \{\mathbf{v}_j^{(i)}\}_{j=1}^{N_i}$, and each component version contains its corresponding audio feature vector $\vec{x}_j^{(i)}$ and rank (or relevance) label $y_j^{(i)}$, i.e., $\{\vec{x}_j^{(i)}, y_j^{(i)}\} \in \mathbf{v}_j^{(i)}$. The rank label y can be binary, numerical, or ordinal score.

Our principle for leaning the ranking model is that we only consider the rank relationship (obtained by $\{y_j^{(i)}\}_{j=1}^{N_i}$) among different versions $\{\mathbf{v}_j^{(i)}\}_{j=1}^{N_i}$ within a query song group $\mathbf{s}^{(i)}$. In other words, we do not care about the audio quality comparison between $\mathbf{v}_p^{(a)}$ and $\mathbf{v}_q^{(b)}$, where $1 \leq p \leq N_a$, $1 \leq q \leq N_b$, and $a \neq b$, because subjects are not asked to compare them.

Within a query song group, versions do not differ from each other in their *musical content*, such as melody, harmony, rhythm, and instrumentation. Live concert recordings of Britney Spear's "Hold it against me," for example, will have similar, if not identical, melody, harmony, and rhythm regardless of where or how the music is recorded. Because audio features extracted from the waveform can be related to musical content as well as audio quality, comparing versions of different songs would disproportionately highlight the differences in musical content. As a result, music content will vastly overpower audio quality in the learning process. Therefore, we only compare different versions of the same song (i.e. within a query song group) to control for the differences in music content features and to ensure that the algorithm learns the most from features related to audio quality. Hypothetically, if the algorithm can exclude features related to musical content, the effect of the discrepancy in audio quality can be enhanced regardless of the data's live-ness and relevance. In the testing phase, we conduct subjective evaluation using the NDB database, which may contain non-live and irrelevant videos alongside live and relevant ones, to verify this hypothesis.

4.4.2 Formulation

The goal of model training is to learn a global ranking function $f(\cdot)$ for the audio feature vector $\vec{x}$ of a live music recording. Given a training dataset $\mathcal{S}$, the learning objective can be generalized to the minimization of the following loss function L with respect to f,

$$L(f; \mathcal{S}) = \sum_{i=1}^{M} p(\mathbf{s}^{(i)}) \cdot l(f; \mathbf{x}^{(i)}, \mathbf{y}^{(i)}) + \lambda \|f\|, \qquad (1)$$

where $p(\mathbf{s}^{(i)})$ is the prior model for $\mathbf{s}^{(i)}$, $l(\cdot)$ is the query-level loss function, $\mathbf{x}^{(i)} = \{\vec{x}_j^{(i)}\}_{j=1}^{N_i}$, $\mathbf{y}^{(i)} = \{y_j^{(i)}\}_{j=1}^{N_i}$, and $\|f\|$ represents the regularizer for f. Such a learning objective can be solved by many existing LTR algorithms.

In the testing phase, given an arbitrary query song group $\mathbf{s}^{(*)}$, the system ranks the component versions by sorting the scores $\{f(\vec{x}_j^{(*)})\}_{j=1}^{N_*}$.

4.4.3 Learning-to-Rank Algorithms

According to their input/output representations and loss functions, learning-to-rank algorithms can be categorized into three groups [25] – pointwise, pairwise, and listwise – all of which can be used in our framework. From each category, we adopt the algorithm with superior performance in our preliminary study.

- **Pointwise (MART):** For this approach, the annotations are converted to numerical scores. MART [11] uses a general gradient descent "boosting" paradigm with multiple additive regression trees to minimize the objective function.
- **Pairwise (SVM-Rank):** For this approach, the annotations in each query song group are converted into the pairwise orders among versions. Support vector machine (SVM) is generally regarded as one of the most powerful supervised learning methods. SVM-Rank [20, 21] minimizes the loss function with respect to a weight vector by considering all the pairwise orders as constraints.
- **Listwise (AdaRank):** For this approach, the annotations in each query song group are converted into a ranked list. Similar to AdaBoost [10], AdaRank learns a number of "weak rankers" and their associated weights and makes prediction by linearly combining them. We utilize AdaRank [36] to directly optimize the ranking evaluation measures with a boosting scheme.

We utilize RankLib [6] for implementing MART and AdaRank and the SVM^{rank} tool [21] for SVM-Rank.

5. OBJECTIVE EVALUATION

The goals of our objective evaluation are twofold. First, we compare the effectiveness of different LTR algorithms in terms of ranking performance based on the overall quality aspect, denoted by LTR method comparison (LTRC). Second, we study the performance with different audio feature sets based on the six perception aspects, denoted by audio features comparison (AFC). For each goal, we use both ADB-H and ADB-S, leading to four evaluation tasks in total, as summarized in Table 3.

5.1 Audio Quality Labels

We exploit three types of ground truth labels in ADB-H and ADB-S: binary, ranking, and numerical. Binary and ranking labels describe the overall quality aspect, while numerical labels rate the other five perception aspects, i.e. instrumental, vocal, environmental, equipment-related, and pleasantness. For the binary labels, "Good" is assigned a value of 1 and "Poor" 0. Recall that ADB-H consists of ranking order annotated by the subjects, with each query song group containing 5 versions, and that ADB-S uses ranking order as shown in Table 3, with each query song group containing 8 versions. Ranking labels are assigned as the reverse of the ranking order within the query song group so that the label of the best-quality version (rank 1) has the highest assigned value (5 for ADB-H or 8 for ADB-S) and vice versa. The numerical label, which is only applicable to ADB-H, takes the average rating of each component version.

5.2 Methods Compared

We implement three LTR algorithms (i.e., MART, SVM-Rank, and AdaRank) and a baseline method (i.e., *Random*) in the experiment. The Random method generates a permutation randomly for ranking the versions in each test

Table 3: Summary table for objective evaluation. The task names, LTRC and AFC, stand for "LTR method comparison" and "audio feature comparison," respectively.

Task	Dataset	Label Type	LTR Methods	Features	Aspect	Result
LTRC 1	ADB-H	Binary Ranking	Random, MART, SVM-Rank, AdaRank	All features	Overall	Fig. 6
LTRC 2	ADB-S	Binary Ranking	Random, MART, SVM-Rank, AdaRank	All features	Synthetic	Fig. 7
AFC 1	ADB-H	Numerical Binary Ranking	SVM-Rank	Low-level, MFCCs, Psychoacoustic, All features	Instrument Vocal Environment Equipment Pleasantness Overall Overall	Fig. 8
AFC 2	ADB-S	Binary Ranking	SVM-Rank	Low-level, MFCCs, Psycho., All features	Synthetic	Fig. 9

query song group without accounting for their audio quality. We repeat the random permutation 10 times for each test query song group and calculate the average baseline performance. For each LTR algorithm, we perform a ten-fold cross-validation and calculate the average performance.

5.3 Performance Measure

The normalized discounted cumulative gain (NDCG) [19], a widely used metric in information retrieval, is adopted to measure the ranking performance. To calculate NDCG, the discounted cumulative gain (DCG) at a particular rank position p is first calculated in a way that penalizes the score gain near the bottom more than those near the top.

$$\text{DCG@}p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log_2(i+1)}, \tag{2}$$

where rel_i is the ground truth label of the recording at position i. It is then normalized so that the performance for each test query can be compared.

$$\text{NDCG@}p = \frac{\text{DCG@}p}{\text{IDCG@}p}, \tag{3}$$

where IDCG@p serves as the normalization term that guarantees the ideal NDCG@p to be 1. We summarize the performance by averaging the NDCGs over the test query set.

5.4 Result and Discussion

5.4.1 Comparison among LTR Algorithms

We first evaluate the effectiveness of each LTR algorithm holistically. All audio feature sets are concatenated into a single vector representation. The LTR models are trained with the binary or the ranking labels (both pertaining to overall quality) of either the ADB-H or the ADB-S dataset.

Figure 6 presents the average NDCG@5 on ADB-H. First, all LTR algorithms significantly ($p < 0.01$) outperform Random in all cases, demonstrating the effectiveness of our proposed approach. Trained with binary labels, MART, SVM-Rank, and AdaRank outdo Random by 11%, 17%, and 8%, respectively; with ranking labels, 16%, 17%, and 15%, respectively. Second, SVM-Rank achieves the best ranking performance of the three LTR algorithms. Interestingly, the performance difference between SVM-Rank and the other two are larger with binary labels than with ranking labels. This may be attributed to the difference of learning criteria

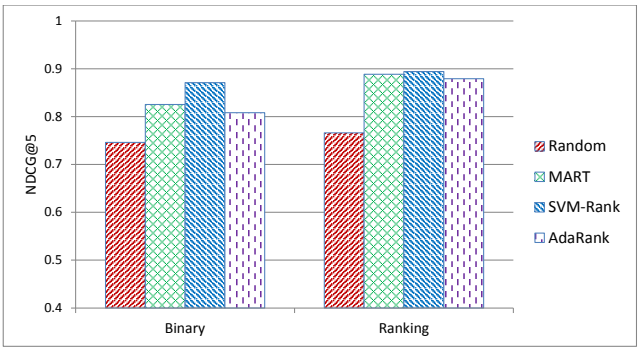

Figure 6: Performance based on overall quality using the binary and ranking labels on ADB-H.

among the three LTR approaches. For pairwise approach, the learning criteria do not compare training examples with the same binary labels, e.g., the approach does not need to discriminate the feature difference between two "Good" quality examples. While for the other two approaches, the feature difference between two "Good" quality examples are still taken into account, leading to certain randomness or contradiction during the learning process. For example, suppose a subject has the same binary labels in the case in Table 1, a listwise approach may convert the binary labels into #4>#5>#1>#6>#3>#2, which is much different from the ranking labels by human, #1>#5>#4>#2>#3>#6.

Figure 7 presents the the average NDCG@8 on ADB-S. MART, SVM-Rank, and AdaRank also significantly ($p < 0.01$) outperform Random with improvements of 31%, 93%, and 71% for binary labels and 31%, 51%, and 41% for ranking labels, respectively. Moreover, SVM-Rank achieves a remarkable performance of about 99% for both binary and ranking labels. This lends support to our hypothesis (Section 4.4.1) that minimizing the discrepancy in musical content features can magnify the learning potential with the audio quality features.

In sum, SVM-Rank achieves the best performance among the three LTR algorithms on both ADB-H and ADB-S. We thus adopt SVM-Rank in the following experiments.

5.4.2 Comparison among Different Audio Features

Using SVM-Rank, we compare the performance among different audio feature sets, namely low-level, MFCCs, psy-

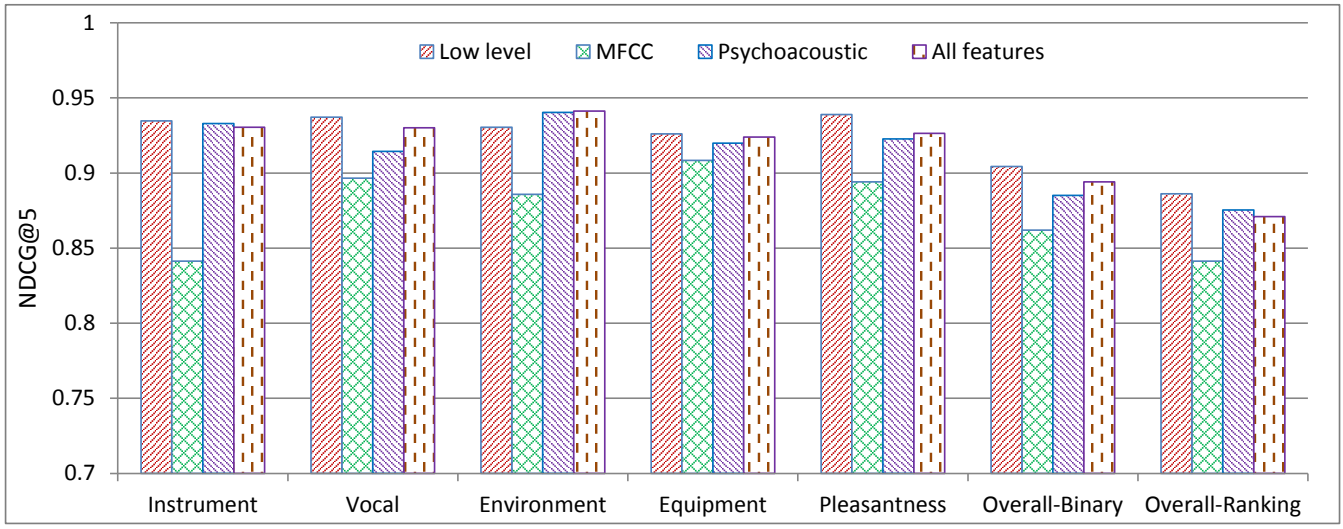

Figure 8: Performance on ADB-H using SVM-Rank with all types of labels and different audio feature sets. Except the overall quality aspect, the others are all numerical labels.

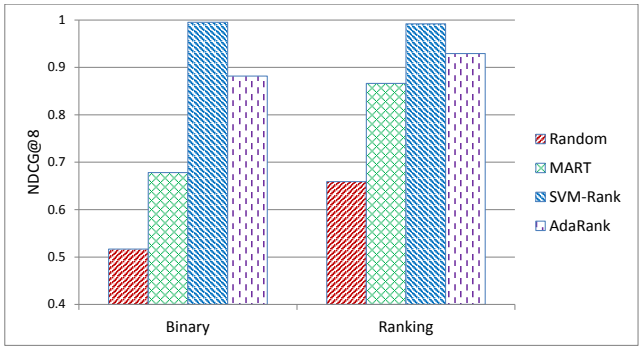

Figure 7: Performance based on overall quality using the binary and ranking labels of ADB-S.

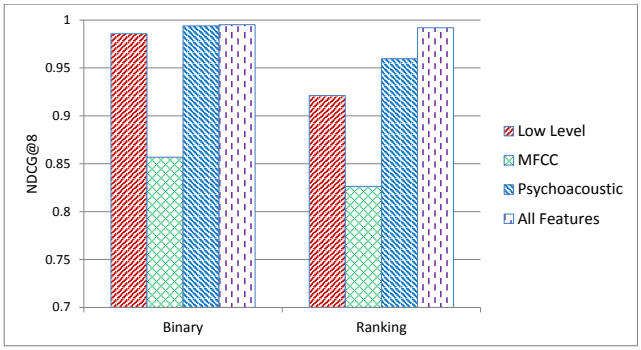

Figure 9: Performance of SVM-Rank on ADB-S using different audio feature sets.

choacoustic, and all three together. For ADB-H, we use all types of labels to train the SVM-Rank models, and Figure 8 shows the result. For ADB-S, we train SVM-Rank with the binary and ranking labels and show the result in Figure 9.

In almost all cases, we observe that both the low-level and psychoacoustic feature sets achieve better performance than MFCCs, as does the all-concatenated feature set. Recalling Section 4.4.1, we hope the learned model could neglect features related to musical content. Since MFCCs are developed to capture the spectral envelope based on human audition system, they tend to carry much more information about the musical content instead of the noise, and hence are less capable of identifying the difference in audio quality we are interested in. As shown in Figure 9, this phenomenon is more evident for ADB-S, in which all the noisy versions are generated directly from the clean one.

From Figure 8, we observe that the performance of the low-level and psychoacoustic feature sets with the numerical labels are very consistent with NDCG@5 $\approx$ 93%. Except in the case of environmental aspect, using low-level features always leads to superior performance, further demonstrating its effectiveness and robustness in this task.

6. SUBJECTIVE EVALUATION

To further evaluate the effectiveness and usability of the proposed approaches, a subjective study was carried out using NDB. Two SVM-Rank models trained with ADB-H and ADB-S (termed as SVM-Rank-H and SVM-Rank-S), respectively, are adopted to rank each query song group in NDB by audio quality. Their performance is evaluated in terms of (1) how users perceive the overall ranking quality of SVM-Rank compared to Random; and (2) how well SVM-Rank ranks the recordings with the best and the worst audio quality in a query song group.

6.1 Methodology

We recruit twenty subjects, all of whom are music lovers without auditory disorders. After a brief introduction to familiarize with the evaluation interface, each subject is asked to evaluate 10 test query song groups that are randomly chosen but represent all four music genres. For each of the first five groups (the first perspective), two ranked lists are presented in randomized order. One list is generated by Random, and the other is generated by SVM-Rank-H half the time and SVM-Rank-S half the time. Subjects listen through the two ranked lists and indicate the better one.

Table 4: Supporting Rates of SVM-Rank-H and SVM-Rank-S when comparing with Random.

	Rock	Country	Electronic	Pop	**Avg**
SVM-Rank-H	0.800	0.800	0.500	0.700	**0.720**
SVM-Rank-S	0.867	0.933	0.600	0.700	**0.800**

For each of the last five groups (the second perspective), they need to identify from a randomly permutated list the versions with the best and the worst audio quality. In summary, both SVM-Rank-H and SVM-Rank-S are evaluated with a total of 1,000 recordings (100 query song groups), half of which for overall ranking evaluation and the other half for best/worst audio quality ranking.

6.2 Performance Measure

The first perspective is measured by the *supporting rate* (SR) $SR_m = n_m/|S|$, where n_m is the number of test song groups on which method m outperforms Random, and $|S|$ is the total number of test song groups.

To measure the performance from the second perspective, we adopt both mean reciprocal rank (MRR) [34] and mean rank position (MRP).

$$MRR = 1/|S| \sum_{i=1}^{|S|} \frac{1}{rank_i}, \tag{4}$$

$$MRP = 1/|S| \sum_{i=1}^{|S|} rank_i, \tag{5}$$

where $rank_i$ denotes the rank of the best-quality version (or the worst-quality version) in the ranked list for the i-th query song group.

We denote the MRR for the best-quality and the worst-quality versions as MRR_b and MRR_w, respectively. A better ranking method would result in MRR_b closer to 1 and MRR_w closer to $1/N$, where $N = 10$ is the number of versions in a query song group. We calculate the MRR_b and MRR_w of Random, respectively, by using 100 random integers randomly generated between 1 and 10 as the ranking positions of all best-quality or worst-quality versions.

6.3 Results and Discussion

Table 4 presents the average SR of SVM-Rank-H and SVM-Rank-S when comparing with Random. The overall SR performance of SVM-Rank-H and SVM-Rank-S are significantly better than that of Random, clearly indicating that subjects are more pleased with the ranked lists generated by our approaches. The SRs in terms of different music genres show that both our approaches perform well for country and rock, but not for electronic. This is possibly due to the fact that electronic music tends to be intrinsically noisy and thus makes it more difficult for subjects to judge the differences in audio quality.

Table 5 presents the MRR performances of SVM-Rank-H and SVM-Rank-S. For Random, the calculated MRR_b and MRR_w are 0.301 and 0.299, respectively. Both SVM-Rank-H and SVM-Rank-S significantly ($p < 0.01$) outperform Random in the task of ranking the best/worst-quality versions. The MPRs of the best-quality and the worst-quality versions are 2.60 and 8.56, respectively, for SVM-Rank-H; and 2.70 and 7.78, respectively, for SVM-Rank-S. For Random, the MRPs are around 5.02 for both quality versions. These results indicate that our proposed approaches are able to rank the best-quality versions at higher positions and the worst-quality versions at lower positions. In terms of the four music genres, we observe that both SVM-Rank-H and SVM-Rank-S still perform very well for country music. Unlike the previous subjective evaluation perspective, however, identifying the best-quality and worst-quality versions of electronic music seems to have become relatively easier for subjects. Moreover, the performance variance among the four genres is also smaller.

Table 5: The performance of SVM-Rank-H and SVM-Rank-S for ranking the best-quality (MRR_b) and the worst-quality (MRR_w) versions.

	SVM-Rank-H		SVM-Rank-S	
	MRR_b	MRR_w	MRR_b	MRR_w
Rock	0.503	0.115	0.588	0.144
Country	0.698	0.116	0.703	0.188
Electronic	0.524	0.126	0.553	0.142
Pop	0.589	0.127	0.541	0.122
Avg	**0.574**	**0.122**	**0.586**	**0.146**

In summary, the results validate the effectiveness of our systems supported by evaluations using human perception and demonstrate the practical usability in ranking live music recordings according to audio quality. An interesting observation is that SVM-Rank-S generally performs slightly better than SVM-Rank-H. We have three plausible explanations for this. First, ADB-S (2,400 pieces) contains more training query song groups than ADB-H (500 pieces), which could benefit the model generalizability of SVM-Rank-H. Second, this result implies that our noise effects can reflect the real factors that degrade the audio quality of live music recordings. Third, the models learned from ADB-H may still involve the influence of musical content features. Since some query song groups in NDB may contain irrelevant recordings, the discrepancy in musical content may confuse their ranking prediction.

7. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel framework to assess audio quality for online live music search. Unlike previous reference-based audio quality assessment methods, our approach is non-referenced. We have established two music datasets of 500 annotated and 2,400 synthetic live music recordings for this study. We believe that the two constructed datasets can also serve as additional benchmark datasets for developing novel learning-to-rank algorithms in the machine learning research. Three LTR models with different audio feature sets were evaluated in terms of ranking performance based on different audio quality aspects. Our objective and subjective experimental results have shown that the proposed approaches can effectively rank live music recordings according to audio quality.

Our future work is fourfold. First, we will enlarge our datasets. Second, we will explore more audio features and learning algorithms to gain the effectiveness of audio quality assessment. Third, we will develop the segment-based approach, because the mean and standard deviation of the whole frame-level features may over-simplify the representation of the audio features of a recording. Fourth, we will in-

tegrate the proposed framework into general music retrieval and recommendation systems.

8. ACKNOWLEDGMENTS

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

9. REFERENCES

[1] J. Barbedo and A. Lopes. A new cognitive model for objective assessment of audio quality. *Journal of Audio Engineering Society*, 53(1/2):22–31, 2005.

[2] K. Brandenburg. A new coding algorithm for high quality sound signals. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 141–144, 1987.

[3] K. Brandenburg and T. Sporer. NMR and masking flag: Evaluation of quality using perceptual criteria. In *Proc. International AES Conference on Audio Test and Measurement*, pages 169–179, 1992.

[4] D. Campbell, E. Jones, and M. Glavin. Audio quality assessment techniques – A review, and recent developments. *Signal Processing*, 89(8):1489–1500, 2009.

[5] C.-Y. Chiu, D. Bountouridis, J.-C. Wang, and H.-M. Wang. Background music identification through content filtering and min-hash matching. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2414–2417, 2010.

[6] V. Dang. Ranklib – A library of learning to rank algorithms. [Online] `http://www.cs.umass.edu/~vdang/ranklib.html`.

[7] A. A. de Lima, F. P. Freeland, R. A. de Jesus, B. C. Bispo, L. W. P. Biscainho, S. L. Netto, A. Said, A. Kalker, R. Schafer, B. Lee, and M. Jam. On the quality assessment of sound signals. In *Proc. IEEE International Symposium on Circuits and Systems*, pages 416–419, 2008.

[8] P. J. O. Doets and R. L. Lagendijk. Extracting quality parameters for compressed audio from fingerprints. In *Proc. International Conference on Music Information Retrieval*, pages 498–503, 2005.

[9] H. Fastl and E. Zwicker. *Psychoacoustics: Facts and models*, volume 22. Springer, 2006.

[10] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[11] J. Friedman. *Greedy function approximation: A gradient boosting machine*, 1999.

[12] S. S. Hemami and A. R. Reibman. No-reference image and video quality estimation: Applications and human-motivated design. *Image Communication*, 25(7):469–481, 2010.

[13] R. Huber and B. Kollmeier. PEMO-Q – A new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions Audio, Speach, and Language Processing*, 14(6):1902–1911, 2006.

[14] International Telecommunications Union Recommendation (ITU-R) BS.1116-1. *Methods for the subjective assessment of small impairments in audio system including multichannel sound systems*, 1997.

[15] International Telecommunications Union Recommendation (ITU-R) BS.1284-1. *General methods for the subjective assessment of sound quality*, 1997–2003.

[16] International Telecommunications Union Recommendation (ITU-R) BS.1387. *Method for objective measurements of perceived audio quality*, 1998.

[17] International Telecommunications Union Recommendation (ITU-R) BS.1534-1. *Methods for the subjective assessment of intermediate quality level of coding systems*, 2003.

[18] International Telecommunications Union Recommendation (ITU-R) P.862. *Peceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speach quality assessment of narrow-band telephone networks and speech codecs*, 2001.

[19] K. Jarvelin and J. Kekalainen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[20] T. Joachims. Optimizing search engines using clickthrough data. In *Proc. the ACM International Conference on Knowledge Discovery and Data Mining*, pages 133–142, 2002.

[21] T. Joachims. Training linear SVMs in linear time. In *Proc. ACM International Conference on Knowledge Discovery and Data Mining*, pages 217–226, 2006.

[22] M. Karjalainen. A new auditory model for the evaluation of sound quality of audio systems. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 10, pages 608–611, 1985.

[23] L. Kennedy and M. Naaman. Less talk, more rock: Automated organization of community-contributed collections of concert videos. In *Proc. International Conference on World Wide Web*, pages 311–320, 2009.

[24] O. Lartillot and P. Toiviainen. A Matlab toolbox for musical feature extraction from audio. In *Proc. International Conference on Digital Audio Effects*, 2007.

[25] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

[26] L. Malfait, J. Berger, and M. Kastner. P.563-8212, the ITU-T standard for single-ended speech quality assessment. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1924–1934, 2006.

[27] A. Rix, J. Beerends, D. Kim, P. Kroon, and O. Ghitza. Objective assessment of speech and audio quality – Technology and applications. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1890–1901, 2006.

[28] M. K. Saini, R. Gadde, S. Yan, and W. T. Ooi. Movimash: online mobile video mashup. In *Proc. ACM International Conference on Multimedia*, pages 139–148, 2012.

[29] T. Sporer. Objective audio signal evaluation–applied psychoacoustics for modeling the perceived quality of digital audio. In *Audio Engineering Society Convention 103*, page 4512, 1997.

[30] T. Sporer, U. Gbur, J. Herre, and R. Kapust. Evaluating a measurement system. In *Audio Engineering Society Convention 95*, page 3704, 1996.

[31] T. Thiede. Perceptual audio quality assessment using a non-linear filter bank. In *PhD thesis, Fachbereich Electrotechnik, Technical University of Berlin*, 1999.

[32] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes. PEAQ - the ITU standard for objective measurement of perceived audio quality. *Journal of Audio Engineering Society*, 48(1/2):3–29, 2000.

[33] W. Treurniet and G. Soulodre. Evaluation of the ITU-R objective audio quality measurement method. *Journal of Audio Engineering Society*, 48(3):164–173, 2000.

[34] E. M. Voorhees. The TREC-8 question answering track report. In *Proc. Text Retrieval Conference*, pages 77–82, 1999.

[35] J.-C. Wang, H.-S. Lee, H.-M. Wang, and S.-K. Jeng. Learning the similarity of audio music in bag-of-frames representation from tagged music data. In *Proc. International Society for Music Information Retrieval Conference*, pages 85–90, 2011.

[36] Q. Wu, C. J. Burges, K. M. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3):254–270, 2010.