UvA-DARE (Digital Academic Repository)

When the Music Stops: Tip-of-the-Tongue Retrieval for Music

Bhargav, S.; Schuth, A.; Hauff, C.

Link to publication

When the Music Stops: Tip-of-the-Tongue Retrieval for Music

Samarth Bhargav*
University of Amsterdam
Amsterdam, Netherlands
s.bhargav@uva.nl

Anne Schuth
Spotify
Amsterdam, Netherlands
aschuth@spotify.com

Claudia Hauff
Spotify
Amsterdam, Netherlands
claudiah@spotify.com

ABSTRACT

We present a study of *Tip-of-the-tongue (ToT)* retrieval for *music*, where a searcher is trying to find an existing music entity, but is unable to succeed as they cannot accurately recall important identifying information. ToT information needs are characterized by complexity, verbosity, uncertainty, and possible false memories. We make four contributions. *(1)* We collect a dataset—ToT_{Music}—of 2,278 information needs and ground truth answers. *(2)* We introduce a schema for these information needs and show that they often involve multiple modalities encompassing several Music IR sub-tasks such as lyric search, audio-based search, audio fingerprinting, and text search. *(3)* We underscore the difficulty of this task by benchmarking a standard text retrieval approach on this dataset. *(4)* We investigate the efficacy of query reformulations generated by a large language model (LLM), and show that they are not as effective as simply employing the entire information need as a query–leaving several open questions for future research.

CCS CONCEPTS

• **Information systems** → **Music retrieval**; **Retrieval models and ranking**; **Document filtering**; *Multimedia and multimodal retrieval*; *Query reformulation*; • **Human-centered computing** → **User studies**.

KEYWORDS

Music Retrieval; Tip-of-the-Tongue Retrieval; Cross Modal Retrieval

1 INTRODUCTION

The *Tip-of-the-tongue (ToT)* retrieval task involves identifying a previously encountered item for which a searcher was unable to recall a reliable identifier. ToT information needs are characterized by verbosity, use of hedging language, and false memories, making retrieval challenging [1, 4]. As a consequence, searchers resort to communities like r/TipOfMyTongue and WatzatSong, where they can post descriptions of items that they know exist but cannot find,

*Work done during an internship at Spotify.

relying on other users for help. Recent research of ToT information needs explored how searchers pose these requests in specific domains like movies [1, 4], or games [24]. Music-ToT, however, is under-explored despite being frequent: it represents 18% of all posts made in a five-year period in the r/TipOfMyTongue community (cf. §3.1). Our work is motivated by the need to understand how such requests are expressed in the music domain.

We examined the r/TipOfMyTongue community, focusing on requests looking for musical entities like albums, artists or songs. We show that these requests often refer to multiple modalities (cf. §4) and thus encompass a broad set of retrieval tasks—audio fingerprinting, audio-as-a-query, lyric search, etc. In our work, we focus on song search. We create ToT_{Music}[1]: the dataset consists of 2,278 *solved* information needs pertaining to a song, each of which is linked to the corresponding correct answer in the publicly available Wasabi Corpus [7]. Using ToT_{Music}, we develop a schema for Music-ToT information needs to reveal what information is contained in them (cf. §3.2). In addition, we are interested in the extent to which standard text retrieval approaches are able to deal with ToT queries. To this end, we benchmark a subset of ToT_{Music} information needs[2] on the Wasabi corpus, as well as Spotify search. Across both settings, the low effectiveness—compared to non-ToT queries—of our evaluated retrieval methods underscores the necessity of novel methods to tackle this task. Lastly, we conduct a preliminary study on reformulating Music-ToT queries using GPT-3 [5]; we find that the task remains very challenging.

2 BACKGROUND

Tip-of-the-tongue (ToT) retrieval is related to known-item retrieval (KIR) or item-re-finding [31], however ToT queries are typically issued only once—not multiple times—and importantly, lack concrete identifiers, instead relying on verbose descriptions, frequently expressed uncertainty and possible false memories [1, 4, 16, 24]. Approaches for simulating such queries [3, 13, 26] may lack realistic phenomena like false memories [19, 20], necessitating the collection of real world data. Data on a large scale is available for only one domain, movies [4]; smaller scale datasets are available for games [24] and movies [1]. Hagen et al. [16] collect a corpus of general known-item queries, including music; however their focus was on general known-item queries and false-memories, and lacked retrieval experiments. Our focus is on the music domain, examining modalities employed by searchers and how they express Music-ToT queries. We build upon Arguello et al. [1] and Bhargav et al. [4], with key differences in (1) the domain—music, (2)

---

[1] ToT_{Music} (along with annotations) will be made available here: https://github.com/spotify-research/tot

[2] Concretely, 1.2K *descriptive* information needs not containing hyperlinks – we aimed to exclude posts where important information is not encoded in the text of the post itself.

the corpus size—millions of items instead of thousands, and, (3) reformulation experiments utilizing an LLM. Music-ToT relates to several research areas in Music IR (MIR).

Lyric- and text-based retrieval involves retrieving a song using lyrics or text [11, 28]. Techniques to handle *misheard* lyrics are common [30, 35–37], including modeling speech sounds [22], which may be insufficient, since ToT queries can contain *descriptions* of lyrics, requiring semantic methods [32], or utilizing the audio itself [38]. Apart from lyrics, Music-ToT queries are frequently free-form natural language queries (cf. §4), requiring methods that can retrieve audio using text, as well as tags, genre or human-generated descriptions [10, 12, 27, 29, 40].

Content-based audio retrieval [14] includes query-by-example (QBE) [21], where the audio is being queried as-is, e.g. audio fingerprinting [17]. Alternatively, users can *imitate* the wanted audio by vocalizing it, termed query-by-vocal-imitation (QBV) [25, 39], which includes query-by-humming (QBH) [15]. ToT queries frequently contain references to user created audio-clips as well as existing media like audio contained in videos (cf. §4).

Other modalities like videos may need to be handled as well, necessitating multi-modal or cross-modal (retrieving one modality using another) methods [33], e.g. retrieving audio using video [23, 34]. Approaches to solve Music-ToT have to account for multiple modalities and free-form natural language including noise, e.g., uncertainty [1] and/or false memories [1, 24].

3 METHODOLOGY

3.1 Data Collection

Gathering ToT_{All}. We gathered posts made across 2017-2021 in the r/TipOfMyTongue community, yielding 503,770 posts (after filtering out posts not marked *Solved* or *Open*), each containing two fields: *title* and *description*. We extracted text categories from the title, e.g. SONG from *"[SONG] Slow dance song about the moon?"*. We manually identified a set of 11 overarching *music-focused* categories (e.g. *Music Video*, *Band*, *Rap Music*). We discarded the remaining non-music posts, resulting in ToT_{All}: 94,363 (60,870 solved and 33,493 unsolved) Music-ToT posts. These posts form a large proportion—18.73%—of the 503K posts we started out with.

Extracting ToT_{Music}. We extracted answers from *Solved* posts following Bhargav et al. [4], retaining *Solved* posts which have a URL as an answer. If the URL points to a *track* on *Spotify*, obtaining the answer was trivial. Otherwise, the `title` portion of the markdown inline URLs, formatted as `[title](url)` (with title often formatted as 'Artist-Song') was used as a query to the *Spotify* search API. Since the API returns multiple results, we created a classifier[3] with 31 features based on the scores of the retriever, the edit distances between `title` and artist name, song title, etc. We used the classifier to predict if a `title` matches the track and artist, scoring 100% on precision on a held out set of 100 samples. Low-confidence candidates were filtered out. This left us with a set of 4,342 posts with *Spotify* tracks as answers. Lastly, we only retained those posts

where the ISRC[4] of the answer track is also present in the Wasabi Corpus [7]: a total of 2,278 posts. We call this collection ToT_{Music}.

Gathering reformulations. We gathered reformulations for all posts in ToT_{Music} by prompting GPT-3 [5][5] with the respective post description and a word count limit: `<description>` *Summarize the query above to* `<N>` *words, focusing on musical elements*. We used $N = \{10, 25, 50\}$.[6] We also employed a prompt without a specific word limit: `<post description>` *Shorten the query above, focusing on musical elements*.

3.2 Music-ToT Schema

Our annotation process involved three steps. We first developed and then refined a schema to describe Music-ToT information needs; in the final step, we annotated 100 samples from ToT_{Music}.

Developing the schema in 2 steps. A preliminary study conducted with one author (self-rated music expertise 7 out of 10) and two volunteers (music expertise 8/10 and 7/10 respectively) involved assigning one or more labels to 78 sentences from 25 randomly sampled posts from ToT_{Music}. We focused on developing new labels specific to Music-ToT, while also re-using labels from Arguello et al. [1]: specifically the *Context* labels, pertaining to the context an item was encountered in (*Temporal Context*, *Physical Medium*, *Cross Media*, *Contextual Witness*, *Physical Location*, *Concurrent Events*), and *Other* annotations (*Previous Search*, *Social*, *Opinion*, *Emotion*, *Relative Comparison*). The latter are generally applicable across ToT information needs. This preliminary study revealed 25 new music labels, in addition to 11 labels from prior work (6 × *Context* and 5 × *Other*). In the second step, the three authors (self-rated musical expertise 7, 6 and 5 respectively) of this paper labeled 110 sentences (20 posts from ToT_{Music}) to validate the schema. Based on our results and discussions, we combined a few finer-grained categories with low support into more general categories, e.g. specific musical elements like *Rhythm* / *Repetition*, *Melody*, *Tempo*, etc., were combined to *Composition*, resulting in **28 labels in total**.

Annotating. Lastly, in step 3, two authors employed the final schema to annotate 536 sentences corresponding to 100 posts. The resulting labels, their frequency, category, inter-rater agreement (Cohen's κ [2, 9]) along with their description and an example, are presented in Table 1.

4 DATA ANALYSIS

We now first discuss Table 1, followed by a brief discussion about the modalities present in the whole collection, ToT_{All}.

Annotation results. Among the music-focused annotations, *Genre* and *Composition*, a description of musical elements and how they fit together, are the two most frequent labels. This is followed by *Music Video Description*, and either direct quotes (*Lyric Quote*) or a description of the lyrics (*Story/Lyric Description*) further highlighting the different information needs that need to be addressed i.e.,

---

[3]Random Forest classifier, parameters selected with grid search on {10, 20, 30, 40, 50} estimators, max depth {2, 3, 4} and min/max scaled features.

[4]The international standard recording code (ISRC) is a standardized code for uniquely identifying recordings.

[5]Model: `text-davinci-003`, with temperature 0.7

[6]Based on manual inspection, we discarded $N = 5$ (too few words for a cohesive query, leading to crucial information being left out) and $N = 100$ (model hallucinations).

Table 1: Annotation Schema: Label, frequency of occurrence in 100 submissions / 536 sentences (F), annotator agreement (κ) and description of label, along with an example for each label.

	Label	F	κ	Description	Example
MUSIC ANNOTATIONS	Composition	87	0.74	Describes (part of) the composition of a piece of music including rhythm, melody, tempo, pitch, chords, notes, and keys; or how they are composed into a cohesive piece of music.	…playing the same *major-key pattern over each chord in a fairly simple repeating loop.*
	Genre	77	0.92	References a genre.	It sounded like a *reggae/ska* type beat
	Music Video Description	75	0.89	Describes a music video associated with a song.	*However, once the music starts, the store is lit up and the tone shifts completely as everything in that store has a pastel colour scheme.*
	Lyric Quote	65	0.89	Directly quotes lyrics that the user overheard, not including sounds / vocalizations	…it wasn't until he said something about the *"just somebody that I used to know"* song that I …
	Story/Lyric Description	60	0.71	Describes either the story conveyed by the lyrics, or the gist of the lyrics instead of directly quoting it.	The song is *a woman singing to/about a man that she was in love with and died, I think he was in the military and got killed and she had a baby at home?*
	Artist Description	54	0.92	Describes the artist.	He was maybe a tad *overweight, shaggy hair, maybe curly.*
	Time Period / Recency	49	0.89	References the time period the user thought the music was produced.	*Late 90s-early 2000s* hip hop song that sounds similar to clip
	Instrument	30	0.86	Mentions instruments that were overheard.	The guy performing was at a *keyboard/piano* …
	Vocals	28	0.69	Describes the voice or vocal type.	*High pitched but kind of floaty female vocals,* a bit …
	Name	23	0.81	Describes a song/artist/album name, what it resembles/contains, or what the searcher remembers of it.	…the name of the song was *brief, one nordic word.*
	Popularity	18	0.83	Describes the popularity of the music, artist, album or music video.	I'm surprised I can't find it since I can remember many specific lyrics, I guess it's *more obscure*
	Recording	15	0.80	A description or reference to user-created content	*I did a vocaroo of the tune, sorry about my voice and any possible background guinea pig noises:* URL
	Language / Region	14	0.92	Either mentions the language of the piece of music and/or references a particular region like state, country, etc.	A *Japanese* song that I don't remember any words to or how the tune goes at all,
	Album Cover	5	1.00	Describes the album cover.	*On the cover there was also a cyan teal line going along the bottom with white text in it.*
	Song Quality / Type	4	0.00	Describes the type of music (original/cover, live/recorded) or the production quality (professional, amateur, etc.)	*Live Cover* of All Along the Watchtower where …
CONTEXT et al. ANNOTATIONS	Uncertainty	162	0.79	Conveys uncertainty about information described.	*I don't know* what genre the song was, it was fairly calming and *I feel like* it couldve been on tiktok but *I don't really know.* Any help appreciated!
	Social	54	0.77	Communicates a social nicety.	*Any help appreciated!*
	Opinion	43	0.44	Conveys an opinion or judgment about some aspect of the music.	I don't remember the lyrics or title, only that *it was a kind of angsty teen* "I want to set the world on fire"
	Temporal Context	36	0.87	Describes when the music was heard, either in absolute terms or relative terms.	…I heard like in a billion videos *6 years ago.*
	Listening Medium	26	0.75	References the medium associated with the item. (e.g., radio, streaming service, etc)	I heard it on the *radio* a couple of times in …
	Embedded Music	26	0.58	References or describes extant media (e.g., Youtube / Twitch URL), including timestamps.	I do have a *video with the song (this video at around minute 4:21:* URL*)*
	Other Cross Media	26	0.19	Describes exposure to the piece of music through different media, excluding other Cross Modal labels	…I'm pretty sure was performed on one of the *early seasons of Glee or maybe Smash.*
	Previous Search	25	0.67	Describes a previous attempt to find the item, including negative results (i.e., it is not song X).	*I've tried humming it into shazam and other sites, looking up the two generic lyrics I remember, even doing those rhythm tapping things and nada*
	Relative Comparison	25	0.77	Describes a characteristic of the music in relative (vs. absolute) terms, by explicitly comparing it with another song / artist / album.	The melody I remember *resembles the beginning of the song "Run to the hills" by Metallica*
	Emotion	25	0.05	Conveys or describes how a piece of music made the viewer feel.	Even talking about it *makes me tear up.*
	Concurrent Events	18	0.09	Describes events relevant to the time period when music was encountered, but excluding descriptions of the music itself.	…when *I was driving down the country* but for the life of me can't remember the name.
	Physical Location	9	0.61	Describes physical location where music was encountered.	…record a 9 second portion of this song at *a Marriott hotel bar in downtown Chicago* …
	Contextual Witness	9	0.49	Describes other people involved in the listening experience.	A few years back, *a friend of mine* showed me an …

lyric search, text search and multi-modal search. However, a simple extraction of *Genre* and metadata such as *Time Period/Recency*, *Instrument*, etc., may not be useful without considering the most frequent label, *Uncertainty*. Search systems therefore would have to handle these elements, as well as consider potential false memories. Furthermore, annotations like *Social*, *Opinion* are also fairly common occurrences in our data, which may have limited utility for retrieval [1], motivating reformulations (cf. §3.1). Searchers also express their queries in terms of other music entities in a *Relative Comparison*, and describe *Previous Search* attempts, explicitly ruling out certain candidates. References to other modalities like user created clips (*Recording*) or existing media (*Embedded Music*) also pose a challenge. We now explore this challenge with a brief study of references to external content in the entire collection, ToT_{All}.

Cross-modal references Music-ToT, like other ToT domains, contains cross-modal and media references [1], where a searcher refers to external content. We here show that Music-ToT posts in particular contain such references frequently. To this end, we gathered frequent websites that appear in ToT_{All}. One author manually labeled these as one of: (1) *User Created*: a clip uploaded by a user, e.g., Vocaroo, Clyp.it, Google Drive, Dropbox, Instaudio, musiclab, Onlinesequencer, Streamable, Speakpipe. (2) *Extant Media*: a clip unlikely

to be uploaded by a user, e.g. an existing clip, corresponding to content/social media websites like Spotify, Twitch, Tiktok, or YouTube. (3) *Other URL*: Not belonging to the previous two categories. We find that *Extant Media* forms a larger proportion of queries (19K, 20.9%) compared to *User Created* queries (14K, 15.3%), with a small number of posts containing references to both types (1.1%). Therefore, Music-ToT information needs are inherently multi-modal. We characterize the remaining 57.7% of queries as *descriptive* queries, which include references to lyrics, or story descriptions (cf. §3.2). In summary, Music-ToT information needs are characterized by uncertainty and multi-modality, requiring methods like text-based audio retrieval, content based audio retrieval/fingerprinting and multi- or cross-modal retrieval.

5 BENCHMARKS

5.1 Experimental Setup

Corpora. We run experiments on two corpora. The first is the `Wasabi` 2.0 Corpus [6, 7]. It consists of 2M commercial songs from 77K artists and 200K albums. Crucially, (1) songs have the ISRC linked, enabling linking to data in *Spotify*; (2) it is an open dataset, consisting of rich information that includes lyrics, extensive metadata, and music snippets. We index the *Song Name*, *Artist Name* and *Lyrics*[7] of all songs using Elasticsearch (BM25 with default parameters). The second corpus corresponds to the *Spotify* US catalog, consisting of hundreds of millions of tracks. The *Spotify* search system [18] utilizes multiple retrieval stages (including lexical- and semantic search) and incorporates historic log data for retrieval purposes.

Queries. We conducted experiments on the 1,256 posts (849 train, 191 validation, and 216 test) from ToT_{Music} that contain no URLs in the post title or post text; we make this choice as in the most extreme case, the entire post may contain just a URL, requiring audio-based search while we focus on text-based methods. From each post, we create different *queries* and label them as follows: (1) `Title`: using the post title only; (2) `Text`: post text; (3) `Title+Text`: title & text concatenated; and finally, (4) `Keywords`: extracting up to ten keywords from the post text[8] with Yake [8]; (5) $Reform_N$: reformulations with $N = \{10, 25, 50, \infty\}$.

Evaluation. We report Recall@K, equivalent to Success@K (i.e., one correct answer) for $K = \{10, 100, 1000\}$ on `Wasabi`. All reported results are on the test set. For *Spotify* search we describe the observed trends (due to the proprietary nature of the system).

5.2 Results

Table 2 provides an overview of our `Wasabi` results.

Post parts as query. The low success across queries and K underscores the difficulty of the task. On `Wasabi`, `Title` queries are more effective than `Text` queries—increased verbosity leads to retrieval failure. However, the text may indeed contain data useful in retrieval, with comparable or higher effectiveness scores for `Title+Text` over `Title` at $K = \{100, 1000\}$, motivating keyword extraction: crucial details might be present in the text, but including

---

[7]We also experimented with other fields like *Album Title*, but saw no improvement in retrieval effectiveness.
[8]Keywords were deduplicated with threshold = 0.2 and algorithm =seqm.

Table 2: Overview of retrieval experiments on `Wasabi`, using Elasticsearch (BM25).

Query	S@10	S@100	S@1000
Title	0.0370	0.0833	0.1389
Keywords	0.0231	0.0463	0.0787
Text	0.0139	0.0648	0.0926
Title+Text	0.0324	0.0833	0.1713
$Reform_{10}$	0.0139	0.0509	0.1204
$Reform_{25}$	0.0278	0.0602	0.1389
$Reform_{50}$	0.0185	0.0741	0.1389
$Reform_{\infty}$	0.0139	0.0741	0.1574

the entire need as a query might harm effectiveness. Our keyword selection method though fails to outperform other queries except for `Text` on S@10.

On *Spotify* search we observe a different trend: `Title+Text` is the most effective query followed by `Title`.

LLM reformulations as query. Examining Table 2, reformulations have limited success compared to `Title` queries. $Reform_{25}$ and $Reform_{50}$ perform as well as `Title` on S@1000, with $Reform_{\infty}$ outperforming it. While `Keywords` beat all but $Reform_{25}$ on S@10, it is outperformed by reformulations on S@100 and S@1000. On *Spotify* search, we find that reformulations fare worse than `Title` queries for S@10, but see limited success on S@100, with $Reform_{25}$ and $Reform_{50}$ achieving higher effectiveness. Most importantly, there is no ideal N on either index, with varying success across metrics. We thus conclude that in our study, reformulations generated using state-of-the-art LLMs have only mixed success.

6 CONCLUSIONS

We explored Tip-of-the-Tongue retrieval for music. Of the 94K posts corresponding to Music-ToT information needs from an online community for ToT requests, we linked 2,278 posts to the corresponding answers in the Wasabi corpus, resulting in ToT_{Music}, thus enabling further research for this challenging task.

We iteratively developed and refined a Music-ToT schema that contains 28 fine-grained labels as shown in Table 1. Labeling 100 posts using this schema, we showed that users express uncertainty frequently, and almost as often refer to other modalities. We benchmarked a subset of 1.2K *descriptive* queries from ToT_{Music}, and highlight the difficulty of the task. Future work should leverage cross- and multi-modal retrieval as well as better approaches for reformulations.

REFERENCES

[1] Jaime Arguello, Adam Ferguson, Emery Fine, Bhaskar Mitra, Hamed Zamani, and Fernando Diaz. 2021. Tip of the Tongue Known-Item Retrieval: A Case Study in Movie Identification. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (Canberra ACT, Australia) *(CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 5–14. https://doi.org/10.1145/3406522.3446021

[2] Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics* 34, 4 (2008), 555–596.

[3] Leif Azzopardi, Maarten de Rijke, and Krisztian Balog. 2007. Building Simulated Queries for Known-Item Topics: An Analysis Using Six European Languages. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Amsterdam, The Netherlands) *(SIGIR '07)*. Association for Computing Machinery, New York, NY, USA, 455–462. https://doi.org/10.1145/1277741.1277820

[4] Samarth Bhargav, Georgios Sidiropoulos, and Evangelos Kanoulas. 2022. 'It's on the Tip of My Tongue': A New Dataset for Known-Item Retrieval. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (Virtual Event, AZ, USA) *(WSDM '22)*. Association for Computing Machinery, New York, NY, USA, 48–56. https://doi.org/10.1145/3488560.3498421

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[6] Michel Buffa, Elena Cabrio, Michael Fell, Fabien Gandon, Alain Giboin, Romain Hennequin, Fabrice Jauvat, Elmahdi Korfed, Franck Michel, Johan Pauwels, Guillaume Pellerin, Maroua Tikat, and Marco Winckler. 2020. *The WASABI Dataset and RDF Knowledge Graph.* https://doi.org/10.5281/zenodo.5603369

[7] Michel Buffa, Elena Cabrio, Michael Fell, Fabien Gandon, Alain Giboin, Romain Hennequin, Franck Michel, Johan Pauwels, Guillaume Pellerin, Maroua Tikat, and Marco Winckler. 2021. The WASABI Dataset: Cultural, Lyrics and Audio Analysis Metadata About 2 Million Popular Commercially Released Songs. In *The Semantic Web*, Ruben Verborgh, Katja Hose, Heiko Paulheim, Pierre-Antoine Champin, Maria Maleshkova, Oscar Corcho, Petar Ristoski, and Mehwish Alam (Eds.). Springer International Publishing, Cham, 515–531.

[8] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018. YAKE! Collection-Independent Automatic Keyword Extractor. In *Advances in Information Retrieval*, Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury (Eds.). Springer International Publishing, Cham, 806–810.

[9] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[10] SeungHeon Doh, Minz Won, Keunwoo Choi, and Juhan Nam. 2022. Toward Universal Text-to-Music Retrieval. *arXiv preprint arXiv:2211.14558* (2022).

[11] J. S. Downie and Sally Jo Cunningham. 2002. Toward a Theory of Music Information Retrieval Queries: System Design Implications. In *International Society for Music Information Retrieval Conference*.

[12] Benjamin Elizalde, Shuayb Zarar, and Bhiksha Raj. 2019. Cross Modal Audio Search and Retrieval with Joint Embeddings Based on Text and Audio. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4095–4099. https://doi.org/10.1109/ICASSP.2019.8682632

[13] David Elsweiler, David E. Losada, José C. Toucedo, and Ronald T. Fernandez. 2011. Seeding Simulated Queries with User-Study Data for Personal Search Evaluation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Beijing, China) *(SIGIR '11)*. Association for Computing Machinery, New York, NY, USA, 25–34. https://doi.org/10.1145/2009916.2009924

[14] Jonathan T Foote. 1997. Content-based retrieval of music and audio. In *Multimedia storage and archiving systems II*, Vol. 3229. SPIE, 138–147.

[15] Asif Ghias, Jonathan Logan, David Chamberlin, and Brian C Smith. 1995. Query by humming: Musical information retrieval in an audio database. In *Proceedings of the third ACM international conference on Multimedia*. 231–236.

[16] Matthias Hagen, Daniel Wägner, and Benno Stein. 2015. A Corpus of Realistic Known-Item Topics with Associated Web Pages in the ClueWeb09. In *Advances in Information Retrieval*, Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr (Eds.). Springer International Publishing, Cham, 513–525.

[17] Jaap Haitsma and Ton Kalker. 2002. A highly robust audio fingerprinting system.. In *Ismir*, Vol. 2002. 107–115.

[18] Helia Hashemi, Aasish Pappu, Mi Tian, Praveen Chandar, Mounia Lalmas, and Benjamin Carterette. 2021. Neural Instant Search for Music and Podcast. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Virtual Event, Singapore) *(KDD '21)*. Association for Computing Machinery, New York, NY, USA, 2984–2992. https://doi.org/10.1145/3447548.3467188

[19] Claudia Hauff, Matthias Hagen, Anna Beyer, and Benno Stein. 2012. Towards Realistic Known-Item Topics for the ClueWeb. In *Proceedings of the 4th Information Interaction in Context Symposium* (Nijmegen, The Netherlands) *(IIIX*

[20] Claudia Hauff and Geert-Jan Houben. 2011. Cognitive Processes in Query Generation. In *Advances in Information Retrieval Theory*, Giambattista Amati and Fabio Crestani (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 176–187.

[21] Marko Helén and Tuomas Virtanen. 2007. Query by example of audio signals using Euclidean distance between Gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, Vol. 1. IEEE, I–225.

[22] Hussein Hirjee and Daniel G Brown. 2010. Solving Misheard Lyric Search Queries Using a Probabilistic Model of Speech Sounds.. In *ISMIR*. 147–152.

[23] Sungeun Hong, Woobin Im, and Hyun S Yang. 2017. Deep learning for content-based, cross-modal retrieval of videos and music. *arXiv preprint arXiv:1704.06761* (2017).

[24] Ida Kathrine Hammeleff Jørgensen and Toine Bogers. 2020. *"Kinda like The Sims... But with Ghosts?": A Qualitative Analysis of Video Game Re-Finding Requests on Reddit.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3402942.3402971

[25] Bongjun Kim and Bryan Pardo. 2019. Improving content-based audio retrieval by vocal imitation feedback. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4100–4104.

[26] Jinyoung Kim and W. Bruce Croft. 2009. Retrieval Experiments Using Pseudo-Desktop Collections. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (Hong Kong, China) *(CIKM '09)*. Association for Computing Machinery, New York, NY, USA, 1297–1306. https://doi.org/10.1145/1645953.1646117

[27] A Sophia Koepke, Andreea-Maria Oncescu, Joao Henriques, Zeynep Akata, and Samuel Albanie. 2022. Audio retrieval with natural language queries: A benchmark study. *IEEE Transactions on Multimedia* (2022).

[28] Meinard Müller, Frank Kurth, David Damm, Christian Fremerey, and Michael Clausen. 2007. Lyrics-Based Audio Retrieval and Multimodal Navigation in Music Collections. In *Research and Advanced Technology for Digital Libraries*, László Kovács, Norbert Fuhr, and Carlo Meghini (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 112–123.

[29] Andreea-Maria Oncescu, A Koepke, Joao F Henriques, Zeynep Akata, and Samuel Albanie. 2021. Audio retrieval with natural language queries. *arXiv preprint arXiv:2105.02192* (2021).

[30] Nicholas Ring and Alexandra L. Uitdenbogerd. 2009. Finding 'Lucy in Disguise': The Misheard Lyric Matching Problem. In *Information Retrieval Technology*, Gary Geunbae Lee, Dawei Song, Chin-Yew Lin, Akiko Aizawa, Kazuko Kuriyama, Masaharu Yoshioka, and Tetsuya Sakai (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 157–167.

[31] Sargol Sadeghi, Roi Blanco, Peter Mika, Mark Sanderson, Falk Scholer, and David Vallet. 2014. Identifying Re-Finding Difficulty from User Query Logs. In *Proceedings of the 2014 Australasian Document Computing Symposium* (Melbourne, VIC, Australia) *(ADCS '14)*. Association for Computing Machinery, New York, NY, USA, 105–108. https://doi.org/10.1145/2682862.2682867

[32] Shoto Sasaki, Kazuyoshi Yoshii, Tomoyasu Nakano, Masataka Goto, and Shigeo Morishima. 2014. LyricsRadar: A Lyrics Retrieval System Based on Latent Topics of Lyrics.. In *Ismir*. 585–590.

[33] Federico Simonetta, Stavros Ntalampiras, and Federico Avanzini. 2019. Multimodal Music Information Processing and Retrieval: Survey and Future Challenges. In *2019 International Workshop on Multilayer Music Representation and Processing (MMRP)*. 10–18. https://doi.org/10.1109/MMRP.2019.00012

[34] K. Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. 2016. A Comprehensive Survey on Cross-modal Retrieval. *ArXiv* abs/1607.06215 (2016).

[35] Xin Xu and Tsuneo Kato. 2012. Robust and Fast Two-Pass Search Method for Lyric Search Covering Erroneous Queries Due to Mishearing. In *Computational Linguistics and Intelligent Text Processing*, Alexander Gelbukh (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 306–317.

[36] Xin Xu, Masaki Naito, Tsuneo Kato, and Hisashi Kawai. 2009. Robust and Fast Lyric Search based on Phonetic Confusion Matrix.. In *ISMIR*. 417–422.

[37] Hongliang Ye, Wanning Zhu, Yue Yu, and Lei Hong. 2020. A Cross-language Music Retrieval Method by Using Misheard Lyrics. In *2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, Vol. 9. 1792–1795. https://doi.org/10.1109/ITAIC49862.2020.9339157

[38] Yi Yu, Suhua Tang, Francisco Raposo, and Lei Chen. 2019. Deep Cross-Modal Correlation Learning for Audio and Lyrics in Music Retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 1, Article 20 (feb 2019), 16 pages. https://doi.org/10.1145/3281746

[39] Yichi Zhang and Zhiyao Duan. 2018. Visualization and interpretation of Siamese style convolutional neural networks for sound search by vocal imitation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2406–2410.

[40] Tiange Zhu, Raphaël Fournier-S'niehotta, Philippe Rigaux, and Nicolas Travers. 2022. A Framework for Content-Based Search in Large Music Collections. *Big Data and Cognitive Computing* 6, 1 (2022). https://doi.org/10.3390/bdcc6010023

'12). Association for Computing Machinery, New York, NY, USA, 274–277. https://doi.org/10.1145/2362724.2362773