EXPECTED RECIPROCAL RANK FOR EVALUATING MUSICAL FINGERING ADVICE

David A. RANDOLPH (drando2@uic.edu)[1], **Barbara DI EUGENIO**[1], and **Justin BADGEROW**[2]

[1]*Department of Computer Science*, **University of Illinois at Chicago**, IL USA
[2]*Department of Music*, **Elizabethtown College**, PA USA

ABSTRACT

We cast the computational modeling of musical fingering as an information retrieval (IR) problem in which the task is to generate an optimally ranked list of fingering suggestions for each phrase in a score. The audience for this list is a set of performers with potentially diverse fingering preferences. Specifically, we adapt the expected reciprocal rank (ERR) metric—proposed by Chapelle and associates as an improved evaluation metric for retrieving documents with graded relevance—to develop a set of novel metrics tailored to the piano fingering IR task. ERR, as originally described, relies on a heuristic function to estimate the probability that a user will be satisfied by a document with a particular graded relevance. For musical fingering, we instead estimate the likelihood that a given performer will deem a suggested fingering sequence sufficient for arriving at a satisfactory solution. Finally, we attempt to validate our specific use of ERR by comparing how it judges several competing models.

1. INTRODUCTION

Pianists typically encounter fingering advice as annotations on a static printed score. In some cases, especially for more difficult repertoire, pianists may own more than one editorial score to obtain a variety of fingering suggestions. A key advantage of an automated advice generation system, therefore, would be its ability to provide a variety of advice on demand. It is also a common feature of existing models to output ranked lists of fingering sequences.

Moreover, assessment methods for the published piano fingering models are inadequate. With few exceptions [1–3], the extrinsic evaluations performed on proposed models consider only a single authoritative source of "correct fingerings." The same can be said of how models of guitar fingering have been evaluated. This is a missed opportunity, as is affirmed by the variability in the domain described qualitatively by [1] and [4]. The existence of multiple ground truths must be acknowledged and accommodated in model evaluation.

Parncutt et al. [1] collect "preferred fingerings" from a set of 28 pianists with unreported hand dimensions or gender, ranking particular fingerings by how many pianists prefer them. Their computer model is deemed a success because, "In most cases, the most popular fingering [among pianists] is in the top 10 [selected by the model]." Jacobs [2] uses an identical evaluation technique, and boasts of an improvement because "more pianist's fingerings are now included in the top 10." Nakamura et al. [3] confront the problem squarely and suggest several methods for comparing automatically generated advice with single or multiple ground truths. These methods rely on perfect matches at each note position and contemplate evaluating only one automatically generated fingering sequence at a time. They also describe a "recombination match rate," which involves calculating a cost of "constructing" a generated fingering sequence by combining the ground truths that are available.

Our approach here is more straightforward and includes multiple system outputs as part of the evaluation framework, as seems appropriate for a domain acknowledged to contain multiple ground truths and for automated systems that should be expected to produce diverse fingering suggestions.

We therefore cast the development of fingering models as an information retrieval (IR) task. Given arbitrary musical input (a "query"), the system generates a list of the most relevant fingerings ("documents"), ordered optimally to satisfy the information need of the pianist ("user.") With this framing of the task, we draw on recent advancements in evaluation measures for IR systems to develop a set of novel evaluation metrics for piano fingering.

Here we demonstrate an application of ERR to piano fingering advice, but the general approach should be appplicable to any instrument where fingering decisions form an important part of skilled performance. At various times, accordianists, string players, and many percussionists may also be eager consumers of fingering (or sticking) advice. Other instruments (e.g., brasses and woodwinds) typically have few fingering choices, making computational models less relevant for them.

To simplify explication, in our application of ERR to the evaluation of piano fingering advice, we reduce the problem space to include only monophonic (melodic) musical phrases played with the right hand.

Open-source implementations of all methods described below are released as part of Pydactyl [5] at `https://github.com/dvdrndlph/pydactyl`. Full release of the corpora used in validation is forthcoming.

2. EXPECTED RECIPROCAL RANK

Specifically, we adapt expected reciprocal rank (ERR), proposed by [6] as an improved metric for search engines when retrieving documents with graded relevance, to the musical fingering IR task. ERR generalizes the Reciprocal Rank (RR) metric of [7], a simple assessment for documents with binary relevance: the quality of a list of documents in which the first relevant document appears at rank r is estimated as $\frac{1}{r}$. To calculate ERR(R) for a returned list of R documents, system evaluators define a function to estimate the probability that a user will be satisfied by a document with a certain graded relevance. Armed with these probabilities, it is then straightforward to determine the likelihood that a user will be satisfied after reviewing a document at each rank. These likelihoods, after being assessed a reciprocal-rank $\frac{1}{r}$ (or similar) discount, are summed to calculate ERR(R) for a list of length R. The basic intuition of ERR is summed up in Equation 1, as defined by [6]:

$$\text{ERR}(R) = \sum_{r=1}^{R} \frac{1}{r} P(\text{user stops at position } r). \quad (1)$$

The details for leveraging the individual probability estimates P_i (that each recommended document in the list will satisfy the user) to determine the likelihood that a user stops the search at rank r are conveyed in Equation 2 (with notation slightly altered from the original). The key point here is that the probability of stopping at a rank r is the probability imputed from the graded relevance of the document at this rank *reduced by the probability that this rank is never reached by the user*—that is, that a document ranked higher has already satisfied the information need. Through this property, the ERR measure reflects a "cascade user model," which has been applied effectively to explain user behavior when interacting with web search results [8]:

$$\text{ERR}(R) = \sum_{r=1}^{R} \frac{1}{r} \prod_{i=1}^{r-1} (1 - P_i) P_r. \quad (2)$$

Chapelle and associates illustrate their metric with a probability estimation function for documents assigned an integer relevance score $g \in \{0, \ldots, g_{max}\}$:

$$\mathcal{P}(g_r) = \frac{2^{g_r} - 1}{2^{g_{max}}}. \quad (3)$$

Thus, the graded relevance of g_r of a document at rank r, the prior probability of the acceptability of document r may be estimated:

$$P_r = \mathcal{P}(g_r). \quad (4)$$

For the musical fingering problem, we adapt the ERR metric to entail distinct methods for estimating the prior probability that a document—that is, a fingering sequence—will be acceptable to the user. Instead of relevance grading, we propose evaluating the quality of a suggestion according to its similarity to a gold-standard fingering sequence.

We assume that any acceptable system advice, in the vast majority of cases, will be *similar* to satisfactory fingering sequences developed by the performer independent of suggestions from third parties or, more generally, to some acceptable advice suggested by more expert pianists. By necessity, fingerings developed by humans form the basis of the expertise being modeled. Crucially, we do not expect performers to require each note's suggested fingering to be identical to the one they ultimately adopt for that note.

The simplest similarity measure between two fingering sequences for a phrase is Hamming distance. In this measure, we simply count the number of individual elements in a sequence that do not match exactly: The fewer mismatches, the more similar the strings. Insofar as previously published fingering models have discussed performance of their models, they appear to refer to "accuracy" as the percentage of exact matches at the level of individual fingerings. This measure makes no assumptions about the process underlying how fingering decisions are made. Normalizing for the length N of the phrase, we have the following grading function g for the fingering at rank r produced by system S:

$$g_r = \frac{\Delta(H, S_r)}{N}, \quad (5)$$

where H is the fingering sequence ultimately adopted by the human pianist (and included in a gold-standard corpus) and Δ is the Hamming edit distance function:

$$\Delta(A, X) = \sum_{n=1}^{N} \delta(A_n, X_n), \quad (6)$$

where

$$\delta(a, x) = \begin{cases} 0 & \text{if } a = x, \\ 1 & \text{otherwise.} \end{cases} \quad (7)$$

Normalizing the Hamming distance conveniently implies $0 < g_r < 1$, so we may use the following to estimate the probability $\mathcal{P}$ that advice at rank r will satisfy the user:

$$\mathcal{P}_r = 1 - g_r. \quad (8)$$

2.1 Modified Unigram Edit Distance

Motivated by the observation that not all fingering deviations are equally significant, however, we pose a modification to simple Hamming distance for use in the piano domain. Clearly, choosing the index finger (2) over the middle finger (3) in many cases might seem arbitrary to the player, or at the very least a substitution readily made as circumstances allow. As such, a 2-3 or 3-2 variation is less likely to damage the overall opinion one holds of a complete sequence. This intuition is supported by pedagogues who place fingers 3 and 2 on near parity. Per [9, p. 115], "The third finger is by nature the skilfullest and strongest. The style of touch which it possesses serves for a time as a standard for the other fingers. . . . The second is the next strongest and skilfullest. Its mobility is probably greater, but in strength it yields to the third."

Selecting the middle finger (3) over the ring finger (4), though perhaps more controversial, is still a relatively minor deviation when compared to substituting more remote

	1	**2**	**3**	**4**	**5**
1	0	1	1	1	1
2	1	0	$\frac{1}{2}$	1	1
3	1	$\frac{1}{2}$	0	$\frac{1}{2}$	1
4	1	1	$\frac{1}{2}$	0	1
5	1	1	1	1	0

Table 1. Confusion matrix for one-handed "adjacent long" weighted edit distance for piano.

fingers. Fingers 2, 3, and 4 are the longest fingers and as such they are adept at playing both black and white keys. While 3 is stronger than 4 typically, these fingers are used interchangeably frequently in certain patterns and reflecting different pianists' approach to standardized structures.

We therefore propose an alternative weighted edit distance, defined as a confusion matrix in Table 1, as a more appropriate $\delta(a, x)$ for piano. We refer to the Δ function applying this confusion matrix as the "adjacent long" edit distance function.

2.2 Trigram Edit Distance

The edit distance functions described above measure deviations between fingerings of individual notes. We have reservations about applying such unigram measures to a problem that is fundamentally about transitions between fingered positions (at least for keyboard and string instruments) and also about planning for future transitions. We therefore here pose alternatives, which measure consistency in two fingering sequences within a sliding window of three-note groups—τ trigram distance functions. The simplest (and most unforgiving) of such measures, described formally in Equation 9, requires exact matches for the finger used to play each note in the group:

$$\tau(A, X, n) = \begin{cases} 1 & \text{if eq}(A^n_{n-2}, X^n_{n-2}), \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

where

$$\text{eq}(abc, xyz) = \begin{cases} 0 & \text{if } a \neq x \text{ or } c \neq z \text{ or } b \neq y, \\ 1 & \text{otherwise.} \end{cases} \tag{10}$$

The added context of the trigram allows us, for piano, to incorporate the concept of adjacent long finger similarity discussed above with more precision. In our "nuanced" formulation, captured in Equation 11, we discount mismatches involving adjacent long fingers by some ε only for the middle note in a trigram and only when the surrounding notes are fingered identically. That is,

$$\tau(A, X, n) = \begin{cases} 0 & \text{if eq}(A^n_{n-2}, X^n_{n-2}), \\ 1-\varepsilon & \text{if sim}(A, X, n), \\ 1 & \text{otherwise,} \end{cases} \tag{11}$$

given equal(abc, xyz) is defined as above in Equation 10,

$$\text{sim}(A, X, n) = \begin{cases} 1 & \text{if eq}(A^n_{n-2}, X^n_{n-2}), \\ 0 & \text{if } A_{n-2} \neq X_{n-2} \text{ or } A_n \neq X_n, \\ 1 & \text{if proxy}(A_{n-1}, X_{n-1}), \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

and where

$$\text{proxy}(a, x) = \begin{cases} 1 & \text{if } a \in \{2, 3\} \text{ and } x \in \{2, 3\}, \\ 1 & \text{if } a \in \{3, 4\} \text{ and } x \in \{3, 4\}, \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

with $0 \leq \varepsilon \leq 1$. Clearly, the decision made for the middle note in a trigram has essentially no bearing on the fingerings outside the scope of the trigram when the outer two notes are fingered identically. A pianist should be quite forgiving of such deviations and be able to substitute their own preferences easily.

We suspect that pianists will be similarly accepting of these similarities, bracketed as they always are by fingerings with which they completely agree, when they appear in other positions (first or third) in other trigrams. We therefore pose one final "relaxed" τ distance function:

$$\tau(A, X, n) = \begin{cases} 0 & \text{if equal}(A^n_{n-2}, X^n_{n-2}), \\ 1-\varepsilon & \text{if simat}(A, X, n-2) \text{ and} \\ & \quad \text{simat}(A, X, n-1) \text{ and} \\ & \quad \text{simat}(A, X, n), \\ 1 & \text{otherwise,} \end{cases} \tag{14}$$

where

$$\text{simat}(A, X, m) = \begin{cases} 1 & \text{if } A_m = X_m \text{ or} \\ & \quad \text{sim}(A, X, m+1), \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

Note that any of these τ functions may be used in place of the δ functions included in Equation 5, like so:

$$\Delta(A, X) = \sum_{n=1}^{N} \tau(A^n_{n-2}, X^n_{n-2}). \tag{16}$$

Note that each note is evaluated in its full trigram context. That is, each individual note contributes to three trigram evaluations, with the first and last two evaluations including referencing positions outside the scope of the note sequence. Such blank or null values are considered equal when compared. To check everything completely, N must be increased by two for all calculations involving trigram measures. Table 2 demonstrates how the summation in Equation 16 aggregates for each iteration (1–9) over the seven-note phrase using the three candidate trigram functions. Here we assume $\varepsilon = 1$, but using a number slightly less than one (such as 0.99, as we do in the experiments described below) affords better differentiation.

3. GENERAL PROBABILITY FORMULATION

We apply this measures into our estimate for the probability that system S advice at position r in a ranked list will

n	**-1**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
H	␣	␣	2	5	3	5	2	3	1	␣	␣
S_r	␣	␣	3	5	4	5	3	4	2	␣	␣
Match	$=$	$=$	$\approx$	$=$	$\approx$	$=$	$\neq$	$\neq$	$\neq$	$=$	$=$
τ_{trigram}			1	2	3	4	5	6	7	8	9
τ_{nuanced}			1	1	2	2	3	4	5	6	7
τ_{relaxed}			0	0	0	0	1	2	3	4	5

Table 2. The calculation of $\Delta(H, S_r)$ using the competing τ trigram distance functions ($\varepsilon = 1$).

satisfy the human who prefers fingering sequence H like so:

$$\mathcal{P}_r = 1 - \frac{\Delta(H, S_r)}{N}. \tag{17}$$

We attempt to validate the utility of Equation 17 with various Δ definitions in the experiments described below.

4. EVALUATING THE EVALUATIONS

Chapelle et al. [6], as a preamble to the evaluation of their ERR metric applied to web search: "The evaluation of new metrics is challenging because there is no ground truth to compare with. Because of that, most papers that propose new metrics do not have direct evaluations." They proceed to defend the application of ERR to web search by leveraging click-through data as a surrogate for what they would ideally have—namely, actual people interacting with lists comprised only of documents with known grades. We find ourselves in a similar position here, as we do not have data from a set of pianists interacting with fingering advice from competing models. What we have are passages fingered by pianists and several model implementations.

4.1 Corpora

Our initial validation efforts here are performed using the fragments from Czerny's challenging *160 Kurze Übungen* [10] that form the basis for evaluation of the original piano fingering model [1]. Parncutt and associates [1] publish fingering data for the opening notes of seven exercises, ranging from four to eight notes in length. Of the 28 pianists who provided fingerings, 25 "performed regularly on a regular basis" and three were enrolled in an "undergraduate music program." They report a "mean total number of years practicing and/or performing = 31" for participants. Data from this cohort constitute what we call the "published" corpus.

In addition, using these same Czerny exercises, we have assembled a more extensive corpus via an online survey and a web application [11] built for the purpose. Subjects were recruited from personal acquaintances of the researchers, email lists published by several music teachers' associations (California, Florida, Georgia, Massachusetts, New Jersey, Ohio, Pennsylvania, and the greater Chicago metropolitan area), and music departments at universities and colleges across the United States and Canada. In all, 5345 recruitment emails were sent, asking recipients to complete the survey and/or forward it to potentially interested students and colleagues. From this, 352 people (6.6%) responded. For the subset of 199 subjects who provided enough information, we determine a median of 36 "years of piano study." Included in this *full* corpus are all complete fingering sequences provided by any participant.

We deem all these data to represent expert fingering advice. Details about each corpus are provided in Table 3.

4.2 Models

To put ERR and our competing probability estimation methods through their paces, we leverage our own implementations [5] of the original piano fingering model described by Parncutt et al. [1] (hereafter referred to as the *parncutt* model) and the enhancement of this model described by Jacobs [2] (the *jacobs* model).

We use the distributions of the *published* and *full* corpora to synthesize two high quality models, which we dub *ideal-p* and *ideal-f*, respectively. The "system" lists produced by such models are composed of the most popular fingering suggestions from a large group of expert pianists. An automated system this good would surely define the state of the art.

Finally, we define two models—*random-p* and *random-f*—that simply apply a random fingering (1–5) to every note in a sequence, generating intentionally sub-optimal ranked "system" lists. These models should clearly underperform any reasonable advice generating system. (We seed the random number generator with a constant arbitrary value to guarantee reproducible results.)

4.3 Approach

For each corpus, we apply four models to each piece to obtain 28 S ranked fingering lists of length five. We then calculate five different ERR scores for each H fingering provided by a human, applying a different Δ function to estimate the probability P_r that the suggestion presented at rank r in the S list will satisfy the pianist. Each of the five estimates is determined by a different method: two that consider unigrams—Hamming distance (as in Equation 7) and "AdjLong" (as captured in Table 1)—and three that compare trigrams (either requiring exact trigram matches (Equation 9), nuanced matches (Equation 11), or yet more relaxed matches (Equation 14).

To summarize our results, we simply average the ERR scores achieved by each method to obtain a mean ERR value,

$$\text{MERR(A)} = \frac{1}{A}\sum_{i=1}^{A} \text{ERR}(H_i), \tag{18}$$

Corpus		*full*			*published*	
Piece	**Notes**	**Annotators**	**Annotations**	**Notes**	**Annotators**	**Annotations**
A (Op. 821 no. 1)	16	202	3232	8	28	224
B (Op. 821 no. 37)	16	201	3216	4	28	112
C (Op. 821 no. 38)	18	198	3564	5	28	140
D (Op. 821 no. 54)	16	190	3040	7	28	196
E (Op. 821 no. 62)	15	195	2925	8	28	224
F (Op. 821 no. 66)	16	192	3072	6	28	168
G (Op. 821 no. 96)	18	195	3510	7	28	196
Totals	115		22559	45		1260

Table 3. Details on Czerny corpora used in evaluation.

where A is the number of human-annotated phrases H in the corpus.

We expect to see ERR methods rank the four models in order from best to worst:

1. *ideal*
2. *jacobs*
3. *parncutt*
4. *random*

We also expect to see an advantage of trigram methods over unigram methods.

We use the *scipy* and *statsmosdels* python packages for statistical analysis.

4.4 Results

The MERR results are displayed in Table 4. The expected ordering of models is present for all four methods employed.

Eight one-way between-subjects ANOVAs conducted to analyze the differences reported between ERR means found all of them to be statistically significant ($p < 0.001$). Follow-up Tukey post-hoc tests indicated statistically significant ($p < 0.05$) differences between all pairs of means except for the *jacobs-parncutt* pairs in both corpora. None of the differences in those means are found to be statistically significant. Thus, none of the methods here allow us to state definitively that this accepted enhancement is clearly better than the original. So we can only safely say this about the relative quality of our models per ERR:

1. *ideal*
2. *jacobs* or *parncutt*
3. *random*

This still suggests that ERR is a valid measure for the piano fingering IR task, in all of its tested guises.

Another noteworthy observation is the apparently superior ability of trigram methods to emphasize differences between very good and very bad models. Consider the Hamming MERR of 0.81098 for *ideal-f* and 0.40306 for *random-f*, a difference of 0.40792. This is a surprisingly narrow advantage for the state of the art over one of the worst models imaginable. The "Relaxed" trigram measure provides a wider spread of $0.73802 - 0.13289 = 0.60513$, which is almost 50% higher. This increased spread is statistically significant ($p < 0.001$) and is a desirable attribute in an evaluation method.

5. CONCLUSIONS

The musical fingering problem is best thought of as an information retrieval task, similar to web search. One of its key advantages over static fingering advice like that available in books is its ability to provide a variety of fingering suggestions on demand. The need for such variety of advice is apparent from the ready acceptance of multiple ground truths by the earliest and latest researchers in the domain [1–3].

There are two fundamental ways to frame the fingering problem. The most straightforward is to model the decisions of a single performer. With this framing, to train and to evaluate models, one need only consider passages fingered by this pianist. With few notable exceptions, all prior research in the domain that has striven for quantitative evaluation has implicitly [12–14] or explicitly [15, 16] focused on this (simplified) formulation. Doing so has allowed the field to ignore what is clearly large disagreement among pianists for even the shortest [1] and most routinized [17] of musical segments. Indeed, the seven exercises in the *published* corpus [1] were selected because "at least two distinct, but arguably equally good, fingerings existed for the opening of each piece."

Presenting the "arguably equally good" as needed should be the charter for future computational models of piano fingering. Note that the "ideal" model we describe above is likely far from ideal. As we try to motivate with our enhanced distance functions, some fingerings differ in trivial ways, in ways that explicitly do not render them "distinct" in the sense that Parncutt et al. use the word. Identifying archetypes of distinct *clusters* of fingerings and presenting these archetypes in an optimal order should produce more optimal lists. Clearly, distinct and diverse suggestions make the best lists. Fortunately, an "extension" to the ERR metric [6, §7.4] provides explicit support for this intuitive notion of "diversity." We expect future work to leverage distance measure like those described here for piano to identify diverse clusters and that such extended ERR metrics will better estimate the utility of musical fingering models.

Corpus	Model	Hamming	AdjLong	Trigram	Nuanced	Relaxed
full	*ideal-f*	0.81098	0.85675	0.70206	0.71696	0.73802
full	*jacobs*	0.66966	0.76734	0.51501	0.54404	0.58154
full	*parncutt*	0.66409	0.75903	0.49974	0.52864	0.5402
full	*random-f*	0.40306	0.53721	0.07596	0.10698	0.13289
published	*ideal-p*	0.83794	0.87434	0.77743	0.78681	0.80017
published	*jacobs*	0.6517	0.72936	0.50254	0.51927	0.54028
published	*parncutt*	0.6028	0.70779	0.45561	0.47757	0.50518
published	*random-p*	0.42696	0.58084	0.17156	0.21516	0.25573

Table 4. Mean ERR score using various Δ functions, all phrases treated equally. Mean pairs in shaded rows are not statistically significant.

Again, we have limited the piano fingering problems supported by the methods described above to monophonic phrases played by the right hand. It is straightforward to include the left hand, applying the logic described symetrically in the obvious ways. Either hands may be evaluated using the Python implementation available at `https://github.com/dvdrndlph/pydactyl`. There is also a pragmatic justification for focusing on melodic fingering: it is arguably where pianists are most in need of good advice. Radicioni and Lombardo [18] convincingly demonstrate that chord fingering for guitarists is much less challenging than melodic fingering because the choices are so severely constrained. If anything, chord fingering in piano is even more highly constrained. Having multiple fretboard locations to play a given note, along with more two-dimensional options for placing fingers, affords the guitarist degrees of freedom that are not at the pianist's disposal. Moreover, unlike for guitar, the fingers in a piano chord played with one hand (as is customary or obligatory in the vast majority of cases) must be applied in an ascending order coinciding with the ascending notes. Fingers may be skipped in building the chord, but the order is constrained.

We also note that the metrics can be applied to polyphonic music and will likely perform reasonably well in the face of limited polyphony. However, we must concede that the piano model metrics described here are not comprehensive.

But broadly speaking, ERR is a promising method for evaluating the now clearly framed musical fingering IR task. Its utility must be validated more comprehensively in future work.

6. REFERENCES

[1] R. Parncutt, J. A. Sloboda, E. F. Clarke, M. Raekallio, and P. Desain, "An ergonomic model of keyboard fingering for melodic fragments," *Music Perception*, vol. 14, no. 4, pp. 341–382, 1997.

[2] J. P. Jacobs, "Refinements to the ergonomic model for keyboard fingering of Parncutt, Sloboda, Clarke, Raekallio, and Desain," *Music Perception*, vol. 18, no. 4, pp. 505–511, 2001.

[3] E. Nakamura, Y. Saito, and K. Yoshii, "Statistical learning and estimation of piano fingering," *Information Sciences*, vol. 517, pp. 68–85, 2020.

[4] E. Clarke, R. Parncutt, M. Raekallio, and J. Sloboda, "Talking fingers: an interview study of pianists' views on fingering," *Musicae Scientiae*, vol. 1, no. 1, pp. 87–107, 1997.

[5] D. A. Randolph, J. Badgerow, C. Raphael, and B. Di Eugenio, "Pydactyl: A Python framework for piano fingering," in *Extended Abstracts for the Late-Breaking Demo Session of the 19th International Society for Music Information Retrieval Conference*, Paris, France, 2018.

[6] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan, "Expected reciprocal rank for graded relevance," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, Hong Kong, 2009, pp. 621–630.

[7] E. Voorhees and D. M. Tice, "The trec-8 question answering track evaluation," *Proceedings of the 8th Text Retrieval Conference*, 11 2000.

[8] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey, "An experimental comparison of click position-bias models," in *Proceedings of the 1st International Conference on Web Search and Data Mining*, Palo Alto, California, USA, 2008.

[9] A. Kullak, *The Aesthetics of Pianoforte Playing*, 5th ed., H. Bischoff, Ed. New York: G. Schirmer, 1860/1893.

[10] C. Czerny, *160 Kurze Übungen, Op. 821*. Leipzig, Germany: C. F. Peters, 1888. [Online]. Available: http://imslp.org/

[11] D. A. Randolph and B. Di Eugenio, "Easy as abcDE: Piano fingering transcription online," in *Extended Abstracts for the Late-Breaking Demo Session of the 17th International Society for Music Information Retrieval Conference*, New York, 2016.

[12] Y. Yonebayashi, H. Kameoka, and S. Sagayama, "Automatic decision of piano fingering based on a hidden Markov models," in *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, 2007, pp. 2915–2921.

[13] M. Balliauw, "A Variable Neighbourhood Search Algorithm to Generate Piano Fingerings for Polyphonic

Sheet Music," Master of Applied Economic Sciences Thesis, University of Antwerp, 2014.

[14] M. Balliauw, D. Herremans, D. P. Cuervo, and K. Sörensen, "Generating fingerings for polyphonic piano music with a tabu search algorithm," in *Proceedings of the International Conference on Mathematics and Computation in Music*. London: Springer, 2015, pp. 149–160.

[15] R. De Prisco, G. Zaccagnino, and R. Zaccagnino, "A differential evolution algorithm assisted by ANFIS for music fingering," in *Swarm and Evolutionary Computation*, ser. Lecture Notes in Computer Science, L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, Eds., vol. 7269. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 48–56.

[16] E. Nakamura, N. Ono, and S. Sagayama, "Merged-output HMM for piano fingering of both hands," in *Proceedings of the 15th International Society for Music Information Retrieval Conference*, Taipei, Taiwan, 2014, pp. 531–536.

[17] O. Beringer and T. F. Dunhill, *Manual of Scales, Arpeggios, and Broken Chords for Pianoforte*. London: The Associated Board of the Royal Schools of Music, 1989.

[18] D. P. Radicioni and V. Lombardo, "A constraint-based approach for annotating music scores with gestural information," *Constraints*, vol. 12, no. 4, pp. 405–428, apr 2007. [Online]. Available: http://dl.acm.org/citation.cfm?id=1295940.1295956