AN EVALUATION OF METHODOLOGIES FOR MELODIC SIMILARITY IN AUDIO RECORDINGS OF INDIAN ART MUSIC

Sankalp Gulati*, Joan Serrà†, and Xavier Serra*

*Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
†Artificial Intelligence Research Institute (IIIA-CSIC), Bellaterra, Barcelona, Spain

ABSTRACT

We perform a comparative evaluation of methodologies for computing similarity between short-time melodic fragments of audio recordings of Indian art music. We experiment with 560 different combinations of procedures and parameter values. These include the choices made for the sampling rate of the melody representation, pitch quantization levels, normalization techniques and distance measures. The dataset used for evaluation consists of 157 and 340 annotated melodic fragments of Carnatic and Hindustani music recordings, respectively. Our results indicate that melodic fragment similarity is particularly sensitive to distance measures and normalization techniques. Sampling rates do not have a significant impact for Hindustani music, but can significantly degrade the performance for Carnatic music. Overall, the performed evaluation provides a better understanding of the processing steps and parameter settings for melodic similarity in Indian art music. Importantly, it paves the way for developing unsupervised melodic pattern discovery approaches, whose evaluation is a challenging and, many times, ill-defined task.

Index Terms— Melodic similarity, Indian art music, evaluation, motifs, melodic patterns, pakads

1. INTRODUCTION

Computational methods to assess melodic similarity between audio excerpts have received considerable attention from the signal processing and the music information retrieval (MIR) communities for a long time [1, 2]. Reliable audio melody extractors allow now to apply these methods for music retrieval in large audio archives [3, 4], as opposed to earlier methods applicable only to MIDI archives [5]. Methods for computing melodic similarity from audio signals are specifically relevant for melody-dominant music traditions, such as Flamenco [6] and Indian art music [7], which are orally transmitted and have large audio archives but few written scores.

In the past, several melodic similarity variants have been proposed. These primarily differ in the choices made for the main processing steps: melody representation, melody segmentation, normalization, and distance measure (see [8] for a recent qualitative survey of such methods). In addition, every variant has a different, specific set of additional choices for selecting the optimal parameters at each processing step. Since the melodic characteristics across music traditions vary considerably, the aforementioned procedure and parameter choices cannot be generalized. Therefore, studies that perform

a comparative evaluation of different variants and analyze the effect of different parameter settings for a specific type of music material are valuable to the community [5, 7, 9, 10]. In this paper, we present a comprehensive evaluation of several such variants for computing similarity between short-time melodic patterns in Indian art music (IAM).

Hindustani and Carnatic music are two traditions of IAM [11, 12]. Both these music traditions are melodically heterophonic in nature, with a predominant melody performed by the lead artist. Melodies in these music traditions are based on the framework of rāg [12]. Rāgs are characterized by their constituent svaras (roughly speaking, notes), their ascending and descending melodic progression and, most importantly, by their melodic motifs or 'catch phrases'. These recurring melodic motifs are the prominent cues used by both the performers, to establish the identity of the rāg, and the listeners, to identify the rāg in a performance. Thus, automatically detecting occurrences of these motifs is key to melodic analyses of IAM, and a crucial step towards meaningful retrieval, recommendation and transcription [7].

The characteristic melodic patterns in IAM constitute the artists' ground for expressing creativity through improvisation. Hence, even when two melodic patterns are perceptually the same for a musician, their surface melodic representation can be drastically different. To illustrate this, we present an example in Fig. 1, where three melodic fragments belonging to the same characteristic phrase are shown.

Recently, several approaches have been proposed for computing melodic similarity in IAM [7, 13, 14, 15]. However, a consensus on the best methodology has not been reached. Thus, to get a deeper understanding of these concepts, we need studies that perform a comparative evaluation of these approaches and a systematic search for the best parameter settings. This paper addresses such a need.

2. METHODOLOGY

The task of computing similarity between melodic patterns involves the extraction of a relevant melodic representation from the audio signal and an appropriate choice of a distance measure for that representation. Furthermore, the melodic similarity measure should be invariant to pitch transpositions and global tempo variations between patterns. Hence, these are the four aspects that we consider in this section. We explore different combinations of the choices for the processing steps and parameter settings.

2.1. Melody Representation

We follow common practice and represent melody by the predominant pitch of an audio signal. For Carnatic music, we use a state-of-the-art melody extraction method proposed by Salamon and Gómez [16]. This algorithm performed very favorably in MIREX,
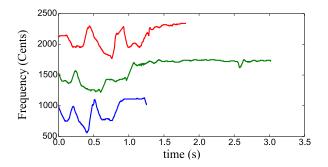
Fig. 1. Melodic fragments corresponding to three renditions of the same characteristic phrase in Hindustani music. For a better visualization, the patterns are transposed by a frequency offset of 600 Cents between them.

an international MIR evaluation campaign focusing on a variety of music genres, including IAM[1]. We use the implementation of this algorithm available in Essentia [17], an open-source C++ library for audio analysis and content-based MIR. We use a frame size of 46 ms and a hop size of 2.9 ms. All other parameters are left to their default values. For Hindustani music we use semi-automatically extracted predominant pitch contours, which have been used in several studies on similar topics [7, 14, 15]. This eases the comparison of results across studies and avoid the effect of pitch errors often present in fully automated melody extraction algorithms. We convert the pitch values from Hertz to Cents in order to make the representation musically more relevant.

Since the automatic assessment of melodic similarity is a computationally expensive task, particularly when done on large audio archives, we desire the minimum possible sampling rate of the melody without compromising the accuracy. We therefore analyze the effect of different sampling rates of the melody representation on melodic similarity. We consider 5 sampling rates 100, 67, 50, 40 and 33 Hz for melody representation, implemented by down-sampling the original melody sequence (mentioned above) by the corresponding factor. We denote these parameter settings by $S_{100}, S_{67}, S_{50}, S_{40}$ and S_{33}.

2.2. Transposition Invariance

In Indian art music the reference frequency for a melody rendition is the tonic of the lead artist [18], which typically varies across artists. Therefore, a meaningful comparison of the melodic patterns across different artists is possible only when the similarity computation is invariant to frequency transposition. It is also required since a melodic pattern may recur in a different octave or tetra-chord within the same recording.

We experiment with five different normalization techniques to achieve frequency transposition invariance. They are as follows. (1) Normalizing the pitch values by the tonic of the lead artist of the recording (N_{tonic}). This is implemented by considering the tonic frequency of the lead artist as the base frequency in Hertz to Cents conversion (see Section 2.1). (2) Zero mean normalization (N_{mean}). (3) Zero median normalization (N_{median}). (4) Z-normalization (N_Z). (5) Median absolute deviation normalization (N_{MAD}). In the case of N_{tonic} the tonic frequency of the lead artist is automatically identified using the method proposed by Gulati et al. [18]. Furthermore, since the reference frequency of the melody is known

for this case, the pitch values can be quantized, as reported in other studies [14]. Hence, we additionally experiment with two quantization levels; semitone level, quantizing pitch values to 100 Cents interval ($N_{tonicQ12}$) and quarter-tone level, quantizing pitch values to 50 Cents interval ($N_{tonicQ24}$). Note that the tonic normalization (N_{tonic}) is helpful only in the scenarios where the frequency transpositions are due to the different tonic frequencies of the lead artists across recordings. It does not handle the cases where a pattern recurs in a different octave or a tetra-chord within the same recording. In total, we consider 8 different normalization variants, including the one without any normalization (N_{off}).

2.3. Uniform Time-Scaling

In order to compensate for global tempo variations across occurrences of a melodic pattern, a typical approach is to consider multiple uniformly time-scaled versions of the patterns [19, 8]. Such tempo differences can significantly degrade the performance in retrieval scenarios where fixed duration patterns are considered. We experiment with two possibilities; first, we do not apply any time-scaling to the patterns (U_{off}) and second, we generate 5 copies of every pattern before similarity computation by uniformly time-scaling it by a factor of 0.9, 0.95, 1, 1.05 and 1.1 (U_{on}). We implement uniform time-scaling using cubic interpolation.

2.4. Similarity Computation

To measure the melodic similarity between two patterns we consider two categories of commonly used distance measures [14, 7, 20]; Euclidean distance (D_{euc}) and dynamic time warping (DTW)-based distance. Euclidean distance is a non-parametric distance measure, whereas, DTW-based distance measure has many variants and parameters to select. In this paper we consider the whole sequence matching DTW variant and two possibilities of the local constraints; first, without any local constraint (D_{DTW_L0}), where the DTW step condition is $\{(1,1),(1,1),(1,1)\}$, and second, with local constraint (D_{DTW_L1}), where $\{(2,1),(1,1),(1,2)\}$ is the DTW step condition. In addition, for both these DTW variants we also apply Sakoe-Chiba global band constraint [21] with the width of the band as 5%, 10% and 90% of the pattern length. Note that 90% band width in the global constraint is to simulate the case of unconstrained DTW. We denote these combinations by $D_{DTW\_L0\_G5}$, $D_{DTW\_L0\_G10}$, $D_{DTW\_L0\_G90}$, $D_{DTW\_L1\_G5}$, $D_{DTW\_L1\_G10}$, and $D_{DTW\_L1\_G90}$, respectively. In total, we consider 7 variants of distance measures for melodic similarity computation.

Since the length of the melodic patterns are different, before computing similarity between two patterns we apply a uniform time-scaling to make their lengths equal. We select the maximum of the lengths of the two patterns as the final length. This operation is a must for the Euclidean distance and has been shown to have a slightly beneficial effect for DTW [22].

3. EVALUATION

3.1. Dataset and Annotations

The music collection used for the evaluation comprises polyphonic vocal music recordings of renowned artists in both Carnatic and Hindustani music. Since the melodic characteristics across these music traditions differ considerably, we perform the evaluation on each music tradition separately. Thus, we use two datasets, Carnatic music dataset (CMD) and Hindustani music dataset (HMD).

Dataset	Rec.	PT	Rāgs	Artists	Duration (hr)
CMD	23	5	5	14	3.82
HMD	9	5	1	7	1.76

Table 1. Details of the datasets in terms of the total number of recordings (Rec.), number of annotated pattern types (PT), number of rāgs, unique number of artists and total duration of the dataset.

Dataset	PT	#Occ	L_{mean}	L_{std}	L_{median}	L_{min}	L_{max}
CMD	C_1	31	1.41	0.24	1.44	0.99	1.94
	C_2	33	1.28	0.21	1.26	0.91	1.91
	C_3	32	1.22	0.25	1.15	0.74	1.71
	C_4	26	1.12	0.17	1.06	0.9	1.6
	C_5	35	0.75	0.09	0.74	0.63	0.98
Overall		157	1.15	0.31	1.12	0.63	1.94
HMD	H_1	41	1.80	1.06	1.44	0.73	5.26
	H_2	139	1.33	0.74	1.22	0.38	5.23
	H_3	21	1.24	0.62	1.16	0.53	2.82
	H_4	61	2.25	1.30	1.74	0.51	5.93
	H_5	78	1.15	0.32	1.13	0.416	2.64
Overall		340	1.51	0.92	1.23	0.38	5.93

Table 2. Details of the annotated melodic patterns. PT: pattern type, #Occ: number of annotated occurrences of patterns of a PT, and L_{mean}, L_{std}, L_{median}, L_{min} and L_{max} are the mean, standard deviation, median, minimum and maximum value of the lengths of the patterns of a PT.

Both these datasets have been used in previous studies for a similar task [13, 14, 7]. In Table 1 we summarize the relevant details for both the datasets.

The melodic patterns are annotated by two professional musicians (one for each music tradition) who have received over 15 years of formal music training. All the annotated melodic patterns are the characteristic melodic phrases of the rāg, which are distinctly recognized by the musicians. Thus, we also minimize the ambiguity involved in the judgment of melodic similarity. In Table 2 we summarize the relevant details for every category of the annotated melodic patterns in both the datasets. From the table we get an idea about the statistics of the length of the patterns across their repetitions for each pattern type. We see that the length of the melodic patterns in Hindustani music have a higher degree of variation compared to Carnatic music.

3.2. Experimental Setup

We consider every annotated pattern as a query and perform an exhaustive search in the target search space comprising of all the annotated patterns in the entire music collection. To make the experimental setup closer to the real world scenario, we add melodic segments other than the annotated patterns in the target search space, which act as noise (referred to as noise candidates). We generate these candidates by randomly selecting short fragments of the melodies in the dataset. The time stamps of the starting of these noise candidates are generated using a uniform distribution, and the lengths are determined using the distribution of the duration values of the annotated patterns. The total number of noise candidates added is 100 times the number of queries for each dataset. For every query, we order the search results by the similarity values and consider top $M = 1000$ nearest neighbours for evaluation. A retrieved pattern is considered as a true hit only if it belongs to the same pattern type

Dataset	MAP	Srate	Norm	TScale	Dist
CMD	0.413	S_{67}	N_{mean}	U_{off}	$D_{DTW\_L1\_G90}$
	0.412	S_{67}	N_{mean}	U_{on}	$D_{DTW\_L1\_G10}$
	0.411	S_{100}	N_{mean}	U_{off}	$D_{DTW\_L1\_G90}$
HMD	0.552	S_{100}	N_{tonic}	U_{off}	$D_{DTW\_L0\_G90}$
	0.551	S_{67}	N_{tonic}	U_{off}	$D_{DTW\_L0\_G90}$
	0.547	S_{50}	N_{tonic}	U_{off}	$D_{DTW\_L0\_G90}$

Table 3. MAP score and the details of parameter settings for the three best performing variants for the CMD and the HMD. Srate: sampling rate of the melody representation, Norm: normalization technique, TScale: uniform time-scaling and Dist: distance measure.

(PT, see Table 2) as the query pattern.

We evaluate all possible combinations of the choices made at each step of the melodic similarity computation discussed in Section 2. We consider 5 different sampling rates of the melody representation, 8 different normalization scenarios, 2 possibilities of uniform time-scaling and 7 variants of the distance measures. In total, we evaluate 560 different variants.

3.3. Evaluation Measures and Statistical Significance

To quantify the performance of a melodic similarity variant considered in this study we use mean average precision (MAP), a typical evaluation measure in information retrieval [23]. MAP is computed by taking the mean of the average precision values of each query in the dataset. This way, we have a single number to evaluate and compare the performance of a variant. In order to assess if the difference in the performance of any two variants is statistically significant, we use the Wilcoxon signed rank-test [24] with $p < 0.01$. To compensate for multiple comparisons, we apply the Holm-Bonferroni method [25]. Thus, considering that we compare 560 different variants, we effectively use a much more stringent criteria than $p < 0.01$ for measuring statistical significance.

4. RESULTS AND DISCUSSION

In this section we present the results of our evaluation of the 560 variants for each of the datasets, CMD and HMD. We order the variants in the decreasing order of their MAP scores and present only the three best performing variants in Table 3 due to space constraints. For complete results see http://compmusic.upf.edu/node/242.

In Table 3 (top half), we show the MAP scores and the details of parameter settings for the CMD. We see that the best performing variant has a MAP score of 0.413. We observe that for the CMD, in the ranked list of the 560 variants, top performing variants consistently use higher sampling rates (either S_{100}, S_{67} or S_{50}). This can be attributed to the rapid oscillatory melodic movements present in Carnatic music, whose preservation requires a higher sampling rate. The top variants invariably use the zero mean normalization (N_{mean}), suggesting that there are several repeated instances of the melodic patterns that are pitch transposed within the same recording. In addition, top variants also use DTW-based distance with local constraint (either $D_{DTW\_L1\_G10}$ or $D_{DTW\_L1\_G90}$), indicating that melodic fragments are prone to large pathological warpings that can significantly degrade the performance. We do not observe any consistency in the usage of uniform time-scaling. However, it strongly correlates with the global constraint in the DTW distance. In majority of the top ranked variants, U_{on} consistently occurs with
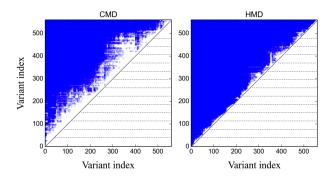
Fig. 2. Matrix indicating the statistical significance of the performance difference between every variant pair for the CMD and the HMD. Variant pairs where the difference in the performance is statistically significant are marked by dots.

Fig. 3. Boxplot of the average precision values for each pattern type (PT) in the CMD and the HMD.

$D_{\text{DTW\_L1\_G10}}$, and U_{off} consistently occurs with $D_{\text{DTW\_L1\_G90}}$. We also performed an analysis of several (small distance) false positives and found that the MAP scores for a number of queries were low because of the spurious errors in the predominant pitch.

The MAP scores and the details of parameter settings for the HMD are shown in Table 3 (bottom half). Compared to the CMD, the best MAP score for the HMD is higher (0.55). All the top ranked variants have the same parameter values except the sampling rate. This suggests that the sampling rates considered in this study have no significant effect on the melodic similarity for the HMD. This can be attributed to the fact that the recordings in the HMD are slow-medium tempo music pieces that do not have fast oscillatory melodic movements, as was the case with the CMD. Furthermore, for the HMD, we observe that the variants using N_{tonic}, N_{tonicQ12} or N_{tonicQ24} perform better than the ones using N_{mean}, which is in contrast to the observation for the CMD. This is primarily because in Carnatic music there are many cases where a pattern recurs in a different octave or tetra-chord within the same recording, whereas, in Hindustani music, such cases are rare. In general, we see that the DTW-based distance performs better than the euclidean distance, and the DTW variant without a global constraint ($D_{\text{DTW\_L1\_G90}}$ or $D_{\text{DTW\_L0\_G90}}$) is preferred. This implies that the repeated instances of melodic patterns in IAM (specifically in Hindustani music) have large non linear timing variations.

To assess the statistical significance of the results we compare every possible pair of the variants ($^{560}C_2 = 156520$ comparisons). The results are shown in Fig. 2, where both the axes are the index of the variants in the ranked list. For every variant pair with index i and j, we mark the pixel (i,j) if the difference is statistically significant. From Fig. 2 we see that a majority of variant pairs have a statistically significant difference in the MAP scores. This indicates that the task of computing melodic similarity is sensitive to the choice of parameters and processing steps, and a small change in the choices made in a variant can lead to a significantly different MAP score. Furthermore, as the marked pixels are higher in number for the HMD, this sensitivity is even higher for the HMD compared to the CMD.

To analyze the consistency in performance across pattern types, we present the boxplot of the average precision values for each pattern type in Fig. 3. For this, we consider only the top performing variant for each dataset. We see that the MAP scores vary considerably across pattern types for both the CMD and the HMD. Furthermore, we observe that the intra pattern type variance of the MAP scores is higher for the CMD as compared to the HMD. In addition, we observe that the pattern types H_2 and H_5 have a higher MAP
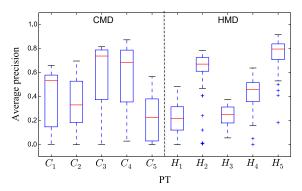
value compared to other pattern types in the HMD. Interestingly, H_2 and H_5 are also the pattern types for which L_{std} is lower and $\#Occ$ is higher than others in the HMD. This correlation is not evident in the CMD, since L_{std} is small for all pattern types.

So far we have considered segmented melodic patterns obtained using annotations, which are rarely available for large audio archives. To simulate a retrieval scenario where the pattern boundaries are not known *a priori*, we consider a simple extension to the experiment, by assuming the target pattern length to be equal to the length of the query pattern. We find that the MAP score for the best performing variant decreases from 0.41 to 0.28 and 0.55 to 0.26 for the CMD and the HMD, respectively. This indicates that the melodic similarity computation task becomes even more challenging in the absence of an accurate melodic segmentation method.

It is worth mentioning that compared to the previous studies [13, 14] we consider more number of pattern types and evaluate the performance of a variant using all annotated patterns as a query. Hence, the results presented are more comprehensive and reliable.

5. CONCLUSIONS

We presented an evaluation of 560 different methodology variants for computing similarity of short-time melodic patterns in IAM. Our results indicate that the task of melodic similarity computation is very sensitive to the choice of parameters and processing steps, specifically to the distance measures and the normalization techniques. A higher sampling rate of the melody representation and mean normalization gives a better retrieval performance for Carnatic music. For Hindustani music, on the other hand, sampling rate has no significant affect on the performance and tonic normalization of the melody results in a better performance. In general, DTW-based distance measure performs better than the euclidean distance, and the usage of local constraint in DTW enhances the performance. The DTW variant without any global constraint is preferred (specially for Hindustani music), which suggests that there are large non-linear timing variations across repeated instances of the melodic pattern in IAM. In the future we plan to include other aspects of melody such as loudness and timbre in the similarity computation and use a bigger dataset consisting of many more melodic patterns for evaluation.

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith, "Query by humming: musical information retrieval in an audio database," in *Proc. of the third ACM Int. Conf. on Multimedia*. ACM, 1995, pp. 231–236.

[2] W. B. Hewlett and E. Selfridge-Field, *Melodic similarity: Concepts, procedures, and applications*, vol. 11, The MIT Press, 1998.

[3] M. Ryynanen and A. Klapuri, "Query by humming of midi and audio using locality sensitive hashing," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing.* IEEE, 2008, pp. 2249–2252.

[4] Y. Zhu and D. Shasha, "Warping indexes with envelope transforms for query by humming," in *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*. ACM, 2003, pp. 181–192.

[5] R. B. Dannenberg, W. P. Birmingham, B. Pardo, N. Hu, C. Meek, and G. Tzanetakis, "A comparative evaluation of search techniques for query-by-humming using the musart testbed," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 5, pp. 687–701, 2007.

[6] A. Pikrakis, J. Mora, F. Escobar, and S. Oramas, "Tracking melodic patterns in Flamenco singing by analyzing polyphonic music recordings," in *Proc. of Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2012, pp. 421–426.

[7] P. Rao, J. C. Ross, K. K. Ganguli, V. Pandit, V. Ishwar, A. Bellur, and H. A. Murthy, "Classification of melodic motifs in raga music with time-series matching," *Journal of New Music Research*, vol. 43, no. 1, pp. 115–131, Jan. 2014.

[8] A. Kotsifakos, P. Papapetrou, J. Hollmén, D. Gunopulos, and V. Athitsos, "A survey of query-by-humming similarity methods," in *Proc. of the 5th Int. Conf. on Pervasive Tech. Related to Assistive Environments*. ACM, 2012.

[9] J. Salamon, J. Serrà, and E. Gómez, "Tonal representations for music retrieval: from version identification to query-by-humming," *International Journal of Multimedia Information Retrieval*, vol. 2, no. 1, pp. 45–58, 2013.

[10] X. Serra, "A multicultural approach to music information research," in *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)*, 2011, pp. 151–156.

[11] T. Viswanathan and M. H. Allen, *Music in South India*, Oxford University Press, 2004.

[12] A. Danielou, *The ragas of Northern Indian music*, Munshiram Manoharlal Publishers, New Delhi, 2010.

[13] V. Ishwar, S. Dutta, A. Bellur, and H. Murthy, "Motif spotting in an Alapana in Carnatic music," in *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)*, 2013, pp. 499–504.

[14] J. C. Ross, T. P. Vinutha, and P. Rao, "Detecting melodic motifs from audio for Hindustani classical music," in *Proc. of Int. Conf. on Music Information Retrieval (ISMIR)*, 2012, pp. 193–198.

[15] J. C. Ross and P. Rao, "Detection of raga-characteristic phrases from Hindustani classical music audio," in *Proc. of 2nd CompMusic Workshop*, 2012, pp. 133–138.

[16] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.

[17] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra, "Essentia: an audio analysis library for music information retrieval," in *Proc. of Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2013, pp. 493–498.

[18] S. Gulati, A. Bellur, J. Salamon, H. G. Ranjani, V. Ishwar, H. A. Murthy, and X. Serra, "Automatic tonic identification in Indian art music: approaches and evaluation," *Journal of New Music Research*, vol. 43, no. 1, pp. 55–73, 2014.

[19] D. Mazzoni and R. B. Dannenberg, "Melody matching directly from audio," in *2nd Annual International Symposium on Music Information Retrieval*. Citeseer, 2001, pp. 17–18.

[20] M. Müller, *Information retrieval for music and motion*, vol. 2, Springer, 2007.

[21] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. on Acoustics, Speech, and Language Processing*, vol. 26, no. 1, pp. 43–50, 1978.

[22] C. A. Ratanamahatana and E. Keogh, "Everything you know about dynamic time warping is wrong," in *Third Workshop on Mining Temporal and Sequential Data*, 2004, pp. 22–25.

[23] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*, vol. 1, Cambridge university press Cambridge, 2008.

[24] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, pp. 80–83, 1945.

[25] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, vol. 6, no. 2, pp. 65–70, 1979.