COMPUTATIONAL MODELS FOR PERCEIVED MELODIC SIMILARITY IN A CAPPELLA FLAMENCO SINGING

N. Kroher, E. Gómez
Universitat Pompeu Fabra
emilia.gomez@upf.edu, nadine.kroher@upf.edu

C. Guastavino
McGill University & CIRMMT
catherine.guastavino@mcgill.ca

F. Gómez
Technical University of Madrid
fmartin@eui.upm.es

J. Bonada
Universitat Pompeu Fabra
jordi.bonada@upf.edu

ABSTRACT

The present study investigates the mechanisms involved in the perception of melodic similarity in the context of a cappella flamenco singing performances. Flamenco songs belonging to the same style are characterized by a common melodic skeleton, which is subject to spontaneous improvisation containing strong prolongations and ornamentations. For our research we collected human similarity judgements from naïve and expert listeners who listened to audio recordings of a cappella flamenco performances as well as synthesized versions of the same songs. We furthermore calculated distances from manually extracted high-level descriptors defined by flamenco experts. The suitability of a set of computational melodic similarity measures was evaluated by analyzing the correlation between computed similarity and human ratings. We observed significant differences between listener groups and stimuli types. Furthermore, we observed a high correlation between human ratings and similarities computed from features from flamenco experts. We also observed that computational models based on temporal deviation, dynamics and ornamentation are better suited to model perceived similarity for this material than models based on chroma distance.

1. INTRODUCTION

The task of modeling perceived melodic similarity among music pieces is a multi-dimensional task whose complexity increases when human judgements are influenced by implicit knowledge about genre-specific musicological aspects and contextual information. Nevertheless, such computational models are of utmost importance for automatic similarity retrieval and recommendation systems in large music databases. Furthermore, analysis of melodic similarity among large amounts of data can provide important clues for musicological studies regarding style classification, similarity and evolution. In the past, numerous approaches have focused on melodic similarity measures, mainly computed from automatically aligned score-like representations. For a complete review of symbolic note similarity measures we refer the reader to [1]. Several previous studies have related computational measures to human ratings. In an extensive study in [14], expert ratings of similarity between western pop songs and generated variants were compared to 34 computational measures. The best correlation was observed for a hybrid method combining various weighted distance measures, which is successfully used to automatically retrieve variants of a given melody from a folk song database. In similar studies, human similarity ratings were compared to transportation distances [16] and statistical descriptors related to tone, interval and note duration distribution [17]. In order to gain a deeper insight into the perception process of melodic similarity, Volk and van Kranenburg studied the relationship between musical features and human similarity-based categorization, where a large collection of folk songs was manually categorized into tune families [15]. Furthermore, human similarity judgement based on various musical facets were gathered. Results indicate that songs perceived as similar tend to show strong similarities in rhythm, pitch contour and contained melodic motifs, whereas the individual importance of these criteria varies among the data. When dealing with audio recordings for which no score is available, it seems natural to focus on the alignment and comparison of the time-frequency representation of the melodic contour. In the context of singing voice assessment, Molina et al. used *dynamic time warping* to align fundamental frequency contours and calculate melodic and rhythmic deviations between them [2].

Despite the growing interest in non-Western music traditions, most algorithms are designed and evaluated on Western commercial music. In a first genre-specific approach to melodic similarity in flamenco music, Cabrera et al. computed melodic similarity among a cappella singing performances from automatic descriptions [3]. The two standard distance measures implemented, the *edit* distance and

the correlation between pitch and interval histograms, obtained rather poor results when compared to expert judgements. As proposed by Mora et al., better results for intra- and inter-style similarity can be obtained for a similarity measure based on manually extracted high-level features (i.e., the direction of melodic movement in a specific part of the performance) [4]. Such studies elucidate the need for exploration of particular characteristics of non-Western music genres and the adaptation of existing music information retrieval systems to such styles.

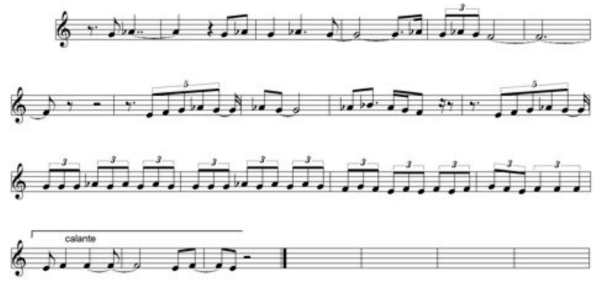
The present study addresses perceived melodic similarity in a cappella flamenco singing from different standpoints: with the aim of gaining insight into the mechanisms involved in perceiving melodies as more or less similar, we gathered similarity ratings among performances of the same style from naïve listeners as well as flamenco experts and analyzed them in terms of intra-subject and intra-group agreement. In order to isolate the melody from other variables such as lyrics, expression and dynamics, we gathered the same ratings for synthesized melodic contours. We furthermore evaluated three computational models for melodic similarity by analyzing the correlation between computed similarity and human ratings. We compared the results to distances computed from manually extracted high-level features defined by experts in the field. The rest of the paper is organized as follows. In Section 2 we provide background information on flamenco music and the *martinete* style, which is the focus of this study. We give a detailed description of the database used in the present experiment in Section 3. Section 4 summarizes the methodology of the listening experiments, the extracted high-level features and the implemented computational similarity models. We give the results of the correlation analysis in Section 5 and conclude our study in Section 6.

2. BACKGROUND

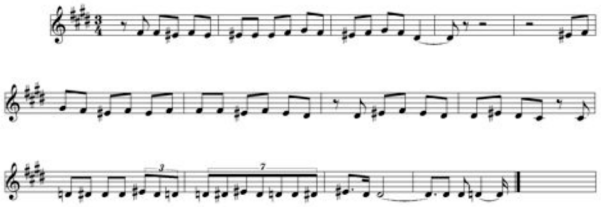
Flamenco is an oral tradition whose roots are as diverse as the cultural influences of its area of origin, Andalusia, a region in southern Spain. Its characteristics are influenced by music traditions of a variety of immigrants and colonizations that settled in the area throughout the past centuries, among them Visigoths, Arabs, Jews and to a large extend gipsies, who decisively contributed to shape the genre as we know it today. For a comprehensive and complete study on history and style, we refer to [5–7]. Flamenco germinated and nourished mainly from the singing tradition and until now the singing voice represents its central element, usually accompanied by the guitar and rhythmic hand-clapping. In the flamenco jargon, songs, but also styles, are referred to as *cantes*.

2.1 The flamenco singing voice

Flamenco singing performances are usually spontaneous and highly improvisational. Songs are passed from generation to generation and only rarely manually transcribed. Even though there is no distinct ideal for timbre and several voice types can be identified, the flamenco singing voice can be generally characterized as matt, breathy, and containing few high frequency harmonics. Moreover, singers usually lack the singer's formant [13]. Melodic movements appear mainly in conjunct degrees within a small pitch range (*tessitura*) of a major sixth interval and are characterized by insistence on recitative notes. Furthermore, singers use a large amount melisma, microtonal ornamentation and pitch glides during note attacks [4].



(a) Performance by *Antonio Mairena*



(b) Performance by *Chano Lobato*

Figure 1. Manual transcriptions of performances a *debla* "En el barrio de Triana"; Transcription: Joaquin Mora

2.2 The flamenco martinete

Martinete is considered one of the oldest styles and forms part of the sub-genre of the *tonás*, a group of unaccompanied singing styles, or *cantes*. As in other *cantes*, songs belonging to *martinete* style are characterized by a common melodic skeleton, which is subject to strong spontaneous ornamentation and expressive prolongations. The untrained listener might perceive two performances of the same *cante* as very different and the fact that they belong to the same style is not obvious at all. To illustrate this principle, Figure 1 shows the transcription of two a cappella performances in Western music notation, both belonging to the same style (*debla*) [4].

Furthermore, the *martinete* is characterized by a solemn performance in slow tempo with free rhythmic interpretation. Traditionally, the voice is accompanied by hammer strokes on an anvil. The tonality corresponds mainly to the major mode, whereas the third scale degree may be lowered occasionally, converting the scale to the minor mode.

3. MUSIC COLLECTION

In consultation with flamenco experts, we gathered 12 recordings of *martinete* performances, covering the most representative singers of this style. This dataset represents

Singer	Percussion
Antonio Mairena	No
Chano Lobato	No
Chocolate	Yes
Jacinto Almadén	No
Jesus Heredia	No
Manuel Simón	Yes
Miguel Vargas	No
Naranjito	No
Paco de Lucia	No
Talegon de Córdoba	Yes
Tomás Pavón	No
Turronero	No

Table 1. Dataset containing 12 *martinete* performances.

a subset of the *tonás* [1] dataset, which contains a total of 56 *martinete* recordings. The average duration of the extracted excerpts containing the first verse is approximately 20 seconds. We limited our study to such a small set, mainly due to the duration of the listening experiment. As an additional stimuli for the listening experiments, we furthermore created synthesized versions of all excerpts. We used the method described in [8] to extract fundamental frequency and energy envelopes and re-synthesize with a sinusoid.

We selected the first verse of each recording, containing the characteristic exposition of the melodic skeleton. Although some *martinete* recordings contain additional accompaniment (guitar, bowing string or wind instruments), we limited our selection to a cappella recordings without rhythmic accompaniment or with very sparse one, as it is found traditionally. We intentionally incorporated a wide range of interpretation characteristics, regarding richness in ornamentation, tempo, articulation and lyrics. Among the singers listed in Table 1, *Tomás Pavón* is to be mentioned as the most influential artist in the a cappella singing styles, performing the *martinete* in an exemplar way. Furthermore, *Antonio Mairena* and *Chocolate* are thought to be the main references for their singing abilities and knowledge of the singing styles. *Chano Lobato* omits some of the basic notes during the melodic exposition and the performance has been included as an example of strong deviation in the melodic interpretation.

4. METHODOLOGY

4.1 Human similarity ratings

In order to obtain a ground truth for perceived melodic similarity among the selected excerpts, we conduct a listening experiment in Montreal (Canada) with 24 naïve listeners with little or no previous exposure to flamenco and in Sevilla (Spain) with 3 experts, as described in [9]. After evaluating various experiment designs (i.e. pair-wise comparison), we decided to collect the similarity ratings in a free sorting task [19]. Using the *sonic mapper* [2] software, subjects were asked to create groups of similar interpretations, leaving the number of groups open. The participants were explicitly instructed to focus on the melody only, neglecting differences in voice timbre, lyrics, percussion accompaniment and sound quality. Nevertheless, in order to isolate the melodic line as a similarity criterion, the experiment had also been conducted with the synthesized versions of the excerpts described above. For each excerpt we extracted the fundamental frequency as described in [8] with a window length of 33 ms and a hop size of 0.72 ms. The pitch contour was synthesized with a single sine wave. A similarity matrix was computed based on the number of times a pair of performances had been grouped together. We compared individual participants' similarity matrices using *Mantel* tests. The *Mantel* test can be considered as the most widely used method to account for distance correlations [12]. We used *zt*, a simple tool for *Mantel* test, developed by Bonnet and Vande Peer [18]., and measured the correlation between participant matrices. We observed that the average correlation for novices is $\mu = 0.0824$, with a $\sigma = 0.2109$ and the average p-value: $\mu = 0.3391$, $\sigma = 0.2139$ (min=0.002). This indicates a very low agreement among them, and indicates differences in perception of melodic similarity depending on the listener's background. Although we should take these results with caution given the small number of experts, we found higher correlation values among them, with an average correlation $\mu = 0.1891$, and $\sigma = 0.1170$. For a detailed description of the procedure and the analysis, we refer to [9].

4.2 Manually extracted high-level features

We manually extracted six high-level features defined by experts in the field. As illustrated above, two *cantes* having the same main notes and different ornamentation would be perceived as the same *cante* by a flamenco aficionado. This fact makes the automatic computation of the features unfeasible. Because of that, we had to rely on manual extraction.

The high-level features were the following.

1. Repetition of the first hemistich. A hemistich is half-line of a verse; the presence of this repetition is important in these *cantes*.
2. Clivis/flexa at the end of the first hemistich. A clivis is a descending melodic movement. Here it refers to a descending melodic contour between main notes. Again, the ornamentation is not taken into account when detecting the presence of the clivis.
3. Highest scale degree in the two first hemistichs. The highest scale degree reached during the *cante* is an important feature.
4. Frequency of the highest degree in the second hemistich. How many times that highest degree is reached is also significant.

[1] http://mtg.upf.edu/download/datasets/tonas

[2] http://www.music.mcgill.ca/ gary/mapper/

5. Final note of the second hemistich.

6. Duration (fast / regular / slow).

A distance matrix was obtained by calculating the Euclidean distance among the feature vectors. The feature vectors were mostly composed of categorical data and we used a standardized Euclidean distance. For a detailed explanation of the descriptors and their musicological background, the reader is referred to [4].

4.3 Computational similarity measures

We implemented three computational measures based on fundamental frequency envelopes and automatic transcriptions and evaluated their suitability for modeling the perceived melodic similarity by analyzing the correlation between computed distance matrices and human judgements. The fundamental frequency contours as well as the automatically generated symbolic note representations were obtained using the system described in [8].

4.3.1 Dynamic time warping alignment

Similar to [2] we used a *dynamic time warping* algorithm to align melodies and estimate their rhythmic and pitch similarity. Since vocal vibrato and microtonal ornamentations strongly influence the cost matrix, we instead align continuous contours of quantized pitch values obtained with the automatic transcription described in [8]. The cost matrix M describes the squared frequency deviation between all possible combinations of time frames between the two analyzed contours $f0_1$ and $f0_2$, where α is a constraint limiting the maximum cost:

$$M_{i,j} = min((f0_1[i] - f0_2[j])^2, \alpha) \quad (1)$$

The *dynamic time warping algorithm* determines the optimal path among the matrix M from first to last frame. The deviation of the slope of the path p with length N from the diagonal path gives a measure for temporal deviation $(DTWtemporal)$,

$$\Delta_{temp} = \frac{\sum_{i=1}^{N}(p[i] - p_{diag}[i])^2}{N} \quad (2)$$

while the average over its elements defines the pitch deviation $(DTWpitch)$:

$$\Delta_{pitch} = \frac{\sum_{i=1}^{N} p[i]}{N}. \quad (3)$$

We used a *MATLAB* implementation [3], which extends the algorithm with several restrictions in order to obtain a musically meaningful temporal alignment. Figure 2 shows the cost matrix and Figure 3 the unaligned and aligned pitch sequences.

[3] http://www.ee.columbia.edu/ dpwe/resources/matlab/dtw/

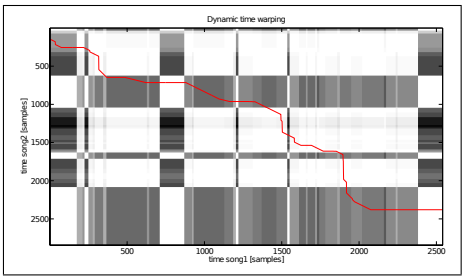


Figure 2. Dynamic time warping: Cost matrix and optimal path.

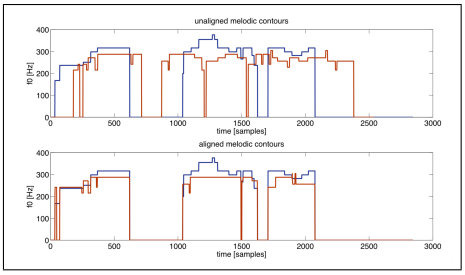


Figure 3. Unaligned (top) and aligned (bottom) melodic contours.

4.3.2 Global performance descriptors

As described in [10], we extracted a total of 13 global descriptors from automatic transcriptions and computed a similarity matrix based on the Euclidean distance among feature vectors. In order to determine the most suitable descriptors for this task, we analyzed the *phylogenetic tree* (Figure 4) computed from the distance matrix of expert similarity ratings. Here, we identify two main clusters, at large distance from each other.

Using these two clusters as classes in a classification task, we perform a support vector machine (SVM) subset selection in order to identify the descriptors that are best suited to distinguish the two clusters. We accordingly extracted the six best ranked descriptors for all songs and computed the similarity matrix from the Euclidean distances among feature vectors. The extracted descriptors are summarized below:

1. **Amount of silence**: Percentage of silent frames.

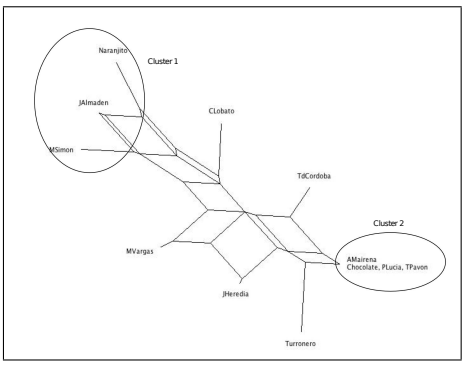


Figure 4. Phylogenetic tree generated from expert similarity judgements.

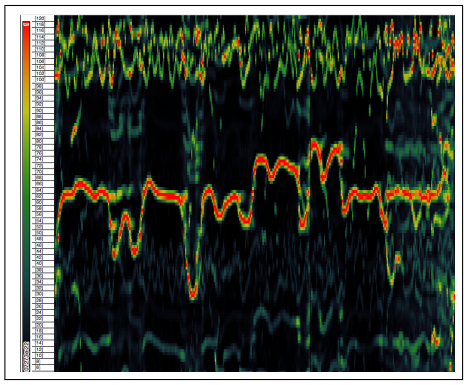


Figure 5. Harmonic pitch class profile for a sung phrase with a resolution of 12 bins per semitone.

2. **Average note duration** in seconds.
3. **Note duration fluctuation**: Standard deviation of the note duration in seconds.
4. **Average volume** of the notes relative to the normalized maximum.
5. **Volume fluctuation**: standard deviation of the note volume relative to normalized maximum.
6. **Amount of ornamentation**: Average per-frame distance in [Hz] between the quantized note value and the fundamental frequency contour.

4.3.3 Chroma similarity

We implemented a similarity measure presented in [11] in the context of cover identification: First, the harmonic pitch class profiles (HPCP) are extracted on a global and a frame basis. The resulting pitch class histogram describes the relative strength of the 12 pitch classes of the equal-tempered scale. HPCPs are robust to detuning as well as variation in timbre and dynamics. After adjusting the key of one sequence to the other, a binary similarity matrix is computed based on the frame-wise extracted HPCPs. Again, dynamic time warping was used to find the best possible path among the similarity matrix. For a detailed description of the algorithm, we refer the reader to [11].

4.4 Evaluation

We evaluated the suitability of the computational models for this task by analyzing the correlation between computed similarity and human ratings. A common method to evaluate a possible relation between two distance matrices is the Mantel test [12]: first, the linear correlation between two matrices is measured with the *Pearson correlation*, which gives a value r between -1 and 1. A strong correlation is indicated by a value significantly different from zero. To verify that a relation exists, the value is compared to correlations to permuted versions of the matrices. Here, 10000 random permutations are performed. The *confidence value* p corresponds to the proportion of permutations giving a higher correlation than the original matrix. Consequently, a confidence value close to zero confirms an existing correlation.

5. RESULTS

Figure 6 shows the comparison of the computed similarity measures by means of correlation r and confidence value p for the different participant groups and stimuli types. We first note that the distance measure obtained from manually extracted high-level descriptors seems to reflect best the perceived melodic similarity for both, expert and naïve listeners. Even though the computed similarity correlates strongly with the expert ratings, the also strong relation with the non-expert similarity judgments is still surprising, given the fact that the descriptors are based on rather abstract musicological concepts. We furthermore find a weaker, but still significant correlation between human ratings and the temporal deviation measure of the *dynamic time warping* algorithm as well as the vector distance among performance descriptors. On the other hand, we find no relation between human ratings and the pitch deviation from the dynamically aligned sequences, nor the chroma similarity measure. Given the fact that the selected performance descriptors are related to dynamic and temporal behavior and ornamentation and the temporal deviation measure does not consider the absolute pitch difference of the aligned sequences, we can speculate that for the given material these factors influence perceived similarity stronger than differences in the pitch progression. *Martinete* presents a particularly interesting case, since the skeleton of the melodic contour and at least its outer envelope is preserved throughout the performances. Notice also that in all cases the found correlation with the similarity ratings of real recordings is stronger than for the synthesized versions. Since none of the computational methods take voice timbre or lyrics into account, we can preclude that these factors influenced human judgement. It is however possible that it was more difficult for the listener to internalize these synthesized sequences compared to real recordings given their artificial nature and consequently judging similarity was more difficult and less precise.

6. CONCLUSIONS

The present study investigates the mechanisms involved in the perception of melodic similarity for the particular case of a cappella flamenco singing. We compared human judgements from experts and naïve listeners for audio recordings and synthesized melodic contours. Computational models are furthermore used to create distance matrices and evaluated based on their correlation with human ratings. We observed a significantly higher agreement among experts and a stronger correlation among computational models and the ratings based on real recordings than when comparing to ratings for synthesized melodies. Furthermore, we discover that models based on descriptors related to rhythm, dynamics and ornamentation are better suited to recreate similarity judgements than models based on absolute pitch distance. We obtained the highest corre-

Subject group	Expert listeners		Naive listener	
Stimuli type	*real*	*synth*	*real*	*synth*
High-level	r=0.585 p=0.001	r=0.424 p=0.001	r=0.429 p=0.000	r=0.202 p=0.054
DTW temporal	r=0.306 p=0.047	r=0.213 p=0.044	r=0.333 p=0.003	r=0.245 p=0.123
DTW pitch	r=-0.118 p=0.256	r=-0.094 p=0.224	r=-0.130 p=0.198	r=-0.096 p=0.204
Perform. descrip.	r=0.431 p=0.011	r=0.207 p=0.044	r=0.308 p=0.0123	r=0.176 p=0.061
Chroma	r=0.108 p=0.239	r=0.107 p=0.187	r=0.090 p=0.247	r=0.102 p=0.193

Figure 6. Correlation between computed similarity and human ratings. Statistically significant results are marked grey.

lation for both expert and non-expert ratings for a similarity measure computed from manually extracted high-level features. The problem of how to compute the high-level features automatically is still open. This problem is equivalent to that of automatically detecting ornamentation and main notes in a flamenco *cante*.

Acknowledgements
The authors would like to thank Joaquin Mora for providing the manual transcriptions and Joan Serrá for computing the chroma similarity measures. This research is partly funded by the COFLA (Proyectos de Excelencia de la Junta de Andalucia, P12-TIC-1362) and SIGMUS (Spanish Ministry of Economy and Competitiveness, TIN2012-36650) research projects as well as the PhD fellowship program of the Department of Information and Communication Technologies, Universitat Pompeu Fabra.

7. REFERENCES

[1] A. Marsden: "Interrogating Melodic Similarity: A Definitive Phenomenon or the Product of Interpretation?" *Journal of New Music Research*, Vol. 4, No. 44, pp. 323–335, 2012.

[2] E. Molina, I. Barbancho, E. Gómez, A. M. Barbanco, and L. J. Tardón: "Fundamental frequency alignment vs. note-based melodic similarity for singing voice assessment," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[3] J. J. Cabrera, J. M. Díaz-Báñez, F. J. Escobar, E. Gómez, F. Gómez, and J. Mora: "Comparative Melodic Analysis of A Cappella Flamenco Cantes," *Proceedings of the Conference on Interdisciplinary Musicology*, 2008.

[4] J. Mora, F. Gómez, E. Gómez, F. J. Escobar, and J. M. Díaz-Báñez: "Melodic Characterization and Similarity in A Cappella Flamenco Cantes," *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, 2010.

[5] J. Blas Vega, and M. Ríos Ruiz: *Diccionario enciclopdico ilustrado del flamenco*, Cinterco,1988.

[6] J. L. Navarro ,and M. Ropero: *Historia del flamenco*, Tartessos, 1995.

[7] J. M. Gamboa: *Una historia del flamenco*, Espasa-Calpe, 2005.

[8] E. Gómez, and J. Bonada: "Towards Computer-Assisted Flamenco Transcription: An Experimental Comparison of Automatic Transcription Algorithms as Applied to A Cappella Singing," *Computer Music Journal*, Vol. 37, No. 2, pp. 73-90, 2013.

[9] E. Gómez, C. Guastavino, F. Gómez, and J. Bonada: "Analyzing Melodic Similarity Judgements in Flamenco A Cappella Singing," *Proceedings of the International Conference on Music Perception and Cognition*, 2012.

[10] N. Kroher: *The Flamenco Cante: Automatic Characterization of Flamenco Singing by Analyzing Audio Recordings*, Master Thesis, Universitat Pompeu Fabra, 2013.

[11] J. Serra, E. Gómez, P Herrera, and X. Serra: "Chroma Binary Similarity and Local Alignment Applied to Cover Song Identification," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 16, No. 6, pp. 1138-1151, 2008..

[12] N. Mantel, and R. S. Valand: "A technique of non-parametric multivariate analysis," *Biometrics*, Vol. 26, pp. 547-558, 1970.

[13] J.Sundberg: "The acoustics of the singing voice," *Scientific American*, Vol. 236 (3), pp.104-116, 1977.

[14] D. Muellensiefen, and K. Frieler: "Modelling experts' notions of melodic similarity," *Musicae Scientiae*, Discussion Forum 4A, pp.183-210, 2007.

[15] A. Volk, and P. van Kranenburg: "Melodic similarity among folk songs: An annotation study on similarity-based categorization in music," *Musicae Scientiae*, 16 (3) pp.317-339, 2012.

[16] R. Typke, and F. Wiering: "Transportation distances and human perception of melodic similarity," *Musicae Scientiae*, Discussion Forum 4A, pp.153-181, 2007.

[17] T. Eerola, T Jaervinen, J. Louhivuori, and P. Toiviainen, : "Statistical Features and Perceived Similarity of Folk Melodies," *Music Perception*, 18 (3), pp.275-296, 2001.

[18] E. Bonnet, and Y. Van de Peer : "zt: a software tool for simple and partial Mantel tests," *Journal of Statistical software*, 7 (10), pp.1-12, 2002.

[19] B. Giordano, C. Guastavino, E. Murphy, M. Ogg, and B.K. Smith: "Comparison of Dissimilarity Estimation Methods". *Multivariate Behavioral Research*, 46, 1-33, 2011.