EVALUATING A COLLECTION OF SOUND-TRACING DATA OF MELODIC PHRASES

Tejaswinee Kelkar
RITMO, Dept. of Musicology
University of Oslo
tejaswinee.kelkar@imv.uio.no

Udit Roy
Independent Researcher
udit.roy@alumni.iiit.ac.in

Alexander Refsum Jensenius
RITMO, Dept. of Musicology
University of Oslo
a.r.jensenius@imv.uio.no

ABSTRACT

Melodic contour, the 'shape' of a melody, is a common way to visualize and remember a musical piece. The purpose of this paper is to explore the building blocks of a future 'gesture-based' melody retrieval system. We present a dataset containing 16 melodic phrases from four musical styles and with a large range of contour variability. This is accompanied by full-body motion capture data of 26 participants performing sound-tracing to the melodies. The dataset is analyzed using canonical correlation analysis (CCA), and its neural network variant (Deep CCA), to understand how melodic contours and sound tracings relate to each other. The analyses reveal non-linear relationships between sound and motion. The link between pitch and verticality does not appear strong enough for complex melodies. We also find that descending melodic contours have the least correlation with tracings.

1. INTRODUCTION

Can hand movement be used to retrieve melodies? In this paper we use data from a 'sound-tracing' experiment (Figure 1) containing motion capture data to describe music–motion cross-relationships, with the aim of developing a retrieval system. Details about the experiment and how motion metaphors come to play a role in the representations are presented in [19]. While our earlier analysis was focused on the use of the body and imagining metaphors for tracings [17, 18], in this paper, we will focus on musical characteristics and study music–motion correlations. The tracings present a unique opportunity for cross-modal retrieval, because a direct correspondence between tracing and melodic contour presents an inherent 'ground-truth.'

Recent research in neuroscience and psychology has shown that *action* plays an important role in perception. In phonology and linguistics, the co-articulation of action and sound is also well understood. Theories from embodied music cognition [22] have been critical to this exploration of multimodal correspondences.

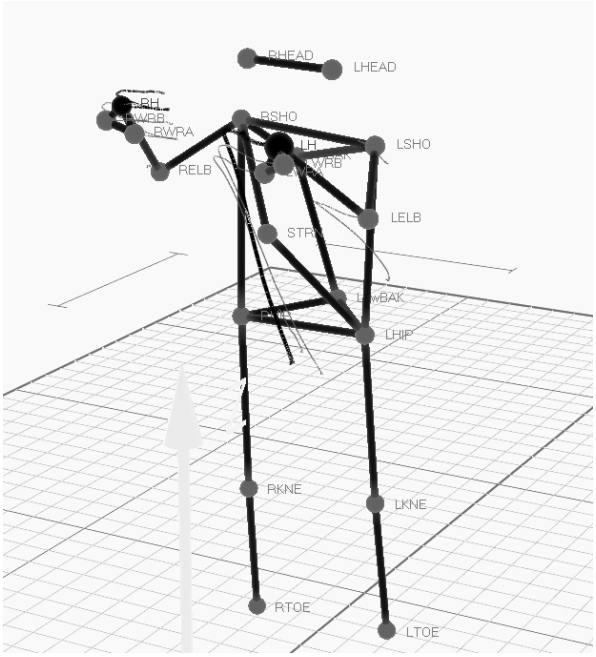

Figure 1. An example of post-processed motion capture data from a sound-tracing study of melodic phrases.

Contour perception is a coarse-level musical ability that we acquire early during childhood [30, 33, 34]. Research suggests that our memory for contour is enhanced when melodies are tonal, and when tonal accent points of melodies co-occur with strong beats [16], making melodic memory a salient feature in musical perception. More generally, it is easier for people to remember the general shape of melody rather than precise intervals [14], especially if they are not musical experts. Coarse representations of melodic contour, such as with drawing or moving hands in the air may be intuitive to capturing musical moments of short time scales [9, 25].

1.1 Research Questions

The inspiration for our work mainly comes from several projects on melodic content retrieval using intuitive and multi-modal representations of musical data. The oldest example of this is the 1975 project titled 'Directory of Tunes and Musical Themes,' where the author uses a simplified contour notation method, involving letters for de-

noting contour directions, to create a dictionary of musical themes where one may look up a tune they remember [29]. This model is adopted for melodic contour retrieval in Musipedia.com [15]. Another system is proposed in the recent project SoundTracer, in which a user's motion of their mobile phone is used to retrieve tunes from a music archive [21]. A critical difference between these approaches is how they handle mappings between contour information and musical information, especially differences between time-scales and time-representations. Most of these methods do not have ground-truth models of contours, and instead use one of several ways of mappings, each with its own assumptions.

Godøy et al. has argued for using motion-based, graphical, verbal, and other representations of motion data in music retrieval systems [10]. Liem et al. make a case for using multimodal user-centered strategies as a way to navigate the discrepancy between audio similarity and music similarity [23], with the former referring to more mathematical features, and the latter to more perceptual features. We proceed with this as the point of departure for describing our dataset and its characteristics, to approach the goal of making a system for classifying sound-tracings of melodic phrases with the following specific questions:

1. Are the mappings between melodic contour and motion linearly related?

2. Can we confirm previous findings regarding correlation between pitch and the vertical dimension?

3. What categories of melodic contour are most correlated for sound-tracing queries?

2. RELATED WORK

Understanding the close relationship between music and motion is vital to understanding subjective experiences of performers and listeners, [7, 11, 12]. Many empirical experiments aimed at investigating music–motion correspondences deal with stimulus data that is made to explicitly observe certain mappings, for example pitched and non-pitched sound, vertical dimension and pitch, or player expertise [5, 20, 27]. This means that the music examples themselves are sorted into types of sound (or types of motion). We are more interested in observing how a variety of these mapping relationships change in the content of melodic phrases. For this we use multiple labeling strategies as explained in section 3.4. Another contribution of this work is the use of musical styles from various parts of the world, including those that contain microtonal inflections.

2.1 Multi-modal retrieval

Multi-modal retrieval is the paradigm of information retrieval used to handle different types of data together. The objective is to learn a set of mapping functions that project the different modalities into a common metric space, to be able to retrieve relevant information in one modality through a query in another. We see that this paradigm is used often in the retrieval of image from text and text from image. Canonical Correlation Analysis (CCA) is a common tool for investigating linear relationships of two sets of variables. In the review paper by Wang et al. for cross modal retrieval [35], several implementations and models are analyzed. CCA is also previously used to show music and brain imaging cross relationships [3].

A previous paper analyzing tracings to pitched and non pitched sounds also used CCA to understand music–motion relationships [25], where the authors describe inherent non-linearity in the mappings, despite finding intrinsic sound-action relationships. This work was extended in [26], in which CCA was used to interpret how different features correlate with each other. Pitch and vertical motion have linear relationships in this analysis, although it is important to note that the sound samples used for this study were short and synthetic.

The biggest reservations in analyzing music–motion data through CCA is that non-linearity cannot be represented, and the dependence of the method on time synchronization is high. The temporal evolution of motion and sound remains linear over time [6]. To get around this, kernel-based methods can be used to introduce non-linearity. Ohkushi et al., present a paper that uses Kernel-based CCA methods to analyze motion and music features together using video sequences from classical ballet, and optical flow based clustering. Bozkurt et al. present a CCA based system to analyze and generate speech and arm motion for prosody-driven synthesis of the 'beat-gesture' [4], which is used for emphasizing prosodically salient points in speech. We explore our dataset through CCA due to the previous successes of using this family of methods. We will analyze the same data using Deep CCA, a neural-network approximation of CCA, to understand better the non-linear mappings.

2.2 Canonical Correlation Analysis

CCA is a statistical method to find a linear combination of two variables $X = (x_1, x_2, ..., x_n)$ and $Y = (y_1, y_2, ..., y_m)$ with n and m independent variables as vectors a and b such that their correlation $\rho = corr(aX, bY)$ of the transformed variables is maximized. Linear vectors a' and b' can be found such that $a', b' = \underset{a,b}{\operatorname{argmax}} \ \operatorname{corr}(a^T X, b^T Y)$. We can then find the second set of coefficients which maximize the correlation of the variables $X' = aX$ and $Y' = bY$ with the additional constraint to keep (X, X') and (Y, Y') uncorrelated. This process can be repeated till $d = min(m, n)$ dimensions.

The CCA can be extended to include non-linearity by using a neural network to transform the X and Y variables as in the case of Deep CCA [2]. Given the network parameters θ_1 and θ_2, the objective is to maximize the correlation $corr(f(X, \theta_1), f(Y, \theta_2))$. The network is trained by following the gradient of the correlation objective as estimated from the training data.

3. EXPERIMENT DESCRIPTION

3.1 Procedure

The participants were instructed to move their hands as if their movement was creating the melody. The use of the term 'creating,' instead of 'representing,' is purposeful, as shown in earlier studies [26,27], to be able to access sound-production as the tracing intent. The experiment duration was about 10 minutes. All melodies were played at a comfortable listening level through a Genelec 8020 speaker, placed 3m in front of the subjects. Each session consisted of an introduction, two example sequences, 32 trials and a conclusion. Each melody was played twice with a 2s pause in between. During the first presentation, the participants were asked to listen to the stimuli, while during the second presentation, they were asked to trace the melody. All the instructions and required guidelines were recorded and played back through the speaker. Their motions are tracked using 8 infra-red cameras from Qualisys (7 Oqus 300 and 1 Oqus 410). We then post-process the data in Qualisys Track Manager (QTM) first by identifying and labeling each marker for each participant. Thereafter, we create a dataset containing Left and Right hand coordinates for all participants.

Six participants in the study had to be excluded due to too many marker dropouts, giving us a final dataset containing 26 participants tracing 32 melodies: 794 tracings for 16 melodic categories.

3.2 Subjects

The 32 subjects (17 females, 15 males) had a mean age of 31 years (SD = 9 years). They were mainly university students and employees, both with and without musical training. Their musical experience was quantized using the OMSI (Ollen Musical Sophistication Index) questionnaire [28], and they were also asked about the familiarity with the musical genres, and their experience with dancing. The mean of the OMSI score was 694 (SD = 292), indicating that the general musical proficiency in this dataset was on the higher side. The average familiarity with Western classical music was 4.03 out of a possible 5 points, 3.25 for jazz music, 1.87 with Sami joik, and 1.71 with Hindustani music. None of the participants reported having heard any of the melodies played to them. All participants provided their written consent for inclusion before they participated in the study, and they were free to withdraw during the experiment. The study design was approved by the National ethics board (NSD).

3.3 Stimuli

In this study, we decided to use melodic phrases from vocal genres that have a tradition of singing without words. Vocal phrases without words were chosen so as to not introduce lexical meaning as a confounding variable. Leaving out instruments also avoids the problem of subjects having to choose between different musical layers in their sound-tracing. The final stimulus set consists of four different musical genres and four stimuli for each genre. The musical genres selected are: (1) Hindustani music, (2) Sami joik, (3) jazz scat singing, (4) Western classical vocalise. The melodic fragments are phrases taken from real recordings, to retain melodies within their original musical context. As can be seen in the pitch plots in Figure 2, the melodies are of varying durations with an average of 4.5 s (SD = 1.5 s). The Hindustani and joik phrases are sung by male vocalists, whereas the scat and vocalise phrases are sung by female vocalists. This is represented in the pitch range of each phrase as seen in Figure 2.

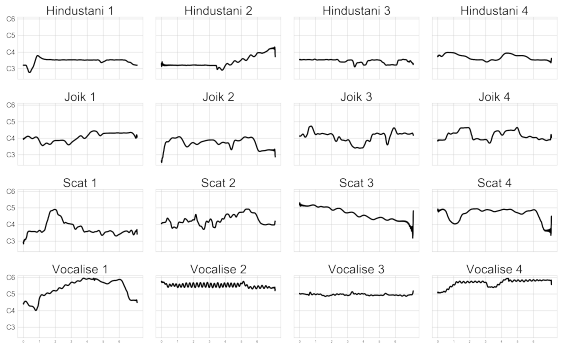

Figure 2. Pitch plots of all the 16 melodic phrases used as experiment stimuli, from each genre. The x axis represents time in seconds, and the y axis represents notes. The extracted pitches were re-synthesized to create a total of 32 melodic phrases used in the experiment.

Seeger	xx	xy	xyy	xyx
Schaeffer	Impulsive	Iterative	Sustained	
Varna	Ascending	Descending	Stationary	Varying
Hood	Arch	Bow	Tooth	Diagonal
Adams	Repetition	Recurrence		

Figure 3. Contour Typologies discussed previously in melodic contour analysis. This figure is representative, made by the authors.

Melodic contours are overwhelmingly written about in terms of pitch, and so we decided to create a 'clean' pitch–only representation of each melody. This was done by running the sound files through an autocorrelation algorithm to create phrases that accurately resemble the pitch content, but without the vocal, timbral and vowel content of the melodic stimulus. These 16 re-synthesized sounds were added to the stimulus set, thus obtaining a total of 32 sound stimuli.

	ID	Description
1	All	16 Melodies
2	IJSV	4 Genres
3	ADSC	Ascending, Descending, Steady or Combined
4	OrigVSyn	Original vs Synthesized
5	VibNonVib	Vibrato vs No Vibrato
6	MotifNonMotif	Motif Repetition Present vs Not

Table 1. Multiple labellings for melodic categories: we represent the 16 melodies using 5 different label sets. This helps us analyze which features are best related to which contour classes, genres, or melodic properties.

3.4 Contour Typology Descriptions

We base the selection of melodic excerpts on the descriptions of melodic contour classes as seen in Figure 3. The reference typologies are based on the work of Seeger [32], Hood [13], Schaeffer [8], Adams [1], and the Hindustani classical Varna system. Through these typologies, we hope to cover commonly understood contour shapes and make sure that the dataset contains as many of them as possible.

3.4.1 Multiple labeling

To represent the different contour types and categories that these melodies represent, we create multiple labels that explain the differences. This enables us to understand how the sound tracings actually map to the different possible categories, and makes it easier to see patterns from the data. We describe these labels as seen in Table 3.4.1. Multiple labels allow us to see what categories does the data describe, and which features or combination of features can help retrieve which labels. Some of these labels are categories, while some are one-versus-rest. Category labels include individual melodies, genres, and contour categories, while one-versus-rest correlations are computed for finding whether vibrato, motivic repetitions exist in the melody, and whether the melodic sample is re-synthesized or original.

4. DATASET CREATION

4.1 Preprocessing of Motion Data

We segment each phrase that is traced by the participants, label participant and melody numbers, and extract the data for left and right hand markers for this analysis, since the instructions asked people to trace using their hands. To analyze this data, we are more interested in contour features and shape information than time-scales. We therefore time-normalize our datasets so that every melodic sample and every motion tracing is the same length. This makes it easier to find correlations between music and motion data using different features.

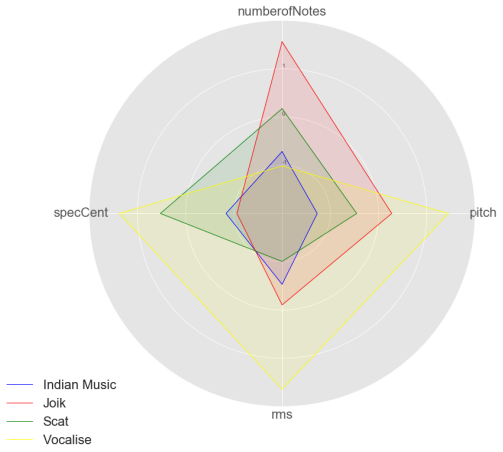

Figure 4. Feature distribution of melodies for each genre. We make sure that a wide range of variability in the features, as described in Table 2 is present in the dataset.

	Feature	Calculated by
1	Pitch	Autocorrelation function using PRAAT
2	Loudness	RMS value of the sound using Librosa
3	Brightness	Spectral Centroid using Librosa
4	Number of Notes	Number of notes per melody

Table 2. Melody features extracted for analysis, and details of how they are extracted.

5. ANALYSIS

5.1 Music

Since we are mainly interested in melodic correlations, the most important feature describing melodies is to extract pitch. For this, we use autocorrelation algorithm available in the PRAAT phonetic program. We use Librosa v0.5.1 [24] to compute the RMS energy (loudness), and the brightness using Spectral Centroid. We transcribe the melodies to get the number of notes per melody. The distribution of these features can be seen for each genre in the stimulus set in Figure 4. We have tried to be true to the musical styles used in this study, most of which do not have written notation as an inherent part of their pedagogy.

5.2 Motion

For tracings, we calculate 9 features that describe various characteristics of motion. We record only X and Z axes, as maximum motion is found along these directions. The derivatives of motion (velocity, acceleration, jerk) and quantity of motion (QoM) which is a cumulative velocity quantity are calculated. Distance between hands, cumulative distance, and symmetry features are calculated as indicators of contour-supporting features, as found in previous studies.

	Feature	Description
1	X-coordinate (X)	Axis corresponding to the direction straight ahead of the participant
2	Z-coordinate (Z)	Axis corresponding to the upwards direction
3	Velocity (V)	First derivative of vertical position
4	Acceleration (A)	Second derivative of vertical position
5	Quantity of Motion	Sum of absolute velocities for all markers
6	Distance between Hands	Sample-wise Euclidean distance between hand markers
7	Jerk	Third derivative of vertical position
8	Cumulative Distance Traveled	Euclidean distance traveled per sample per hand
9	Symmetry	Difference between the left and right hand in terms of vertical position and horizontal velocity

Table 3. Motion features used for analysis. 1-5 are for the dominant hand, while 6-9 are features for both hands.

5.3 Joint Analysis

In this section we present our analysis on our dataset with these two feature sets. We analyze the tracings for each melody as well as utilize the multiple label sets to discover interesting patterns in our dataset which are relevant for a retrieval application.

5.3.1 Dynamic Time Warping

Dynamic Time Warping (DTW) is a method to align sequences of different lengths using substitution, addition and subtraction costs. It is a non-metric method giving us the distance between two sequences after alignment.

In recent research, vertical motion has been shown to correlate with pitch in the past for simple sounds. Some form of non-alignment is also observed between the motion and pitch signals. We perform the same analysis on our data: compute the correlation between pitch and motion in the Z axis before and after alignment with DTW for the 16 melodies and plot their mean and variance in Figure 5.

5.3.2 Longest Run-lengths

While observing the dataset, we find that longest ascending and descending sequences in the melodies are most often reliably represented in the motions, although variances in stationary notes, and ornaments is likely to be much higher. To exploit this feature in tracings, we use "Longest Run-lengths" as a measure. We find multiple subsequences following a pattern which can possess discriminative qualities. For our analysis, we use the ascending and descending patterns, thus finding the subsequences

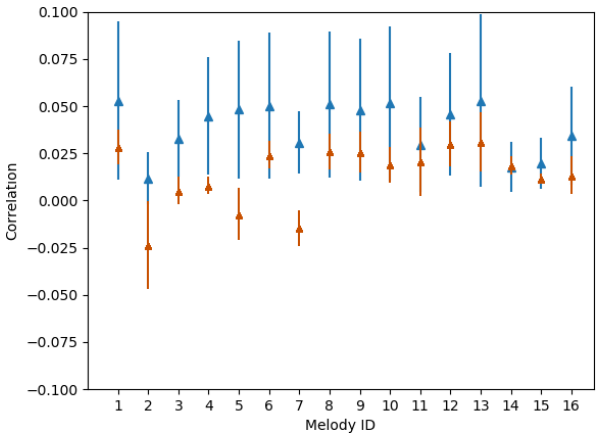

Figure 5. Correlations of pitch with raw data (red) vs after DTW-alignment (blue). Although a DTW alignment improves the correlation, we observe that correlation is still low suggesting that vertical motion and pitch height are not that strongly associated.

from the feature sequence which are purely ascending or descending. We then rank the subsequences and build a feature vector from the lengths of the top N results. This step is particularly advantageous when comparing features from motion and music sequences as it captures the overall presence of the pattern in the sequence remaining invariant to the mis-alignment or lag between the sequences from different modalities. As an example, if we select the Z-axis motion of the dominant hand and the melody pitch as our sequences and retrieve top 3 ascending subsequence lengths. To make the features robust, we do a low pass filtering of the sequence as a preprocessing step.

We analyze our dataset by computing the features for few combinations of motion and music features for ascending and descending patterns. Thereafter, we perform CCA and show the resulting correlation of first transformed dimension in Table 4. We utilize the various label categories generated for the melodies, and show the impact of the features on the labels from each category in Tables 4 and 5. We select the top four run lengths as our feature for each music–motion feature sequence. For Deep CCA analysis, we use a two layered network (same for both motion and music features) with 10 and 4 neurons. A final round of linear CCA is also performed on the network output.

6. RESULTS AND DISCUSSION

Figure 5 shows correlations with raw data and after DTW alignment between the vertical motion and pitch for each melody. Overall, the correlation improves after DTW alignment, suggesting phase lags and phase differences between the timing of melodic peaks and onsets, and those of motion. We see no significant differences between genres, although the improvement in correlations for the vocalize examples is the least pre and post DTW. This could be because of the continuous vibrato in these examples, causing people to use more 'shaky' representations which are most

Motion	**Music**	**All**		**ADSC**		**IJSV**	
Ascend Pattern		**CCA**	**Deep CCA**	**CCA**	**Deep CCA**	**CCA**	**Deep CCA**
Z	Pitch	0.19	0.23	0.25 0.16 0.09 0.05	0.24 0.17 0.12 0.13	0.16 -0.13 0.01 0.37	0.19 0.21 0.08 0.36
Z + V	Pitch	0.21	0.27	0.26 0.09 0.15 0.10	0.30 0.03 0.05 0.17	0.22 -0.13 -0.01 0.35	0.24 0.25 0.15 0.34
All	All	0.33	0.44	0.31 0.14 0.19 0.29	0.44 0.29 0.01 0.36	0.30 0.28 0.23 0.42	0.38 0.43 0.27 0.52
Descend Pattern							
Z	Pitch	0.18	0.21	0.16 -0.11 0.15 0.20	0.17 0.19 0.09 0.19	0.22 0.21 -0.04 0.23	0.22 0.18 0.08 0.28
Z + V	Pitch	0.21	0.31	0.23 0.03 0.14 0.22	0.28 0.28 0.30 0.32	0.26 0.23 0.10 0.24	0.42 0.18 0.34 0.17
All	All	0.35	**0.44**	0.39 0.12 0.20 0.25	0.38 0.02 0.37 0.37	0.35 0.25 0.12 0.36	0.40 0.22 0.14 0.52

Table 4. Correlations for all samples in the dataset and the two major categorizations of music labels, using ascend and descend patterns as explained in Section 5.3.2, and features from Tables 3 and 2

Motion	**Music**	**MotifNonMotif**		**OrgSyn**		**VibNonVib**	
Ascend Pattern		**CCA**	**Deep CCA**	**CCA**	**Deep CCA**	**CCA**	**Deep CCA**
Z	Pitch	0.05 0.23	0.13 0.26	0.19 0.19	0.22 0.25	0.33 0.07	0.33 0.13
Z + V	Pitch	0.10 0.24	0.17 0.31	0.19 0.22	0.24 0.31	0.33 0.09	0.32 0.20
All	All	0.29 0.34	0.36 0.47	0.30 0.35	0.42 0.45	0.38 0.29	0.49 0.40
Descend Pattern							
Z	Pitch	0.20 0.17	0.19 0.21	0.20 0.16	0.23 0.18	0.20 0.17	0.24 0.18
Z + V	Pitch	0.22 0.22	0.32 0.29	0.24 0.20	0.35 0.26	0.22 0.22	0.14 0.34
All	All	0.25 0.40	0.37 **0.45**	0.38 0.33	0.45 0.44	0.33 0.35	**0.54** 0.35

Table 5. Correlations for two-class categories, using ascend and descend patterns as explained in Section 5.3.2 with features from Tables 3 and 2

consistent between participants. The linear mappings of pitch and vertical motion are limited, making the dataset challenging. This also means that the associations between pitch and vertical motion, as described in previous studies, are not that clear for this stimulus set, especially as we use musical samples that are not controlled for being isochronous, nor equal tempered.

Thereafter, we conduct CCA and Deep CCA analysis as seen in Tables 4, 5. Overall, Deep CCA performs better than its linear counterpart. We find better correlation with all features from Table 3, as opposed to just using vertical motion and velocity. With ascending and descending longest run-lengths, we are able to achieve similar results for correlating all melodies with their respective tracings. However, descending contour classification does not have similar success. There is more general agreement on contour with some melodies than others, with purely descending melodies having particularly low correlation. There is some evidence that descending intervals are harder to identify than ascending intervals [31], and this could explain a low level of agreement in this study amongst people for descending melodies. Studying differences between ascending and descending contours requires further study.

While using genre-labels (IJSV) for correlation, we find that scat samples show the least correlation, and the least improvement. Speculatively, this could be related to the high number of spoken syllables in this style, even though the syllables are not words. Deep CCA also gives an overall correlation of 0.54 for recognizing melodies containing vibrato from the dataset. This is an indication that sonic textures are well represented in such a dataset.

With all melody and all motion features, we find an overall correlation of 0.44 with Deep CCA, for both the longest ascend and longest descend features. This supports the view that non-linearity is inherent to tracings.

7. CONCLUSIONS AND FUTURE WORK

Interest in cross-modal systems is growing in the context of multi-modal analysis. Previous studies in this area include shorter time scales or synthetically generated isochronous music samples. The strength of this particular study is in using musical excerpts as are performed, and that the performed tracings are not iconic or symbolic, but spontaneous. This makes the dataset a step closer to understanding contour perception in melodies. We hope that the dataset will prove useful for pattern mining, as it presents novel multimodal possibilities for the community and could be used for user-centric retrieval interfaces.

In the future, we wish to create a system to synthesize melody–motion pairs based on training a network to this dataset, and conducting a user evaluation study, where users evaluate system generated music–motion pairs in a forced–choice paradigm.

8. ACKNOWLEDGMENTS

Partially supported by the Research Council of Norway through its Centres of Excellence scheme (262762 & 250698), and the Nordic Sound and Music Computing Network funded by the Nordic Research Council.

9. REFERENCES

[1] Charles R Adams. Melodic contour typology. *Ethnomusicology*, pages 179–215, 1976.

[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255, 2013.

[3] Nick Gang Blair Kaneshiro Jonathan Berger and Jacek P Dmochowski. Decoding neurally relevant musical features using canonical correlation analysis. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, Souzhou, China*, 2017.

[4] Elif Bozkurt, Yücel Yemez, and Engin Erzin. Multimodal analysis of speech and arm motion for prosody-driven synthesis of beat gestures. *Speech Communication*, 85:29–42, 2016.

[5] Baptiste Caramiaux, Frédéric Bevilacqua, and Norbert Schnell. Towards a gesture-sound cross-modal analysis. In *International Gesture Workshop*, pages 158–170. Springer, 2009.

[6] Baptiste Caramiaux and Atau Tanaka. Machine learning of musical gestures. In *NIME*, pages 513–518, 2013.

[7] Martin Clayton and Laura Leante. Embodiment in music performance. 2013.

[8] Rolf Inge Godøy. Images of sonic objects. *Organised Sound*, 15(1):54–62, 2010.

[9] Rolf Inge Godøy, Egil Haga, and Alexander Refsum Jensenius. Exploring music-related gestures by sound-tracing: A preliminary study. 2006.

[10] Rolf Inge Godøy and Alexander Refsum Jensenius. Body movement in music information retrieval. In *10th International Society for Music Information Retrieval Conference*, 2009.

[11] Anthony Gritten and Elaine King. *Music and gesture*. Ashgate Publishing, Ltd., 2006.

[12] Anthony Gritten and Elaine King. *New perspectives on music and gesture*. Ashgate Publishing, Ltd., 2011.

[13] Mantle Hood. *The ethnomusicologist*, volume 140. Kent State Univ Pr, 1982.

[14] David Huron. The melodic arch in western folksongs. *Computing in Musicology*, 10:3–23, 1996.

[15] K Irwin. Musipedia: The open music encyclopedia. *Reference Reviews*, 22(4):45–46, 2008.

[16] Mari Riess Jones and Peter Q Pfordresher. Tracking musical patterns using joint accent structure. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 51(4):271, 1997.

[17] Tejaswinee Kelkar and Alexander Refsum Jensenius. Exploring melody and motion features in sound-tracings. In *Proceedings of the SMC Conferences*, pages 98–103. Aalto University, 2017.

[18] Tejaswinee Kelkar and Alexander Refsum Jensenius. Representation strategies in two-handed melodic sound-tracing. In *Proceedings of the 4th International Conference on Movement Computing*, page 11. ACM, 2017.

[19] Tejaswinee Kelkar and Alexander Refsum Jensenius. Analyzing free-hand sound-tracings of melodic phrases. *Applied Sciences*, 8(1):135, 2018.

[20] M Kussner. Creating shapes: musicians and non-musicians visual representations of sound. In *Proceedings of 4th Int. Conf. of Students of Systematic Musicology, U. Seifert and J. Wewers, Eds. epOs-Music, Osnabrück (Forthcoming)*, 2012.

[21] Olivier Lartilot. Soundtracer, 2018.

[22] Marc Leman. *Embodied music cognition and mediation technology*. Mit Press, 2008.

[23] Cynthia Liem, Meinard Müller, Douglas Eck, George Tzanetakis, and Alan Hanjalic. The need for music information retrieval with user-centered and multimodal strategies. In *Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 1–6. ACM, 2011.

[24] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. 2015.

[25] Kristian Nymoen, Baptiste Caramiaux, Mariusz Kozak, and Jim Torresen. Analyzing sound tracings: A multimodal approach to music information retrieval. In *Proceedings of the 1st International ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies*, MIRUM '11, pages 39–44, New York, NY, USA, 2011. ACM.

[26] Kristian Nymoen, Rolf Inge Godøy, Alexander Refsum Jensenius, and Jim Torresen. Analyzing correspondence between sound objects and body motion. *ACM Trans. Appl. Percept.*, 10(2):9:1–9:22, June 2013.

[27] Kristian Nymoen, Jim Torresen, Rolf Godøy, and Alexander Refsum Jensenius. A statistical approach to analyzing sound tracings. *Speech, sound and music processing: Embracing research in India*, pages 120–145, 2012.

[28] Joy E Ollen. *A criterion-related validity test of selected indicators of musical sophistication using expert ratings*. PhD thesis, The Ohio State University, 2006.

[29] Denys Parsons. *The directory of tunes and musical themes*. Cambridge, Eng.: S. Brown, 1975.

[30] Aniruddh D Patel. *Music, language, and the brain*. Oxford university press, 2010.

[31] Art Samplaski. Interval and interval class similarity: Results of a confusion study. *Psychomusicology: A Journal of Research in Music Cognition*, 19(1):59, 2005.

[32] Charles Seeger. On the moods of a music-logic. *Journal of the American Musicological Society*, 13(1/3):224–261, 1960.

[33] Sandra E Trehub, Judith Becker, and Iain Morley. Cross-cultural perspectives on music and musicality. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 370(1664):20140096, 2015.

[34] Sandra E Trehub, Dale Bull, and Leigh A Thorpe. Infants' perception of melodies: The role of melodic contour. *Child development*, pages 821–830, 1984.

[35] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.