Landscapes of coarticulation: The co-structuring of gesture-vocal dynamics in Karnatak music performance

Authors:

Lara Pearson[1], Thomas Nuttall[2], Wim Pouw[3]

1. Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany
2. Universitat Pompeu Fabra, Barcelona, Spain
3. Donders Institute for Brain, Cognition, and Behaviour, Radboud University Nijmegen, The Netherlands

Author Note: Correspondence should be addressed to,
Dr Lara Pearson
Max Planck Institute for Empirical Aesthetics
Grüneburgweg 14, 60322 Frankfurt am Main, Germany
Email: lara.pearson@ae.mpg.de

Keywords: vocal performance; gesture; multimodality; coarticulation; South Indian music; motion tracking

Acknowledgments: We would like to thank the Karnatak vocalists, Akkarai Subhalakshmi, Hemmige Prashanth, and Brindha Manickavasakan, who performed for this study. In addition we would like to acknowledge the contribution of the two animators, Mayuri Sajnani and Tushar Malik, who created the performer annotations for our interactive visualisation. Many thanks also to Rainer Polak for his significant assistance in recording the audiovisual and motion capture data on which this study is based, and to Nikita Kudakov for his work in logging the data.

Abstract

In music performance contexts, vocalists tend to gesture in ways that show both similarities and idiosyncrasies across performers. We present a quantitative analysis and visualisation pipeline that characterises the multidimensional codependencies of spontaneous body movements and vocalisations in vocal performers. We apply this pipeline to a dataset of performances within the Karnatak music tradition of South India, including audio and motion tracking data, openly published with this report. Our results show that time-varying features of head and hand gestures tend to be more similar when the concurrent vocal time-varying features are also more similar. While for each performer we find clear co-structuring of sound and movement, they each show their own characteristic salient dimensions (e.g., hand position, head acceleration) on which movement is coarticulated with singing. Our analyses thereby provide a computational characterisation of each performer's unique multimodal coarticulations with singing. The results support our conceptual contribution of widening the conception of coarticulation, from within a 'modality' (e.g., speech articulator positions, joint angles in reaching), to a multimodal coarticulation constrained by both physiological and aesthetic 'control parameters' that reduce degrees of freedom of the multimodal performance such that motifs that sound alike tend to co-structure with gestures that move alike.

Keywords: vocal performance; gesture; multimodality; coarticulation; South Indian music; motion tracking

1. Introduction

Across a range of performance contexts worldwide, vocalists tend to gesture with hand and head movements while they sing. Such gesturing can be understood as part of performers' expressive interaction with the audience and also with the music itself (Davidson, 2001). Qualitative studies have noted correspondences between gestures and musical motifs (Pearson, 2013; Rahaim, 2012), but systematic analyses of this relationship remain few. In this study, we focus on relationships between sung vocalisation and co-occurring gesture, seeking to characterise the multidimensional codependencies of vocalisations and body movement and investigate how this varies across performers. Based on our empirical analysis of the co-structuring of sound and movement in vocal performance, we also make a conceptual contribution to understanding these codependencies, theorising vocalisations as coarticulatory with co-occurring body movement. Here we develop the concept of *multimodal coarticulation*, constrained by physiological and aesthetic 'control parameters' that structurally reduce the degrees of freedom of the multimodal performance.

We explore this in the context of a South Indian musical style known as Karnatak music, in which vocalists typically gesture as they sing (see https://youtu.be/INk1KvYOf8U). Karnatak music (Karṇāṭaka Saṅgīta) is a style of art and devotional music, originating in the royal courts and temples of South India and still performed today in concert halls and at temple festivals (Subramanian, 2006). In both this and related North Indian vocal styles, the majority of performers gesture spontaneously as they sing, producing simultaneous strands of body movement and sound. Their gestures are neither planned ahead nor based on any formal system (Rahaim, 2012), but rather they are spontaneous and unreflective in a way that is similar to much of our co-speech gesturing (Cooperrider, 2019; Gallagher, 2005). However, as with co-speech gesturing (Cooperrider, 2019; Feyereisen, 2017), certain gestural forms and tendencies are picked up through enculturation, and so some similarities can be seen across performers (Rahaim, 2012). Discussion of co-singing gesture can be found in the growing body of work on gesture in Indian music contexts, largely focussing on the North Indian, Hindustani style (Clayton et al., 2024; Fatone et al., 2011; Leante, 2009; Mani, 2017; Paschalidou, 2017; Paschalidou et al., 2016; Pearson & Pouw, 2022; Rahaim, 2012), in addition to research on gesture in Western art music and choral contexts (Brunkan & Bowers, 2021; D'Amario et al., 2023; Nafisi, 2013; Prové, 2022).

This article presents a quantitative analyses and visualisation pipeline for characterising the multidimensional codependencies of body movement and vocalisations in vocal performers.

We apply this pipeline to a dataset of 3.79 hours of Karnatak (South Indian) vocal performances (audio, video, motion-capture), investigating how performer gestures (hand and head movements) co-structure with short musical patterns, referred to here as motifs. The Karnatak music term most closely related to motif is sañcāra, with both terms referring to short musical phrases that have a sense of coherence. Such units are musically meaningful in Karnatak music, acting as building blocks of compositions and extemporisations (Viswanathan, 1977). Indeed, motifs are of structural significance in most musical styles, and are a common focus of research in both music analysis and cognition (Eitan & Granot, 2009; Zbikowski, 1999).

In this study we ask whether there is a systematic relationship between sonic similarity of motifs and kinematic similarity of the co-occurring gestures. In addition, we seek to better characterise the multidimensional codependencies of body movement and vocalisation, asking how the various sonic and kinematic features examined (audio features: f0, Δf0, loudness, spectral centroid; gesture kinematics: 3d position, acceleration, velocity of hand and head motion) differ in the extent to which they co-structure. A further key question concerns the differences between individual performers' co-structuring of sound and movement.

This research builds on work in gesture studies showing that semantically related gestures move alike (Pouw et al., 2021), where it was found in a two-part study that silent or co-speech gestures have similar kinematic trajectories when they convey a similar concept. Specifically, in the first part it was shown that in silent gestures the word2vec-based *semantic* distance between the conveyed concepts had a weak but reliable correlation with the dynamic-time-warping-based *kinematic* distance of the silent gestures. In the second part, using a different dataset, it was shown that when two people were communicating complex visual shapes, the semantic distance between the names for different shapes they arrived at by the end of the conversation were correlated with the kinematic distance between gestures they produced for the respective shapes during the conversation. This study provides an important grounding for the idea that gestures, while often highly unconventionalised, are structurally interrelated such that their kinematic differences inform about semantics (and vice versa). In fact, Pouw et al. (2021) argued from this that gestural semantics may have similar contextual constraints that ground meaning as the principle of distributional semantics in text, where we can glean semantic dissimilarity between words simply by assessing differences in the context with which such words structurally associate. Hagoort and

Özyürek (2024) have recently further explored this idea of distributed gesture semantics suggested in Pouw et al. (2021), demonstrating its current interest amongst researchers.

The question of whether gestures are structurally interrelated with the vocal modality immediately emerges in the context of Karnatak vocal performances, which would have the implication that bodily gestures and vocalisations (loosely) form a structurally connected repertoire, co-patterning or fused together in vocal performance. While preliminary investigations were already reported in Pearson et al. (2023), recent research using 2D video-based tracking has provided additional evidence that bodily gestures and vocal performances indeed co-pattern to the extent that three stereotypical melodic motifs chosen by the researchers could be distinguished based on the degree of kinematic distance between the concomitant gesture (Nadkarni et al., 2023). However, the question of whether gestures structurally interrelate with vocal performance in general (rather than in only three motifs) requires examination with a wider range of more freely varying vocal units and high-resolution kinematic analysis, the approach taken in the present article. In addition, the analyses presented below afford insight into which precise dimensions of gestures structurally interrelate with the vocal modality, and how this varies across individual performers. Through this more comprehensive approach, we problematise the conceptual juxtaposition of vocalisations and gestures as isolated units that are structurally combined from a static library of stereotypical types. Instead, we explore the concept of coarticulation as a gradient phenomenon to understand the broader structural coherence of gesture and vocalisation in performance, as discussed further below.

Note that the structural combination of gesture and vocalisation is not the same as asking whether gestures and vocalisations are coupled in time within a gesture-vocal event (Pearson & Pouw, 2022). For example, it is possible that some particular gestures tend to co-occur with specific vocal motifs, but that those gestures and their concomitant vocalisation need not be synchronised in their concurrent activity. Pearson and Pouw (2022) find evidence of gesture-vocal synchrony, especially for acceleration and f0, whereby acceleration peaks tend to co-occur with peaks in vocal f0, which are also correlated in their magnitude. Acceleration is an important kinematic marker of force-transfers onto the body, and they therefore suggest that this coupling might be in part biomechanical, in line with research on co-speech gestures (for an overview see Pouw & Fuchs, 2022). The study also reveals a high degree of performer variability and concludes that while concurrent gesture-vocal coupling is most stable across performers in the acceleration-

f0 dimension, there are likely to be many more constraints that structure the multimodal performance. One such possible source is the systematic coarticulation of particular movement qualities with particular vocal qualities, explored in the following section.

2. Towards a multimodal coarticulation

Coarticulation has been most commonly explored in the context of linguistics, where the term refers to how the production of a phoneme is influenced by those that precede and follow it (Kühnert & Nolan, 1999). This type of temporal coarticulation has also been examined in body motion, for example in American Sign Language finger spelling (Jerde et al., 2003) and other sign languages (Segouat, 2009). In musical contexts, coarticulation has been studied in pianists' hand movements (Engel et al., 1997) and violin fingering (Wiesendanger et al., 2006). Rolf Inge Godøy extended the concept of coarticulation beyond body movement in musical contexts to the musical sounds produced by such movement (Godøy et al., 2010), noting that "Coarticulation happens because effector motion takes time, so that there is a constraint-based temporal smearing from one motion event to another, hence often also from one output sound to another" (Godøy, 2022, p. 3). This conceptual approach later formed the basis of an examination of melodic coarticulation in Karnatak music performance (Pearson, 2016a).

While the above examples are of temporal coarticulation, simultaneous coarticulation across effectors, often referred to as spatial coarticulation or synergies (Latash, 2008), has also been explored in a number of contexts. A good demonstration of synergies comes from research in speech articulation where it is shown that the lips move to make a labial sound when the jaw cannot (Kelso et al., 1984). As such a secondary articulator (lips) in the task co-adjusts to maintain a functional unity when one of the primary articulators (jaw) is perturbed. Spatial coarticulations are also seen in piano performance, where, for example, the wrist, elbow, shoulder and even the whole torso may move along with the finger movements that press down on the piano keys (Godøy et al., 2010). The extent to which such spatial coarticulation can be considered functionally necessary in order to play is variable. Pianists' must move their arms and wrists in order to position their fingers at the correct point of the piano keyboard and exert downward force onto the keys, but the manner and extent of this arm and wrist movement is a matter of individual technique and expressivity. Indeed, it is helpful here to highlight that in musical performances the performance variable is not only (bio)mechanical, but as Latash (2008, p. 359) argues in *Synergy*, "meaning"

becomes the performance variable for movements aiming towards expression, where necessary (bio)mechanical functions are fused with expressive functions.

In the present article's example of multimodal coarticulation between vocalisation and gesture there is also a necessity to move the body to a degree when singing – for example, it would be challenging to sing without any head or torso movement (Pettersen & Westgaard, 2004). The greater than strictly necessary gestural movement that contributes to expression in many vocal styles may be viewed as fused or *coarticulated* with such minimally necessary body motion in a way that is coherent with such motion. This coarticulation includes biomechanical influences of gesture-voice coupling (Pouw & Fuchs, 2022) but it is clearly not fully determined by it, as we know that gesture-vocal coordination varies considerably across individual performers (Pearson & Pouw, 2022). Constraints on gesture-vocal coarticulation include the individual's musical life and learning experience (the gesturing tendencies picked up from their teacher and other performers) and their own interaction with the music and audience in the immediate moment of performing. Following from such bodily, sociocultural and musical constraints, the coarticulated gesture-vocal flow can attain semiotic functions through, for example, contiguity relations and recurring gestures, as discussed below.

3. Semiotic dimensions relating to gesture-vocalisation coarticulation

3.1. Contiguity

Contiguity is a relation with semiotic potential, where meaning is formed through systematic bordering (spatial and/or temporal) of one against the other, such as in pointing (Mittelberg & Hinnell, 2023). Such systematic bordering between gesture and vocalisation is a clear implication of existing findings on gesture-vocalisation coupling where peaks in hand acceleration and change in pitch have been found to synchronise (Pearson & Pouw, 2022). As a result of such coupling, vocalists' gestures can sometimes appear as a form of pointing, in which vocalists index melodic movement with their body movements. Depending on the perspective taken, contiguity can be considered as both a constraint on gesture, perhaps with the intent of meaning-making (for example, a performer pointing to the pitch movement in order to highlight a repeated pattern), and as the *result* of constraints such as biomechanical connections between vocalisations and upper body movement, or gesturing habits picked up through the learning process. In both cases, through extended stretches of contiguity between gestural movement and melodic movement, the two

modalities become associated and multimodal semiosis is evident. It should be noted that this semiotic interpretation does not preclude the significance of body-voice connections. Indeed, it is likely that vocalists' gestures also index their own vocal production movements since their musical vocalisations are directly related to the physical movement required to create them (Pearson, 2016b, pp. 112-113). Movement in general has been theorised as an important aspect of musical meaning (Clarke, 2001; Tagg, 2012), and so when considering multimodal semiosis in vocal performance we should allow for the possibility that bodily movement is itself a focus of meaning formation, indexed by both sound and gesture.

3.2. Recurring gestures

In qualitative research on Indian vocal music practices, connections have been noted between performers' sung musical phrases and their co-occurring hand gestures (e.g., Pearson, 2013; Rahaim, 2012). For example, Matthew Rahaim (2012) notes that in one Hindustani vocal performance, 30 out of the 34 occurrences of a particular melodic shift were accompanied by a gesture where the vocalist's hands curl around a small empty space. Rahaim suggests that such recurring gestural patterns can be considered catchments, conceptualised in co-speech context as regions of recurring gestures that index underlying discourse themes (McNeill, 2000). A related concept in co-speech gesture research is that of recurrent gestures (Harrison & Ladewig, 2022; Ladewig, 2014; Mortimer & Pereira, 2023; Müller, 2018), which show a stable form-meaning pairing within individual and/or culturally shared repertoires (Müller, 2018, p. 277), implying a stability beyond a single interaction or performance.

While it is clear from existing qualitative studies on Hindustani and Karnatak vocal performance that there are cases where particular motifs co-occur repeatedly with gestures that have highly similar forms, such recurring motif-gesture combinations may be infrequent relative to the entire set of gesture-vocal utterances within any given performance. Furthermore, the recurrences may be significantly more gradient and imperfect, whereby we should understand them as chaotic recurrences much like a complex mutli-variable system visiting similar regions in a space of possibilities but never repeating the exact states (Favela, 2020); or as "repetition without repetition" (Bernstein, 1967). To understand this gradient recurrence at a larger scale, in this article, instead of focusing only on a few stereotypical motifs, we explore gesture-vocalisation relations

across a wide range of vocal segments, located automatically using machine learning methods rather than chosen by the researchers.

3.3. Current study

The research questions and goals of this study are as follows:

- Is there a systematic relationship between sonic similarity of motifs and kinematic similarity of the co-occurring gestures? Do motifs that sound similarly, move similarly?
- What are the multidimensional codependencies of body movement and vocalisations? How do the various sonic and kinematic features examined differ in the extent to which they co-structure? Are head and hand movements part of the systematic relationship with sonic features, or is there systematicity with vocalisation in one movement modality only?
- How does the co-structuring of sonic and kinematic features vary across different performers? What can we learn about the different ways that individuals co-structure sound and movement in their performances?

The overarching goal of the research is to gain insight into why vocalists gesture as they do. In addition, we seek to contribute to theory building in ways that simultaneously acknowledge biomechanical, musical and sociocultural constraints on performance gesture-vocalisation relations. The analysis pipeline reported here also contributes to reproducible methods that can be applied to investigate systematic but gradient contiguity in many different types of multimodal communication, including music making, but also animal multimodal communication (Partan & Marler, 1999), and thereby broadens the scope relative to earlier work on conventionalised speech semantics and gradient gesture kinematics (Pouw et al., 2021).

4. Methods

4.1. Performers and performances

In total, 44 recorded performances of a Karnatak musical format known as *rāga ālāpana* were analysed, across 8 different ragas (melodic frameworks), with actual singing time in the performances lasting *M* (*SD*) duration = 310.00s (118.72s) seconds, min-max duration = 100s -

586s and a total singing time of 3.79 hours. It should be noted that *rāga ālāpana* does not have a musical metre or steady beat, as can be found in many other musical styles, and therefore the question of entrainment to a repetitive steady beat does not arise in this musical format. 3 right-handed vocalists participated in this study (1 male and 2 female, *M* age = 35.7, *SD*age = 5.8), having given their written informed consent. These vocalists, based in Chennai and Bengaluru, are all experienced and currently active performers within the South Indian, Karnatak music community, each having combined studying and performing experience of between 22 and 37 years.

4.2. Measurements and equipment

4.2.1. Motion tracking: An inertial measurement system, Xsens MVN Awinda (60Hz sampling) was used to track the upper body in terms of position, velocity, and acceleration in 3D space (Roetenberg et al., 2009). From amongst the recorded body points, left hand, right hand and head segments were used for the gesture kinematic analyses. The entire upper body motion data was exported from Xsens as FBX files to create the animations used in the interactive visualisation.

4.2.2. Audio recording: Audio was recorded at 48 Khz using Neumann KM184 condenser microphones placed directly in front of the performers, as in a typical Karnatak concert.

4.2.3. Video recording: Video was recorded with GoPro Hero4 cameras at 50fps. Video recordings were used for qualitative cross-checking of our quantitative research.

4.2.4. Synchronisation: Synchronisation of the different streams was achieved using the :pulse timecode system and checked manually for peak deceleration at clapperboard closure. Clapperboards were performed by the vocalist and recorded at the start and end of each performance.

4.2.5: Karnatak alapana multimodal dataset: The above described synchronised audio and motion capture data, together with animations created from the motion capture data are included in the Karnatak alapana dataset, made open with this publication: [link placed here on publication].

4.3. Analyses overview

Our analyses pipeline begins with motif identification followed by feature extraction and dynamic time warping (DTW) computation (see Figure 1 for a visualisation of the processes). This forms the basis of subsequent correlation and regression analyses, as described in the overview below. All code necessary to reproduce the analysis and visualisations are provided here: [link placed here on publication].

4.3.1. Automated repeated melodic motif extraction

We use a machine learning methodology tailored for Karnatak music to locate regions of repeated melodic patterns across the dataset (Nuttall et al., 2022), implemented as part of the compIAM package (Plaja-Roglans, Nuttall, & Serra, 2023). The process uses self-similarity computations of autoencoder embeddings of constant-Q transforms (CQT) of the raw audio to identify regions of consistent melodic similarity. To be returned by the process, patterns have to repeat at least once in a performance and have a minimum length of 1.5 seconds. Across all performances, 595 unique, non-overlapping, regions are identified, each corresponding to a melodic motif that is repeated at least once.

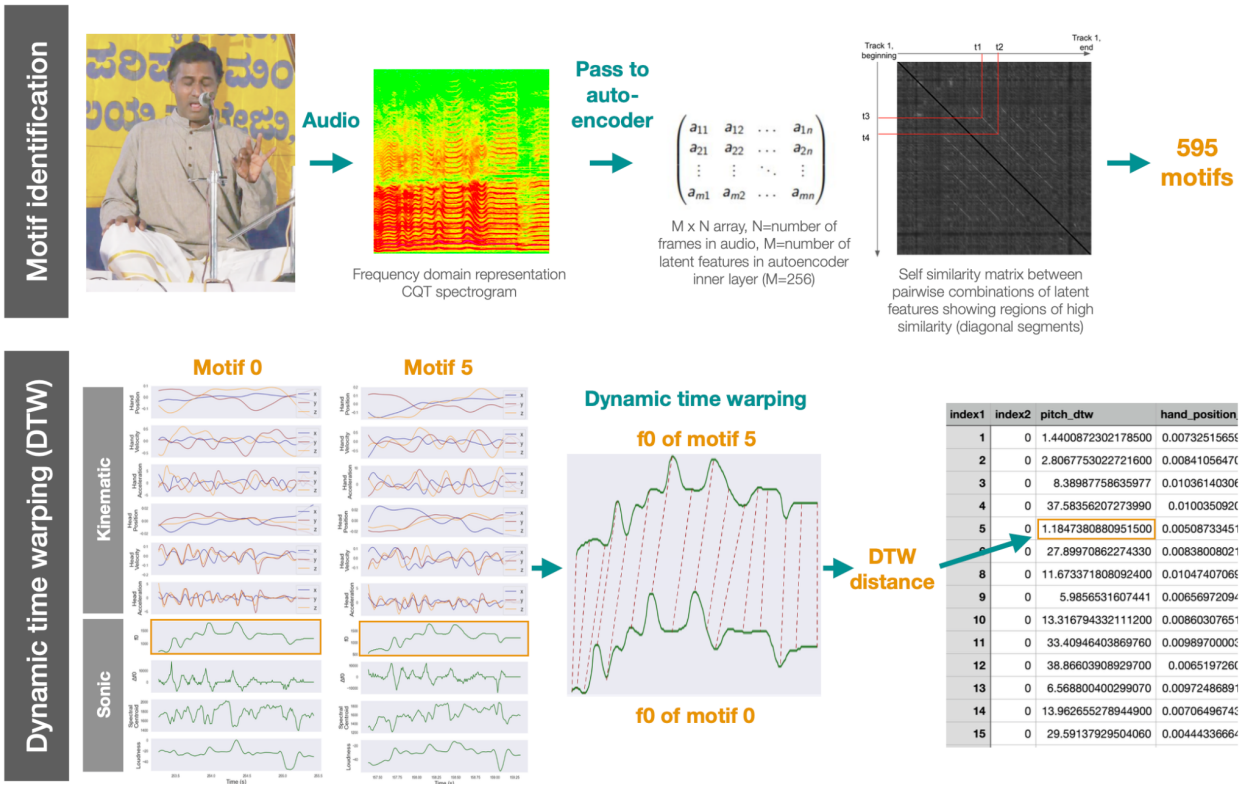

Figure 1: An overview of the first two stages in the analysis pipeline. The upper row shows the motif identification process, wherein pairwise regions of consistently high melodic similarity are identified as repeated motifs using features learnt by an autoencoder. The lower row visualises the dynamic time warping process, in which DTW distances are calculated for pairs of all 10 sonic and kinematic features and placed in the DTW distance dataframe. The photograph shows the Karnatak vocalist, Hemmige S Prashanth, performing on stage in Mangalore in 2014.

4.3.2. Feature extraction and processing

For each of the 595 motifs, we extract a time series corresponding to the following features: the **f0** of the predominant sung melody, using a machine learning methodology tailored for Karnatak music (Plaja-Roglans, Nuttall, Pearson, et al., 2023), also implemented as part of the compIAM package; **Δf0**, an approximation of the first derivative of the f0 curve as outlined in (Keogh & Pazzani, 2001); **loudness**, $L = 10\ log_{10}\left(\frac{S}{ref}\right)$, where S is the power spectrum of the raw audio signal and ref is its maximum value; the **spectral centroid** of the raw audio signal, a representation of timbre that has been found to correspond to ratings of perceptual salience (Schultz et al., 2021); and the 3-dimensional **position**, **velocity** and **acceleration** of the hand and head event trajectories as captured using the Xsens MVN Awinda system.

In total, for each of the 595 motifs, 10 time series are extracted (f0, Δf0, loudness, spectral centroid, position hand, position head, velocity hand, velocity head, acceleration hand, acceleration head). The movement time series are n-dimensional (containing multiple effectors and dimensions). Each time-series is smoothed using a 2nd order Savitzky-Golay filter with a window length of 125ms. For the f0 time series, silences of 350ms or less are linearly interpolated to account for incorrectly annotated silence that occur within ornaments, which can be due to various reasons including glottal closure or other rapid vocal movements.

To account for slight differences in performer position between performances, before extraction, each gesture curve is rotated such that the line between the positional centroid of the left and right shoulders is parallel to the x-axis to ensure that the performer is facing the "front of the stage", and the origin of the gesture space set to the centroid of the pelvis position.

For the gestural hand data, the predominant hand of the performer is selected for each motif based on that which has the most kinetic energy, KE, as computed across the entire duration of the motif: from the velocity curve, v, $KE = \frac{1}{2}mv^2$, where m is the mass of the body part in question and is assumed equal for both sides. The x-axis values are mirrored for the right hand such that all motifs occur in the same "left-hand space". This allows us to compare gestures that are recurrent but mirrored due to hand preference differences. 89.0% of motifs are identified as left-handed and 11.0% are identified as right-handed. The proportion of motifs with a ratio between the dominant-hand energy and the non-dominant hand energy of greater than 1.2 is 97.0%, indicating that there is almost always a very clear dominant hand.

For each pairwise combination of motifs, the dynamic time warping distance (DTW) between each of these 10 time series is computed, whereby each pair has one DTW value for f0, one for Δf0, one for loudness, and so on. To account for slight variances in segmentation point between pairs of identical motifs, we use a custom, dependent DTW implementation that allows for each extreme of the warping path to begin within 0.1*L of the start and end of each pattern, where L is the length of the longest motif in the pair. The Sakoe-Chiba window size is also equal to 0.1*L. The resulting dataframe has 176,715 rows (one for each pair of motifs excluding identical pairs), and 10 columns containing DTW distances for each of the 10 features.

4.3.3. Relationship between Dynamic Time Warping and Perceived Melodic Similarity

Since we are interested in studying whether melodically similar motifs co-occur with similar performer gestures, it is important that the DTW calculations effectively capture similarity; similar motifs should have a lower DTW distance between them than dissimilar motifs. In addition to checking qualitatively that both sonic and and gestural similarity indicated by DTW distance aligned with our human perception of similarity, we also conducted a more systematic check for melodic similarity. To validate the use of DTW as a proxy for melodic similarity between pitch motifs we asked a professional Karnatak vocalist to annotate the 800 most similar motif pairs in the DTW distance dataframe (lowest 800 DTW values between f0 curves). The vocalist was presented with the audios of each motif pair and asked to label them as "same" or "different". Definitions were provided to the vocalist beforehand, and can be found in the supplemental materials Box S1.

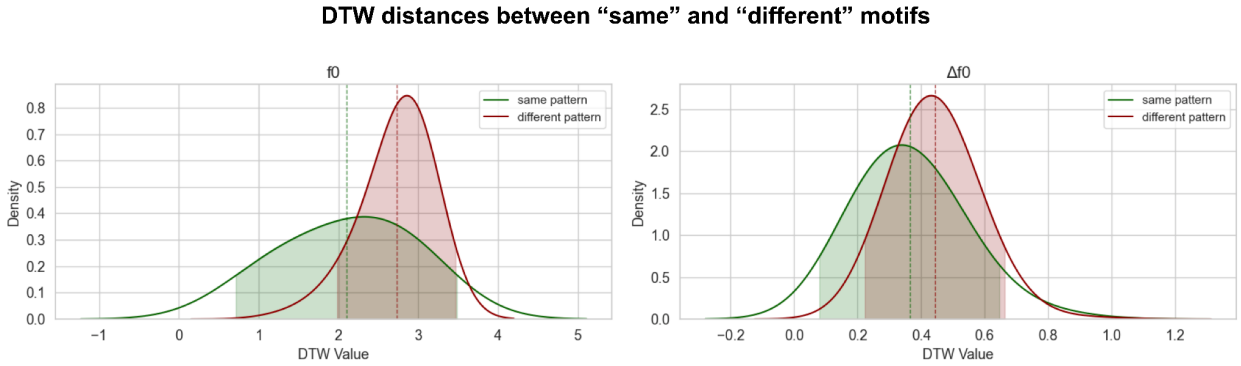

Figure 2: The plots display the kernel density estimates of DTW distances for the f0 and Δf0 time series corresponding to motif pairs in the "same" and "different" groups.

We perform independent t-tests to compare the means of the same and different samples for both f0 ($M_{different} = 2.73$, $SD_{different} = 0.38$; $M_{same} = 2.10$, $SD_{same} = 0.72$) and Δf0 ($M_{different} = 0.44$, $SD_{different} = 0.11$; $M_{same} = 0.36$, $SD_{same} = 0.15$). For f0, we obtain a statistically significant difference between same and differently categorised motifs, $t\,(798) = 15.84$, $p < .0001$. A statistical difference was also found for Δf0, $t\,(798) = 8.05$, $p < .0001$ (see Figure 2 for plots displaying the kernel density estimates). Furthermore we compute the point biserial correlation coefficient, r_{pb} between the DTW distances and the categorical label of "same" and "different": for f0 ($r_{pb} = 0.49$, $p < .0001$) and Δf0 ($r_{pb} = 0.28$, $p < .0001$) . We conclude that although DTW distance does not fully align with the expert categorisation of same and different motifs, the distributions are sufficiently

distinct to warrant using DTW as a proxy for melodic similarity. It should be noted that assessments of this type made by an expert musician will be coloured by their knowledge of the style. For example, two highly similar pitch curves that are audibly in different ragas (melodic types) are likely to be defined as different by an expert musician, notwithstanding their high degree of melodic similarity. Such subtleties contribute to the overlap seen in the results.

4.3.4. Regression and correlation analyses

Four analyses are performed on the dataframe containing the Dynamic Time Warping distances of the four sonic features (f0, Δf0, loudness and spectral centroid) and six gestural features (position hand, position head, velocity hand, velocity head, acceleration hand, acceleration head), as outlined below.

Overview Analysis 1: Do sonic motif DTW distances covary with spatiotemporal patterns of gesture? For each pairwise combination of 4 sonic feature columns and 6 gestural feature columns, we compute the Spearman's rank correlation coefficient between the column values to ascertain whether there is a systematic relationship between sonic similarity of motifs and kinematic similarity of the co-occurring gestures across all performers. This constitutes 24 computations; we further compute these values on subsets of the DTW distance dataframe corresponding to individual performers. For the "all performer" test, the number of patterns from each is subsampled such that each performer is represented equally. We will report test results as compared to a Bonferroni corrected significance value (α), where α is divided by the number of tests - 96 - before comparison.

Overview Analysis 2: Can sonic features be predicted from combined gesture features? We hypothesise that sonic features can be predicted from gesture features, with head and hand features combined being more predictive than simply one or the other. To investigate this we train a Gradient Boosting regressor on all 6 gestural features to predict each individual sonic feature (4 models in total). The hyperparameters of the model are selected using a 3-fold cross validation grid search in a sensible hyperparameter space. We evaluate our model at training time using the R2 score and report our results on a holdout subset of the DTW distance dataframe corresponding to 20% of the entire dataset (not used at all during training).

We further investigate the effect of leaving out certain gestural features, repeating the process four times: once on randomised data, once on the gestural features corresponding to the head, once on the gestural features corresponding to the hand, and once on both hand/head features. As before, we also repeat this analysis on subsets of the DTW distance dataframe corresponding to individual performers. The accompanying github repository includes the pitch time series of the 595 motifs; the dataframe with metadata and the code to create it; the analyses results and the code to reproduce them.

4.3.5. Dynamic Visualisation pipeline

A key conceptual contribution of the current work is that performers each have their own coarticulation landscape that captures to some extent the overall qualities of the multimodal performances. As such we believe that our analyses that describe these coarticulation landscapes can also contribute to qualitative investigation of multimodal performance from a musicological perspective. For this purpose, intuitive dynamic dashboards can be valuable, where quantitative static data points are linked with the original dynamic data. We therefore offer animations of the performances with an integrated dashboard that visualises the DTW distance dataframe using dimensionality reduction techniques.

4.3.5.1. Animations: Animations were created by exporting FBX files of the motion capture data recorded using the Xsens Awinda system. These FBX file data were then retargeted to a human base mesh using 3DS Max, cloth simulation was created in Marvellous Designer and rendered using Unreal Engine.

4.3.5.2. Dashboard: In Python, using Plotly Dash we developed a dashboard that linked the animated audiovisual recordings of each identified motif with a UMAP representation of the gesture kinematic distances and sonic distances. This application allows users to identify whether similar gestures also have similar motifs, and vice versa. The application can be used for further explorations of possible structural combinations of gestural and sonic features in this dataset (e.g., performer/performance clusters, codependencies, boundary gestures). The dashboard (see Figure 3) can thus be used as an exploratory hypotheses-generating tool that will increase the usability of the current dataset, which is also made open access with this publication as a contribution (link

provided on publication). We host the application on a Apache2-supported server which will be online indefinitely: https://tsg-131-174-75-200.hosting.ru.nl/karnatak/. The code for recreating the dashboard and running it locally can be found here (open on publication).

4.3.5.3. Data preparation for the dashboard: We used `r` to prepare the datasets as input for the dashboard. In the DTW distance dataframe, for each motif ($N = 595$) we have DTW distances for particular sonic and kinematic variables (e.g., loudness, hand position etc.) relative to every other motif. In essence we have a high-dimensional NxN embedding space where each motif has a location in that space which is defined relative to the (dis)similarity of all other motifs. For visualisation purposes we can represent this high-dimensional embedding space on a 2D plane using dimensionality reduction. We used `r`-pacakge '`UMAP`' for this. The resulting output is x, y coordinates for each motif where points closer in space are more similar. It should be noted that 2D representations of high-dimensional spaces are often distorted and may not properly reflect the actual global structure of the data. This is why we do not perform statistics on dimensionality reduced data with UMAP, and instead use it as a tool to visualise highly similar sonic or co-vocal gestures for all the motifs in an efficient way.

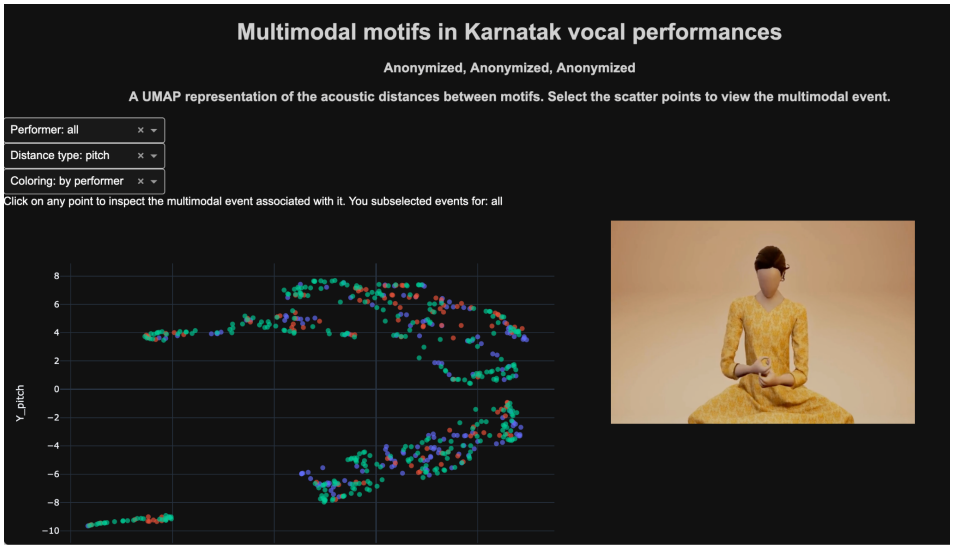

Figure 3. A dynamic dashboard for data exploration. The interface of the dashboard is shown: https://tsg-131-174-75-200.hosting.ru.nl/karnatak/. Users can click on any point in the 2D embedding space and set filters to adjust whether all performers or only one performer is shown, and whether to colour the points by performer, raga or performance, as well as set the variable of interest (e.g., pitch (f0) or hand position). When clicking on a point the animated audiovisual

recording of that gesture-motif is shown. By clicking on nearby points the user can explore highly similar gesture-motifs, and by looking at other regions or clusters in the space the user is able to assess those that are more dissimilar, which can be cross-checked visually by inspecting the animations.

5. Results

Using the dynamic visualisations we qualitatively observe that some recurring motifs show a degree of recurring gesture structure. The following analyses provide a deeper investigation into this phenomenon.

Analysis 1: Do sonic motif DTW distances covary with spatiotemporal patterns of gesture?

Across all performers and performances, we find a significant positive correlation between all kinematic distances and f0, Δf0, loudness and spectral centroid distances (up to 0.42 *r*'s, *p*s < .0001). For individual performers, these correlations are greater (up to 0.53 *r*'s, *p*s < .0001), with notable individual differences observed. Figure 4 shows these correlation coefficients in the form of a heatmap. Non-significant correlations ($p < .0001$) are excluded from the heatmaps and displayed as grey squares.

Overall, loudness correlates most strongly across the kinematic features, with head velocity being most prominent ($r = 0.42$), followed closely by all other kinematic features ($0.31 < r < 0.38$). The relationship between the kinematic features and spectral centroid are much weaker ($0.096 < r < 0.20$), driven largely by Performer 3, with other performers exhibiting either no or very weak correlation.

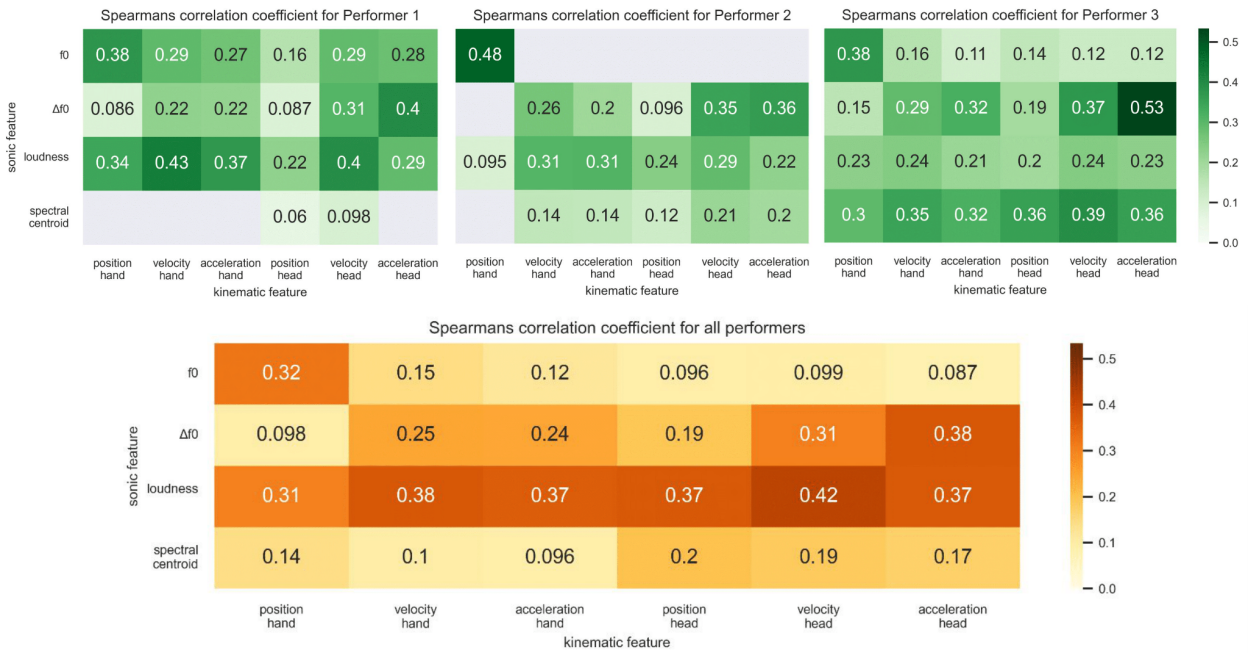

Figure 4: Spearman's correlation coefficient for each sonic and kinematic feature. Absent squares represent tests with *p*-value above the Bonferroni corrected significance level of (0.0001/96). It can be seen that each performer has distinct patterns of correlation strengths. A full table of results is provided in the supplemental materials Table S2.

We observe a notable overall relationship between pitch and the position of the hand (r=0.32). Amongst individual performers this is most pronounced for Performer 2 (r=0.48), but evident across all performers ($0.38 < r < 0.48$). However, with the exception of Performer 1, the relationship between pitch and all other kinematic features is considerably weaker. Of the two pitch-related features, it is in fact with Δf0 that we observe the strongest and most consistent relationship with kinematics. The correlation magnitudes and order of importance between Δf0 and each of the kinematic features is consistent across all performers, the strongest of which is with head velocity/acceleration ($0.31 < r < 0.53$), followed closely by hand velocity/acceleration ($0.22 < r < 0.32$) and finally, more weakly with head/hand position ($0.096 < r < 0.19$).

Analysis 2: Can sonic features be predicted from combined gesture features?

Given that each performer has their own coarticulation landscape that relates to the constraints of movement and vocalisation, we seek to learn whether and how the combined movement features can predict the audio features.

Across all performers and performances, we observe a significantly better than chance prediction of all individual sonic features when using the combined kinematic features; Figure 5 shows the R2 values for these regression models on the test dataset. For models trained on all kinematic features, the most predictable sonic features overall are Δf0 and loudness (both R^2 =0.19), followed closely by f0 (R^2 = 0.14) and spectral centroid (R^2 = 0.11). On a performer level, we observe greater values for the pitch based sonic features: f0 ($0.19 < R^2 < 0.34$), Δf0 ($0.21 < R^2 < 0.30$) and more variation for loudness ($0.11 < R^2 < 0.28$) and spectral centroid ($0.08 < R^2 < 0.25$). Overall and for each of the individual performers, we notice that the predictive power of all head and hand kinematic features combined surpasses either the head or hand features when considered alone.

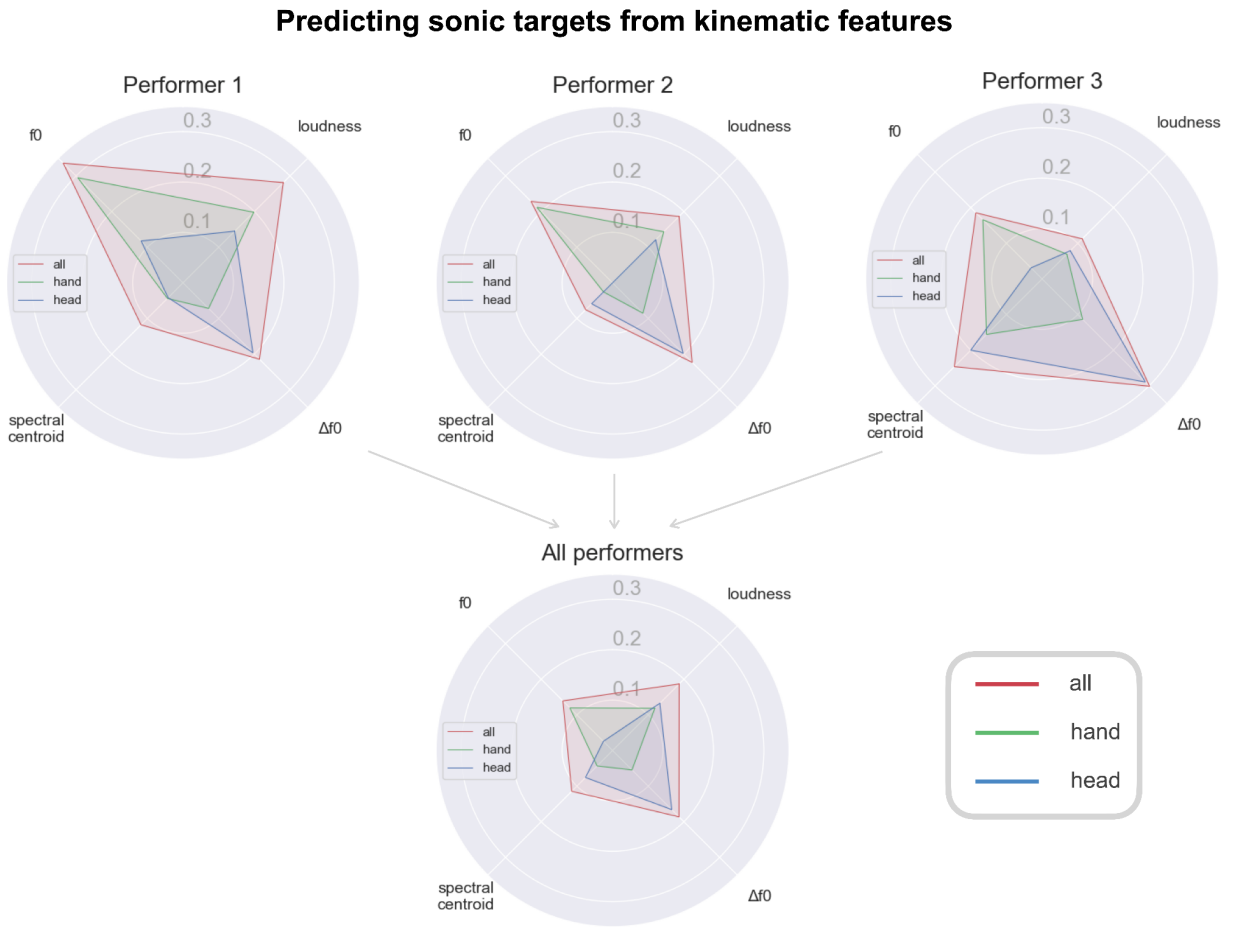

Figure 5: A visualisation of test R^2 values for regression models trained on kinematic features (hands only, head only, and "all" combined) to predict each sonic target for individual and all performers. It can be seen that the combined head and hand movement features have higher R^2 for

predicting a single sonic feature than either head and hand alone. A table of numerical results can be found in the supplemental materials, table S3.

6. Discussion

The results show that sound and body movement are systematically related at the motif level, which suggests the potential for multimodal meaning-formation through contiguity. We note that the various sonic and kinematic features examined differ in the extent to which they co-structure; loudness correlates most strongly across kinematic features, with notable correlations also between pitch (f0) and hand position. However, of the two pitch-related features, it is change in pitch (Δf0) that has the strongest and most consistent relationship with kinematics across all performers. We see that individual performers reliably co-structure sound and movement using differing characteristic salient dimensions (e.g., prioritising co-structuring between pitch and hand position, or change in pitch and head acceleration). The regression results demonstrate that sound-gesture relationships are better understood when hand and head motion is combined, indicating a more whole-body coordination with vocalisation. In sum it is clear from our investigation that performers structurally copattern head and hand gestures with vocalisations, and they exploit different dimensions of possible couplings which likely capture the rich variability between performers that determines their style.

6.1. Individual performer landscapes of coarticulation

It can be seen from Figure 4 that each performer has their own particular landscape[1] of coarticulation. The strongest correlations for performer 1 are between loudness and head/hand movement, with a few other correlations at similar levels, for example between head acceleration and change in pitch. In this way, for Performer 1 the stronger correlations are quite broadly distributed across features, whereas for Performer 2, there is a clearer strongest correlation between pitch and hand position. Performer 3 also has a clear strongest correlation, but in this case between pitch and head acceleration. The regression analyses (see Figure 5) highlight some of the same relationships: for example, showing that Performer 3 is the only one for whom kinematics are

[1] This formulation is inspired by de Haan et al. (de Haan et al., 2013) on the landscape of affordances.

predictive of spectral centroid. The regression analyses also show that in general, head movement is more predictive for change in pitch, while hand movement is more predictive of pitch.

This study thereby opens up opportunities for perception studies (e.g., Huang et al., 2017; Luck et al., 2010; Morrison et al., 2014; Trujillo et al., 2018) as they relate to landscapes of coarticulation. For example, we can ask whether particular performer-specific coupling dimensions, as well as the overall dimensionality of coupling (how diverse is the coupling), resonates with audience members' experience of the performance. This analytical approach is of course not limited to music making, and the pipeline can be applied in behavioural biology to assess the potential mate-selection consequences of multimodal performances in birds (Soma & Shibata, 2023), or the audience's experience of public speaking performances (Chollet & Scherer, 2017).

We hope that methods used in the current article provide a more multidimensional characterisation of multimodal performances that can be used as a basis to understand how such landscapes of coarticulation relate to the expressive and semiotic qualities of different performances and performers. We can see from watching vocalists perform that they have different gesturing styles; the analytical approaches presented here provide insight into why, and allow a quantification of where those differences lie. The analyses also help us understand which features of the musical sound are more strongly indexed by performers (either individually or across individuals) through co-structuring of gesture and sound, which has semiotic implications for the performance and indeed for the musical style. With the opening of the current dataset and a dashboard made for data exploration we invite further research on these topics.

6.2. Conceptualising gesture-vocalisation as coarticulation across effectors

While it is possible to observe in our dataset recurring gestural forms that co-occur with particular types of melodic movement, we argue that both these and the larger array of gesture-vocalisations found in performances are not sufficiently characterised by the idea of isolated sound-gesture units that are structurally combined from a mental library of stereotypical gesture types. In this section we discuss why, with reference to our results above.

Recurrent gestures observed through qualitative analysis of the performances in our dataset include small circular hand motions co-occurring with oscillating ornaments known as *kampita* (see https://youtu.be/FKoucCIcmtM) and two-handed stretching motions co-occurring with a

range of moderately emphatic melodic movements, often involving ornaments that briefly touch on a higher pitch before pulling down onto a lower pitch (see https://youtu.be/9zQd17uSQdY). To better understand vocalists' production of these two recurring gestures, using manual annotation in ELAN video annotation software (Lausberg & Sloetjes, 2009), we analysed their appearance in a subset of the dataset, comprising two ragas (anandabhairavi and atana) each performed by the three performers: a total of 6 performances (results are provided in the supplemental materials, Table S4). The analysis reveals great variability in the way that these recurring gestures are used across performers and performances. Some vocalists produce these recurring gestures more than others. For example, Performer 1 often uses the two-handed stretching gestures (34 times across the two performances), but the gesture is entirely absent in Performers 2 and 3. Instead, for motifs where Performer 1 uses two-handed stretching gestures, Performer 3 often uses a hand gesture that pushes out towards the audience and then back towards his body (see https://youtu.be/ky0uQXINAWY).

Even within performances by a single performer, such recurring gestures do not have a one-to-one relationship to particular motifs. Stretching gestures made by Performer 1 in raga anandabhairavi can be seen accompanying a wide array of somewhat emphatic melodic movements, across at least ten different motifs. Therefore, in these performances, any stable meaning, such as should be apparent in a recurrent gesture (Müller, 2018), appears to be broader than a particular motif, indicating instead a more general melodic/sonic quality, such as emphasis or oscillation. Meanwhile the same performer may produce the same melodic movement with either a stretching gesture, or an entirely different hand gesture. For example, in a performance of raga atana, Performer 1 uses the stretching gesture five times in the first 30 seconds, and then abandons it completely for the remainder of the performance, even when singing the motifs with which the stretching gesture originally co-occurred. Finally, the borders between definitions of recurring gestures can be fuzzy, for example, the "circular" gesture that co-occurs with oscillating melodic movement tends to look like a circle in Performer 1 (see https://youtu.be/FKoucCIcmtM), but in Performer 3, it appears more as a repeated pulsing or pushing motion, without much circular trajectory (e.g., https://youtu.be/sW4LQUAF9xA).

Due to these issues discussed above, we suggest that the gesture-vocal coherence that an audience and performer might experience is best understood as a continuous coarticulated affair, where some aspects are allowed to vary while other degrees of freedom are more co-constrained

for a particular expressive quality or idea. Thus, while it might seem evident that there are undeniable cases of similarity in gestures or vocal sequences that indeed expert annotators can reliably judge upon, there is also more continuous coherence that cannot be clearly expressed in categorical terms but rather in terms of coarticulation landscapes.

The analyses in this article reveal individual performers' landscapes of coarticulation expressed as correlations between dimensions. These landscapes provide insight into individual performers' gesturing habits, which in addition to tendencies towards producing stronger relationships between specific gestural and sonic dimensions, might also include a tendency to produce, or not produce, two-handed stretching gestures, for example. The recurring gestures observed (stretching and circular hand motions) are both the results of constraints (bodily, musical and sociocultural) as well as acting themselves as constraints on what the performer is likely to produce at any given moment, due to their habitual aspect. Individual bodily tensegrity (the way in which different bodies have different action potentialities such as myofascial-skeletal pre-stressedness: (Caldeira et al., 2021; Profeta & Turvey, 2018)) mediating between vocal production and upper body movement may be one factor amongst many that constrain such gesturing habits. Other important constraints include performers' ideation of the music while they sing, their expressive goals in relation to the audience, and also their music learning and life experience (for example, the impact of learning over a period of many years from a teacher who gestures in particular ways). All of these constraints act upon the performer in the moment, resulting in a particular landscape of coarticulation.

6.3. Limitations of the research

One feature of this study that may be considered a limitation in certain musicological contexts is that the motifs found using the machine learning approach are not always segmented at points that would likely be chosen by an expert human annotator. This is the result of the process used, which prioritises the identification of pairwise combinations of regions of high similarity, with no explicit information regarding what might constitute a boundary considered musically meaningful by an expert. For the purposes of this study, this approach is justified as we specifically want to include patterns that have a high degree of sonic similarity. However, as a result of this process, some of the motifs would, from a musicological perspective, be considered incomplete. To improve motif completeness, either an updated automated segmentation approach or manual annotations could

be considered for future research. In order to generalise out to the Karnatak music style as a whole, more performers should be included in future studies. Future research could include studies examining changes in performers' coarticulation landscapes over several years of their training, to learn how the landscapes develop during that process over time.

6.4. Conclusions

It has been argued that meaning in multimodal language can be understood by analysing the neighbouring context, much like how methods in natural language processing can glean semantic information from text using the principle of distributional semantics (Boleda, 2020). In some performance-based styles, such as the present Karnatak music context, gesture and vocalisation appear entangled to an even greater extent than in everyday co-speech gesturing. Here we show that there is indeed a continuous co-structuring of gesture-vocal performances at multiple dimensions. We suggest that this multidimensionality forms a landscape of coarticulation that captures the style of the performer, and when viewed across performers, provides insight into the multimodal semiotic potential of the musical style more broadly. We thereby contribute to the wider project of understanding multimodal meaning making.

References

Bernstein, N. (1967). *The Co-ordination and Regulations of Movements*. Pergamon Press.

Boleda, G. (2020). Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics*, *6*(Volume 6, 2020), 213–234. https://doi.org/10.1146/annurev-linguistics-011619-030303

Brunkan, M. C., & Bowers, J. (2021). Singing with Gesture: Acoustic and Perceptual Measures of Solo Singers. *Journal of Voice*, *35*(2), 325.e17-325.e22. https://doi.org/10.1016/j.jvoice.2019.08.029

Caldeira, P., Davids, K., & Araújo, D. (2021). Neurobiological tensegrity: The basis for understanding inter-individual variations in task performance? *Human Movement Science*, *79*, 102862. https://doi.org/10.1016/j.humov.2021.102862

Chollet, M., & Scherer, S. (2017). Assessing public speaking ability from thin slices of behavior. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 310–

316. https://doi.org/10.1109/FG.2017.45

Clarke, E. (2001). Meaning and the specification of motion in music. *Musicae Scientiae*, *5*(2), 213–234. https://doi.org/10.1177/102986490100500205

Clayton, M., Li, J., Clarke, A., & Weinzierl, M. (2024). Hindustani raga and singer classification using 2D and 3D pose estimation from video recordings. *Journal of New Music Research*, *0*(0), 1–16. https://doi.org/10.1080/09298215.2024.2331788

Cooperrider, K. (2019). Foreground gesture, background gesture. *Gesture*, *16*(2), 176–202. https://doi.org/10.1075/gest.16.2.02coo

D'Amario, S., Ternström, S., Goebl, W., & Bishop, L. (2023). Body motion of choral singers. *Frontiers in Psychology*, *14*, 1220904. https://doi.org/10.3389/fpsyg.2023.1220904

Davidson, J. W. (2001). The role of the body in the production and perception of solo vocal performance: A case study of Annie Lennox. *Musicae Scientiae*, *5*(2), 235–256. https://doi.org/10.1177/102986490100500206

de Haan, S., Rietveld, E., Stokhof, M., & Denys, D. (2013). The phenomenology of deep brain stimulation-induced changes in OCD: An enactive affordance-based model. *Frontiers in Human Neuroscience*, *7*, 653. https://doi.org/10.3389/fnhum.2013.00653

Eitan, Z., & Granot, R. Y. (2009). Primary versus secondary musical parameters and the classification of melodic motives. *Musicae Scientiae*, *13*(1_suppl), 139–179. https://doi.org/10.1177/102986490901300107

Engel, K. C., Flanders, M., & Soechting, J. F. (1997). Anticipatory and sequential motor control in piano playing. *Experimental Brain Research*, *113*(2), 189–199. https://doi.org/10.1007/bf02450317

Fatone, G., Clayton, M., Leante, L., & Rahaim, M. (2011). Imagery, Melody and Gesture in Cross-cultural Perspective. In A. Gritten & E. King (Eds.), *New perspectives on music and gesture* (pp. 203–220). Ashgate Publishing.

Favela, L. H. (2020). Dynamical systems theory in cognitive science and neuroscience. *Philosophy Compass*, *15*(8), e12695. https://doi.org/10.1111/phc3.12695

Feyereisen, P. (2017). *The cognitive psychology of speech-related gesture*. Routledge.

Gallagher, S. (2005). *How the Body Shapes the Mind*. Oxford University Press. http://www.oxfordscholarship.com/view/10.1093/0199271941.001.0001/acprof-9780199271948

Godøy, R. I. (2022). Thinking rhythm objects. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.906479

Godøy, R. I., Jensenius, A. R., & Nymoen, K. (2010). Chunking in Music by Coarticulation. *Acta Acustica United with Acustica*, *96*(4), 690–700. https://doi.org/10.3813/aaa.918323

Hagoort, P., & Özyürek, A. (2024). Extending the Architecture of Language From a Multimodal Perspective. *Topics in Cognitive Science*. https://doi.org/10.1111/tops.12728

Harrison, S., & Ladewig, S. H. (2022). Recurrent gestures throughout bodies, languages, and cultural practices. *Gesture*, *20*(2), 153–179. https://doi.org/10.1075/gest.21014.har

Huang, Y.-F., Coleman, S., Barnhill, E., MacDonald, R., & Moran, N. (2017). How do conductors' movements communicate compositional features and interpretational intentions? *Psychomusicology: Music, Mind, and Brain*, *27*(3), 148–157. https://doi.org/10.1037/pmu0000186

Jerde, T. E., Soechting, J. F., & Flanders, M. (2003). Coarticulation in Fluent Fingerspelling. *Journal of Neuroscience*, *23*(6), 2383–2393. https://doi.org/10.1523/JNEUROSCI.23-06-02383.2003

Kelso, J. A., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C. A. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. *Journal of Experimental Psychology. Human Perception and Performance*, *10*(6), 812–832. https://doi.org/10.1037//0096-1523.10.6.812

Keogh, E. J., & Pazzani, M. J. (2001). Derivative Dynamic Time Warping. In *Proceedings of the 2001 SIAM International Conference on Data Mining* (pp. 1–11). Society for Industrial and Applied Mathematics. https://epubs.siam.org/doi/10.1137/1.9781611972719.1

Kühnert, B., & Nolan, F. (1999). The origin of coarticulation. In W. J. Hardcastle & N. Hewlett (Eds.), *Coarticulation: Theory, data and techniques* (pp. 7–30). Cambridge University Press.

https://www.cambridge.org/core/product/identifier/CBO9780511486395A012/type/book_part

Ladewig, S. H. (2014). 118. Recurrent gestures. In *118. Recurrent gestures* (pp. 1558–1574). De Gruyter Mouton. https://doi.org/10.1515/9783110302028.1558

Latash, M. L. (2008). *Synergy*. Oxford University Press.

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods*, *41*(3), 841–849. https://doi.org/10.3758/BRM.41.3.841

Leante, L. (2009). The Lotus and the King: Imagery, Gesture and Meaning in a Hindustani Rāg. *Ethnomusicology Forum*, *18*(2), 185–206. https://doi.org/10.1080/17411910903141874

Luck, G., Toiviainen, P., & Thompson, M. R. (2010). Perception of expression in conductors' gestures: A continuous response study. *Music Perception*, *28*(1), 47–57. https://doi.org/10.1525/mp.2010.28.1.47

Mani, C. (2017). Gesture in musical declamation: An intercultural approach. *Musicologist*, *1*(1), 6–31. https://doi.org/10.33906/musicologist.373122

McNeill, D. (2000). Catchments and contexts: Non-modular factors in speech and gesture production. In D. McNeill (Ed.), *Language and gesture* (pp. 312–328). Cambridge University Press.

Mittelberg, I., & Hinnell, J. (2023). Gesture Studies and Semiotics. In J. Pelkey & P. Cobley (Eds.), *Bloomsbury Semiotics Volume 4: Semiotic Movements*. Bloomsbury Academic. https://doi.org/10.5040/9781350139435

Morrison, S. J., Price, H. E., Smedley, E. M., & Meals, C. D. (2014). Conductor gestures influence evaluations of ensemble performance. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00806

Mortimer, E. F., & Pereira, R. R. (2023). Recurrent gestures in organic chemistry in tertiary education: Creating emblems through material and embodied actions. *Research in Science & Technological Education*, *0*(0), 1–19. https://doi.org/10.1080/02635143.2023.2287062

Müller, C. (2018). How recurrent gestures mean: Conventionalized contexts-of-use and embodied motivation. *Gesture*, *16*(2), 277–304. https://doi.org/10.1075/gest.16.2.05mul

Nadkarni, S., Roychowdhury, S., Rao, P., & Clayton, M. (2023). Exploring the correspondence of melodic contour with gesture in raga alap Singing. *Proceedings of the 24th International Society for Music Information Retrieval Conference*, 21–28. https://doi.org/10.5281/zenodo.10265213

Nafisi, J. (2013). Gesture and body-movement as teaching and learning tools in the classical voice lesson: A survey into current practice. *British Journal of Music Education*, *30*(03), 347–367. https://doi.org/10.1017/S0265051712000551

Nuttall, T., Plaja-Roglans, G., Pearson, L., & Serra, X. (2022). In Search of Sañcāras: Tradition-Informed Repeated Melodic Pattern Recognition in Carnatic Music. *Proceedings of the 23rd International Conference on Music Information Retrieval (ISMIR), Bengaluru, India*, 337–344. https://repositori.upf.edu/handle/10230/56440

Partan, S. R., & Marler, P. (1999). Communication Goes Multimodal. *Science*, *283*(5406), 1272–1273. https://doi.org/10.1126/science.283.5406.1272

Paschalidou, P.-S. (2017). *Effort in gestural interactions with imaginary objects in Hindustani Dhrupad vocal music* [PhD, Durham University]. http://etheses.dur.ac.uk/12308/

Paschalidou, P.-S., Eerola, T., & Clayton, M. (2016). Voice and movement as predictors of gesture types and physical effort in virtual object interactions of classical Indian singing. *Proceedings of the 3rd International Symposium on Movement and Computing*, 1–2. https://doi.org/10.1145/2948910.2948914

Pearson, L. (2013). Gesture and the sonic event in Karnatak music. *Empirical Musicology Review.*, *8*(1), 2–14. https://doi.org/10.18061/emr.v8i1.3918

Pearson, L. (2016a). Coarticulation and gesture: An analysis of melodic movement in South Indian raga performance. *Music Analysis*, *35*(3), 280–313. https://doi.org/10.1111/musa.12071

Pearson, L. (2016b). *Gesture in Karnatak Music: Pedagogy and Musical Structure in South India* [PhD, Durham University]. http://etheses.dur.ac.uk/11782/

Pearson, L., Nuttall, T., & Pouw, W. (2023). *Motif-Gesture Clustering in Karnatak Vocal Performance: A Multimodal Computational Music Analysis*. ICMPC17-APSCOM7, the Joint Conference of the

17th International Conference on Music Perception and Cognition (ICMPC) and the 7th Conference of the Asia- Pacific Society for the Cognitive Sciences of Music (APSCOM), College of Art, Nihon University, Japan, August 24-28, 2023. https://hdl.handle.net/21.11116/0000-000E-AFCD-7

Pearson, L., & Pouw, W. (2022). Gesture–vocal coupling in Karnatak music performance: A neuro–bodily distributed aesthetic entanglement. *Annals of the New York Academy of Sciences*, *n/a*(n/a). https://doi.org/10.1111/nyas.14806

Pettersen, V., & Westgaard, R. H. (2004). Muscle activity in professional classical singing: A study on muscles in the shoulder, neck and trunk. *Logopedics Phoniatrics Vocology*, *29*(2), 56–65. https://doi.org/10.1080/14015430410031661

Plaja-Roglans, G., Nuttall, T., Pearson, L., Serra, X., & Miron, M. (2023). Repertoire-Specific Vocal Pitch Data Generation for Improved Melodic Analysis of Carnatic Music. *Transactions of the International Society for Music Information Retrieval*, *6*(1), Article 1. https://doi.org/10.5334/tismir.137

Plaja-Roglans, G., Nuttall, T., & Serra, X. (2023). *compIAM* (0.3.0) [Computer software]. https://mtg.github.io/compIAM/

Pouw, W., de Wit, J., Bögels, S., Rasenberg, M., Milivojevic, B., & Ozyurek, A. (2021). Semantically Related Gestures Move Alike: Towards a Distributional Semantics of Gesture Kinematics. In V. G. Duffy (Ed.), *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Body, Motion and Behavior* (Vol. 12777, pp. 269–287). Springer International Publishing. https://doi.org/10.1007/978-3-030-77817-0_20

Pouw, W., & Fuchs, S. (2022). Origins of vocal-entangled gesture. *Neuroscience & Biobehavioral Reviews*, *141*, 104836. https://doi.org/10.1016/j.neubiorev.2022.104836

Profeta, V. L. S., & Turvey, M. T. (2018). Bernstein's levels of movement construction: A contemporary perspective. *Human Movement Science*, *57*, 111–133. https://doi.org/10.1016/j.humov.2017.11.013

Prové, V. (2022). Measuring embodied conceptualizations of pitch in singing performances: Insights from an OpenPose study. *Frontiers in Communication*, *7*. https://doi.org/10.3389/fcomm

Rahaim, M. (2012). *Musicking Bodies: Gesture and Voice in Hindustani Music*. Wesleyan University Press.

Roetenberg, D., Luinge, H., & Slycke, P. (2009). Xsens MVN: full 6DOF human motion tracking using miniature inertial sensors. *Xsens Motion Technologies BV, Tech. Rep*, 1–7.

Schultz, B. G., Brown, R. M., & Kotz, S. A. (2021). Dynamic acoustic salience evokes motor responses. *Cortex*, *134*, 320–332. https://doi.org/10.1016/j.cortex.2020.10.019

Segouat, J. (2009). A study of sign language coarticulation. *ACM SIGACCESS Accessibility and Computing*, *93*, 31–38. https://doi.org/10.1145/1531930.1531935

Soma, M., & Shibata, M. (2023). Dancing in singing songbirds: Choreography in Java sparrows. In Y. Seki (Ed.), *Acoustic Communication in Animals: From Insect Wingbeats to Human Music (Bioacoustics Series Vol.1)* (pp. 95–111). Springer Nature. https://doi.org/10.1007/978-981-99-0831-8_6

Subramanian, L. (2006). *From the Tanjore Court to the Madras Music Academy: A social history of music in South India*. Oxford University Press.

Tagg, P. (2012). *Music's meanings: A modern musicology for non-musos* (Vol. 1). The Mass Media Music Scholars' Press. http://tagg.org/mmmsp/NonMusoInfo.htm

Trujillo, J. P., Simanova, I., Bekkering, H., & Özyürek, A. (2018). Communicative intent modulates production and comprehension of actions and gestures: A Kinect study. *Cognition*, *180*, 38–51. https://doi.org/10.1016/j.cognition.2018.04.003

Viswanathan, T. (1977). The Analysis of Rāga Ālāpana in South Indian Music. *Asian Music*, *9*(1), 13–71. https://doi.org/10.2307/833817

Wiesendanger, M., Baader, A., & Kazennikov, O. (2006). Fingering and bowing in violinists: A motor control approach. In E. Altenmüller, M. Wiesendanger, & J. Kesselring (Eds.), *Music, motor control and the brain* (pp. 109–123). Oxford University Press.

https://doi.org/10.1093/acprof:oso/9780199298723.003.0007

Zbikowski, L. M. (1999). Musical Coherence, Motive, and Categorization. *Music Perception: An Interdisciplinary Journal*, *17*(1), 5–42. https://doi.org/10.2307/40285810

Supplemental materials for "Landscapes of coarticulation: The co-structuring of gesture-vocal dynamics in Karnatak music performance"

Same *-the two audio clips are highly similar, defined as melodically at least 80% the same. There may, however, be some fleeting pitch differences (e.g., gamaka differences), minor start and end point differences, or differences in loudness or speed. If sa is placed on a different pitch but the clips are otherwise the same, the pair should also be considered in this 'same' category.*

Different *- the two are either entirely different, or that less than 80% of one audio clip is found in the other (less than 80% the same). This may sometimes be difficult to assess. If it seems to be a borderline case, please place it in the 'same' category.*

Box S1. Definitions provided to the professional Karnatak vocalist for the same/different annotation process. Feedback from the vocalist indicated that when using these definitions, the majority of motif pairs were clearly either the same or different, but that there were a number of motif pairs (25 out of 800) where it was more difficult to decide.

		hand position	hand velocity	hand acceleration	head position	head velocity	head acceleration
All	f0	0.3232	0.1478	0.1193	0.096	0.099	0.0866
	Δf0	0.0976	0.2477	0.2441	0.1926	0.3058	0.3765
	loudness	0.3055	0.3763	0.3665	0.3739	0.4198	0.372
	spectral centroid	0.1431	0.1006	0.0957	0.2017	0.1854	0.171
Performer1	f0	0.3775	0.2926	0.2696	0.1633	0.2877	0.2774
	Δf0	0.0864	0.2215	0.2189	0.0874	0.3067	0.3996
	loudness	0.3435	0.4285	0.3743	0.2241	0.3952	0.2887
	spectral centroid				0.0602	0.0976	
Performer2	f0	0.4832					
	Δf0		0.2627	0.2038	0.0956	0.3483	0.3621
	loudness	0.0948	0.3072	0.3119	0.2393	0.2883	0.2215
	spectral centroid		0.1442	0.141	0.1197	0.2101	0.1956
Performer3	f0	0.3758	0.164	0.1081	0.1439	0.1236	0.1231
	Δf0	0.1533	0.2908	0.3241	0.1934	0.3711	0.5332
	loudness	0.2272	0.2419	0.2101	0.1987	0.2418	0.2277
	spectral centroid	0.2953	0.3494	0.3202	0.3554	0.3916	0.3577

Table S2: Spearman's correlation coefficient for each sonic and kinematic feature. Empty cells represent tests with *p*-value above the Bonferroni corrected significance level of 0.0001/96. Sample sizes are as follows: All, 61425; Performer 1, 11781; Performer 2, 6786; Performer 3, 52326

target	level	body_part	test_score	n_samples
f0	all	hand	0.1195	176715
		head	0.0258	176715
		all	0.1393	176715
	performer 1	hand	0.2950	11781
		head	0.1175	11781
		all	0.3360	11781
	performer 2	hand	0.2118	6786
		head	-0.0012	6786
		all	0.2285	6786
	performer 3	hand	0.1661	52326
		head	0.0310	52326
		all	0.1856	52326
Δf0	all	hand	0.0544	176715
		head	0.1661	176715
		all	0.1863	176715
	performer 1	hand	0.0718	11781
		head	0.1962	11781
		all	0.2145	11781
	performer 2	hand	0.0853	6786
		head	0.1980	6786
		all	0.2234	6786
	performer 3	hand	0.1136	52326
		head	0.2894	52326
		all	0.3013	52326

target	level	body_part	test_score	n_samples
loudness	all	hand	0.1186	176715
		head	0.1325	176715
		all	0.1868	176715
	performer 1	hand	0.1981	11781
		head	0.1449	11781
		all	0.2813	11781
	performer 2	hand	0.1434	6786
		head	0.1207	6786
		all	0.1864	6786
	performer 3	hand	0.0694	52326
		head	0.0791	52326
		all	0.1124	52326
spectral centroid	all	hand	0.0433	176715
		head	0.0757	176715
		all	0.1144	176715
	performer 1	hand	0.0438	11781
		head	0.0417	11781
		all	0.1178	11781
	performer 2	hand	0.0255	6786
		head	0.0587	6786
		all	0.0754	6786
	performer 3	hand	0.1558	52326
		head	0.2001	52326
		all	0.2462	52326

Table S3: Test R^2 values for regression models trained on kinematic features to predict each sonic target.

		Performer 1	Performer 2	Performer 3
Anandabhairavi	Total singing duration (s)	224	237	380
	Total gesturing duration (s)	178.679	171.2	259.103
	Number of circle gestures	21	10	17
	Total circle gesture duration (s)	17.559	6.95	15.931
	Number of stretch gestures	29	0	0
	Total stretch gesture duration (s)	28.361	0	0
Atana	Total singing duration (s)	101	210	314
	Total gesturing duration (s)	85.084	164.173	218.85
	Number of circle gestures	6	6	10
	Total circle gesture duration (s)	4.19	4.161	12.694
	Number of stretch gestures	5	0	0
	Total stretch gesture duration (s)	5.214	0	0

Table S4: Recurring gesture analysis results. Circle gestures were defined as small repeating circular or pulsing hand gestures. Stretch gestures were defined as a two handed gesture where the hands start together and then pull apart, as though stretching something.