Computer Analysis of Sentiment Interpretation in Musical Conducting

Kelly Karipidou[1⋆] Josefin Ahnlund[1⋆] Anders Friberg[2]
Simon Alexanderson[2] Hedvig Kjellström[1∗]
[1] Robotics, Perception, and Learning, KTH, Sweden, {kellyk,jahnlund,hedvig}@kth.se
[2] Speech, Music, and Hearing, KTH, Sweden, {afriberg,simonal}@kth.se
⋆ KK and JA have contributed equally to the paper ∗ HK is corresponding author

***Abstract*— This paper presents a unique dataset consisting of 20 recordings of the same musical piece, conducted with 4 different musical intentions in mind. The upper body and baton motion of a professional conductor was recorded, as well as the sound of each instrument in a professional string quartet following the conductor. The dataset is made available for benchmarking of motion recognition algorithms.**

An HMM-based emotion intent classification method is trained with subsets of the data, and classification of other subsets of the data show firstly that the motion of the baton communicates energetic intention to a high degree, secondly, that the conductor's torso, head and other arm conveys calm intention to a high degree, and thirdly, that positive vs negative sentiments are communicated to a high degree through other channels than the body and baton motion – most probably, through facial expression and muscle tension conveyed through articulated hand and finger motion.

The long-term goal of this work is to develop a computer model of the entire conductor-orchestra communication process; the studies presented here indicate that computer modeling of the conductor-orchestra communication is feasible.

I. INTRODUCTION

Classical music sound production is structured by an underlying manuscript, the *sheet music*, that specifies into some detail what will happen in the music, e.g. in terms of rhythm, tone heights, and rough dynamics (i.e., changes in sound level). However, the sheet music specifies only up to a certain degree how the music sounds when performed by an orchestra; there is room for considerable variation in terms of timbre, texture, balance between instrument groups, tempo, local accents, and dynamics [16], [21].

In larger ensembles, such as symphony orchestras, the interpretation of the sheet music is done by the *conductor*. The role of the conductor is extremely important; even though the orchestra is an organism in itself, consisting of individuals with opinions on interpretation, the conductor has a major responsibility for the interpretation, and executes this to a higher or lower degree. The variability of the musical outcome dependent on the conductor can be observed by listening to recordings of two generations of the same orchestra (Wiener Philharmoniker) playing the same music piece (Mozart's Symphony #40) with two different conductors (Karl Böhm[1] in 1977 and Daniel Barenboim[2] in 2012). Among other things, the later interpretation is "lighter" and faster than the former. The difference is due to the different interpretations these two conductors have made of the piece – of course, under influence of the traditions about how to perform Mozart at the time of each recording, and the effect the different musicians have on each other during the performance [5], [14]. Our point is that *it is possible for the conductor to inflict these large differences in the way the music sounds to the audience*.

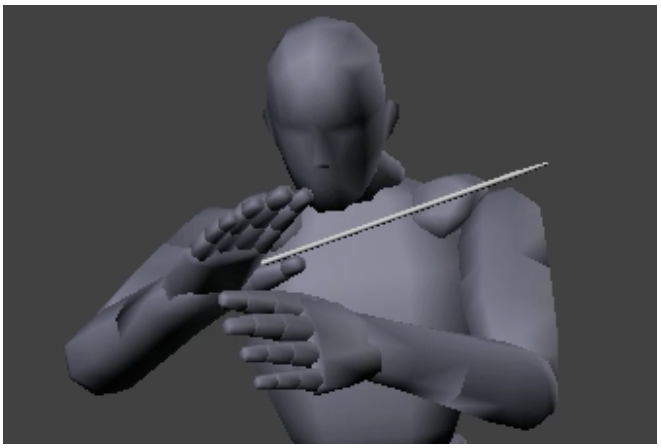


Fig. 1. A novel motion capture data set of musical conducting is presented, along with an experimental analysis of how information about musical intention is communicated from a conductor to an orchestra. The dataset can be used for benchmarking of state-of-the art methods for human motion recognition, and the findings of the analysis are valuable in the design of automatic conducting recognition systems (i.e., gaming applications where you conduct a virtual orchestra, and conductor education systems for analysis of recorded conducting sessions).

Conducting is essentially a complex musical sign language, which has been developed during the last 300 years in order to as efficiently as possible communicate the conductor's musical intentions and help coordinate the music production of an orchestra or ensemble [16]. The conductor has prior to the interaction with the orchestra formulated a clear goal function, a vision on how the musical score should be interpreted. During the rehearsals, the conductor then communicates the interpretation to the orchestra, partly through verbal explanations, but to a very high degree through body language. This effect of the conductor's non-verbal signals on the music performance has been verified scientifically [13].

The conductor can be seen as having two functions; firstly, to coordinate the music production of all groups of the

[1]Clip at www.youtube.com/watch?v=sZHKJQdB_Ng
[2]Clip at www.youtube.com/watch?v=K8Uc5vv4D70

978-1-5090-4023-0/17/$31.00

orchestra and steer the pace of the music; secondly, to provide an interpretation of the music, and communicate this interpretation to the orchestra. In this paper, we focus on the second function, and investigate in a controlled setting how different aspects of the interpretation are communicated non-verbally. The findings are compared to those of Gallops [13], obtained with a very different research method.

Our findings are intended to be used to guide the design of computer models of conductor-orchestra communication. We propose to use a Machine Learning approach; a simplified generative model of the entire music production process of the orchestra in interaction with the conductor is modeled and learned from recorded data: the conductor's articulated body motion in combination with the produced orchestra sound. Sec. III goes further into detail about the process.

This model can be exploited for two applications; the first is "home conducting" systems, i.e., conductor-sensitive music synthesizers, which can, in different levels of sophistication, be used 1) as an orchestra simulator for conducting students [24] whose training is extremely resource-consuming, as the student must spend extensive time practicing with a pianist or an orchestra; 2) for gaming, e.g., Wii or Xbox 360 games where the user conducts a virtual orchestra. This is further discussed in Sec. II.A.

The second application is tools for analyzing conductor-orchestra communication, where latent states in the conducting process are inferred from recordings of conducting motion and orchestral sound. These can be used 1) as a tool for development of conductors and conducting students [24], which would give the possibility of feedback about aspects of the conducting not visible to the human (teacher) eye; 2) as a tool for analysis of human non-verbal communication in general. This is further discussed in Sec. II.B.

The contributions of this paper are twofold (Fig. 1):

Firstly, we present a novel *conducting motion dataset* consisting of 20 recordings of the same musical piece, conducted with 4 different musical intentions in mind. The upper body and baton motion of a professional conductor was recorded, as well as the sound of each instrument in a professional string quartet following the conductor. The data collection is described in Sec. IV.[3]

Secondly, we conduct an *analysis of the dataset*, which shows firstly that the motion of the baton communicates energetic intention to a high degree, secondly, that the conductor's torso, head and other arm conveys calm intention to a high degree, and thirdly, that positive vs negative sentiments are communicated to a high degree through other channels, such as facial expression and muscle tension conveyed through articulated hand and finger motion. The data analysis method is described in Sec. V, and the experiments are described in Sec. VI. Based on these experimental findings, recommendations are given as to how a conducting motion analysis/synthesis system should be designed.

[3]The dataset along with example videos are available at drive.google.com/drive/folders/0B0EA64_cC-3abjVYdlJhRVk0aHc

II. RELATED WORK

A. Computer simulation of the conducting process

The orchestra's conductor-decoding and music production process has been simulated, e.g., by Friberg et al. [10], [11] who present a music performance system that can reflect variations such as speed, articulation, phrasing, and dynamics in a users gestural patterns. Argueta et al. [1] present a similar system that allows control of tempo and the relative emphasis on different instruments. In another line of work, Borchers et al. [4] present a "virtual orchestra" simulator which was on public display at the House of Music in Vienna. The music production is example based, combining recordings of the Wiener Philharmoniker.

In our system design proposal, we instead suggest that the modes of variation in the orchestral sound, as well as the rules for combining them, will be learned from data. This would constitute a major contribution; the current state of the art is "far away from capturing the whole nature of conducting" [16]. The data needed for training would be of the same type as the data presented here – motion capture of articulated conductor motion and the simultaneous orchestra sound – but with a larger number of data samples, and a wider variety of music, conductors, and interpretations. The dataset and analysis presented here is a first step on the way towards such a system.

B. Analysis of the conducting process

During the last decades, there has been a rapid development of methods for machine analysis of human spoken language. However, the information communicated between interacting humans is only to a small part contained in the spoken language; humans transfer huge amounts of information through non-verbal cues: Facial expressions, body motion, muscle tensions, and also the prosody and tone of voice [2], [18]. It is thus of great benefit for the usability of social robots and other intelligent artificial systems to give them capability to interpret human non-verbal behavior, and to produce human-like non-verbal behavior.

However, a great challenge in the study and modeling of human non-verbal behavior is that the underlying factors, as well as the effect of the communication on the dialogue partner, is unobservable. The problem of learning generative computer models from observed training data is thus, in the general case, ill-posed. One approach to address this problem is to study human non-verbal communication in constrained settings, and extrapolate findings from there to the general case of human non-verbal communication.

Musical conducting is such a special case. A conductor has, by years of training, *taken control of his or her communicative face and body language*, and uses it to reach specific outcomes, which the conductor is aware of [16]. An educated conductor is aware of their own encoding process; studies [25] have shown that conductors can recognize their own conducting motion among the motion of other conductors, to a higher degree than they can recognize their subconscious motion patterns, like walking style. Moreover, although different conductors have vastly different conducting

styles, musicians can correctly interpret the intentions of a conductor they have not met before [13]. Although musicians have trained interpretation of conducting motion for a long time, the consistency over different individuals indicates that there are aspects of the perception and production of musical communicative gestures that are common between many different individuals, thus intrinsic to human cognition.

The conductor is aware of certain aspects in his or her musical intentions, and can verbally express them [13]. This means that it is possible to retrieve, in parallel with the conductor's gestures, some aspects of the intention of the conductor. Furthermore, we can observe the result of the orchestral encoding process in the form of sound. The conducting application thus makes it possible to learn about the encoding and decoding processes of human non-verbal communication in a more constrained and observable setting.

Moreover, computer analysis of the conducting process is also interesting from a music-theoretical and music-education perspective. The conductor develops his or her style by practicing. This process is naturally extremely resource-consuming, as the conducting student requires access to a pianist or a small orchestra for several hours a week during their entire education. Siang et al. [24] suggest that the kind of system proposed in Sec. III would be of great benefit in the education of conducting students, both for post-analysis of recorded sessions with a live orchestra, and as a simulator system, where the orchestra's conductor-decoding and music production process is simulated.

A very ambitious study was conducted by Gallops [13]. He recorded video of the hands of a number of conductors conducting the same music piece with different musical interpretations. A number of musicians were then recorded performing the piece under guidance from the recorded conducting hand motions. Baseline recordings without conductor were also done. A set of musical experts then labeled both the conducting sequences and the musical performances according to a music-theoretically informed protocol. Statistical analysis showed high correlation between conducting and musical performance.

In the study presented in this paper, we repeat this experiment in a smaller setting with a more controlled labeling protocol. Since we want to capture the entire conductor-musician interaction in an as natural setting as possible, we record the conductor and musicians during live interaction.

However, since our intended application is computer modeling of the conductor-orchestra communication process, we employ a different scientific method. Instead of letting human experts represent and classify the conducting motion and music sound, we have set up the recording so that the conductor, so to speak, labels each recording session. The recording procedure is described in Sec. IV. Representative features are then automatically extracted from the recorded conducting motion (Sec. V.A) and music sound (Sec. V.B) signals. The degree of correlation between the conductor's musical intention and different aspects of the conducting motion and music sound is then investigated in terms of how well an automatic, HMM-based classifier (Sec. V.C) can distinguish different underlying intentions given the observed motion and sound features.

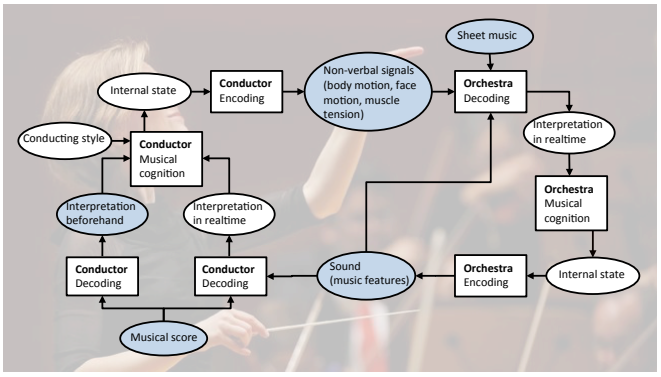


Fig. 2. A flow model of the conductor-orchestra interaction process, intended as a basis for computer implementation. White squares indicate cognitive processes, with incoming variables indicated by incoming arrows, and outgoing variables indicated by outgoing arrows. White ellipses indicate latent variables, where neither the value, nor the state-space can be measured or observed. Blue ellipses indicate observable variables, that can be measured in an objective manner (with more or less uncertainty).

The findings and the labeling protocol from the study by Gallops are interesting for the further development of our computer modeling, and also for the setup of recordings of training data for our computer models. This is further discussed in Sec. VII.

III. THE CONDUCTING PROCESS

A. Modeling the conducting process

For the purpose of computer implementation, we propose to model the conducting and orchestra music production process as outlined in Fig. 2. The configuration of encoding and decoding processes is based on the model by MacKay [19], pp. 16–21, and adapted for the conducting case with support in [13], [16] and discussions with professional conductors and musicians (see Sec. VIII). We have here omitted the verbal communication that takes place beforehand.

A few things can be noted about the model. Firstly, the choreography of a conductor is to some degree determined by an agreed-upon set of gestures, that follow from the music noted down in the score. However, there are enormous style variations between different conductors. The conductor style is static over the temporal extent of the interaction. The musical intention beforehand can be observed by talking to the conductor before the interaction. However, the resulting non-verbal signals depend also to a very high degree on the dynamics during the conductor-orchestra interaction – the conductor reacts to the orchestra sound and adjust the communication in order to, e.g., correct aspects of the orchestra interpretation that differ from the conductor's original intention, such as further emphasizing a crescendo or bringing out an instrumental group.

Secondly, the orchestra musicians decode the conductor's signals, and generate music from those and the sheet music (which are parts of the score relating to their instrument). The decoding and interpretation process of an orchestra is

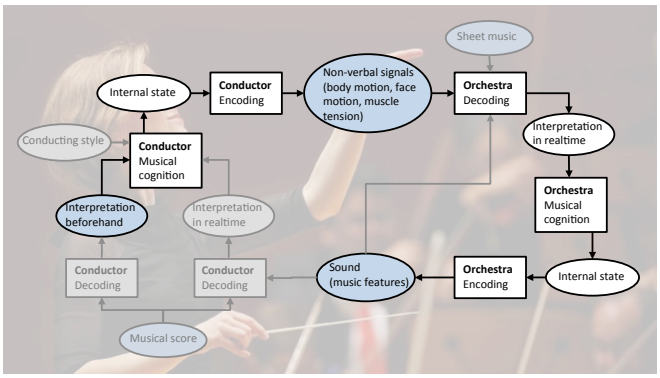


Fig. 3. A simplified version of the conducting process, modeled in this paper. Certain variation was removed, mainly due to limitations in recording resources: one conductor, one piece of music. Interpretation of the music was structured according to a sentiment classification protocol: Each recording was done with one of four sentiments in mind. The sentiment space and the sentiment labels for each recording were only known to the conductor, not to the musicians.

extremely complex, with a very intricate pattern of interaction between different instruments. The proposed computer model in Fig. 2 does not capture all details of this process, as the focus of the model is on the analysis of the conducting motion.

B. Modeling communication of sentiment from conductor to orchestra

As described in the introduction, our long-term aim is to model the conducting process in Fig. 2, in order to create computer applications where the orchestra encoding, musical cognition, and decoding processes are automated; or to provide analysis of recorded conductor-orchestra interaction.

In this paper, we take a first step towards this goal by training a simplified version of the model that encompasses transfer of interpretation from conductor to orchestra, as shown in Fig. 3. To train the sentiment recognition system, we recorded a sequence of conductor-orchestra interactions with controlled variation of the sentiment (interpretation) that the conductor had in mind for each recording. The data collection is described in the next section, while the sentiment classification is described in Sec. V. In Sec. VI.A, the system is evaluated in terms of how well sentiments are communicated through different parts of the signal.

In addition, our training data enables analysis of to what degree the musicians in reality perceived the conductor's intended sentiment. As discussed in Sec. II.B, this experiment is along the same lines as the study by Gallops [13], but with a different research method. It is presented in Sec. VI.B.

IV. DATA COLLECTION

A. Recording setup

We brought a professional conductor and a string quartet consisting of four professional musicians to our motion capture studio, and captured 20 recordings of the first 28 bars of *Ludvig Norman, Andante Sostenuto in F major, op. 65*. See footnote [3] for information on how to obtain the data.

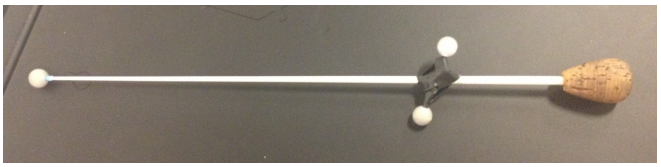


Fig. 4. The baton with three markers, tracked as a 6D rigid object in the OptiTrack motion capture system.

The conductor's upper body pose (a kinematic skeleton with 18 joints), and the baton pose (a rigid rigid body, see Fig. 4), were recorded with a time frequency of 120 Hz using an OptiTrack Prime 41 3D motion capture system. The skeleton data was transformed to 3D joint positions to provide a better metric for classification according to [22].

Each instrument in the string quartet was recorded in a separate sound channel using DPA d:vote microphones attached to the instruments.

B. Sentiment labeling

To enable an objective ground truth of conductor intent, each of the 20 recordings got a sentiment label before the recording was made. The conductor then interpreted how the sentiment label should affect the performance of the 1 min piece, and communicated this interpretation to the musicians through non-verbal gestures. (This is in contrast to the study by Gallops [13] where the recordings and the conductor motion were labeled by musical experts after the recording was made.) There were four labels, and five instances of each interpretation were recorded. The interpretations were arranged in a random (computer-generated) order only known to the conductor, and no instructions were given to the orchestra other than that they should ignore all dynamics notation in the score, and follow the conductor.

The vocabulary of sentiment labels was developed according to the logic shown in Fig. 5. The four selected sentiments, *sad*, *neutral*, *passionate*, and *angry*, represent a crude segmentation of the 2D sentiment space into equally sized regions. The lower left quadrant, *sad*, was not used since it is difficult for the conductor to convey a calm and negative interpretation of a music piece in a major ("happy") key; it would be virtually impossible for the musicians to

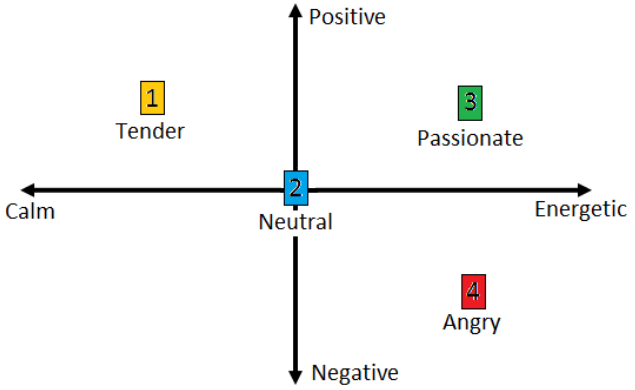


Fig. 5. The 2D valence-activation space of sentiments used in the data collection.

discern it from the calm and positive interpretation, *tender*. The semantic class names were determined together with the conductor.

The use of a 2D sentiment space spanned by a valence dimension (positive vs negative) and an activation dimension (intensity) is motivated by the findings of Russel [23], who shows that such a space is more in accordance with how humans self-report and describe their emotional state, than the more common discrete emotion spaces with 6-12 perpendicular emotion dimensions such as "anger", "sadness", "fear", etc, e.g., as suggested by Ekman and Rosenberg [9].

It should be noted that this labeling system is overly simplified – in reality, the conductor's emotional intent can be described as a time-varying continuum over the 2D sentiment space. Moreover, the emotional intent is not only conveyed in terms of pose and motion locally in time. Timing in the conducting motion is also a means for the conductor to change the emotional expression of the music; in addition to changing the timing of the music (which it naturally does), it also changes how the musicians express the music, thus affecting the emotional expression.

However, while removing some interesting variation and detail of musical expression from the recordings, our discrete and time-static labeling protocol makes the data suitable for a quantitative evaluation of the emotional information content in different parts of the conducting signal.

The same conductor appears in all recordings, which means that the sentiment classifier trained with this data, evaluated in Sec. VI.A is person-specific. Thus, much additional work is needed in order to create a robust sentiment recognition method for home conducting systems – the present work can be seen as a feasability study, which indicates what non-verbal features need to be recorded in order to communicate effectively. Moreover, the dataset presented in this paper can serve as a benchmark dataset for state-of-the-art motion recognition methods.

However, note that the musicians were not trained to interpret this particular individual, nor were they aware of the sentiment space or labels. The sentiment classification from sound is thus person-independent. In the experiment in Sec. VI.B, we show that emotional intent was successfully transfered from conductor to orchestra.

In the rest of the paper, we will use the following mathematical notation: Upper body sequences are denoted $\{X_1^{\mathrm{ub}}, X_2^{\mathrm{ub}}, ..., X_{20}^{\mathrm{ub}}\}$ and baton sequences are denoted $\{X_1^{\mathrm{b}}, X_2^{\mathrm{b}}, ..., X_{20}^{\mathrm{b}}\}$, where X_i^{ub} is a $48 \times T_i$ matrix where each column $\mathbf{x}_{i,t}^{\mathrm{ub}}$ of length 48 (3D positions of joints) correspond to one frame $t \in [1, T_i]$, and X_i^{b} is a $12 \times T_i$ matrix where each column $\mathbf{x}_{i,t}^{\mathrm{b}}$ of length 12 (3D positions of markers) correspond to one frame $t \in [1, T_i]$. Concatenated upper body and baton sequences are of size $60 \times T_i$ and denoted X_i^{ubb}. Only the sound of one of the instruments, violin 1, is used in this study.[4] The violin 1 sound sequences are denoted $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_{20}\}$ where $\mathbf{s}_i$ is a vector of length τ_i. Moreover, sentiment labels for the 20 sequences are denoted $\{y_1, y_2, ..., y_{20}\}$, where $y_i \in [tender, neutral, passionate, angry]$.

[4]Violin 1 is the leader of the string quartet, thus the musician with the most responsibility in interpreting the conductor, and also the instrument with the most melody, i.e., most information in its sound signal.

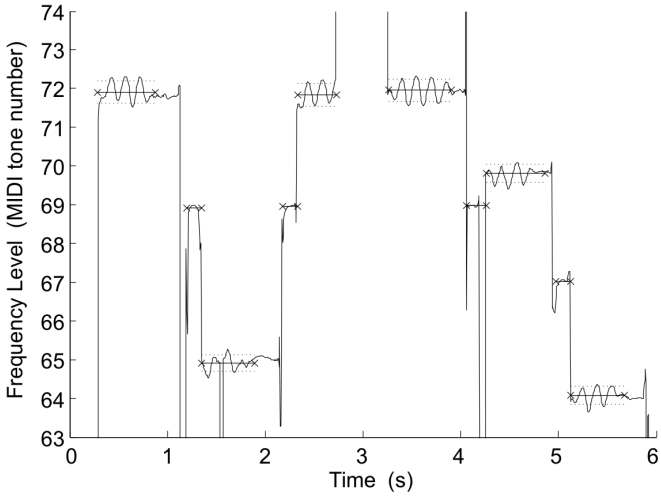


Fig. 6. An example of some of the extracted parameters in CUEX, from [12]. It shows the pitch parameters (curve), the detected notes (x-x), and the vibrato extent (dashed lines) for a short melody.

V. DATA ANALYSIS

A. Conductor motion representation

For the purpose of sentiment classification, a straightforward representation of a motion sequence is used, Z, where each column $\mathbf{z}_t$ correspond to one frame $t \in [1, T]$. The representation at time frame t is:

$$\mathbf{z}_t = \begin{bmatrix} \mathbf{x}_t \\ \mathbf{x}_{t+1} - \mathbf{x}_{t-1} \\ 2\mathbf{x}_t - \mathbf{x}_{t+1} - \mathbf{x}_{t-1} \end{bmatrix} \quad (1)$$

i.e., the concatenation of pose and its first and second derivative (velocity and acceleration). The baton representation $\mathbf{z}_t^{\mathrm{b}}$ of a particular frame t is thus 36D, while the upper body + baton representation $\mathbf{z}_t^{\mathrm{ubb}}$ of a particular frame t is 180D.

Moreover, low-dimensional approximations of the baton and upper body + baton signals are also extracted using PCA. The signal containing the 3 dominant modes of variation in the baton signal is denoted $\mathbf{z}_t^{\mathrm{b3}}$. The signals containing the 3 and 10 dominant modes of variation in the upper body + baton signal are denoted $\mathbf{z}_t^{\mathrm{ubb3}}$ and $\mathbf{z}_t^{\mathrm{ubb10}}$, respectively. More sophisticated low-dimensional representation techniques are discussed in Sec. VI.A and VII.

B. Violin 1 sound representation

The sound signals $\mathbf{s}$ are represented in terms of their musical content as analyzed by the CUEX algorithm [12]. The algorithm is designed for a homophonic sound signal, containing a single melody. For details regarding the parameter extraction, see [12].

Fig. 6 illustrates how the parameters are extracted. The signal pitch (frequency level) $\mathbf{p}$, sound level (volume) $\mathbf{v}$ and spectral balance (ratio of high vs low frequency content) $\mathbf{sb}$ are first extracted from $\mathbf{s}$. In music, notes correspond

to parts of the signal with similar pitch over time. Notes can be followed by silence or by another note. Based on this, the algorithm then detects the times for onset (start) t_n^{on} and offset (end) t_n^{off} of each note n in the sound signal. Onsets and offsets are marked with $\times$ in the graph.

Each note n in the sound signal $\mathbf{s}_i$ is then characterized with a vector $\mathbf{c}_{i,n}$, consisting of 6 parameters:

- Average sound level v between t_n^{on} and t_n^{off},
- Articulation $\frac{t_n^{\text{off}} - t_n^{\text{on}}}{t_{n+1}^{\text{on}} - t_n^{\text{on}}}$,
- Onset velocity, i.e., how fast the tone is attacked, i.e., sound level derivative v' at t_n^{on},
- Average spectral balance sb between t_n^{on} and t_n^{off}.
- Average vibrato rate, i.e., how fast the pitch $\mathbf{p}$ oscillates between t_n^{on} and t_n^{off}.
- Average vibrato extent, i.e., with what amplitude $\mathbf{p}$ oscillates between t_n^{on} and t_n^{off}.

Together, these parameters capture a large portion of the expressive differences in the sound of a violin. Thus, the CUEX signal C_i is a rich and adequate representation of the sound signal s_i from violin 1 during recording i.

The challenges with representing polyphonic music, i.e., music with several melodies (as that emitted from an orchestra) are discussed in Sec. VI.B and VII.

C. Sentiment classification

For classification we use a standard HMM approach [8]:

Let $\{W_i\}$, be a set of training sequences (matrices where each column $\mathbf{w}_{t,i}$ correspond to one frame t) and $\{y_i\}$, $y \in [1, L]$ their corresponding labels. A set of k Gaussians are fitted to all frames $\{\mathbf{w}_{t,i}\}$, providing a GMM representation of the distribution over $\mathbf{w}$. These Gaussians are the hidden states of the HMM. Each frame $\{\mathbf{w}_{t,i}\}$ is then approximated with a multinomial distribution $\{\mathbf{g}_{t,i}\}$ of size k, representing the proportional density of hidden states for that frame; a "soft" version of the usual HMM discrete hidden state representation. For each class $l \in [1, L]$, the portion of sequences G_i where $y_i = l$ are taken out and used to learn a transition matrix:

$$\mathcal{T}_l \propto \sum_{i:y_i=l} \sum_{t=2}^{T_i} \mathbf{g}_{t-1,i} * \mathbf{g}_{t,i}^T \tag{2}$$

The most probable class $\hat{y}^*$ of an unseen sequence W^* can now be found with the Viterbi algorithm, approximating W^* by the GMM approximation G^* (i.e., "soft" state assignments for all frames in W^*), and comparing the probabilies of Viterbi paths through each transition matrix $\mathcal{T}_l$.

A 0:th order approximation of this is to learn a typical hidden state distribution for each class $l \in [1, L]$:

$$\mathcal{g}_l \propto \sum_{i:y_i=l} \sum_{t=1}^{T_i} \mathbf{g}_{t,i} \tag{3}$$

The most probable class $\hat{y}^*$ of an unseen sequence W^* can now be found by approximating W^* by the GMM approximation G^* (i.e., "soft" state assignments for all frames in W^*), and comparing the average hidden state $\bar{\mathbf{g}}^*$ of that sequence to each typical hidden state $\mathcal{g}_l$. This classification requires less training data and is more robust but less accurate than the full HMM classifier.

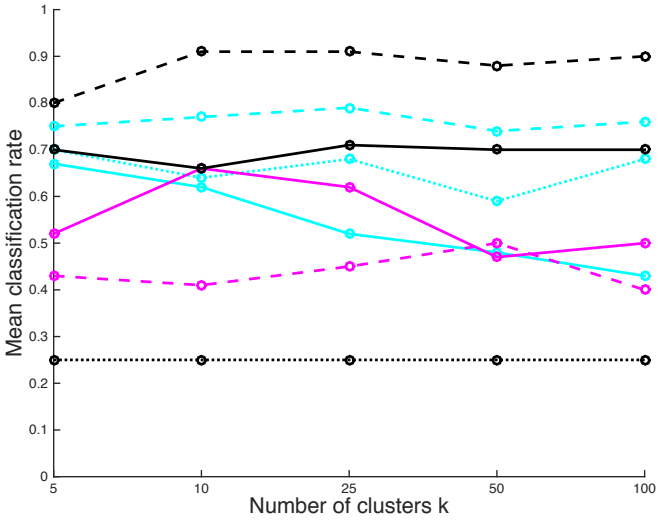


Fig. 7. Mean classification rates as a function of number of clusters k. Solid cyan curve = Z^{ubb}, raw 180D upper body + baton. Dashed cyan curve = Z^{ubb3}, PCA 3D projection of upper body + baton. Dotted cyan curve = Z^{ubb10}, PCA 10D projection of upper body + baton. Solid magenta curve = Z^{b}, raw 36D baton. Dashed magenta curve = Z^{b3}, PCA 3D projection of baton. Solid black curve = C, 6D violin 1 sound features. Dashed black curve = C^0, 0:th order classification, 6D violin 1 sound features. Dotted black curve = baseline corresponding to chance.

The next section describes how this classification method is applied to the different kinds of data described in Sec. V.A ($W_i = Z_i$) and V.B ($W_i = C_i$), in order to evaluate to what degree they correlate with the emotional labels.

VI. EXPERIMENTS

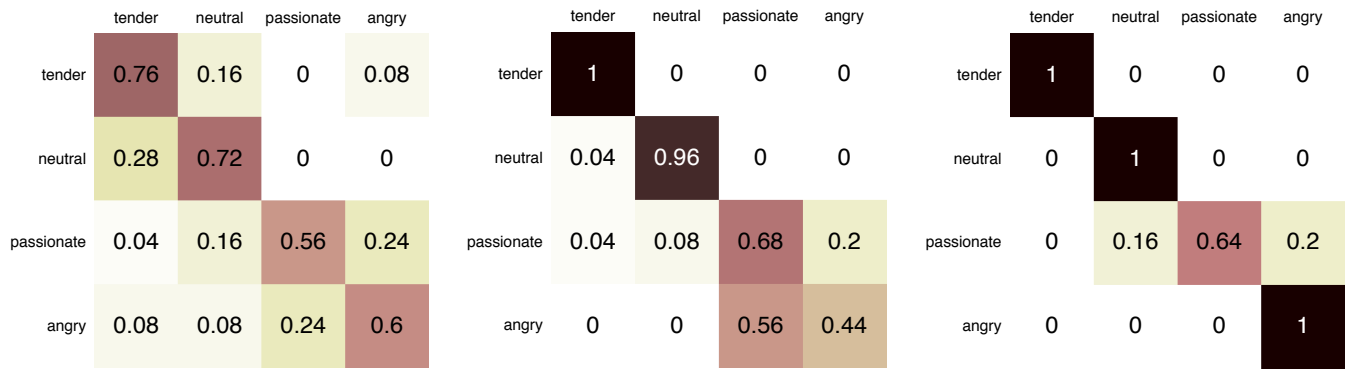
A. Evaluation of automatic sentiment recognition

The goal of the first experiment is to evaluate the emotional information content in the conducting motion, to learn about what parts of the signal carry different aspects of the emotional intention. This is intended, firstly, as a feasibility study for a "home conducting" or conducting analysis system which is the long term goal of our research. Moreover, the classification experiments in this subsection are also intended as a baseline for future benchmarking of state-of-the-art motion classification methods.

We use a stratified cross-validation procedure, where 1 recording of each sentiment is taken out as test sequence and the other 4 for training, to generate 5 training-test configurations for each sentiment. Since the GMM clustering step of the HMM training is random to a certain degree, we run the classification for each configuration 5 times, and calculate the mean of the 25 test classifications of each of the 4 sentiments.

Fig. 7 shows the mean classification rates for the 5 different motion signals, and the sound signal, for different numbers of clusters k in the HMM's pose vocabulary. From this we can draw the following two conclusions:

Firstly, there are significant correlation between the conductor's intended sentiment and the conductor's upper body



(a) Z^{b}, baton, $k = 10$ (b) Z^{ubb3}, upper body + baton, $k = 10$ (c) C^{0}, violin 1 sound, $k = 10$

Fig. 8. Confusion matrices for baton (a), upper body + baton (b), and violin 1 sound (c).

and baton motion (cyan curves). Moreover, only observing the motion of the baton is sufficient to a certain degree (magenta curves). This has implications for the design of home conducting gaming and conducting training systems, since motion capture of the entire upper body is more cumbersome than just capturing the motion of a single rigid body, the baton.

Secondly, linear dimensionality reduction of the upper body + baton motion helps removing spurious variation in the different recordings of the same sentiment; the classification is persistently better over different values of k. It should be noted that the clustering into a discrete state space in the HMM classification also is a representation which addresses the curse of dimensionality; thus the raw data with possible linear dimensionality reduction suffices in this classification, if there are large enough clusters so that spurious variation is removed in the clustering process, but this discrete representation also fails to capture many aspects of the variation in the data. The fact that the upper body + baton data benefit from linear dimensionality reduction but not the baton data is in accordance with knowledge about the curse of dimensionality; the raw upper body + baton data spans 5 times as many dimensions as the raw baton data. The true mapping function between the conductor's emotional intent and the conductor's articulated motion is probably non-linear in such a way that the poses lie on very wrinkled and entangled manifolds [3]. This is indicated by the fact that for the very articulated upper body + baton space, the best is to use only the three largest eigenmodes (and remove much entanglement but also much information), while for the baton space, the best is to use all dimensions.

We proceed by studying qualitative results. Fig. 8(a) and (b) shows the confusion matrices for classification for the baton signal and upper body + baton signal, respectively, for $k = 10$. Two things can be observed:

Firstly, if the upper body is not observed, just the baton (Fig. 8(a)), there is more confusion between the intents *tender* and *neutral* than if the upper body is also captured (Fig. 8(b)). This is probably because cues about calmness intention were communicated with the conductor's left (non-baton) hand, torso and/or head. Domain knowledge about musical conducting indicates that the left hand is very important for communicating this information.

Secondly, in both Fig. 8(a) and (b) there is a high degree of confusion between the intents *passionate* and *angry*, which are distinguished in the sentiment space by the degree of positivity. There could be two reasons for this: Either the conductor successfully was able to convey the difference between these two intents to the musicians, but the recorded upper body + baton motion failed to capture this difference, or the conductor failed to express the difference. In order to answer this question, we now proceed to study how the violin 1 sound signal correlate with the conductor's emotional intentions.

B. Evaluation of musicians' sentiment recognition

To verify to what degree the conductor's intent was communicated to the musicians, the CUEX signals C_i extracted from the violin 1 sound were also classified in the same manner as the conductor motion. Fig. 7 (black curves) and show the results of both full HMM classification (Eq. 2) and 0:th order classification (Eq. 3). The 0:th order classification (black dashed curve) works better, probably due to scarce training data, and we will in the following consider this.

As as can be seen in Fig. 7, the conductor's intent was indeed communicated with over 90% accuracy, even though the musicians were not aware of the sentiment space, or even of the fact that emotional intent was varied in a controlled fashion and that each recording had a time-constant sentiment label. Moreover, the musicians had not trained to interpret this particular conductor (although they all know him).

Looking closer at a confusion matrix in Fig. 8(c), we see that all intents except *passionate* are perfectly detected. Most notably, the musicians spotted all *angry* recordings immediately. This indicates that the differences between positive and negative emotional intention were in fact communicated successfully to the musicians, but that the conductor used cues other than the full body motion, such as facial expressions and hand and finger articulation, which were not captured in the conductor signal recording.

The results in Fig. 8 concord with the studies of Gallops [13] in that the conductor can communicate emotional intent accurately without verbal communication, and also with the thesis of Russel [23] about the topology of the sentiment space; using any signal, there is a higher degree of confusion between intentions that are close in this space, and virtually no confusion between *tender* and *angry* which are far apart.

VII. CONCLUSIONS

In this paper, we reported about the recording of a dataset consisting of 20 recordings of the same musical piece, conducted with 4 different musical intentions in mind. The upper body and baton motion of a professional conductor was recorded, as well as the sound of each instrument in a professional string quartet following the conductor, as detailed in Sec. IV. The dataset is made available [3] for benchmarking of motion recognition algorithms.

As described in Sec. V and VI, a conductor intention classification method was trained with subsets of the data, and classification of other subsets of the data showed firstly (from Fig. 8(b)), that the motion of the baton communicates energetic intention to a high degree, secondly (from differences between Fig. 8(a) and (b)), that the conductor's torso, head and other arm conveys calm intention to a high degree, and thirdly (from differences between Fig. 8(a,b) and (c)), that positive vs negative sentiments are communicated to a high degree through other channels than the body and baton motion; most probably, through facial expression and muscle tension conveyed through articulated hand and finger motion.

The long-term goal of this work is to develop a computer model of the entire conductor-orchestra communication process, as described in Sec. III. In order to reach this goal, a number of research efforts are needed:

More sophisticated representations of the conductor communication should be developed. Results in Sec. VI.A indicated that the mapping from conductor emotional intent to pose and motion is non-linear with a high degree of entanglement. To disentangle the representation, we propose a deep generative model learned from data, such as a deep autoencoder [17], [20] or a deep Gaussian process model [6], [7]. The same kinds of models can be used to model all parts of the conductor-orchestra communication process (white boxes in Fig. 2).

In order to train such models, an extensive data collection effort is needed. We are planning further data collections with a wider range of music, more elaborate intent labeling, more conductors, more motion features such as face and hand articulation, and larger orchestra settings.

Moreover, while this study focused on communication of emotional intent, the other function of the conductor in the orchestra, to coordinate and control timing, should also be addressed in the modeling. We are currently developing a conducting gesture tracking method based on LSTM [15], which allows control of the speed of the music (but not timbre, attack etc). The method will be the core machinery in a demonstrator in which a user can control the play-back speed of a recorded orchestra.

VIII. ACKNOWLEDGMENTS

The authors gratefully acknowledge the input from and discussions with the professional conductors David Björkman, Simon Crawford-Phillips, and Daniel Harding, and the professional orchestra musicians Christian Bergqvist, Jakob Ruthberg, and Malin Broman.

REFERENCES

[1] C. R. Argueta, C.-J. Ko, and Y.-S. Chen. Interacting with a music conducting system. In *Human-Computer Interaction*, 2009.
[2] M. Argyle. *Bodily Communication*. Routledge, 1988.
[3] Y. Bengio, A. Courville, and P. Vincent. *Representation Learning: A Review and New Perspectives*. arXiv:1206.5538v3, 2014.
[4] J. Borchers, E. Lee, W. Samminger, and M. Mühlhäuser. Personal orchestra: A real-time audio/video system for interactive conducting. *Multimedia Systems*, 9:458–465, 2004.
[5] A. Camurri, B. Mazzarino, M. Rcchetti, R. Timmers, and G. Volpe. Multimodal analysis of expressive gesture in music and dance performances. In *Gesture-Based Communication in Human-Computer Interaction*. 2004.
[6] Z. Dai, A. Damianou, J. González, and N. Lawrence. Variational auto-encoded deep Gaussian processes. In *International Conference on Learning Representations*, 2016.
[7] A. Damianou and N. Lawrence. Deep Gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, 2013.
[8] A. Davies, C. H. Ek, C. J. Dalton, and N. W. Campbell. Generating 3D morphable model parameters for facial tracking: Factorising identity and expression. In *International Conference on Computer Vision Theory and Application*, 2012.
[9] P. Ekman and E. L. Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS)*. Oxford University Press, 1998.
[10] A. Friberg. Home conducting – control the overall musical expression with gestures. In *International Computer Music Conference*, 2005.
[11] A. Friberg. pDM: an expressive sequencer with real-time control of the KTH music performance rules. *Computer Music Journal*, 30(1):37–48, 2006.
[12] A. Friberg, E. Schoonderwaldt, and P. N. Juslin. CUEX: An algorithm for automatic extraction of expressive tone parameters in music performance from acoustic signals. *Acta Acustica united with Acustica*, 93(3):411–420, 2007.
[13] R. W. Gallops. *The Effect of Conducting Gesture on Expressive-Interpretive Performance of College Music Majors*. PhD Thesis, University of South Florida, 2005.
[14] D. Glowinski, M. Manchini, R. Cowie, A. Camurri, C. Chiorri, and C. Doherty. The movements made by performers in a skilled quartet: a distinctive pattern, and the function that it serves. *Frontiers in Psychology*, 4, article 841, 2006.
[15] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber. *LSTM: A Search Space Odyssey*. arXiv:1503.04069, 2015.
[16] G. Johanssen and T. Marrin Nakra. Conductors' gestures and their mapping to sound synthesis. In *Musical Gestures: Sound, Movement, and Meaning*. 2009.
[17] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.
[18] M. L. Knapp, J. A. Hall, and T. G. Horgan. *Nonverbal Communication in Human Interaction*. Wadsworth, 2013.
[19] D. MacKay. Formal analysis of communicative processes. In *Non-Verbal Communication*. Cambridge U. Press, 1972.
[20] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. In *International Conference on Learning Representations*, 2016.
[21] C. Palmer. Music performance. *Annual Review of Psychology*, 48(1):115–138, 1997.
[22] J. Romero, H. Kjellström, C. H. Ek, and D. Kragic. Non-parametric hand pose estimation with object context. *Image and Vision Computing*, 31(8):555–564, 2013.
[23] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 1980.
[24] G. P. L. Siang, N. A. Ismail, and P. Y. Yong. A study on potential of integrating multimodal interaction into musical conducting education. *Journal of Computing*, 2(5):48–52, 2010.
[25] C. Wöllner. Self recognition of highly skilled actions: A study of orchestral conductors. *Consciousness and Cognition*, 21(3):1311–1321, 2012.