



HAL
open science

Analysing Gesture and Sound Similarities with a HMM-based Divergence Measure

Baptiste Caramiaux, Frédéric Bevilacqua, Norbert Schnell

► **To cite this version:**

Baptiste Caramiaux, Frédéric Bevilacqua, Norbert Schnell. Analysing Gesture and Sound Similarities with a HMM-based Divergence Measure. *Sound and Music Computing*, Jul 2010, Barcelona, Spain. pp.1-1. hal-01161273

HAL Id: hal-01161273

<https://hal.science/hal-01161273v1>

Submitted on 8 Jun 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Analysing Gesture and Sound Similarities with a HMM-based Divergence Measure

Baptiste Caramiaux, Frédéric Bevilacqua, Norbert Schnell

UMR STMS IRCAM - CNRS

1, place Igor Stravinsky

75004 Paris, FRANCE

baptiste.caramiaux@ircam.fr

ABSTRACT

In this paper we propose a divergence measure which is applied to the analysis of the relationships between gesture and sound. Technically, the divergence measure is defined based on a Hidden Markov Model (HMM) that is used to model the time profile of sound descriptors. We show that the divergence has the following properties: non-negativity, global minimum and non-symmetry. Particularly, we used this divergence to analyze the results of experiments where participants were asked to perform physical gestures while listening to specific sounds. We found that the proposed divergence is able to measure global and local differences in either time alignment or amplitude between gesture and sound descriptors.

1. INTRODUCTION

Our research is concerned with the modelling of the relationships between gesture and sound in music performance. Several authors have recently shown the importance of these relations in the understanding of sound perception, cognitive musical representation and action-oriented meanings ([1], [2], [3]), which constitutes a key issue for expressive virtual instrument design ([4], [5]).

A gesture is described here as a set of movement parameters measured by a motion capture system. In turn, a sound is described as a set of audio descriptors representing musical properties such as audio energy, timbre or pitch. Specifically, our goal is to propose a computational model enabling the measure of the similarities between the gesture parameters and sound descriptors.

Previous works on the quantitative analysis of the gesture-sound relationship often deal with variance-based statistical methods as principal correlation analysis (PCA) ([6]) or canonical correlation analysis (CCA) ([7]). PCA allows for the determination of principal components that models the variation of the gesture parameters. Analyzing these components together with musical features (as tempo or metric) enabled to understand how listeners try to synchronize their movements on music beats ([6], [8]). In [7] the

CCA method is used as a selection tool for mapping analysis. In this work, the authors showed that this method can return the gesture and sound predominant features. However, both variance-based methods suffer from a lack of temporal modeling. Actually, these models assume as stationary both gesture parameters and audio descriptors, in the sense that statistical moments (mean, variance, etc.) do not depend on the ordering of the data. As a matter of fact, these models return a global static similarity measure without considering intrinsic dynamic changes.

To overcome these limitations, it is necessary to model the time profiles of the parameters. A large number of works dealing with time series modelling are based on hidden Markov models. HMM-based methods indeed allow for the temporal modeling of a sequence of incoming events, and have been used in audio speech recognition [9], gesture recognition ([10], [11]) and multimodal audio-visual speech recognition [12]. The common classification task generally considers a sequence as a unit to be classified and returns a decision once completed based on the computation of likelihood values. In [11] the authors present a HMM method designed for continuous modeling of gesture signals, that allows for the real-time assessment of the recognition process. Moreover, this method allows for the use of a single example for the learning procedure.

We propose to use in order to provide a measurement tool in a cross-modal fashion. HMM were already employed in cross-modal contexts [13], [14]. Here the novelty is to use HMM methods to model relationships between non-verbal sounds and hand gesture of passive listeners. More precisely, we propose here to use this method to further define a statistical distance between two time profiles, typically called a divergence measure (see for instance [15]) in information processing. Specifically, we report here that this HMM-based divergence measure has properties, induced by its underlying Markov process [16], that makes it suitable to study the time evolution of the similarity between gesture parameters and sound descriptors.

This paper is structured as follows. First, we describe the general method and context of this work. Second, we present the theoretical framework of hidden Markov modeling (section 3). In section 4 we detail the divergence measure based on this framework and a specific learning process. Third, we report an experiment and the results that illustrate a possible use of our method (section 5). Fi-

nally, we conclude and present future works in section 6.

2. CONTEXT AND GOAL

Consider the following experiment: a participant listens to a specific sound several times, and then proposes a physical gesture that “mimics” the sound. The gesture is then performed (and captured) while the participant listens to the sound. Our general aim is to answer the following question: how can we analyse the gesture in relation to the sound ?

In this experiment, the gestures can be considered as a “response” to a “stimulus”, which is actually the sound. In our framework, we will thus consider the sound as the “model” and the gestures as the “observations”, as if they were generated by the model.

For each participant’s gestures, as illustrated in figure 1, our model should allow us to compute a divergence measure between each gesture and the corresponding sound (or in other words, to quantify similarity/dissimilarity between the gesture and sound). In the next section, we describe the mathematical framework enabling the computation of such a divergence measure.

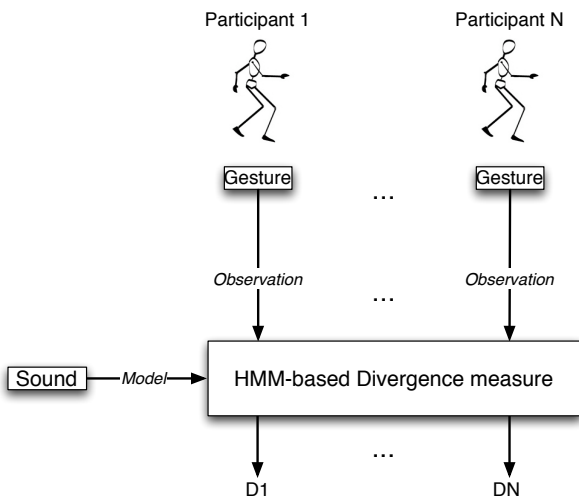


Figure 1. Methodology: Each participant’s trials are taken as input and a selected sound is taken as model. We measure the divergence between each trial and the sound.

3. HIDDEN MARKOV MODELING

In this section we briefly report the theoretical HMM framework used to further define the divergence measure in section 4.

3.1 Definition

Hidden Markov modeling can be considered as two statistically dependent families of random sequences O, X ([17], [16], [9]). The first family corresponds to the observations $\{O_t\}_{t \in \mathbb{N}}$ which represent measurements of a natural phenomenon. A single random variable O_t of this stochastic process takes value in a continuous finite dimensional

space \mathcal{O} (e.g \mathbb{R}^P). The second family of random process is the underlying state process $\{X_n\}_{n \in \mathbb{N}}$. A state process is a first-order time-homogenous Markov chain and takes values in a state space denoted by $\mathcal{X} = \{1, 2, \dots, N\}$. If we note T the length of O , statistical dependency between the two processes can be written as

$$P(O_1 \dots O_T | X_1 \dots X_T) = \prod_{t=1}^T P(O_t | X_t) \quad (1)$$

We define a hidden Markov model as

$$\lambda = (A, B, \pi)$$

Where A is the time-invariant stochastic matrix, or transition matrix, $P(X_{t+1} = j_1 | X_t = j_2), (j_1, j_2) \in \mathcal{X}^2$; B is the time invariant observation distribution $b_j(o) = P(O_t = o | X_t = j), j \in \mathcal{X}$; and π is the initial state probability distribution $P(X_0 = j), j \in \mathcal{X}$. The HMM structure is reported in figure 2.

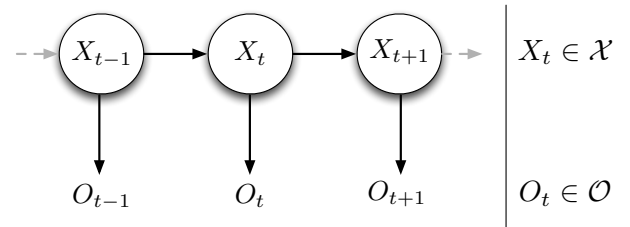


Figure 2. A general schema of HMM. $\{X_t\}_{t \in \mathbb{N}}$ is the model state random process where each state emits an observation O_t with a probability defined by B

In our case, $\{X_0 \dots X_T\}$ corresponds to an index sequence of audio descriptor samples and $\{O_1 \dots O_T\}$ a sequence of vector of samples from gesture parameter signals.

3.2 Topology

A and π have to be fixed according to a modeling strategy. π describes where in the sequence model we start to decode. A is dedicated to constrain the neighborhood of state j , taken at time t , in which a model state must be taken at the next time step $t + 1$. This data has a great influence on the resulting decoding computation. Lets consider two extreme situations for the Markov chain topology as illustrated in figure 3.

In the first case, if current state is j we constrain to look forward until $j + 1$ for the best state emitting O_{t+1} whereas in the second case we allow to look forward until the last state N to find this closest state. The first situation is more discriminative but occults possible fine corrections permitted by the stochastic modeling.

3.3 Learning

Here we present how λ is learned using a new approach presented in [11]. The parameters A (transition probability matrix) and π (initial probability) are fixed according to

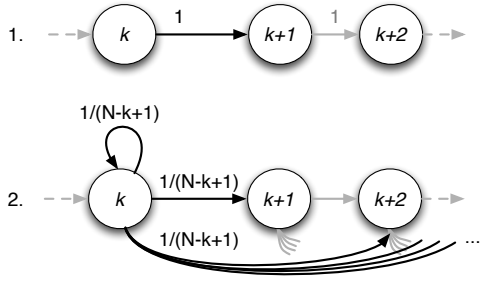


Figure 3. Two extreme cases of topology. First, one step forward is permitted in the state space. Second, each state from the current to the last one can be caught

user's choice of topology. B is such that time invariant observation distributions are gaussian, i.e

$$b_j(o) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[\frac{-1}{2} \frac{(o - \mu_j)^2}{\sigma^2} \right] \quad (2)$$

Gaussian functions are centered on the model signal samples and the standard deviation σ can be adjusted by the user (see figure ??). In our case, model signal samples are the audio feature samples computed from the chosen sound. A single example, namely the model, is needed for the learning procedure.

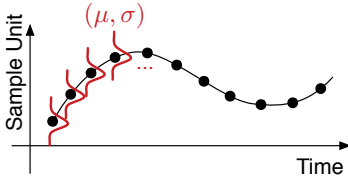


Figure 4. Learning phase. Gaussian functions are centered on the model signal samples and the standard deviation σ is *a priori* defined as a tolerance parameter.

3.4 Decoding

Given an input sequence O and a HMM λ , one of the interesting problems is to compute the probability $P(O|\lambda)$. As mentioned in [9], in practice we usually compute the logarithm of this probability as

$$\log [P(O|\lambda)] = \sum_{t=1}^T \log \left[\sum_{j=1}^N \alpha_t(j) \right] \quad (3)$$

Where $\alpha_t(i)$ is called the forward variable and is defined as $\alpha_t(i) = P(O_1 O_2 \dots O_t, X_t = i|\lambda)$, namely the probability of having the observation sequence $O_1 \dots O_t$ and the current state i . Also, this variable can be computed recursively providing an incremental method to find the desired probability [9], i.e $\forall j \in \llbracket 1, N \rrbracket$

$$\begin{aligned} t = 1 \quad & \alpha_1(j) = \pi_j b_j(O_1) \\ t > 1 \quad & \alpha_t(j) = \left(\sum_{i=1}^N \alpha_{t-1}(i) a_{ij} \right) b_j(O_t) \end{aligned} \quad (4)$$

This forward inference allows for real time applications in which input signal is decoded inductively.

4. DIVERGENCE MEASURE

In this section we define the divergence measure based on the HMM framework and the learning method described in section 3.3. Three main properties of this divergence are proved below: non-negativity; global minimum; non-symmetry.

4.1 Divergence Measure Definition

We consider two uniformly sampled signals: a model $M = \{M_1, \dots, M_N\}$ and an observation $O = \{O_1, \dots, O_T\}$. We define here the divergence measure between the observation O and a HMM learned from signal M as in section 3.3, based on decoding presented in section 3.4. We denote $\lambda_M = (A_M, B_M, \pi_M)$ the HMM learned from M . As mentioned in 3.3, we fix A_M and π_M for the divergence independently to M . Observation distributions b_j^M are defined as equation (2) with $\mu_j = M_j$. Hence we have $\lambda_M = (A, B_M, \pi)$. We define the divergence measure as

$$D_{A,\pi}(O||M) = -\log [P(O|\lambda_M)] \quad (5)$$

In the following, for convenience $D_{A,\pi}$ will be noted D . Divergence measure corresponds to the logarithm of the likelihood of having the sequence of observations O given a model λ_M learned from a signal M . More precisely, $D(O||M)$ measures the divergence between the input observation and a sequence of model states generating the observations. This sequence respects temporal structure of the model thanks to the underlying Markov chain. The result is a temporal alignment of model states on observations with a probabilistic measure evaluating how the alignment fits the observation sequence in terms of time stretching and amplitude (cf figure 5).

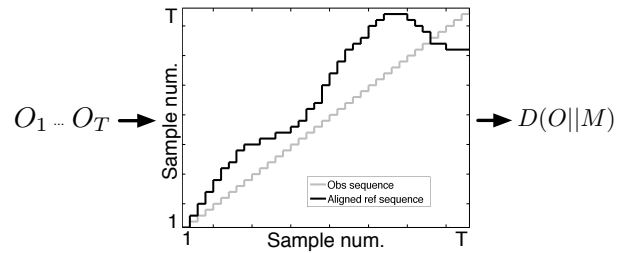


Figure 5. The HMM takes as input the sequence of observations $O_1 \dots O_t$. A sequence of model states (whose likelihood of emitting observations is maximum) approximates the observations. The quality of modeling is returned and defines $D(O||M)$.

4.2 Divergence Properties

In this section, we present that divergence measure between observation O and model M defined by (5) satisfies important properties. We refer the reader to the appendix for more details.

Non-negativity Divergence $D(O\|M)$ is always positive. Theoretically, the divergence measure does not have to be finite. Actually, $D(O\|M)$ is finite because signals considered have a finite length ($T, N < +\infty$) and infinite values are theoretically impossible, due to numerical precision. The log of very small values can be either considered as zero or disregarded.

Lower bound. The most important corollary of non-negativity is the existence of a lower bound i.e a global minimum for our divergence measure which varies according to parameters A, π, σ . Moreover, the global minimum is explicit. Depending on A and π , the minimum $D(M\|M)$ is not necessarily zero. Minimum analysis returns how close the HMM learned from M can generate O . In section 5.3 we will show that extremum analysis is pertinent in the analysis of the similarities between a sound and a gesture performed while listening to it.

For brevity, explicit global minimum is not reported here and its analytic formulation will not be explicitly used in the following.

Non-symmetry. The measure is not symmetric. Strategies to symmetrize divergence measures can be found in the literature (see for instance [18] for the well known Kullback-Liebler divergence), but we are interested here in the analysis of the divergence from an observed gesture to a fixed sound model and there is *a priori* no reason why their relation should be symmetric.

4.3 Temporal evolution of the measure

The considered sample-based learning method trains an HMM that closely models the time evolution of the signal. Moreover, from forward decoding we can find at each time t which model state emits the considered observation. Thus, at each time step the model can inform us on the close relation between both signals in terms of time evolution and amplitudes. This aims to an explicit temporal evolution of the divergence measure. Let any truncated observation signals be denoted by $O|_t = \{O_1 \dots O_t\}$ and the whole model λ_M . Hence D is defined as a function of time by,

$$D(O|_t\|M) = - \sum_{k=1}^t \log \left[\sum_{j=1}^N \alpha_k(j) \right] \quad (6)$$

5. EXPERIMENTS

In this section we present an evaluation of the previously defined divergence measure to gesture and sound data. The measure returns an overall coefficient of the similarity between descriptors of both sound and performed gesture. Temporal evolution of this measure allows for the analysis of temporal coherence of both signals. We discuss the results at the end of this section.

5.1 Data Collection

The data was collected on May 2008 in the University of Music in Graz. For the experiment 20 subjects were invited

to perform gestures while listening to a sequence of 18 different recorded sound extracts of a duration between 2.05 and 37.53 seconds with a mean of 9.45 seconds. Most of the sound extracts were of short duration. Since the experience was explorative, the sound corpus included a wide variety of sounds: environmental and musical of different styles (classical, rock, contemporary).

For each sound, a subject had to imagine a gesture that he or she performed three times after an arbitrary number of rehearsals. The gestures were performed with a small hand-held device that included markers for a camera-based motion capture system recording its position in Cartesian coordinates. The task was to imagine that the gesture performed with the hand-held device produces the listened sound. A foot-pedal allowed the beginning of the movement to be synchronized with the beginning of the playback of the sound extract in the rehearsal as well as for the recording of the final three performances.

5.2 Data Analysis

We refer the reader to the previously introduced method in figure 1. We first select a sound as a model. This sound is *waves*. It is a sequence of five successive rising and falling ocean's waves at different amplitudes and durations. According to the sound model, we consider the three trials performed by each candidate while listening to it.

The divergence measure parameters are set as follows. The chosen transition matrix corresponding to the Markov chain topology is illustrated in figure 6 (see [11] for further explanations). The initial probability distribution π is such that $\pi_1(O_1) = 1$ and $\forall i \neq 1, \pi_i(O_1) = 0$. The states of the Markov chain are the index of the audio descriptor samples (see section 3.3).

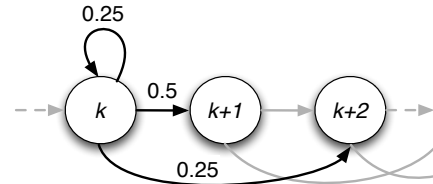


Figure 6. The chosen topology gives the predominant weight to a transition to the next state. An equal weight is given to the self-transition and to the transition above the next state.

The choice of audio description and gesture variables are based on our previous works (cf. [7]). We have shown that the predominant features when participants have performed gestures while listening to a wave sound is the audio loudness and gesture velocity. As we present some results based on the same data, we consider these two unidimensional signals for describing the data.

In the whole set of data captured, some trials had data missing; for others gesture and sound were not synchronized and finally some trials were missing for some participants. A selection is performed based on these criteria. Among the 20 participants, a set of 14 are kept. For all

of the 14 participants, we measure the divergence between each trial and the selected sound. Gesture sequence for participant s and trial p is noted $O^{s,p}$, loudness signal is noted M . Figure 7 reports the results.

Result analysis is divided into three main points illustrating the contribution of applying divergence measure between gesture and sound data.

1. *Divergence Extrema*. Participant performances for which the divergence measure is the lowest and the highest

$$\arg \min_{O^{s,p}} [D(O^{s,p} \| M)]$$

$$\arg \max_{O^{s,p}} [D(O^{s,p} \| M)]$$

2. *Temporal Evolution*. Evolution of divergence measure for the same selected participant performances as above.

$$D(O^{s,p}_t \| M)$$

3. *Gesture Variability*. Participant performances for which the standard deviation of resulting divergences is low or high.

5.3 Results and Discussion

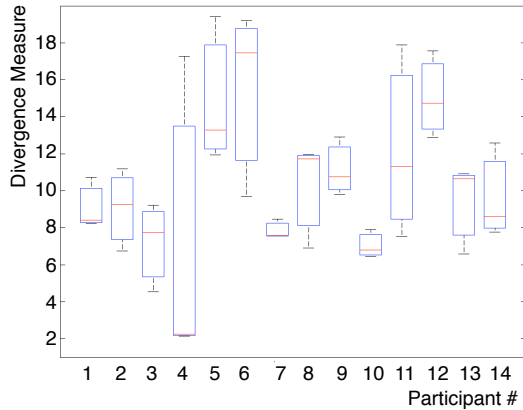


Figure 7. The figure reports statistics on divergence measures between each participant’s trial and the sound waves. The figure reports each quartile.

Divergence Extrema. Consider first the global minimum and maximum for divergence results obtained on the whole set of data (cf. figure 7). It reveals that participant 7 holds the minimum 2.44 for the second trial. In the same way, participant 9 holds the maximum 9.20 for the second trial. As reference for the reader, the global minimum is: $D(M \| M) = 0.01$. Lets see more precisely the data that return these extrema. In figure 8, the gesture trial minimizing the divergence measure is plotted on the top-left together with the model. On the top-right of figure 8, we report the gesture trial maximizing the divergence together with the model. It reveals that participant 7’s gesture is more synchronized to the sound than participant 9’s and the variations in velocity amplitude fit the best loudness proper variations. Actually, participant 7 tends to increase

his arm’s velocity synchronously with each wave falling. Otherwise, participant 9’s gesture performance is less synchronized with sound and the velocity curve shows variations in amplitude between high variations in loudness meaning that he used in-between hand movements with lower velocity variations which are not directly linked to sound but reflect subjective body control.

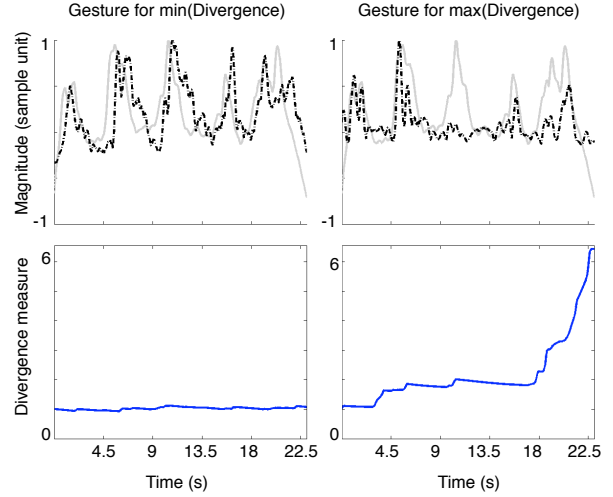


Figure 8. At the top, both gesture velocity signals are plotted in dashed line for both participant 7 (left) and participant 9 (right). The model (*waves* loudness) is also plotted in solid gray line. The bottom is divergence measure at each t between the respective signals above the plot.

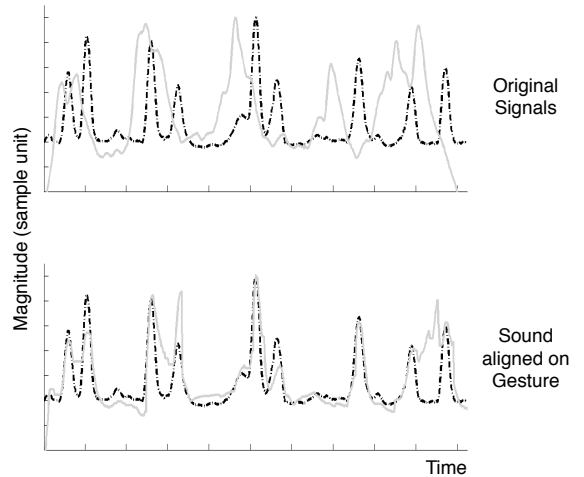


Figure 9. TODO

Temporal Evolution. As explained in section 4.3, the quality of model state sequence according to observation signal can be measured at each time t . At the bottom of figure 8 is the divergence measures evolving over time for the second trial of both participant 7 (left) and participant 9 (right). On the one hand, lets analyze bottom left plot corresponding to participant 7’s performance (see figure 10

for a better view of the divergence curve). The first samples of O are different from those of M so divergence increases until a better matching between both signals (less than 1 second). During the next 4 seconds, divergence decreases meaning that signals are synchronous and amplitudes are close (relatively to σ). Around 6 seconds, an interesting peak occurs (of magnitude 0.6). It happens during a raise in the sound loudness (wave falling). A velocity raise also occurs but its amplitude and its duration are lower than for loudness. Such a peak informs us at which time a divergence occurs and its magnitude permits us to evaluate the degree of mismatching. In this example, a magnitude of 0.6 does not involve a huge mismatching as illustrated in figure 8 (top-left part). Thanks to the underlying stochastic structure, the state sequence corrects itself according to the new inputs. Indeed, the divergence measure is then decreasing but not abruptly since the sum over time (from 1 to t , see equation 3) of the log-probabilities induces a memory of the past signals' mismatching. Global shape presents sawtooth-type variations interpreted as local mismatching (peak) and correction (release) (see figure 10)

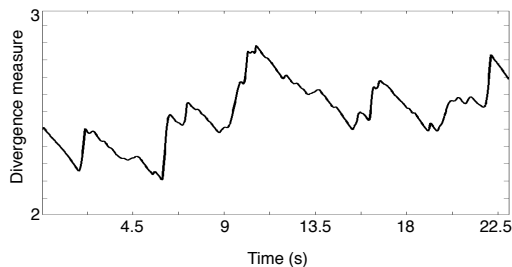


Figure 10. Zoom on divergence measure curve for participant 7. Zooming into this curve illustrates sawtooth-type behavior of the divergence.

Consider now gesture performed by participant 9, shown in the right part of figure 8. The global evolution of the divergence measure is increasing indicating that they globally diverge, contrary to the previous behavior, and its magnitude is higher. The temporal shape shows constant parts (as around 4sec, 7sec, 12sec, 17sec and 21sec). During these intervals, mismatching has less impact because amplitude of both signals is lower. The peaks occur for non-synchronized peaks meaning highly divergent amplitude values. Contrary to the respective bottom-left plot, no decreasing can be seen due to the overall past divergence values that are not good enough to involve a decrease in the divergence: as seen before, the sum propagates past mismatching.

Thereby, two different dynamic behaviors for the divergence measure have been highlighted. Locally mismatching induced a saw shape for $D(O_t \| M)$ whereas globally mismatching induced an ascending temporal curve which can roughly be approximated as piecewise constant. These behaviors give us useful hints to understand dynamic relationships between gesture and the sound which was listened to highlighting relevant parts of the signals where both signals are coherent or really distinct. Otherwise, since the method considers a global model correspond-

ing to the whole sound signal, we hypothesize from this discussion that it should be interesting to analyze gesture-sound relationship at an intermediate temporal scale between the sample and the global signal.

Gesture Variability. Illustration of standard deviation between trial divergences in figure 8 reflects the tendency of each participant to perform similar trials in terms of temporality and amplitude. Participant 4 performed very consistent trials compared to participants 5 or 6. Their respective divergence medians suggest that participant 5 has performed three different gesture performances whereas participant 6 has only one divergent gesture compared to the remaining two. Figure 11 illustrates this result reporting the three trials performed by participants 4, 5, and 6.

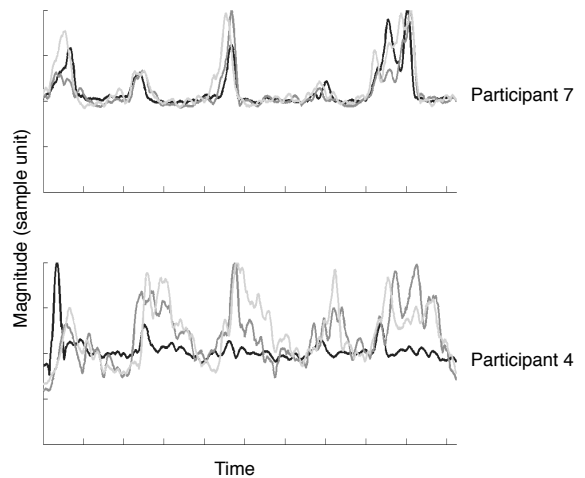


Figure 11. At the top are the trials for which variance in divergence measure is the lowest. Below we plot trials performed by participant 5 and 6 corresponding to the highest variance. Divergence median for participant 5 is roughly the mean (three different trials) of divergence values whereas divergence median for participant 6 is very low (one very different trial from the others)

6. CONCLUSIONS

In this paper we have presented a divergence measure based on a HMM that is used to model the time profile of sound descriptors. Gestures are considered as observations for the HMM as if they were generated by the model. The divergence measure allows similarity/dissimilarity between the gesture and sound to be quantified. This divergence has the following properties: non-negativity; global minimum; non-symmetry. Experiments on real data have shown that the divergence measure is able to analyze either local or global relationships between physical gesture and the sound which was listened to in terms of time stretching and amplitude variations. Some constraints (changing parameters A , π or σ) could be added in order to reinforce or relax softness of the measure.

Otherwise, this theoretical work could easily be used in

several applications. For instance, it permits us to design a gesture-driven sound selection system whose scenario is as follows. First, we build a corpus of distinct sounds with specific dynamic, timbre or melodic characteristics. Then we choose an interface allowing physical gesture capturing. Finally one can perform a gesture and the system will automatically choose the sound for which the divergence measure returns the minimal value. Such application could be useful for game-oriented systems, artistic installations or sound-design software.

Finally, we have introduced a suitable future direction in section 5.3 that proposes the design of computational models focusing on in-between temporal scale analysis as curve segments.

7. ACKNOWLEDGMENTS

We would like to thank the COST IC0601 Action on Sonic Interaction Design (SID) for their support in the short-term scientific mission in Graz.

8. REFERENCES

- [1] M. Leman, *Embodied Music Cognition and Mediation Technology*. Cambridge, USA: Massachusetts Institute of Technology Press, 2008.
- [2] R. I. Godoy, “Gestural-sonorous objects: embodied extensions of schaeffer’s conceptual apparatus,” *Organised Sound*, vol. 11, no. 2, pp. 149–157, 2006.
- [3] F. Varela, E. Thompson, and E. Rosch, *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, USA: Massachusetts Institute of Technology Press, 1991.
- [4] D. Van Nort, “Instrumental listening: sonic gesture as design principle,” *Organised Sound*, vol. 14, pp. 177–187, August 2009.
- [5] N. H. Rasamimanana, F. Kaiser, and F. Bevilacqua, “Perspectives on gesture-sound relationships informed from acoustic instrument studies,” *Organised Sound*, vol. 14, no. 2, pp. 208 – 216, 2009.
- [6] J. MacRitchie, B. Buck, and N. Bailey, “Visualising musical structure through performance gesture,” in *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, 2009.
- [7] B. Caramiaux, F. Bevilacqua, and N. Schnell, “Towards a gesture-sound cross-modal analysis,” *Lectures Notes in Computer Science, Springer-Verlag*, 2009.
- [8] G. Luck and P. Toiviainen, “Ensemble musicians’ synchronization with conductors’ gestures: An automated feature-extraction analysis,” *Music Perception*, vol. 24, no. 2, pp. 189–200, 2006.
- [9] L. R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, pp. 257–286, 1984.
- [10] A. F. Bobick and A. D. Wilson, “A state-based approach to the representation and recognition of gesture,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 12, pp. 1325–1337, 1997.
- [11] F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guédy, and N. Rasamimanana, “Continuous realtime gesture following and recognition,” in *Gesture in Embodied Communication and Human-Computer Interaction: Lecture Notes in Computer Science (LNCS)*, Springer Verlag, 2009.
- [12] M. Gurban, *Multimodal Feature Extraction and Fusion for Audio-Visual Speech Recognition*. PhD thesis, École Polytechnique Fédérale de Lausanne, 2009.
- [13] Y. Li and H.-Y. Shum, “Learning dynamic audio/visual mapping with input-output hidden markov models,” *IEEE Trans. on Multimedia*, vol. 8, no. 3, pp. 542–549, 2006.
- [14] M. Sargin, E. Erzin, Y. Yemez, A. Tekalp, A. Erdem, C. Erdem, and M. Özkan, “Prosody-driven head-gesture animation,” in *ICASSP’07*, 2007.
- [15] I. Csizsár, “Information-type measures of difference of probability distributions and indirect observation,” *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 229–318, 1967.
- [16] Y. Ephraim and N. Merhav, “Hidden markov processes,” *IEEE Trans. on Info. Theory*, vol. 48, no. 6, pp. 1518–1569, 2002.
- [17] J. Silva and S. Narayanan, “Upper bound kullback-leibler divergence for hidden markov models with application as discrimination measure for speech recognition,” in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, 2006.
- [18] D. H. Johnson and S. Sinanović, “Symmetrizing the kullback-leibler distance,” *IEEE Trans. on Info. Theory*, 2001.

A. APPENDIX

DIVERGENCE MEASURE PROPERTIES

Non-negativity Divergence $D(O\|M)$ is always positive.

$$\forall t \in \llbracket 1, T \rrbracket, \sum_{i=1}^N \alpha_t(i) = P(O_1 \dots O_t | \lambda_M) \in [0, 1]$$

Hence,

$$D(O\|M) = - \sum_{t=1}^T \log \left[\sum_{j=1}^N \alpha_t(j) \right] \in [0, +\infty] \quad (7)$$

Theoretically, the divergence measure does not have to be finite. Actually, $D(O\|M)$ is finite because signals considered have a finite length ($T, N < +\infty$) and infinite values are theoretically impossible, due to numerical precision. The log of very small values can be either considered as zero or disregarded.

Lower bound. The most important corollary of non-negativity is the existence of a lower bound i.e a global minimum for our divergence measure which varies according to parameters A, π, σ . Moreover, the global minimum is explicit.

Function $b_j^M(o)$ holds a global maximum in \mathbb{R}^p for

$$\forall j \in \llbracket 1, N \rrbracket, M_j = \arg \max_x b_j^M(x)$$

For brevity, the whole demonstration is not reported here, but it can be shown that this global maximum aims to a global maximum for $\alpha_t(j)$ leading to a global minimum for the divergence measure $D(O\|M)$ considering any inputs different from the model.

$$\forall O \neq M, D(O\|M) \geq D(M\|M) \quad (8)$$

Depending on A and π , the minimum $D(M\|M)$ is not necessarily zero. Minimum analysis returns how close the HMM learned from M can generate O . In section 5.3 we will show that extremum analysis is pertinent in the analysis of the similarities between a sound and a gesture performed while listening to it.

For brevity, explicit global minimum is not reported here and its analytic formulation will not be explicitly used in the following.

Non-symmetry. From equation (4), let $\alpha_t(j)$ be rewritten as

$$\forall t \geq 1, \alpha_t(j) = C_{t,j} b_j(O_t)$$

Where $C_{1,j} = \pi_j$ and $C_{t,j} = \sum_{i=1}^N \alpha_{t-1}(i) a_{ij}$. From respective expression of $D(O\|M)$ and $D(M\|O)$, we have $\forall t \geq 1$,

$$\sum_{j=1}^N \frac{C_{t,j}}{\sigma \sqrt{2\pi}} e^{-\frac{(O_t - M_j)^2}{2\sigma^2}} \neq \sum_{j=1}^N \frac{C_{t,j}}{\sigma \sqrt{2\pi}} e^{-\frac{(M_t - O_j)^2}{2\sigma^2}}$$

Meaning that the divergence is not symmetric.

$$D(O\|M) \neq D(M\|O) \quad (9)$$

Strategies to symmetrize divergence measures can be found in the literature (see for instance [18] for the well known Kullback-Liebler divergence), but we are interested here in the analysis of the divergence from an observed gesture to a fixed sound model and there is *a priori* no reason why their relation should be symmetric.