Interactive Sound Texture Synthesis Through Semi-Automatic User Annotations

Diemo Schwarz, Baptiste Caramiaux

▶ **To cite this version:**

Diemo Schwarz, Baptiste Caramiaux. Interactive Sound Texture Synthesis Through Semi-Automatic User Annotations. Springer International Publishing; Aramaki, M., Derrien, O., Kronland-Martinet, R., Ystad, S. Sound, Music, and Motion, Lecture Notes in Computer Science, Vol. 8905, pp.372-392, 2014. hal-01161076

HAL Id: hal-01161076
https://hal.science/hal-01161076

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.

Interactive Sound Texture Synthesis through Semi-Automatic User Annotations

Diemo Schwarz[1] and Baptiste Caramiaux[2]

[1] Ircam–CNRS–UPMC, Paris, France
schwarz@ircam.fr
[2] Goldsmiths College, University of London, UK
b.caramiaux@gold.ac.uk

Abstract. We present a way to make environmental recordings controllable again by the use of continuous annotations of the high-level semantic parameter one wishes to control, e.g. *wind strength* or *crowd excitation level*. A partial annotation can be propagated to cover the entire recording via cross-modal analysis between gesture and sound by canonical time warping (CTW). The annotations serve as a descriptor for lookup in corpus-based concatenative synthesis in order to invert the sound/annotation relationship. The workflow has been evaluated by a preliminary subject test and results on canonical correlation analysis (CCA) show high consistency between annotations and a small set of audio descriptors being well correlated with them. An experiment of the propagation of annotations shows the superior performance of CTW over CCA with as little as 20 s of annotated material.

Keywords: sound textures, audio descriptors, corpus-based synthesis, canonical correlation analysis, canonical time warping

1 Introduction

Environmental sound textures or atmospheres, such as rain, wind, traffic, or crowds, are an important ingredient for cinema, multi-media creation, games and installations.

In order to overcome the staticality of fixed recordings, we propose a method to make these recordings controllable again via a high-level semantic parameter so that they can be adapted to a given film sequence, or generated procedurally for games and installations. We have the two following use cases in mind:

Use Case 1 — Film Post-Production: A film sound designer or producer works on editing the sound track for a windy outdoor scene. On the rushes used for the scene, the sound is not available or unusable for some reason, but other rushes capture the intended sound atmosphere of the scene well. The sound designer annotates 30 s of these audio tracks for "wind strength" and the system propagates that annotation automatically to the rest of the recording. The sound designer is then able to directly and interactively create, by moving one slider, the

evolution of the sound to match the wind in the edited scene, observable on trees and objects. This relieved him from having to find contiguous audio sequences with the right temporal evolution, and to cut and splice them together.

Use Case 2 — Computer Games: In a sports game, the stadium crowd has to react to actions made by the players. Instead of preparing several sound samples or loops, and specifying the allowed transitions and superpositions, the game sound designer annotates a small corpus of stadium sounds and has it controlled by the game engine with one single parameter. The sound designer then stores the corpus to be reused for the next version of the game.

The method we propose in this article makes use of a 1D continuous manual annotation of an environmental recording that describes the sound quality one wishes to control, e.g. *wind strength* or *crowd excitation level*, or even a totally subjective parameter. The time/annotation relationship is then inverted to retrieve segments of the original recording via target annotation values by means of corpus-based concatenative synthesis [14]. The idea is to automatically retrieve the relationship between the sound and the annotation allowing for a subjective understanding of "strength" or "excitation level" in sound. The basic workflow is summarised in Fig. 1.

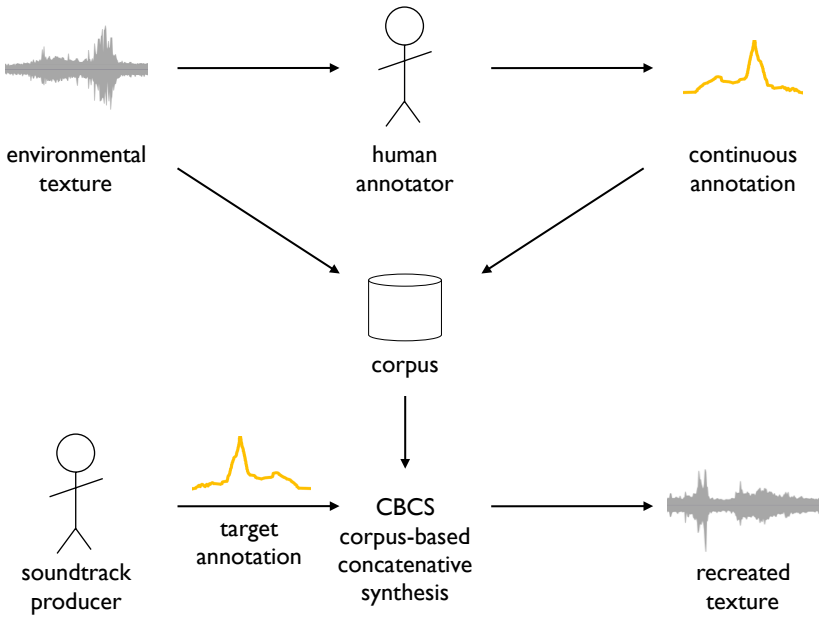

Fig. 1. Overview schema of the basic retexturing workflow.

In addition to this automatic analysis, we aim at providing a tool that allows the annotator to be able to annotate only a part of the sound and that propagates the annotation to the remaining recording. The term *semi-automatic* hence refers to the ability of the method to draw upon a user's incomplete annotation and to propagate it accurately to the whole sound. The workflow how the annotation propagation is enriching the corpus is shown in Fig. 2, the resynthesis staying the same as in Fig. 1.

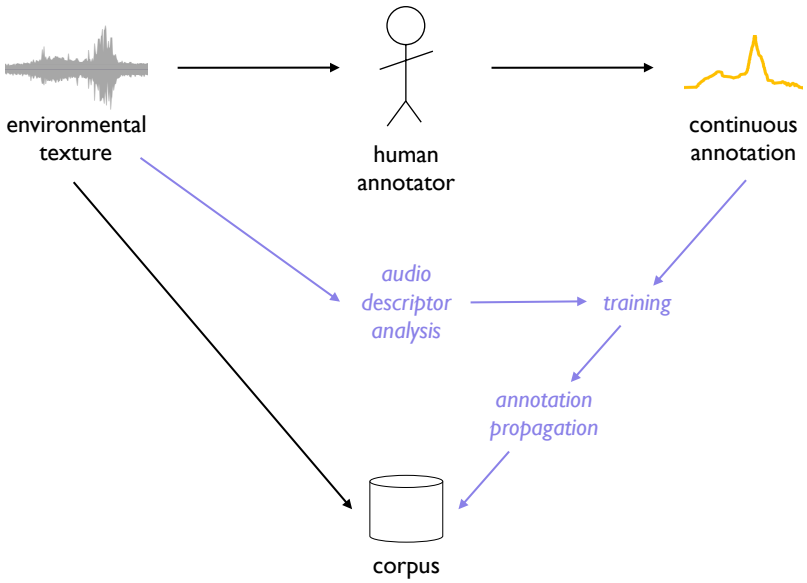

Fig. 2. Overview schema of annotation propagation workflow (light grey/blue parts in italics).

In this article we report on the use of Canonical Correlation Analysis and Canonical Time Warping for retexturing and annotation propagation. The article is structured as follows. We will present our *ReTexture* method in section 3 and a preliminary evaluation subject test, that allowed us to collect data from annotations from several users. These data enabled us to learn in section 4 how the percept of e.g. *wind strength* is correlated to sound descriptors by using canonical correlation analysis (CCA). This might allow in the future to automatise the inversion process. For long recordings, the recording does not need to be annotated over its entire length, since we developed a method to propagate a partial annotation to the rest of the recording by canonical time warping (CTW), that is presented, evaluated, and compared to CCA in section 5.

2 Previous and Related Work

The present research draws on previous work on corpus-based sound texture synthesis [17, 16]. A state-of-the-art overview on sound texture synthesis in general can be found in [15].

Gestural responses to sound have been studied from different perspectives. In psychology of music, prior work investigated how people would represent controlled stimuli made of pure tones on a 2-dimensional surface [7]. In ecological psychology, other authors investigated the relationship between the gestural description of environmental sounds and the perception of their sound source [1], or in musicology, bridging theoretical concepts of sound study and body motion [10]. These methods draw upon embodied music condition theory [8] to provide insights on the relationships between physical motion properties and sound characteristics.

Based on this theoretical background, other prior work tried to use statistical techniques to automatically investigate the link between sound stimuli and gestural responses. The motivation behind the use of automatic techniques borrowed from data mining is to be able to find the function that links the sound to the gestures in order to invert it for gesture-based control of sound. Caramiaux et al. [4] make use of Canonical Correlation Analysis to automatically retrieve the implicit mapping between listened sound and synchronously performed arm movements. The set of sounds comprised various types such as musical excerpt, environmental sound, etc. Later, Nymoen et al. [9] used the same method to find the implicit mapping in terms of correlation between controlled sounds and movement of a stick in 3-dimensional space. Unfortunately, a major drawback of this technique, already mentioned in [4, 2], is that the canonical analysis is based on correlation, meaning that both the sound and the gesture signals must be temporally aligned in order to be analyzed. An approach has been to use a different gesture–sound metric than the correlation, namely a probabilistic alignment based on hidden Markov models [3]. However these methods do not allow for feature selection but rather both the motion and the sound features must be chosen beforehand.

We base our work on the same methodology as above, but use a simpler and more focused task (annotate one specific sound quality) and input modality (a 1D slider), serving our concrete application in post-production.

Extensions of CCA have been proposed in order to overcome its inherent shortcomings. A kernel-based CCA is used to overcome the linear constraint between datasets [6] and has been recently proposed and used in an application of music recommendation based on body motion [11]. On the other hand, the synchronicity constraint has been shown to be relaxed by introducing a dynamical temporal alignment together with CCA. This model is called Canonical Time Warping and has been proposed by Zhou and De la Torre [18, 19] for spatio-temporal alignment of human motion. The methods proceed in alternating between solving the time warping using DTW and computing the spatial projection using CCA, by the mean of a modified Expectation–Maximization (EM) method.

3 Interactive Sound Texture Synthesis through Inversion of Annotations

We will now explain the *ReTexture* method of interactive sound texture synthesis through subjective user annotations (section 3.1), followed by a subject experiment (section 3.2) that allowed to evaluate the method, and to gather a database of annotations (section 3.3).

3.1 The ReTexture method

We collect a continuous one-dimensional annotation of a subjective quality of an environmental recording via a slider on a computer screen (see Fig. 3) or on an external MIDI controller. The slider is moved by a human annotator while listening to the recording.

The collected 1D break-point function is then used as a descriptor for corpus-based concatenative synthesis (CBCS), i.e. as an index to retrieve sound segments to be concatenated by annotation value. The index is implemented efficiently using a kD-tree.

For interactive recreation of a new texture, the user moves the same annotation slider that now controls the target value for concatenative resynthesis: Lookup is performed by choosing the 9 segments of length 800 ms around the annotated values closest to the given target value. One of the segments is chosen randomly (avoiding repetition), cosine windowed, and played with 400 ms overlap.

The prototype annotation application is implemented in MAX/MSP using the MUBU extensions for data management, visualisation, granular and corpus-based synthesis [13]. Examples of recreated evolutions of wind sounds can be heard online[3].

3.2 Subject Test

We performed a pre-test with 5 expert subjects with good knowledge of sound synthesis and gestural control to validate the annotation interface (Fig. 3), get first feedback on the workflow and control efficacy, and start collecting annotations. These provided us with the knowledge on what audio descriptors correlate best with the chosen sounds (section 4) and ground truth data used for training of the propagation of annotations (section 5).

The test corpus consisted of 4 sound files to be annotated: two wind recordings of 1:33 made by sound designer Roland Cahen, and two stadium crowd recordings of about 40s length from a commercial sound library. These recordings and the annotation data are available online for reference.[3]

The test procedure was as follows: After an explanation of the aim of the test, the qualities to annotate (wind strength and crowd excitation level, respectively) were made clear. Then, for each of the 4 sound files to be annotated, the subject

[3] `http://imtr.ircam.fr/imtr/Sound_Texture_Synthesis`

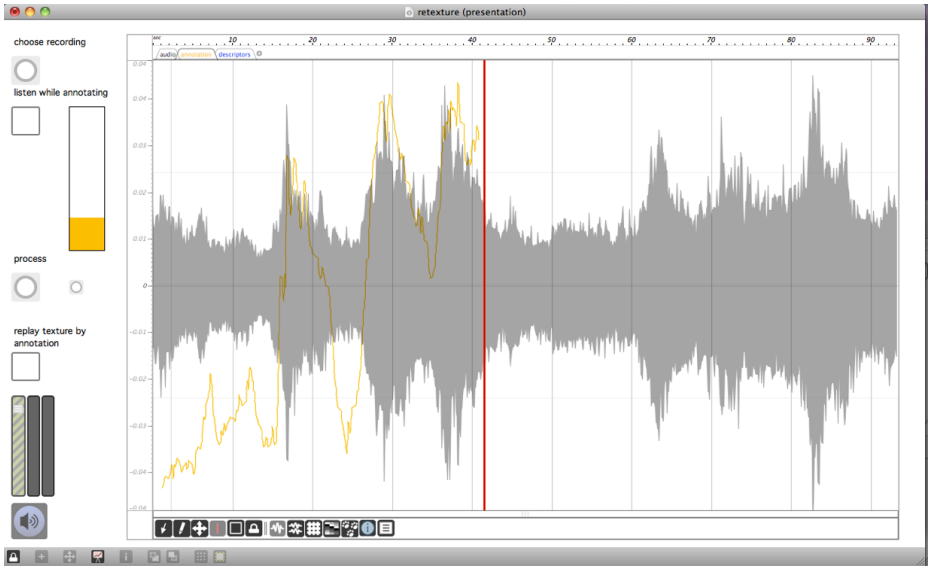

Fig. 3. Retexturing annotation interface.

could first explore the recording by starting playback at any time (by clicking on a point in the waveform), in order to familiarise herself with the different sound characters and extrema present. Then, the waveform was hidden in order not to bias annotation by visual cues on energy peaks, and recording of the annotation input via an on-screen slider started in parallel to playback of the whole sound file. At the end of this process, the raw annotation data was saved to a file, and the subject could then re-control the sound file using the annotation slider as control input, after which first impressions and free feedback were gathered.

After the whole corpus was annotated, the subject answered a questionnaire about the process with 6 questions, soliciting responses on a 5-point Likert scale (strongly disagree = 1, disagree = 2, neither agree nor disagree = 3, agree = 4, strongly agree = 5). The questions and mean ratings over the 5 subjects are given in Table 1.

3.3 Annotation Results

Figure 4 shows, for each of the 4 sound files, the annotation curves of the 5 subjects, individually normalised to zero mean and unit standard deviation. We can see that there is a high concordance between subjects, except sometimes at the very beginning of some sounds, presumably due to a start-up effect (the slider was left in the previous position, and some subjects needed a fraction of a second to home in to the value they intended for the beginning of the sound).

Table 1. Questionnaire and mean and standard deviation of response values.

Question	μ	σ
Q1: It was easy to annotate the sound quality during listening	4.0	0.71
Q2: It was often not clear which character of sound or type of events should be annotated with what value.	1.8	0.45
Q3: It was easy to recognise the annotated sound qualities when replaying via the control slider.	3.6	0.89
Q4: It was easy to recreate a desired evolution in the sound with the control slider.	4.0	0.71
Q5: One can precisely control the desired sound character via the annotations.	3.0	0.71
Q6: The created sound sequence is natural.	4.4	0.55

4 Correlation between Annotations and Descriptors

In this section, we will investigate if there is a correlation between the collected annotations and some audio descriptors. More precisely, we will extract the audio descriptors (or linear combinations of audio descriptors) that better correlate with the user-generated annotations.

4.1 Audio Descriptor Analysis

The 20 annotations collected in the preliminary subject test described in section 3.2 were correlated with a large set of audio descriptors [12], covering temporal, spectral, perceptual, and harmonic signal qualities.

The descriptors were calculated with the IrcamDescriptor library outputting 47 descriptors of up to 24D in up to 6 scaling variations in instantaneous and median-filtered versions, resulting in 437 numerical features. Figure 5 shows a subset of the descriptors for each of the 4 audio files.

4.2 Correlation Analysis

In order to get a first hint on what descriptors best represent the annotated quality of the sounds, canonical correlation analysis (CCA) was applied to the two data sets.

CCA is a common tool for investigating the linear relationships between two sets of variables in multidimensional reduction. In our case, the first (mono variate) set is the annotation (resampled and interpolated to the time base of the descriptors), and the second are the descriptors. Formally, if we let $\mathbf{X}$ and $\mathbf{Y}$ denote two datasets, CCA finds the coefficients of the linear combination of variables in $\mathbf{X}$ and the coefficients of the linear combination of variables from $\mathbf{Y}$ that are maximally correlated. The coefficients of both linear combinations are called *canonical weights* and operate as projection vectors. The projected variables are called *canonical components*. The correlation strength between canonical components is given by a correlation coefficient ρ. CCA operates similarly to Principal

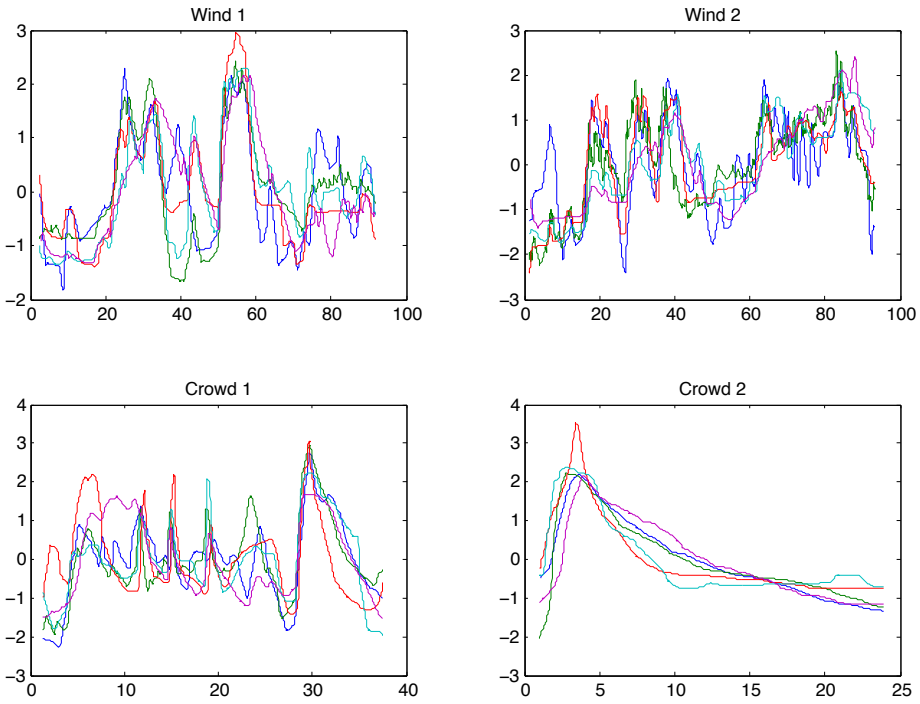

Fig. 4. All normalised subject annotations for each sound file over time [s].

Component Analysis (PCA) in the sense that it reduces the dimension of both datasets by returning N canonical components for both datasets where N is equal to the minimum of dimensions in $\mathbf{X}$ and $\mathbf{Y}$. In other words, CCA finds two projection matrices $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_N]$ and $\mathbf{B} = [\mathbf{b}_1 \dots \mathbf{b}_N]$ such that for all h between $1 \dots N$, the correlation coefficients $\rho_h = correlation(\mathbf{X}\mathbf{a}_h, \mathbf{Y}\mathbf{b}_h)$ are maximised and ordered $(\rho_1 > \dots > \rho_N)$.

Finally, a closer look at the projection matrices allows us to interpret the mapping, i.e. to extract the most correlated audio descriptors in the mapping with the annotation.

4.3 Correlation Results

We applied correlation analysis to our data in order to find the descriptors most correlated with the annotations. We collect the rank of each feature in the sorted vector of correlation coefficients ρ of all 20 annotations in a matrix $\mathbf{R}$(437 x 20), i.e. if $r_{ij} = k$, feature i has the k^{th} best correlation with annotation j.

We then coalesce the different scaling variations of the 47 descriptors, i.e., for each descriptor, we conserve only the best ranked feature for each annotation in a coalesced rank matrix $\mathbf{R}$'(47 x 20), i.e. if $r'_{ij} = k$, *some* feature of descriptor i has the k^{th} best correlation with annotation j.

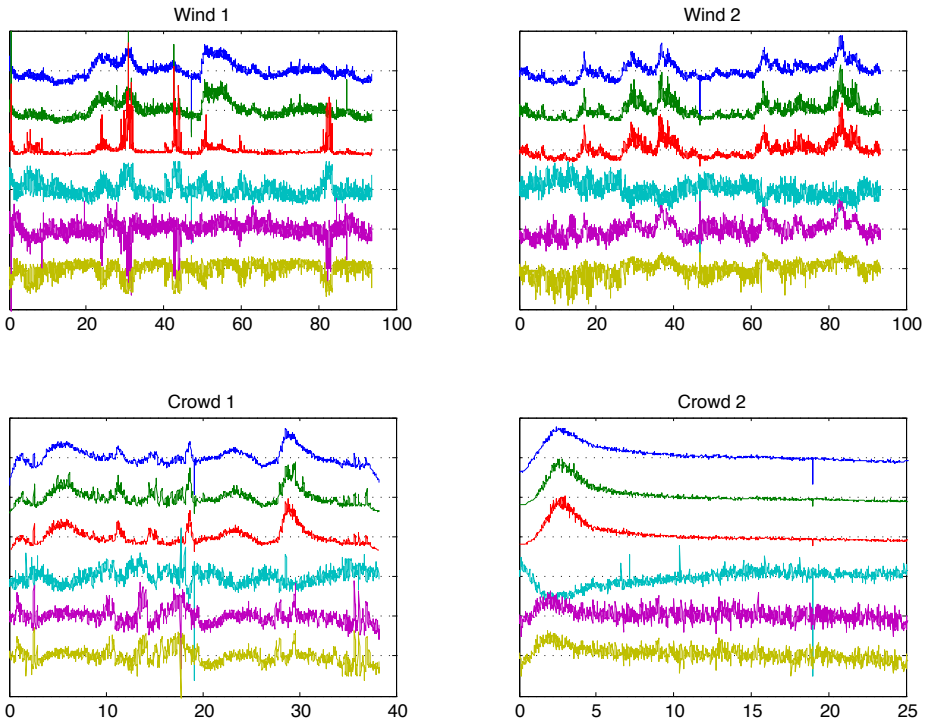

Fig. 5. Normalised descriptors for each sound file over time [s]. From top to bottom: *Loudness, Harmonic Energy, Noise Energy, Spectral Spread, Perceptual Spectral Slope, Spectral Centroid.*

Table 2. Best 8 ranked correlations between the 20 annotations and audio descriptors.

Descriptor	Mean Rank	Stddev of Rank	Min Rank	Max Rank
Loudness	1.95	1.47	1	5
Relative Specific Loudness	5.50	2.01	3	12
Harmonic Energy	6.30	5.90	1	19
Noise Energy	14.30	15.00	3	53
Harmonic Tristimulus	23.45	22.27	3	71
Perceptual Tristimulus	29.25	31.82	3	105
Perceptual Spectral Spread	33.10	24.15	13	113
Perceptual Spectral Slope	41.45	30.44	11	107

Applying basic statistics on **R'** over the 20 annotations, given in Table 2, shows us that the 3 highest-ranked descriptors have a consistently high correlation with the annotations across the different audio files and different subjects, confirming Caramiaux's findings for environmental sounds [5].

5 Propagation of Annotations

When the environmental recording is long, as in use case 1, where rushes were used, annotating can be time consuming. To speed up the annotation process, we developed a method whereby a partial annotation can teach the system how to automatically annotate the rest of the recording.

This propagation uses the partial annotation as training data to derive a mapping between annotation and audio descriptors, and applies the inverted mapping to the descriptors of the rest of the recording to reconstruct an annotation that can then be used to resynthesise the texture interactively as in section 3. Of course, the part chosen by the human annotator should be representative of the whole recording.

5.1 Temporally Aligned Correlation Analysis

In our aim of propagating annotations, we have to take into account the fact that annotations and audio stimuli are not synchronised. This would consequently affect the correlation analysis and propagate errors over time (note that we will come back to this issue in the presentation of our experimentations). To deal with this issue, we propose here the use of an extension of CCA that includes a time warping in between both datasets. The technique is called *Canonical Time Warping* (CTW) [18, 19].

Similarly to CCA, running canonical time warping returns projection matrices $\mathbf{V}_x, \mathbf{V}_x$. In addition CTW returns the alignment matrices $\mathbf{W}_y, \mathbf{W}_x$ from dynamic time warping, that are used for the reconstruction of the aligned CTW annotation from the descriptors: $\mathbf{R}_{CTW} = D\mathbf{V}_y\mathbf{V}_x^{-1}$

5.2 Training of Annotation Propagation

We base the automatic propagation of the partial human annotation T, used to train the propagation method, and D, a normalised and median-smoothed subset of the audio descriptor data of the recording. The subset has been determined by the correlation analysis in section 4 and contains 76 features from the 12 descriptors *Loudness, Harmonic Energy, Total Energy, Relative Specific Loudness, Noise Energy, MFCCs, Spectral Spread, Perceptual Spectral Spread, Spectral Slope, Perceptual Spectral Slope, Perceptual Spectral Centroid, Spectral Flatness*. D_t are the descriptors corresponding to the annotated segment.

We run CTW on the data and compute the reconstruction $\mathbf{R}_{CTW}$ as explained above. For the evaluation in section 5.3, we also use CCA to reconstruct the annotation from descriptors: $\mathbf{R}_{CCA} = D\mathbf{B}\mathbf{A}^{-1}$, where $\mathbf{A}, \mathbf{B}$ are the projection matrices from CCA.

The reconstruction and evaluation is done per individual annotation of one sound by one user, since this is closest to the use case where one expert sound designer needs to work on one specific sound. Nevertheless, we'll examine in the following the statistical influence of various parameters of the reconstruction over all our 20 examples, to obtain recommendations of minimum training segment length and robustness.

5.3 Evaluation of Annotation Propagation

We will in this section evaluate the power of annotation propagation by cross-validation on the annotation data we collected, and compare it with CCA as baseline method.

We split each of the 20 annotations into a training segment T, apply CCA and CTW training on T and D_T, and reconstruct the annotations R_{CCA} and R_{CTW} from the audio descriptors D as described above.

This procedure is performed for 5 different lengths l_i of T of 5, 10, 20, 30 seconds and the whole length of the recording $l_5 = L$, and with 5 equally distributed starting positions s_i between 0 and $L - l_i$ for each length l_i, except for the whole length $l_5 = L$, where there is only $s_1 = 0$.

Two examples of original and reconstructed annotations are given in Figs. 6–11. Figures 6–8 for recording *Wind 2* show the robustness of CCA and CTW when the segment length is $l_3 = 20$ s or more. They also show the "start-up effect" (see section 3.3) of this particular annotator, that makes the reconstruction be less stable when trained at the beginning s_1. Figures 9–11 show the more difficult example *Crowd 2*, where the prominent peak at the beginning makes the reconstruction over- or undershoot when trained outside of it, presumably because the annotated slope does not have the right relation between the peaky part and the decreasing part. That is a pathological case of not picking a representative segment to train the annotation. We can also see that CTW is clearly more robust to this difficulty than CCA.

For a quantitative evaluation, the reconstructed annotations are then compared to the whole annotation taken as ground truth using three comparison methods: the absolute global correlation c, the euclidean distance e, and the total DTW cost d.

First we aim at examining the influence of the factors segment length and start position on the reconstruction errors (given by c, e, d) from CCA and CTW. Figure 12 illustrates the influence of the training segment length and start position on these measures. We can see that the variability of the segment start diminishes or disappears from $l_3 = 20$ s onwards, and that CTW is always better or at least equal to CCA in terms of correlation.

To determine the effect of the two factors segment start position s_i and size l_i on the metrics between the reconstruction and the annotation under the two conditions CCA and CTW, we performed an ANOVA. The analysis shows that start position does not affect the correlation of the reconstruction with the ground truth annotation, but segment size influences significantly the correlation metric ($F(3, 792) = 97.5$, $p < 0.01$), and stabilises with the 20 s length (no signifiant difference between 20 and 30 seconds).

We further show the inequality between CCA and CTW for each segment size via Student's T-test. This test allows us to quantify if each mean metric differs significantly ($alpha = 0.05$) between the two conditions CCA or CTW. Its results are given in Table 3 and show that CTW is better except for the shortest segment.

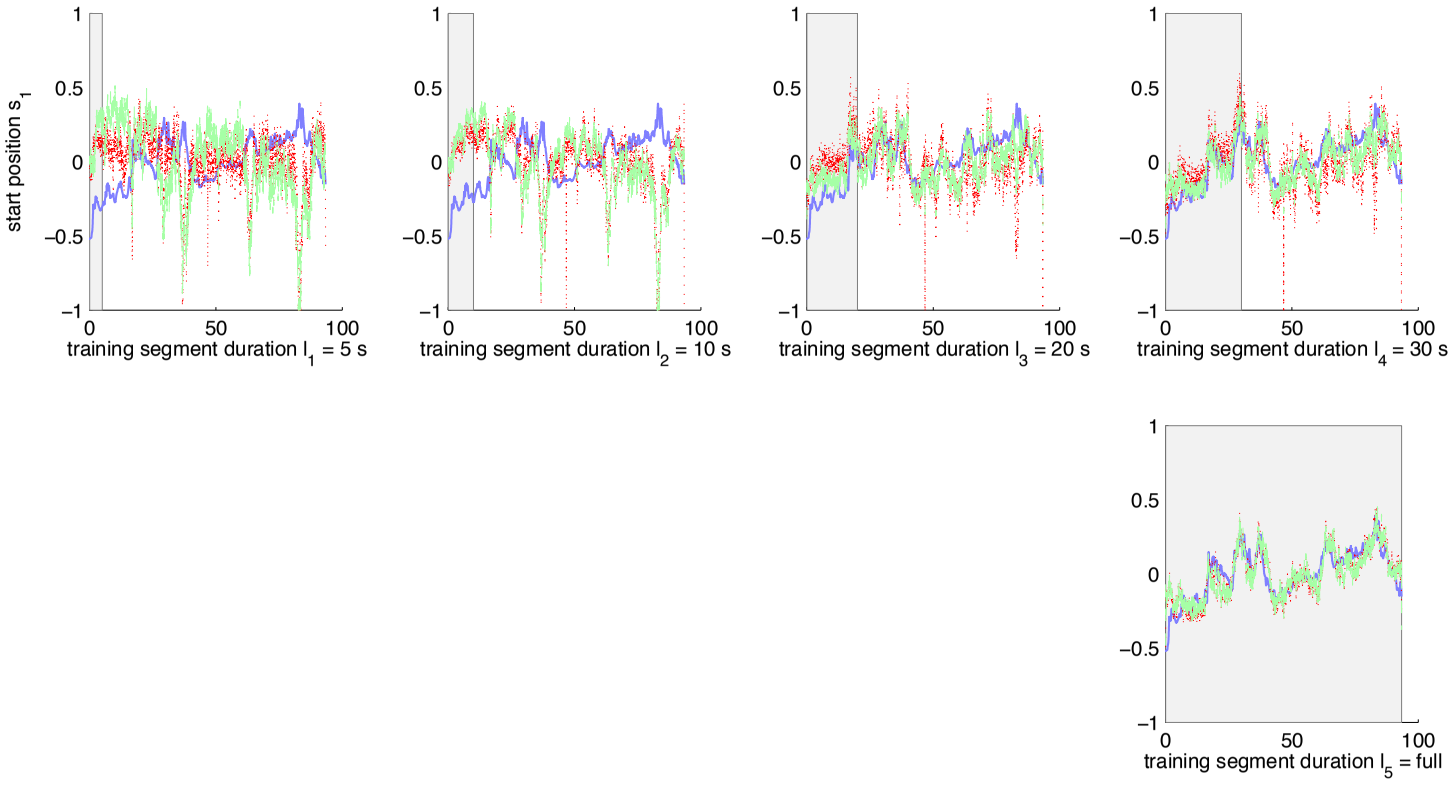

Fig. 6. CCA (dotted dark grey/red line) and CTW (dashed light grey/green line) reconstructions of one human annotation (solid middle grey/blue line) of the *Wind 2* example for different durations l_i and start positionss_1 of the training segment (visualised as grey rectangle). X-coordinates are in seconds, Y are normalised coordinates.

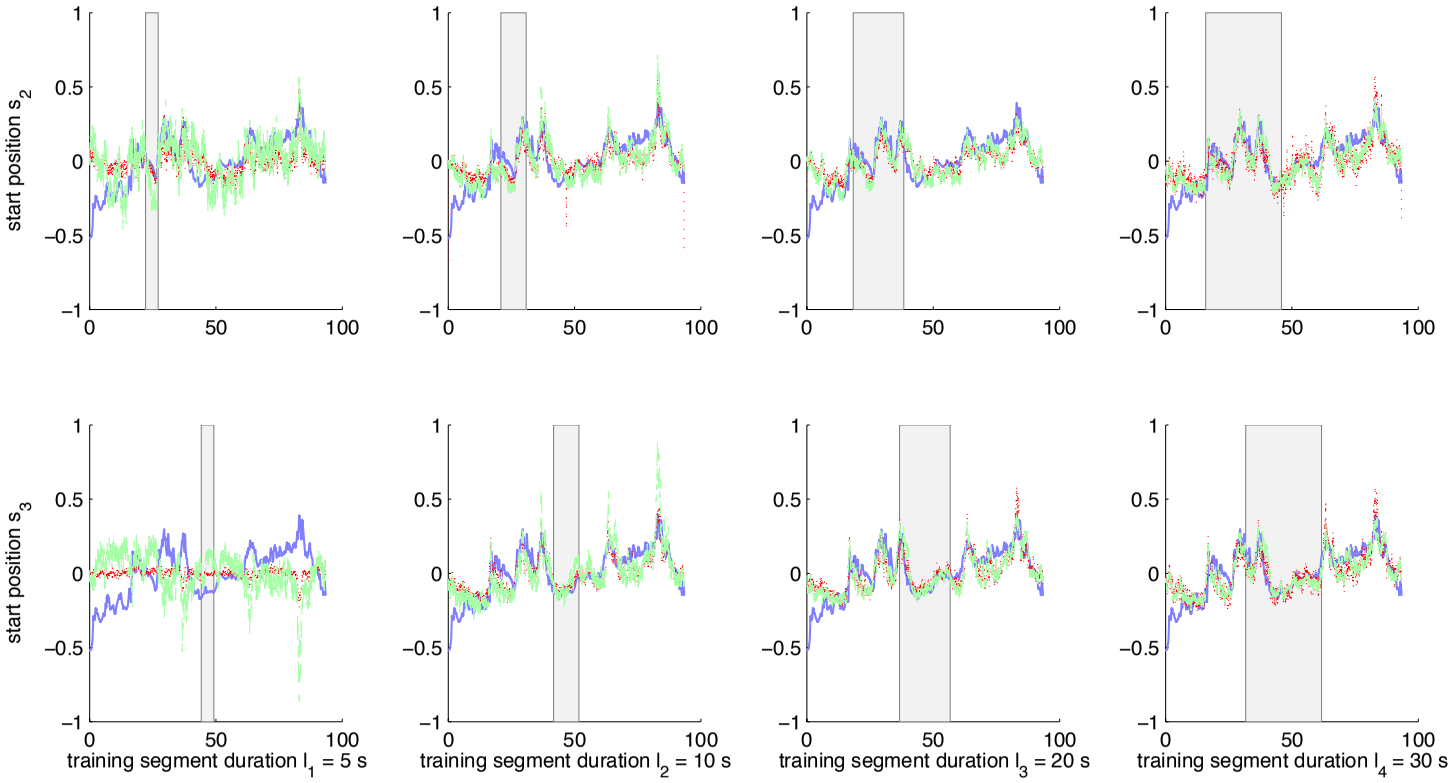

Fig. 7. CCA (dotted dark grey/red line) and CTW (dashed light grey/green line) reconstructions of one human annotation (solid middle grey/blue line) of the *Wind 2* example for different durations l_i and start positions s_2 and s_3 of the training segment (visualised as grey rectangle). X-coordinates are in seconds, Y are normalised coordinates.

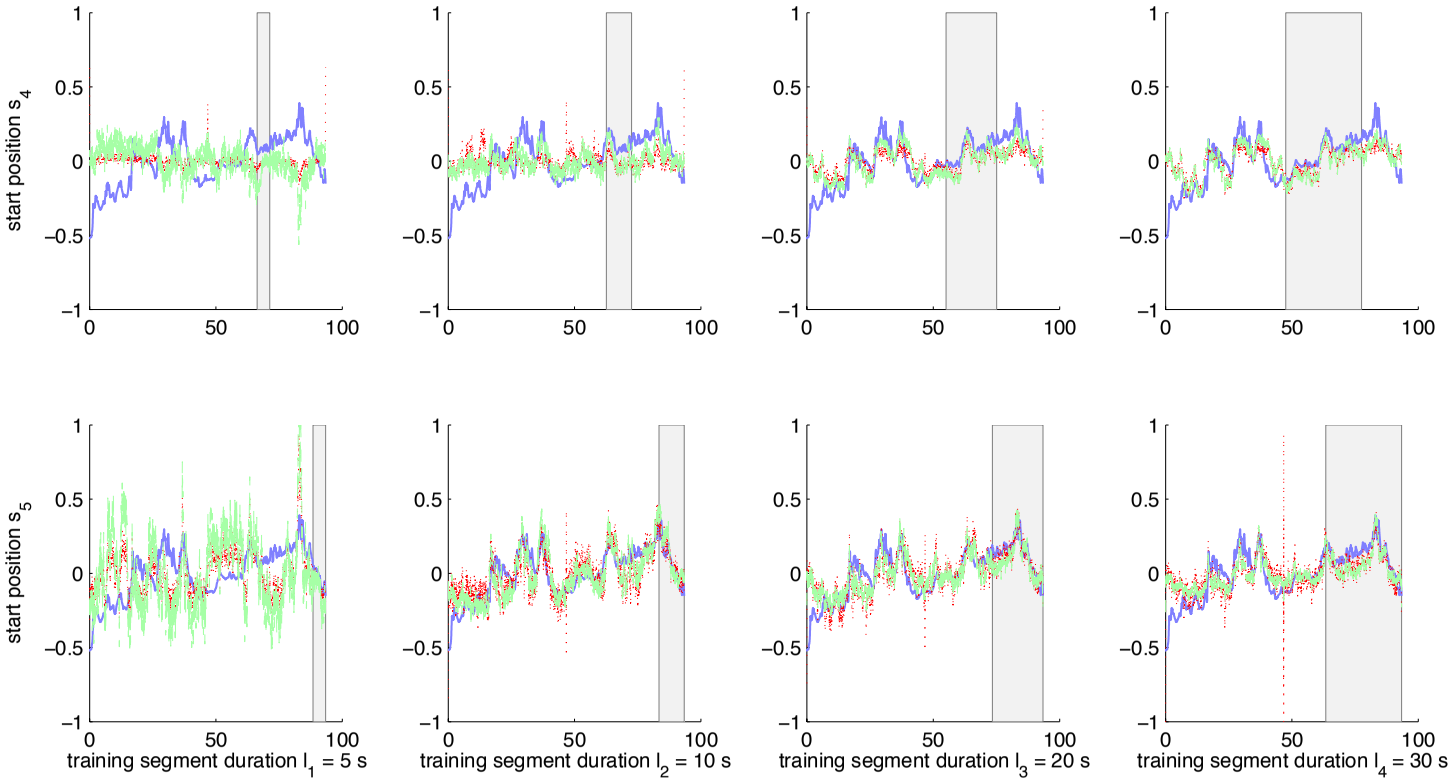

Fig. 8. CCA (dotted dark grey/red line) and CTW (dashed light grey/green line) reconstructions of one human annotation (solid middle grey/blue line) of the *Wind 2* example for different durations l_i and start positions s_4 and s_5 of the training segment (visualised as grey rectangle). X-coordinates are in seconds, Y are normalised coordinates.

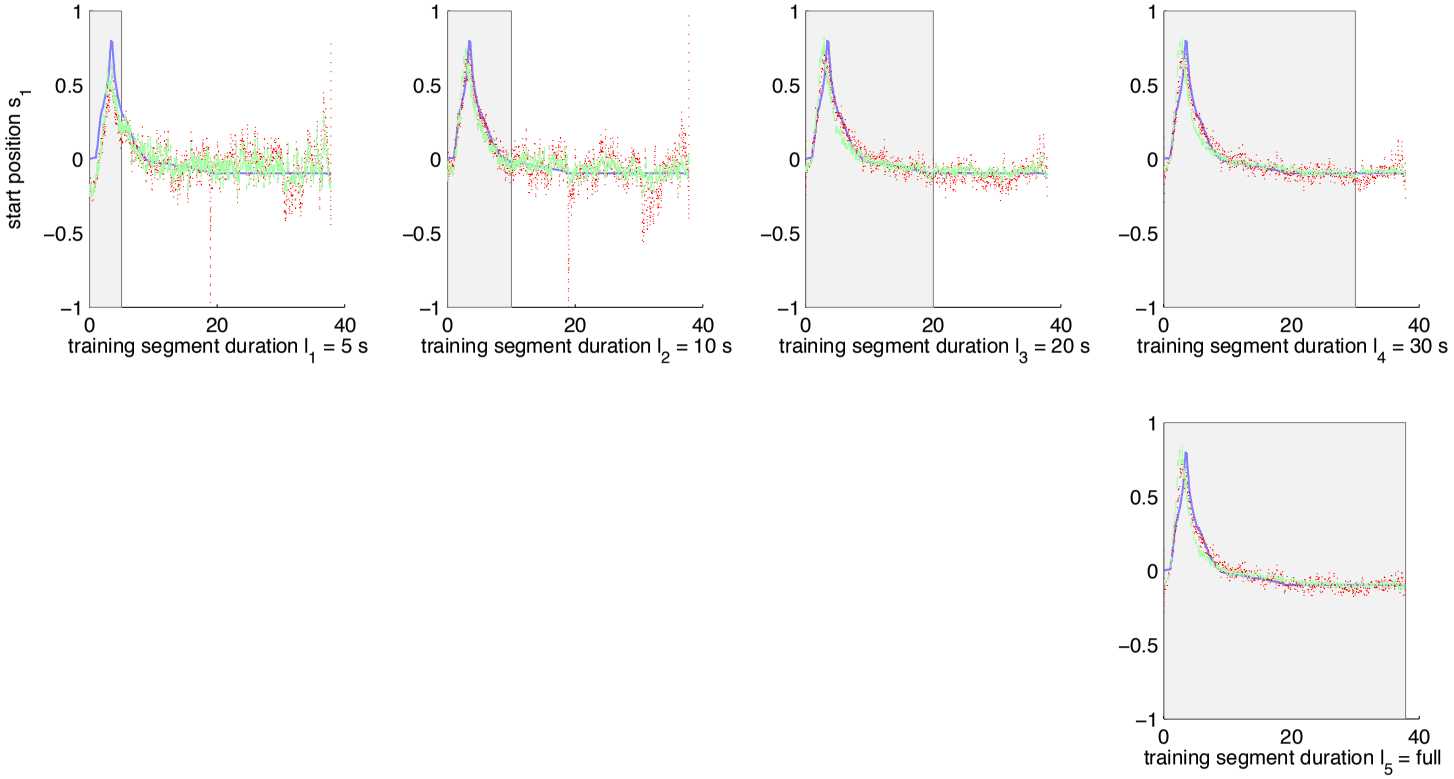

Fig. 9. CCA (dotted dark grey/red line) and CTW (dashed light grey/green line) reconstructions of one human annotation (solid middle grey/blue line) of the *Crowd 2* example for different durations l_i and start positionss_1 of the training segment (visualised as grey rectangle). X-coordinates are in seconds, Y are normalised coordinates.

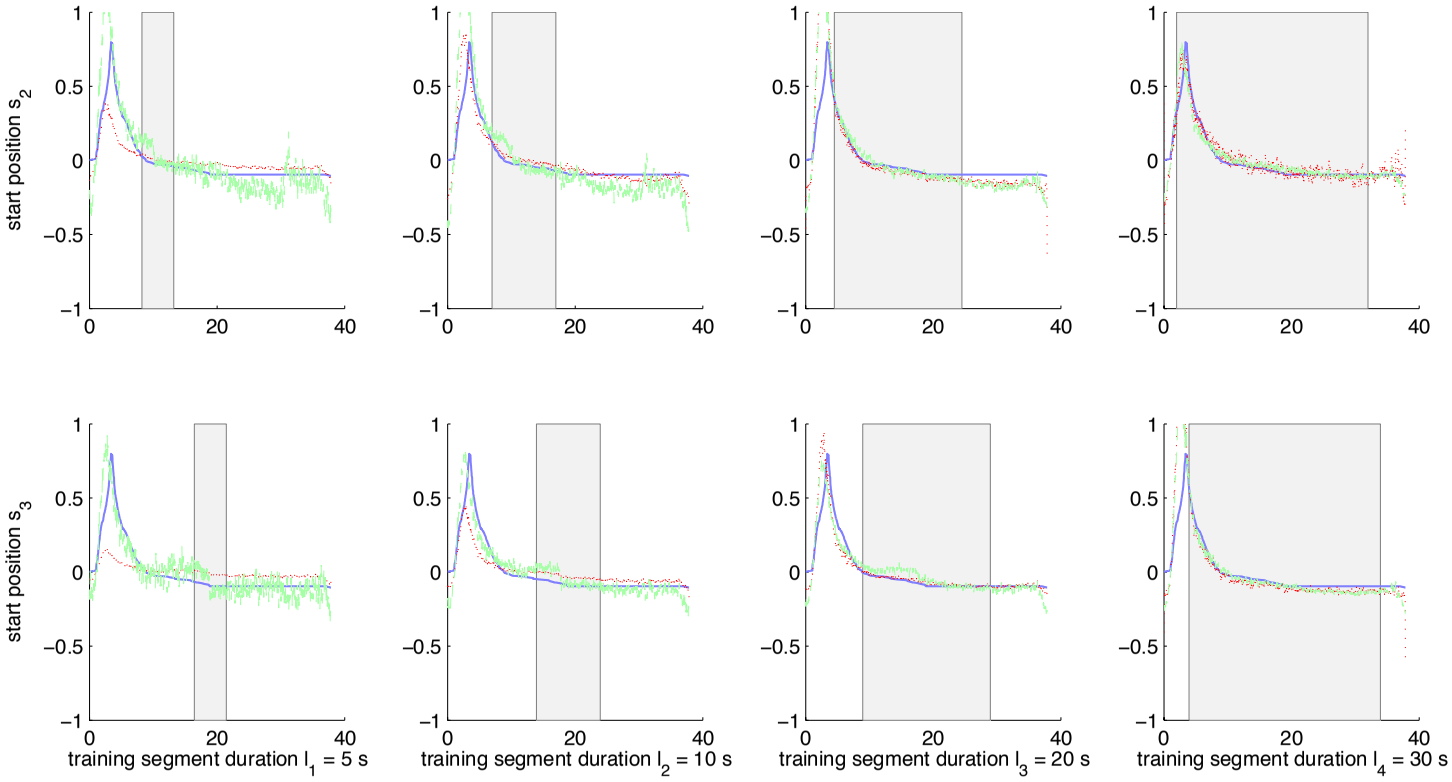

Fig. 10. CCA (dotted dark grey/red line) and CTW (dashed light grey/green line) reconstructions of one human annotation (solid middle grey/blue line) of the *Crowd 2* example for different durations l_i and start positions s_2 and s_3 of the training segment (visualised as grey rectangle). X-coordinates are in seconds, Y are normalised coordinates.

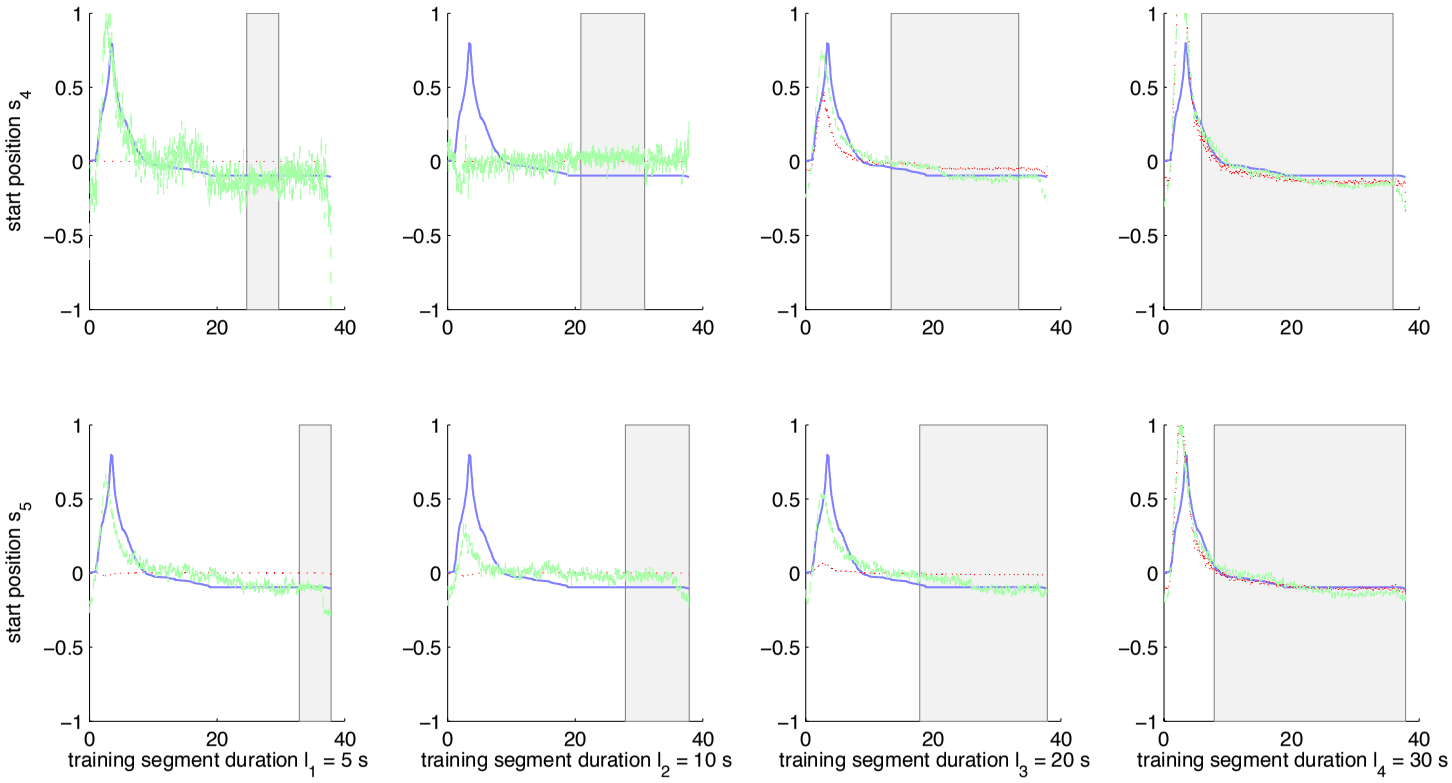

Fig. 11. CCA (dotted dark grey/red line) and CTW (dashed light grey/green line) reconstructions of one human annotation (solid middle grey/blue line) of the *Crowd 2* example for different durations l_i and start positions s_4 and s_5 of the training segment (visualised as grey rectangle). X-coordinates are in seconds, Y are normalised coordinates.

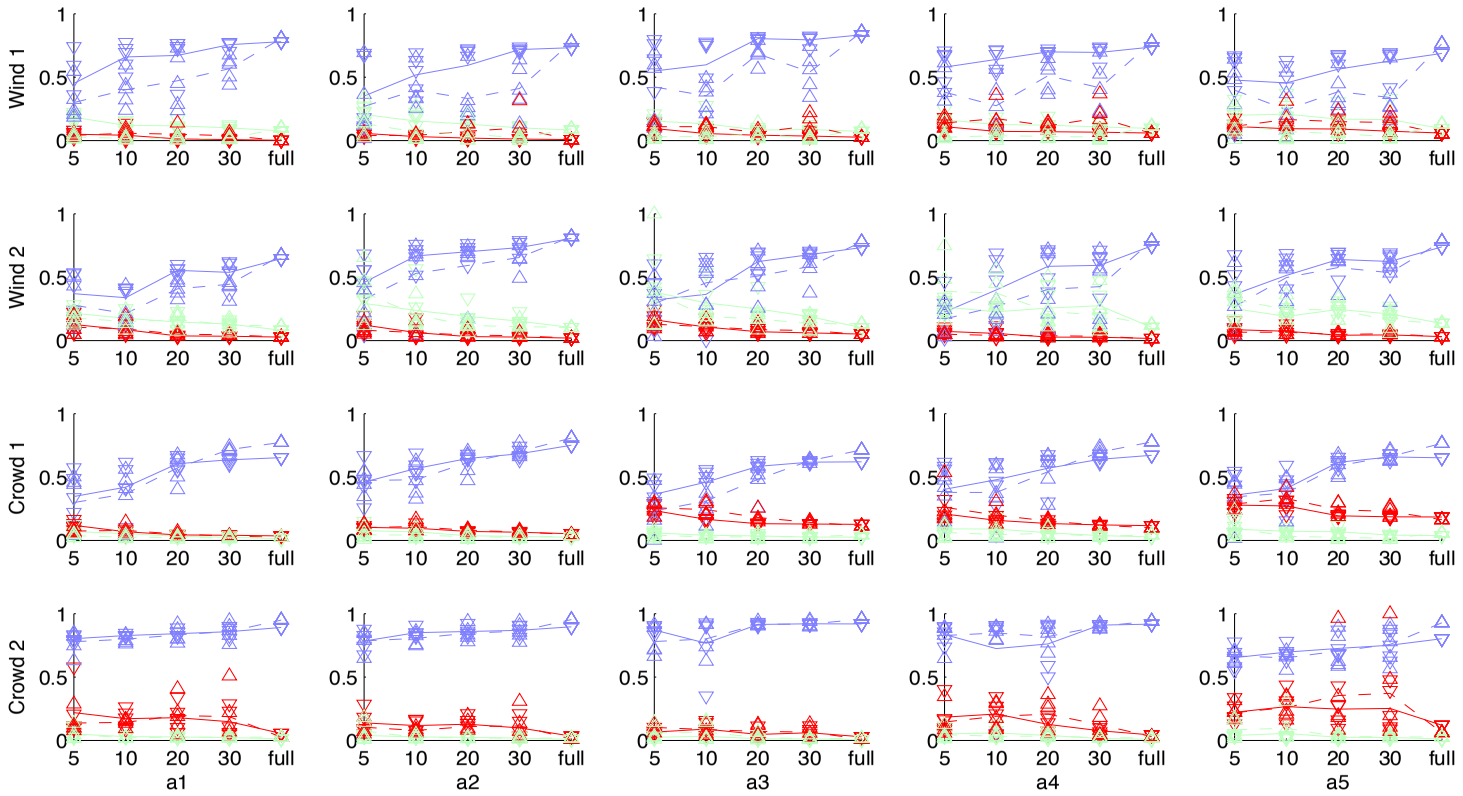

Fig. 12. Scatterplot and mean of absolute global correlation c (medium grey/blue), normalised euclidean distance e (dark grey/red), and normalised total DTW cost d (light grey/green) of CCA ($\triangle$, dashed line) and CTW ($\triangledown$, straight line) for training segment length l_i [s] (x-axis) and all segment start positions s_i (scatter) for all 20 annotations of the 4 sounds (rows) and by annotators a1...a5 (columns).

Table 3. Mean metrics (global correlation c, euclidean distance e, total DTW cost d) of reconstruction over all data sets per training segment length l_i in seconds. Best values are in bold. The column h expresses the result of the t-test whether the null hypothesis (the means are equal) can be rejected with 5% confidence ($\alpha = 0.05$). Note that the table only reports the h values (1 if $p < 0.05$, 0 otherwise).

l_i	c_{CCA}	c_{CTW}	h	e_{CCA}	e_{CTW}	h	d_{CCA}	d_{CTW}	h
5	0.45	**0.47**	0	**0.0042**	0.0044	0	484.96	**474.63**	0
10	0.46	**0.55**	1	0.0043	**0.0037**	0	**347.06**	397.17	0
20	0.59	**0.68**	1	0.0038	**0.0030**	1	**276.07**	319.76	0
30	0.65	**0.72**	1	0.0035	**0.0028**	1	**248.75**	289.53	0
all	0.84	**0.99**	1	0.0019	**0.0003**	1	232.63	**197.17**	0

5.4 Discussion

All these tests conclude that 20 s is a viable and surprisingly short length of annotation that allows to propagate the annotation to the rest of the recording. In addition, the starting position does not affect the reconstruction which means that it can be the beginning of the sound. This is in favour of our application context: a sound designer could annotate the 20 first seconds of a sound, and the method would propagate the annotation accurately.

6 Conclusion and Future Work

The evaluation and results reported here showed promising first results, encouraging us to believe that the method presented in this article could make a significant contribution to sound texture authoring for moving image or in interactive settings.

The results and feedback gathered in the subject test showed us a number of points to improve: First, one could observe a systematic and consistent lag in the annotation compared to the events in the sound (gusts of wind or climaxes in the crowd cheers), presumably due to the reaction time taken by the brain to go from perception to action, and also by a possible persistence of momentum (when to inverse the direction the annotation takes). In future work, this lag could be measured and corrected for in the lookup, thus improving control accuracy on resynthesis of textures.

In the resynthesis phase, several subjects were very apt in discovering what they considered "errors" in their annotation. This hints at giving the possibility to edit and refine the annotation interactively.

In future work, to streamline the exploration phase of the sound to be annotated, we could think of automatic selection of the n most diverse excerpts, to convey an idea of the extreme points for the annotation. This will be especially important for much longer recordings to be annotated.

The comparison of automatic propagation of the annotation by CCA versus CTW showed that CTW is more robust and needs less annotated training mate-

rial than CCA. Already with surprisingly short 20 s human annotation can the rest of the recording be reliably automatically annotated.

Further questions concern the dimensionality of annotation. The one dimension of "strength" asked for was intuitively clear to every subject, but some tried to linearly include special features into the annotation (e.g. reserve a region of annotation values for the presence of horns in the crowd examples).

In a future version of the subject test, certain improvements should be applied: The questions of the subject test should all be the same scale (higher is better) or present randomised scales. Cross-subject listening tests on the resynthesised sound could remove a possible bias of the subject having produced the resynthesis.

Finally, a more systematic study of the most appropriate input device for annotation should be carried out. Candidates (with maximum dimensionality in parentheses) are the mouse on a slider or free movement (2D), digitizer tablets (2D and pressure) hardware faders (n x 1D), game controllers such as joystick or Wiimote (2D).

Acknowledgements

The work presented here is partially funded by the French *Agence Nationale de la Recherche* within the project *PHYSIS*, ANR-12-CORD-0006. The authors would like to thank Sarah, Alejandro, Jules, and Louis.

References

1. B Caramiaux, F Bevilacqua, T Bianco, N Schnell, O Houix, and P Susini. The role of sound source perception in gestural sound description. *ACM Trans. on Applied Perception*, 11(1), 2014.
2. Baptiste Caramiaux. *Studies on the Gesture–Sound Relationship for Musical Performance*. Phd thesis, Université Université Pierre et Marie Curie (Paris 6), Paris, 2012.
3. Baptiste Caramiaux, Frédéric Bevilacqua, and Norbert Schnell. Analysing gesture and sound similarities with a hmm-based divergence measure. In *Proceedings of the Sound and Music Computing Conference, SMC*, volume 11, 2010.
4. Baptiste Caramiaux, Frédéric Bevilacqua, and Norbert Schnell. Towards a gesture-sound cross-modal analysis. In *Gesture in Embodied Communication and Human-Computer Interaction: LNAI 5934*, pages 158–170. Springer Verlag, 2010.
5. Baptiste Caramiaux, Patrick Susini, Tommaso Bianco, Frédéric Bevilacqua, Olivier Houix, Norbert Schnell, and Nicolas Misdariis. Gestural embodiment of environmental sounds: an experimental study. In *New Interfaces for Musical Expression (NIME)*, 2011.
6. David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.
7. Mats B Küssner and Daniel Leech-Wilkinson. Investigating the influence of musical training on cross-modal correspondences and sensorimotor skills in a real-time drawing paradigm. *Psychology of Music*, page 0305735613482022, 2013.

8. Marc Leman. *Embodied music cognition and mediation technology*. Mit Press, 2008.
9. Kristian Nymoen, Baptiste Caramiaux, Mariusz Kozak, and Jim Torresen. Analyzing sound tracings - a multimodal approach to music information retrieval. In *ACM Multimedia, Workshop MIRUM*, 2011.
10. Kristian Nymoen, Rolf Inge Godøy, Alexander Refsum Jensenius, and Jim Torresen. Analyzing correspondence between sound objects and body motion. *ACM Transactions on Applied Perception (TAP)*, 10(2):9, 2013.
11. Hiroyuki Ohkushi, Takahiro Ogawa, and Miki Haseyama. Music recommendation according to human motion based on kernel cca-based relationship. *EURASIP Journal on Advances in Signal Processing*, 2011(1):1–14, 2011.
12. Geoffroy Peeters. A large set of audio features for sound description (similarity and classification) in the Cuidado project. Technical Report version 1.0, Ircam – Centre Pompidou, Paris, France, April 2004.
13. Norbert Schnell, Axel Röbel, Diemo Schwarz, Geoffroy Peeters, and Ricardo Borghesi. MuBu & friends – assembling tools for content based real-time interactive audio processing in Max/MSP. In *Proceedings of the International Computer Music Conference (ICMC)*, Montreal, Canada, August 2009.
14. Diemo Schwarz. Corpus-based concatenative synthesis. *IEEE Signal Processing Magazine*, 24(2):92–104, March 2007. Special Section: Signal Processing for Sound Synthesis.
15. Diemo Schwarz. State of the art in sound texture synthesis. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx)*, Paris, France, September 2011.
16. Diemo Schwarz, Roland Cahen, Ding Hui, and Christian Jacquemin. Sound level of detail in interactive audiographic 3D scenes. In *Proceedings of the International Computer Music Conference (ICMC)*, Huddersfield, UK, 2011.
17. Diemo Schwarz and Norbert Schnell. Descriptor-based sound texture sampling. In *Proceedings of the International Conference on Sound and Music Computing (SMC)*, pages 510–515, Barcelona, Spain, July 2010.
18. Feng Zhou and Fernando De la Torre. Canonical time warping for alignment of human behavior. In *Advances in Neural Information Processing Systems Conference (NIPS)*, December 2009.
19. Feng Zhou and Fernando De la Torre. Generalized time warping for multi-modal alignment of human motion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.