INFERRING THE HAND CONFIGURATION FROM HAND CLAPPING SOUNDS

Antti Jylhä and Cumhur Erkut

Dept. of Signal Processing and Acoustics,
Helsinki University of Technology, TKK
Espoo, Finland
antti.jylha@tkk.fi, Cumhur.Erkut@tkk.fi

ABSTRACT

In this paper, a technique for inferring the configuration of a clapper's hands from a hand clapping sound is described. The method was developed based on analysis of synthetic and recorded hand clap sounds, labeled with the corresponding hand configurations. A naïve Bayes classifier was constructed to automatically classify the data using two different feature sets. The results indicate that the approach is applicable for inferring the hand configuration.

1. INTRODUCTION

Humans are used to interact with each other by sound, and our everyday listening skills are well-developed for extracting information from our environment [1]. Similarly, in a fluent sonic interaction between a human and a computer, the computer must be able to recognize the sonic control signals the user invokes, and to distinguish them from other sounds in the environment. Such a sonic interaction may occur also using everyday sounds as the conveyor of information instead of speech or music. In this paper, we use hand claps as a test case for the feasibility of such an interaction. Hand claps are a ubiquitously familiar phenomenon and easy to capture by relatively cheap equipment. Therefore, they could be widely applied in different applications.

Automatic recognition of the hand clap type would be interesting in human-computer interaction not only because it would enable more ways of exploiting hand claps as conveyors of information, but also because it can potentially allow personified control interfaces and clapper identification. Assuming that the hand clap sounds of individual clappers are systematically different, they could be applied to control applications, which are only desired to be controlled by a specific person.

In this paper, we will discuss some possibilities of estimating the system-specific parameters from the audio signals for a hand clap model. We focus on offline methods in this exploratory research. It is also a long-term objective in our research to make an online algorithm for estimating the parameters of a physics-based sound synthesis system.

2. PERCEPTION, ANALYSIS, AND SYNTHESIS OF HAND CLAPS

Hand claps are a relatively primitive conveyor of sonic information, yet they are widely applied for different purposes [2]. In different cultures hand claps are used in a musical context, and we are used to give feedback of a performance by applause [3], by indicating different levels of enthusiasm to the performers. Hand claps are an essential part of flamenco music, in which rhythmic patterns of soft and sharp claps are used as an accompaniment. Hand claps have also been used to call for service.

Previous work on hand clap sounds and human-computer interaction includes for example a hand clap language as a common means of communication between humans and robots [4]. This implementation does not consider different hand clap types, however. A recent work has also investigated the identification of synchronous vs. asynchronous applause using Mel-frequency cesptral coefficients and a genetic algorithm [5].

As sound events, hand claps of an individual are very short in time. In anechoic conditions, a hand clap sound lasts typically around 5 ms, and it is difficult to pinpoint any systematic differences between different kind of hand claps. However, according to Repp [2], human observers are able to deduce the hand configuration of a clapper with good accuracy in a single-clapper setting by listening to the hand clap sound. Based on spectral analysis, Repp has proposed eight different hand configurations which have audible differences. These clapping modes are presented in Fig. 1.

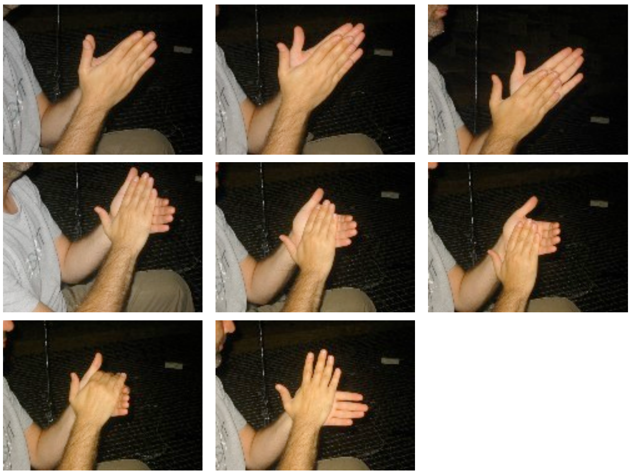

Figure 1: Hand clap types reproduced after [2] and [6].

A short description of the clapping modes is as follows. In P (parallel) modes in the first row of Fig. 1, the hands are kept parallel and flat, whereas in A (angle) modes in the second row of Fig. 1 they form a right angle with a natural curvature. The following numerical index indicates the position of the right hand relative to the left hand: from palm-to-palm (1) to fingers-to-palm (3), (2) corresponding to about the midpoint. Finally, in the third row, the curvature of hands vary in A1 mode to result a flat (A1-) or a very cupped (A1+) configuration, compared to A1.

Both Repp [2] and Peltola et al. [3] have noticed that in a hand clap sound there is a pronounced resonance, whose center frequency, Q-value, and magnitude depend on the hand configuration, namely, on the air cavity between the hands. For example, a smaller air cavity will cause the pronounced resonance to occur on a higher frequency than a bigger air cavity. Based on these ideas, a hand clap synthesis system has been implemented and will be described shortly in the following section.

2.1. Hand clap synthesis: Overview of ClaPD

ClaPD is a stochastic model in realm of [7, 8] and it is implemented as a PureData [9] library[1]. ClaPD contains low-level synthesis and higher-level control blocks, and primitive agents for event generation, which are fine-tuned by hand-clapping statistics. It can produce expressive, human-like synthetic applause of a single clapper with adjustable hand configuration, or asynchronous or synchronous applause of a clapper population (*audience*). As a part of a more complex auditory display representing a listening environment, artificial reverberation has been optionally included in ClaPD. ClaPD is discussed in detail in [3, 10].

ClaPD can be used to synthesize hand clapping sequences with varying virtual hand configurations. A single clap event is synthesized by a second order resonant filter excited by a burst of enveloped noise. The parameters of the resonant filter depend on the virtual hand configuration of the synthetic clapper and are based on the work described in [3] and [6]. The parameters applied for each clapping mode are presented in Table 1.

We can present the filter parameters in Table 1 graphically by assuming a Gaussian distribution with the listed mean and standard deviation, conditional on the corresponding clap type. The characteristics of the distributions are qualitatively visualized in Fig. 2 by ellipsoids, whose centers are defined by the mean values of the class-dependent parameters and the radii by the corresponding standard deviations. We notice that there are big differences between the illustrated distributions. Only in two cases there seems to be significant overlap in the distributions, i.e., with clap types P2 and A1, and P1 and A2.

3. CLASSIFICATION TECHNIQUE

To perform automatic classification of hand configurations based on hand clap sounds, a simple probabilistic method was chosen. Since the previous work in [2] and [3] indicate that the spectral characteristics of different clap types are systematically different, we approach the classification problem in the spectral domain based on the same principles as the synthesis in ClaPD.

3.1. Feature selection

As features for the classification, to obtain a reference point for other features, we chose to apply the magnitude bins of the Fast Fourier Transform (FFT). To reduce the computational load, we applied a two-level strategy. First, the sounds were downsampled from 44100 Hz by the factor of 9 to 4900 Hz. In this procedure, no essential information was lost since the cavity resonances always occur well below 2000 Hz. Second, an analysis window of 10 ms (49 samples) was applied and so a zero-padded 128-bin FFT gave sufficient resolution, the one-sided FFT consisting of 64 bins. These bins were used as features in the reference classification model.

Table 1: Synthesis filter parameters. f is the average center frequency in Hz, B is the average -3 dB bandwidth, and G is the average filter gain. The d values are the deviations used to randomize the parameters.

	P1	P2	P3	A1	A2	A3	A1+	A1-
f	1364	774	1562	775	1355	1452	500	984
d_f	333	448	257	39	57	12	178	396
B	188	139	375	137	222	183	91	137
d_B	54	7	94	15	53	40	27	36
G	27.2	13.8	19.5	22.2	24	13.8	33.8	11.3
d_G	9.76	3.5	6.31	3.55	8.03	4.45	5.57	1.42

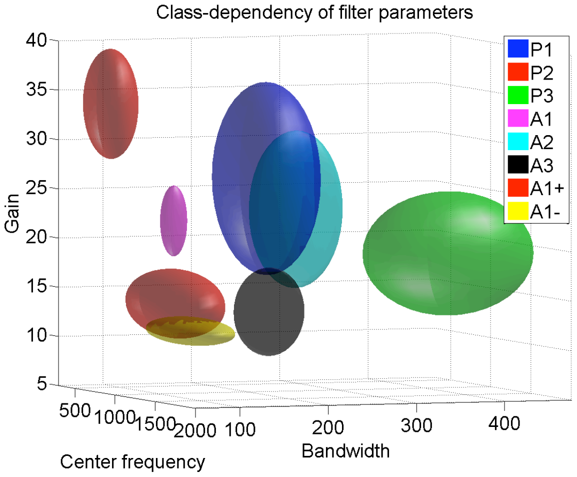

Figure 2: Illustration of the clap type dependency of the estimated filter parameters.

To achieve a lower data dimension, we chose to experiment also with the coefficients of a low-order IIR filter fit to a hand clap signal as features. Assuming a single prominent resonance in the spectrum of a clap event due to the cavity between hands, we chose to use a second-order all-pole filter, which would model the resonance. The clap response was windowed with the Blackman-Harris window to emphasize the prominent resonance. The filter was fit to the windowed clap response by the Steiglitz-McBride algorithm [11], which proved to provide better classification results than linear prediction [12], which we also experimented with. We used as features the numerator (gain) coefficient and both of the non-unity denominator coefficients, i.e., a total of three features.

3.2. Classification

Using Repp's taxonomy [2] (see Fig. 1), we have a classification $C = \{$P1,P2,P3,A1,A2,A3,A1+,A1-$\}$ of hand configurations. We assumed a conditionally independent model for the features resulting from each class, i.e., the probability distribution

$$p(C, Y_1, ..., Y_N) = P(C)P(Y_1|C)...P(Y_N|C), \qquad (1)$$

where Y_i denotes the i:th feature and N is the number of features. In practice, this is equivalent to the naïve Bayes classifier [13].

Given the naïve model, we assumed normal distribution for the features given the class C. That is, for each feature we have

[1] ClaPD is released as a free software under the GNU Public License (GPL) and it can be downloaded from `http://www.acoustics.hut.fi/software/clapd`

the conditional distribution

$$p(Y_i|C=j) \sim N(\mu_{i,j},\sigma^2_{i,j}), \tag{2}$$

where μ and σ^2 are the mean and variance of the distribution, respectively.

To train such a model, a training set of labeled data can be used. When the class $c \in C$ and the starting time of a clap event are known, it is straightforward to evaluate the features for that clap instance. This way, a conditional set of data is obtained for each class c. From this data, it is possible to obtain the parameters for the conditional distributions of the features presented in Eq. 2.

Once the model is trained with the labeled data, it can be used for classifying new sets of data. From the naïve Bayes model, we can derive the conditional probability of class C given the observed features Y. According to the Bayes rule, we have

$$p(C|Y_1,Y_2,...,Y_N) = \frac{p(Y_1,Y_2,...,Y_N|C)p(C)}{p(Y_1,Y_2,...,Y_N)}. \tag{3}$$

Now, given the observations $Y = y_1, y_2, ..., y_N$, we can compute the log-likelihood of each conditional distribution $p(Y|C = c)$ and select the maximum likelihood class to be the most likely class for these observations.

4. EXPERIMENTS AND RESULTS

To evaluate the classification technique, several experiments were conducted. We first evaluated the classification approach with synthetic data, and then proceeded with real hand clap recordings.

4.1. Generating test data with ClaPD

To test the classification technique presented in Section 3, we generated synthetic data sets with the ClaPD synthesis engine presented in Section 2.1. As a training set, we used a 60 second sequence of synthetic claps without reverberation. The set consisted of 190 claps of randomized clap types. We also generated four 30 second sequences of randomized claps without reverberation and two 30 second sets with artificial reverberation (`freeverb~`) for testing. The training set was not part of the test data.

4.2. Gathering real hand clap data

We recorded sequences of real hand claps in an ITU-R BS.116 standard listening room, with reverberation time of 0.3 s. Two male subjects A and B performed 20 claps of each type. In addition, a sequence of flamenco type claps was recorded by one of the subjects, with hand configurations resembling the clap types A1+ and A3 in Fig. 1, with two different strengths. The data was labeled manually.

4.3. Results

To provide a reference for the classifier performance, the results for the synthetic data are presented in Table 2 for both the FFT bins and the IIR coefficients as features. In the table, the rows correspond to the actual hand configuration, and the columns correspond to the automatic classification result. The numbers explicate the portion of instances of one class classified into each class. The diagonal elements show the success ratio of each clap type being labeled correctly into its own class.

From the Table 2, we can see that the classification accuracy of different classes varies. The best results are obtained with clap type A1+, which is classified correctly over 90 % of all instances in both cases. Also the clap types P3, A3, and A1- reach the accuracy of more than 80 % in the FFT bin case. There is systematic misclassification of class P1 as A2, and vice versa. This is in line with the original inspection of overlaps in the classes in Fig. 2.

Taking a closer look at the results shows that if classes P1 and A2 were clustered to one class, and P2 and A1 to another class, the results would be better. Indeed, even for the synthetic data, these classification results suggest a different kind of taxonomy for the hand configurations. We leave the construction of such a taxonomy for the future.

The overall performance of the magnitude spectrum bin classification was 71.7 %, and that of the filter coefficient classification was 69.9 %. The overall performance of the filter-coefficient classification was affected by windowing the analysis frame. Without windowing, the performance was 64.4 %.

The artificial reverberation did not affect the results much in any of the cases. The results for the reverberant signals were quite well aligned with the results of the cases without reverberation. This is a promising result, considering the fact that any real-life environment does incorporate some degree of reverberation.

For real claps, we performed a randomized cross-validation procedure, in which the recorded data was divided into separate training and validation sets, with the probability of a clap event belonging to the test set being 0.33. The classifier was trained with the training data and tested with the validation data in 20 successive runs with differently selected training and validation sets, and the obtained results were averaged. The results are presented in Table 3.

We notice that although these results are worse than for the synthetic data, they still are well above chance level. For test subject 2 the results are good, with 0.64 % correct classification rate. Comparison with the synthetic data results shows that the systematic overlaps between different classes differ from those of the synthetic data. This result must be because the synthesis filter

Table 2: Relative classification results for the synthetic hand claps. Overall correct classification performance is 71.7 % for the FFT bin classifier and 69.9 % for the filter coefficient classifier..

	FFT bin classifier							
Label	P1	P2	P3	A1	A2	A3	A1+	A1-
P1	**0.30**	0	0.13	0	0.51	0.05	0	0
P2	0	**0.70**	0	0.16	0	0	0.03	0.11
P3	0.05	0	**0.81**	0	0.14	0	0	0
A1	0	0.28	0	**0.63**	0	0	0.04	0.04
A2	0.21	0	0.05	0	**0.67**	0.07	0	0
A3	0.08	0	0.01	0	0.03	**0.88**	0	0
A1+	0	0.07	0	0	0	0	**0.90**	0.02
A1-	0	0.13	0	0.06	0	0.01	0	**0.81**
	Filter coefficient classifier							
Label	P1	P2	P3	A1	A2	A3	A1+	A1-
P1	**0.68**	0	0.01	0	0.12	0.17	0	0.01
P2	0	**0.61**	0	0.34	0	0	0.01	0.04
P3	0.03	0	**0.76**	0	0.02	0.19	0	0
A1	0	0.34	0	**0.60**	0	0	0	0.06
A2	0.44	0	0.02	0	**0.39**	0.16	0	0
A3	0.19	0	0.19	0	0.03	**0.59**	0	0
A1+	0	0.06	0	0	0	0	**0.94**	0
A1-	0.03	0.04	0	0.01	0	0	0	**0.92**

Table 3: Relative classification results for the real hand claps with the filter coefficient classifier. Overall correct classification rates were 48 % for subject 1, 64 % for subject 2, and 79 % for the two-class problem.

	Filter coefficient classification, subject A							
	P1	P2	P3	A1	A2	A3	A1+	A1-
P1	**0.17**	0.27	0	0	0	0	0.10	0.46
P2	0.06	**0.69**	0.01	0	0	0	0	0.24
P3	0.05	0.08	**0.38**	0	0	0	0.14	0.34
A1	0	0	0	**0.39**	0.42	0.10	0.08	0.10
A2	0	0	0	0.10	**0.71**	0.19	0	0
A3	0	0	0	0.13	0.32	**0.55**	0	0
A1+	0.15	0.05	0.11	0.02	0	0	**0.56**	0.10
A1-	0.26	0.41	0	0	0	0	0.01	**0.32**

	Filter coefficient classification, subject B							
	P1	P2	P3	A1	A2	A3	A1+	A1-
P1	**0.52**	0.19	0	0.03	0	0	0.02	0.24
P2	0.02	**0.88**	0	0	0	0	0	0.10
P3	0	0.12	**0.76**	0	0	0	0.11	0.02
A1	0	0	0.01	**0.69**	0.05	0	0.25	0
A2	0	0	0	0	**0.86**	0.14	0	0
A3	0	0	0	0.04	0.21	**0.74**	0.01	0
A1+	0.09	0	0	0.31	0	0.08	**0.50**	0.02
A1-	0.66	0.17	0	0	0	0	0	**0.17**

	Filter coefficient classification, Flamenco claps							
	A3	A1+						
A3	**0.82**	0.18						
A1+	0.23	**0.77**						

parameters were based on the analysis of another clapper's claps.

As an easier classification task, we experimented with the flamenco type of data, labeled as consisting of two hand configurations. For this kind of two-class problem the classification approach seems to work well. It should be noted that the FFT bin classifier worked well for the two-class flamenco type claps, yielding almost 100 % classification rates. Instead, for the eight-class cases, the FFT bin classifier did not perform very consistently.

5. CONCLUSIONS AND FUTURE WORK

As the results of this research indicate, it is possible to make inference of the hand configuration of a clapper given the resulting sound. The results also suggest that the claps of individual clappers may incorporate systematic differences from other people's claps, which would enable personified control interfaces. In the future, both the personification and clapper-independency of the proposed system should be studied.

Another future step is to apply some better feature selection method to the features such as the genetic algorithm applied in [5], and to test the usefulness of more features. In the running of this research, experiments with several different features were made, but the results obtained so far were inconclusive. Obviously, the naïve Bayes assumption does not hold for all feature sets, as assuming conditional independence between some features may be unreasonable.

A related problem is the inference of the resonator filter parameters directly from the clapping sounds. For this, we plan to extend the model to a hierarchical Bayesian model, with the features conditioned on the filter parameters. This approach would also be better suited for coping with the continuous deviations in the hand configurations. It will also be an interesting task to include temporal parameters in the model, namely the interval between the claps and the rhythmic deviation, to complete the single clapper model. This would enable the identification of rhythmic flamenco patterns, for example.

An important step in the future is to try out a real-time implementation of the model. The current algorithm is light enough for real-time inference. Of course, a real-time system will also require an automatic method for clap onset detection, and the robustness of the system to noise and distortion must be verified.

6. ACKNOWLEDGEMENTS

This work is supported by the Academy of Finland (project 120583 Schema-SID). Our thanks to Leevi Peltola for the measurement material and the initial release of the ClaPD, and to Jukka Ahonen for assistance with the recordings.

7. REFERENCES

[1] W.W. Gaver, "What in the World Do We Hear?: an Ecological Approach to Auditory Event Perception," *Ecological Psychology*, vol. 5, no. 1, pp. 1–30, 1993.

[2] B.H. Repp, "The sound of two hands clapping: An exploratory study," *The Journal of the Acoustical Society of America*, vol. 81, no. 4, pp. 1100, 1987.

[3] L. Peltola, C. Erkut, P.R. Cook, and V. Välimäki, "Synthesis of hand clapping sounds," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1021–1029, 2007.

[4] K. Hanahara, Y. Tada, and T. Muroi, "Human-robot communication by means of hand-clapping (preliminary experiment with hand-clapping language)," *Proc. IEEE Intl. Conf. Systems, Man and Cybernetics*, pp. 2995–3000, Jan 2007.

[5] J. Olajec, C. Erkut, and R. Jarina, "GA-based feature selection for synchronous and asynchronous applause detection," in *Proc. Finnish Signal Processing Symposium (Finsig'07)*, Oulu, Finland, August 2007.

[6] L. Peltola, "Analysis, Parametric Synthesis, and Control of Hand Clapping Sounds," M.S. thesis, Helsinki University of Technology, 2004.

[7] P.R. Cook, "Physically Informed Sonic Modeling (PhISM): Synthesis of Percussive Sounds," *Computer Music Journal*, vol. 21, no. 3, pp. 38–49, 1997.

[8] D. Rocchesso, "Physically-based sounding objects, as we develop them today," *J. New Music Research*, vol. 33, no. 3, pp. 305–313, September 2004.

[9] M. Puckette, "Pure data: another integrated computer music environment," in *Proc. Second Intercollege Computer Music Concerts*, Tachikawa, Japan, 1996, pp. 37–41.

[10] C. Erkut, "Towards physics-based control and sound synthesis of multi-agent systems: Application to synthetic hand clapping," in *Proc. Nordic Music Technology Conf.*, Trondheim, Norway, October 2006.

[11] K. Steiglitz and L. McBride, "A technique for the identification of linear systems," *Automatic Control, IEEE Transactions on*, vol. 10, no. 4, pp. 461–464, 1965.

[12] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[13] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, Singapore, 2006.