Sonic Handprints: Person Identification with Hand Clapping Sounds by a Model-Based Method

Antti Jylhä[1], Cumhur Erkut[1], Umut Şimşekli[2], and A. Taylan Cemgil[2]

[1]*Aalto University School of Electrical Engineering, Dept. Signal Processing and Acoustics, P.O. Box 13000, FI-00076 Aalto*

[2]*Department of Computer Engineering, Boğazici University, 34342 Bebek, Istanbul, Turkey*

Correspondence should be addressed to Antti Jylhä (antti.jylha@aalto.fi)

ABSTRACT

Sound-based person identification has largely focused on speaker recognition. However, also non-speech sounds may convey personal information, as suggested by our previous studies on hand clap recognition. We propose the use of a probabilistic model-based technique for person identification based on their hand clapping sounds. The method is based on a Hidden Markov Model which uses spectral templates in its observation model. The technique has been evaluated in an experiment with 16 subjects, resulting in an overall correct classification rate of 64 %. The algorithm runs in real-time, making it suitable also for interactive systems.

1. INTRODUCTION

Person identification has many practical applications, ranging from security systems to non-security critical real-time applications, such as multi-user video games. Several approaches to person identification have been presented especially in the field of security systems, such as iris scanning [18], acoustic cues and face recognition [2], biometric sensor fusion [5], and biological motion [19]. The use of multiple information sources has been found beneficial. However, many of these systems require either expensive specialized hardware or heavy algorithms, which may be unsuitable for less serious real-time applications, such as multiplayer video games or multi-user musical systems, with multiple people performing simultaneously.

A remedy on desktop systems can be found in temporality: just as with handwritten signatures, the way an individual types on a keyboard can be unique enough to identify her [11]. In the last decade, research on keystroke dynamics, i.e., the study of extracted keystroke timing features, has been an active area of research (see [7] for an overview). More recently, keystroke dynamics have been combined with affective computing to infer the emotional states of people during their interaction with desktop computers [6]. These kind of studies that combine feature extraction, classification, and matching with self reports are promising in manipulating the interfaces and systems within close proximity. For longer distances, sound-based solutions that rely on the metaphor "sound is touch at distance" may work better.

To date, sound-based person identification has mainly focused on speaker recognition. However, also other sounds may convey personal information. For example, it has been suggested that humans are capable of recognizing their own hand clapping sound from that of others [15]. Our previous work on the recognition of percussive non-speech sounds has indicated that an algorithm taught on one person's hand clapping sounds does not function well with those of another person [12], suggesting that the recognition of personal information is extendable also to computational systems.

In this work, we propose the use of a probabilistic model-based technique for person identification from their hand clapping sounds. We have previously utilized the technique in the recognition of different hand clapping types and percussion instrument strokes [4], and now aim at examining its utility on person identification.

In the following section, a brief review of related work on percussive sound recognition and sound-based person identification is presented, followed by a description of our recognition algorithm in Section 3. The experiment

and results of the clapper identification problem are presented in Section 4, and finally conclusions and indications for future work are drawn in Section 5.

2. RELATED WORK

The problem setting in this study relates to detecting subtle differences in percussive sounds. While there are numerous techniques applicable for percussive event recognition, one prominent approach in the recent past has been the use of template-based techniques. Computing a characteristic spectral template for the sound events has been utilized for example by Yoshii, Goto, and Okuno in an automatic description system for bass and snare drum sounds [20]. The technique functions in two steps of template matching and template adaptation, making it possible to tune the templates to a given musical excerpt.

The spectral template can also be characterized by features of lower dimensionality, such as Mel-frequency cepstral coefficients (MFCC), as proposed by Paulus and Klapuri [14] for automatic drum transcription. They use MFCCs in conjunction with musical meter analysis and probabilistic modeling. More recently, the same authors have proposed the use of networked Hidden Markov Models (HMM) for drum sound detection in polyphonic music, using MFCCs and their first differences as characteristic features.

Related to detecting subtle differences in percussive sounds, Gillet and Richard have studied the automatic recognition of Tabla strokes using a combination of Gaussian Mixture Models (GMM) and HMMs [9]. This line of study differs from typical instrument sound recognition in that the instrument remains the same, and the differences in sound arise from the stroke, and therefore resembles the hand configuration inference problem of [12] and [4].

For a more comprehensive overview of percussive sound identification (in the context of musical transcription), the reader is referred to [10] and [8].

3. MODEL-BASED TECHNIQUE FOR HAND CLAP RECOGNITION

Şimşekli and Cemgil have presented two probabilistic models for online pitch tracking [17]. The models are template-based and do not heavily depend on the application, which makes the models fairly easy to apply to percussive events. Instead of detecting pitch labels from streaming audio data, in [4], one of the probabilistic models has been adapted to percussive event tracking, aiming at inferring a predefined set of short, percussive events. Here, a summary of the model is presented; for more details, see [4].

In the model, the audio signal is represented by its magnitude spectrum that is computed via the fast Fourier transform (FFT). $x_{\nu,\tau}$ is defined as the magnitude spectrum of the audio data with frequency index ν and τ as the time frame index. Here, $\nu \in \{1,2,...,F\}$ and $\tau \in \{1,2,...,T\}$. For each time frame τ, an indicator variable r_τ is defined on a discrete state space D_r, determining the label we are interested in. D_r consists of event labels, i.e., in this study, clapper identities. The indicator variables r_τ are hidden since they are not observed directly.

In the model, the main idea is that each event has a certain characteristic spectral shape which is assumed to be rendered by a specific hidden *scaling* variable, v_τ. The spectral shapes, referred to as *spectral templates*, are denoted by $t_{\nu,i}$. The ν index is again the frequency index and the index i indicates the event labels. Here, i takes values between 1 and I, where I is the number of different spectral templates. The scaling variables v_τ define the overall amplitude factor, by which the whole template is multiplied. The probabilistic model is formally defined as follows:

$$\begin{aligned} r_0 &\sim p(r_0) \\ r_\tau | r_{\tau-1} &\sim p(r_\tau | r_{\tau-1}) \\ v_\tau &\sim \mathcal{G}(v_\tau; a_v, b_v) \\ x_{\nu,\tau} | v_\tau, r_\tau &\sim \prod_{i=1}^{I} \mathcal{PO}(x_{\nu,\tau}; t_{\nu,i} v_\tau)^{[r_\tau = i]}. \end{aligned} \quad (1)$$

Here $[x] = 1$ if x is true, $[x] = 0$ otherwise and the symbols $\mathcal{G}$ and $\mathcal{PO}$ represent the Gamma and the Poisson distributions respectively, where

$$\begin{aligned} \mathcal{G}(x; a_v, b_v) = \exp((a_v - 1)\log x - b_v x - \log\Gamma(a_v) \\ + a_v \log(b_v)) \end{aligned} \quad (2)$$

$$\mathcal{PO}(y; \lambda) = \exp(y \log\lambda - \lambda - \log\Gamma(y+1)), \quad (3)$$

where Γ is the Gamma function.

The Poisson distribution, recently applied in non-negative matrix factorization problems [3], is chosen as the observation model. Also, a Gamma prior on v_τ is

chosen to preserve conjugacy and make use of the scaling property of the Gamma distribution. The conjugate prior grants an analytic closed-form solution to the posterior.

Moreover, a Markovian prior is chosen on the indicator variables, r_τ which means r_τ depends only on $r_{\tau-1}$. A single state is used in order to represent the acoustic events (i.e. hand claps).

One advantage of this model is that the scaling variables v_τ can be integrated out analytically. It is easy to check that once this is done, provided the templates $t_{v,i}$ are already known, the model reduces to a standard Hidden Markov Model (HMM) with a Compound Poisson observation model as shown below:

$$
\begin{aligned}
&p(x_{1:F,\tau}|r_\tau = i) \\
&= \int dv_\tau \exp\Big(\sum_{v=1}^{F} \log \mathscr{P}\mathscr{O}(x_{v,\tau}; v_\tau t_{v,i}) + \log \mathscr{G}(v_\tau; a_v, b_v)\Big) \\
&= \exp\left(\log\Gamma\left(\sum_{v=1}^{F} x_{v,\tau} + a_v\right) - \sum_{v=1}^{F} \log\Gamma(x_{v,\tau}+1)\right. \\
&\quad + \sum_{v=1}^{F} x_{v,\tau} t_{v,i} - \left(\sum_{v=1}^{F} x_{v,\tau} + a_v\right) \log\left(\sum_{v=1}^{F} t_{v,i} + b_v\right) \\
&\quad \left. + a_v \log b_v - \log\Gamma(a_v)\right) \qquad (4)
\end{aligned}
$$

As we have a standard HMM from now on, the well-known forward algorithm can be run in order to perform inference on the model. Also, the most probable state sequence can be estimated by running the Viterbi algorithm, which lets us identify the clappers on streaming audio. A benefit of having a standard HMM is that the inference algorithm can be made to run very fast. Therefore, the inference scheme can be implemented in real-time without any approximation [1]. Furthermore, for learning the spectral templates, the well-known Expectation - Maximization (EM) algorithm is utilized (for details, see [4] and [16]).

Our previous work on recognizing different hand clapping types has indicated that reverberation and room reflections can cause degradation in classification accuracy with hand clapping sounds [4]. As a remedy, the algorithm output is post-processed by skipping the 10 frames (23.2 ms) after the initial non-silent frame for each detected event to reduce the degradation effect.

4. RESULTS

Monophonic hand clapping sounds of 16 people were recorded in a quiet room (concrete walls, reverberation time around 0.7 s on low frequencies, 0.5 s at 250 Hz - 2000 Hz, and less than 0.3 s on high frequencies). The sounds were captured with the built-in microphone of a Dell laptop computer in order to keep the audio hardware similar to a realistic use scenario. The distance between the hands of the seated subjects and the microphone was 40-60 cm. The subjects were instructed to clap freely with a constant tempo for a minimum of 30 seconds, resulting in a sound bank of 78 claps per subject on average (ranging from 44 claps of subject 2 to 138 claps of subject 12). The clapping style was not strictly controlled, since we wanted to capture as natural clapping as possible.

The individual sound files of each clapper were segmented into four segments of equal duration, and two segments were randomly assigned as training data and two as test data. In other words, the training and the test data were both recorded in the same environment. The model was then trained with the training data from all 16 subjects, resulting in spectral templates. Fig. 1 illustrates the templates learned by the algorithm after the training. The differences in frequency distribution between different people can be clearly observed, but also some very similar templates can be pinpointed, such as those of subjects 9, 12, and 16, portraying a prominent resonance around the lower frequencies, and 2, 3, and 11, having a concentration of energy within a common wider frequency range. It should be noted, however, that in some cases (such as with subject 2) the hand configuration evolved throughout the clapping sequence, which is also audible in the resulting sound file. As the template represents the overall spectral shape, it can thus contain information from a number of fluidly changing hand configurations.

The classification accuracy was evaluated with the test data from all 16 subjects. The classification results are presented in Table 1. The overall performance of correctly classifying different people's clapping is 64 %. While the performance of the system varies between subjects, it performs way beyond chance level (6.25 %) in all cases. For people whose spectral templates clearly contain unique high-energy regions, the results are very good. Some subjects were very consistent with their hand configuration during the recording, for example

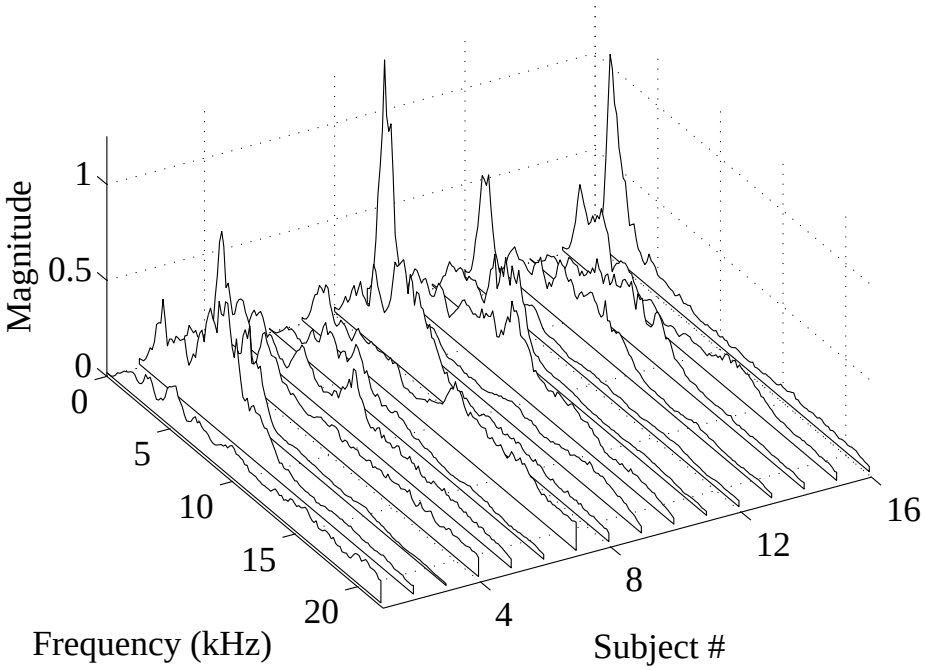

Fig. 1: Spectral templates of 16 subjects.

Table 1: Relative classification results for recognizing the clapping of 16 subjects. Row labels indicate target class and columns the percentage (rounded) of classifying the claps of target class to all classes. Overall correct classification rate is 64 %.

	Subject A															
Label	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16
S1	**96**	0	0	0	0	0.05	0	0	2	0	0	0	0	0	0	0
S2	4	**22**	4	0	0	4	0	4	0	0	22	4	35	4	0	0
S3	5	9	**45**	0	0	0	9	2	2	0	18	0	0	0	9	0
S4	0	0	4	**79**	4	0	7	0	0	0	0	4	0	0	0	4
S5	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0
S6	2	4	0	0	6	**77**	2	0	0	6	0	0	0	2	0	0
S7	12	0	0	11	5	2	**60**	0	2	7	0	0	0	0	2	0
S8	0	10	13	3	0	0	0	**40**	0	0	7	0	20	7	0	0
S9	0	0	0	14	3	0	8	0	**44**	0	0	25	0	0	0	6
S10	6	0	4	0	2	0	12	2	0	**42**	0	2	13	15	2	0
S11	3	13	8	0	3	3	0	3	0	0	**68**	0	0	0	3	0
S12	0	0	1	9	1	1	3	0	4	0	0	**76**	0	0	1	3
S13	2	9	0	9	5	9	9	0	0	2	2	0	**49**	2	0	0
S14	5	0	0	0	20	0	3	3	5	3	0	0	3	**57**	3	0
S15	0	0	0	10	5	2	2	2	0	2	0	2	0	0	**74**	0
S16	0	0	0	0	0	0	0	0	2	0	0	2	0	0	2	**93**

subjects 5 and 16. Notably, subject 5 has a unique resonance structure observable from the template, yielding 100 % classification accuracy with this data set, while the frequency distribution of subject 1 is very wideband compared to most others with a couple of prominent resonances, giving 96 % correct classification.

On the other hand, for people, whose hand configuration varied more during the recording, the classification results are less consistent. A systematic overlap can also be observed between people, whose natural clapping sounds according to the learned spectral templates are similar, such as with subjects 2, 3, and 11.

The suggestion of Repp [15] that humans could identify their own hand clapping sound has previously not been utilized in computational systems. The results of this study show that the claim is plausible and that there indeed are machine-identifiable spectral differences between the hand clap sounds of different people.

These results on identifying the person based on free clapping sounds are promising and indicate that hand claps could be used as one type of person identification, but possibly not in systems with high security requirements. On the other hand systems such as multiplayer video games or multi-user musical software could potentially utilize hand claps as one identification mechanism. Combining the spectral template-based classification with temporal patterning of the events could be beneficial.

5. CONCLUSIONS

We proposed the use of a model-based machine learning technique for person identification based on hand clap sounds. The probabilistic model is generic and suitable for the recognition of percussive sounds in real-time, and has now been evaluated on the hand clapping sounds of 16 people. The results of clapper recognition are encouraging, and given the overall performance, we can assume that clapping could be used as one identifier in applications requiring person identification. For example, in multi-person games there is typically a limited number (2-4) of players, and therefore the proposed method can achieve arguably a higher classification rate than in this study with 16 people.

One curious question is how well human listeners would perform in a similar identification task, compared to the 64 % overall classification rate of our algorithm among 16 subjects. While we have not specifically tested this, Repp's results reported in [15] can provide some insight. In his tests, while the participants identify their own claps with a score of 46 % among 20 subjects, their overall performance considering the whole group is 11 % when self-recognition scores are excluded and 13 % when included. One should note, however, that our experimental design was very different from that of [15], where for instance the subjects had three guesses about the identity of the clapper.

In this experiment some of the clappers varied their hand configuration during the test, so it is reasonable to assume that had they been instructed to keep their hand configuration constant, the results would be even better. While we did not examine the utilization of temporal data in this experiment, time-related information from multiple claps could be used in combination with the spectral information for improved accuracy. Clapper identification based on a sequence of claps is likely more robust than identifying the person based on a single clap, so fusing the plain clap recognition with simple rhythmic patterns can be considered a promising direction.

As discussed, reverberation and varying acoustic conditions can degrade the recognition accuracy. If the room response has strong resonances overlapping the strong modes of a certain clapper, the algorithm can make classification errors. A future study is planned to quantify the effects of varying acoustic conditions.

The model-based technique used in this study has been designed for monophonic signals. A considerable challenge is to perform person identification in polyphonic signals, such as dense, applause-type signals [13], akin to how pitch detection is performed in polyphonic musical signals [17]. If this challenge can be tackled, many applications of interactive spatial audio coding, for instance point-of-view renderings of multichannel concert recordings, will be possible.

6. REFERENCES

[1] Ethem Alpaydın. *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, Cambridge, MA, USA, 2004.

[2] R. Brunelli and D. Falavigna. Person identification using multiple cues. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17(10):955–966, 1995.

[3] A. T. Cemgil. Bayesian inference in non-negative matrix factorisation models. *Computational Intelligence and Neuroscience*, (4):1–17, 2009.

[4] U. Şimşekli, A. Jylhä, C. Erkut, and A.T. Cemgil. Real-time recognition of percussive sounds by a model-based method. *EURASIP J. Advances in Signal Processing*, 2011:1–14, 2011. Special Issue on Musical Applications of Real-Time Signal Processing.

[5] U. Dieckmann, P. Plankensteiner, and T. Wagner. SESAM: A biometric person identification system using sensor fusion. *Pattern Recognition Letters*, 18(9):827–833, 1997.

[6] C. Epp, M. Lippold, and R. L. Mandryk. Identifying emotional states using keystroke dynamics. In *Proc. Conf. Human Factors in Computing Systems (CHI '11)*, pages 715–724, Vancouver, Canada, May 2011.

[7] S. H. Fairclough. Fundamentals of physiological computing. *Interacting with Computers*, 21(1):133–145, Jan 2009.

[8] D FitzGerald and J Paulus. Unpitched percussion transcription. In A. Klapuri and M. Davy, editors, *Signal Processing Methods for Music*. Springer, New York, NY, USA, Jan 2006.

[9] O Gillet and G Richard. Automatic labelling of tabla signals. In *Proc. Intl. Conf. Music Information Retrieval (ISMIR)*, Baltimore, MD, USA, Oct 2003.

[10] P Herrera, A Dehamel, and F Gouyon. Automatic labeling of unpitched percussion sounds. In *Proc. Audio Eng. Soc. 114th Convention*, pages 1–14, Amsterdam, Netherlands, Jan 2003.

[11] R. Joyce and G. Gupta. Identity authentication based on keystroke latencies. *Comm. ACM*, 33(2):168–176, Feb 1990.

[12] A Jylhä and C Erkut. Inferring the hand configuration from hand clapping sounds. In *Proc. Intl. Conf. Digital Audio Effects (DAFx)*, pages 301–304, Espoo, Finland, Sep 2008.

[13] M.-V. Laitinen, F. Kuech, S. Disch, and V. Pulkki. Reproducing applause-type signals with directional audio coding. *J. Audio Eng. Soc.*, 59(1):29–43, 2009.

[14] J. Paulus and A. Klapuri. Model-based event labeling in the transcription of percussive audio signals. In *Proc. Intl. Conf. Digital Audio Effects (DAFx)*, pages 73–77, London, UK, Sep 2003.

[15] B.H. Repp. The sound of two hands clapping: An exploratory study. *J. Acoust. Soc. Am.*, 81(4):1100–1109, 1987.

[16] Umut Şimşekli. Bayesian methods for real-time pitch tracking. Master's thesis, Boğaziçi University, Istanbul, Turkey, 2010.

[17] Umut Şimşekli and Ali Taylan Cemgil. A comparison of probabilistic models for online pitch tracking. In *Proc. 7th Sound and Music Computing Conf. (SMC)*, pages 502–509, Jul 2010.

[18] C. Tisse, L. Martin, L. Torres, and M. Robert. Person identification technique using human iris recognition. In *Proc. Vision Interface*, pages 294–299, Calgary, Canada, May 2002.

[19] N.F. Troje, C. Westhoff, and M. Lavrov. Person identification from biological motion: Effects of structural and kinematic cues. *Attention, Perception, & Psychophysics*, 67(4):667–675, 2005.

[20] K Yoshii, M Goto, and HG Okuno. Automatic drum sound description for real-world music using template adaptation and matching methods. In *Proc. Intl. Conf. Music Information Retrieval (ISMIR)*, pages 184–191, Barcelona, Spain, Oct 2004.