CLASSIFICATION SCHEMES FOR STEP SOUNDS BASED ON GAMMATONE-FILTERS

Robert Anniés[1]
[1]Neural Information Processing Group
Technische Universität Berlin
robokopp@cs.tu-berlin.de

Elena Martínez Hernández[1,2]
[2] Music Technology Group
Universitat Pompeu Fabra, Barcelona
emartinez@iua.upf.edu

Kamil Adiloğlu[1]
kamil@cs.tu-berlin.de

Hendrik Purwins[1,2]
hpurwins@iua.upf.edu

Klaus Obermayer[1]
oby@cs.tu-berlin.de

Abstract

In this study the classification performance of 2 machine learning methods and 2 sound representations schemes are compared, having the focus on short impact like sounds: Footsteps have been classified according to the material of the floor and the shoe type. The gamma-tone auditory filter-bank is a spectral analyser, that converts a given signal into a multi-channel simulation of the basilar membrane motion. Combinations of the gamma-tone auditory filter bank with the Hilbert transform and with the Meddis Inner Hair-cell model have been evaluated and compared in classification tasks. The experiments show that the gammatone based representation techniques yield in general promising results in the classification tasks of impact like everyday sounds. The support vector machines outperform the hidden Markov models, where both in general perform equal or better using the inner hair cell model representation.

1 Introduction

Everyday sounds play a significant role in communication, localisation and interaction in everyday life. The question is whether and how it is possible to detect properties of the sound generating process and/or listeners perception, i.e. what is the sound source, what materials are involved, and what impressions are received by a listener. We investigate this question in a classification framework. In our approach, we classify only everyday sounds based on the function they fulfil, their material or shape, or the objects they interact with.

In this particular study we focused on impact like sounds emitted when objects of different matetrials collide, here represented by a sound corpus of footsteps. The acoustics of footsteps and walking have been extensively analysed [1]. Spectra of various floor types and the resulting subjective loudness of foot steps have been compared [2]. In [3], it has been shown that biologically inspired representations allow for better classification of step sounds than widely used representations such as mel-frequency cepstrum coefficients (MFCCs) or a combination of low-level features (such as zero crossings, spectral centroids, roll-off).

In this paper, we investigate the biologically motivated sound descriptors, starting with gammatone filter banks followed by analysis of the temporal modulation envelopes or the application of a inner hair cell model.

We compare how two classifiers, namely support vector machines (SVM) and hidden Markov models (HMM), perform on sperating sounds of footsteps with different materials. The results of each descriptor-classifier pair are compared with the other descriptor-classifier pairs, in order to find the optimal representation and the classification model for this corpus.

Organisation of the paper: The following section describes the methods of preprocessing the audio signals with the aim to reduce their dimensionality without loosing classification relevant information. Section 3 gives information about the applied machine learning methods, section 4 shows details about the collection of sounds that were used followed by section 5 that presents the expriments and their results. The discussion and conclusions close the paper.

2 Representation

In order to perform classfication on audio signals an appropriate method has to be chosen to represent the signal such that the dimensionality of the signal is reduced without loosing class relevant information. In this study we compare 2 algorithms – SVM and HMM – using 2 ways of representation – spectrum oriented and time series oriented.. This leads to 4 cases:

	SVM	HMM
time	1) one vector encodes temporal structure of whole signal	2) one vector per time frame encodes (micro-) temporal content of this frame
spectrum	3) one vector encodes spectral content	4) vector per time frame encodes spectral content of this frame

HMM has, in opposite to SVM, inherently the ability to analyse a sequence of time frames. A preprocessing is therefore done on a windowed signal. In case 2) this leads to a 2 level temporal analysis, whereas in case 3) there is no temporal content at all[1].

Mel Frequency Cepstrum Coefficients (MFCC's) [4] are well established representation scheme, which dominate applications in speech recognition and music processing. However, for everyday sounds it has been shown [3, 5, 6, 7] that gamma-tone auditory filter-banks yield better classification results than MFCC's.

Figure 1 describes the 2 paths of preprocessing the input signal. Common is the beginning with a gammatone filter bank and the last step to compress a sequence of values into 4 values by using the Mean-Variance feature intergration scheme.

2.1 Spectral content

2.1.1 Gamma-tone Auditory Filter-bank

A gamma-tone auditory filter-bank [8] is the first step of the cochlea simulation, where the basilar membrane motion is simulated in a filter bank. The impulse response of a gamma-tone filter is highly similar to the magnitude characteristics of a human auditory filter, which makes the gamma-tone auditory filter bank a biologically plausible representation. With increasing centre frequency, the spacing and the bandwidth of the gamma-tone filters increase, however the overlapping of each consecutive filter stays the same (equivalent rectangular bandwidth or shortly ERB [9]). ERBs are similar to the Bark or the Mel scale.

As a pre-processing of the everyday sounds, we use the gamma-tone filter implementation in Malcolm Slaney's Auditory Toolbox [10]. We use 18 gamma-tone filters in total. Therefore, for each given sound, we obtain 18 filter responses from the gamma-tone filter bank. A frequency of $f_{low} = 3$ lies in the audible frequency range captured by the Basilar membrane.

[1] temporal information was encoded by using very low frequencies of the spectrum ($< 20\,\mathrm{Hz}$) as 'rythmic' content. Depending on the point of view this might be seen as spectral and/or temporal content

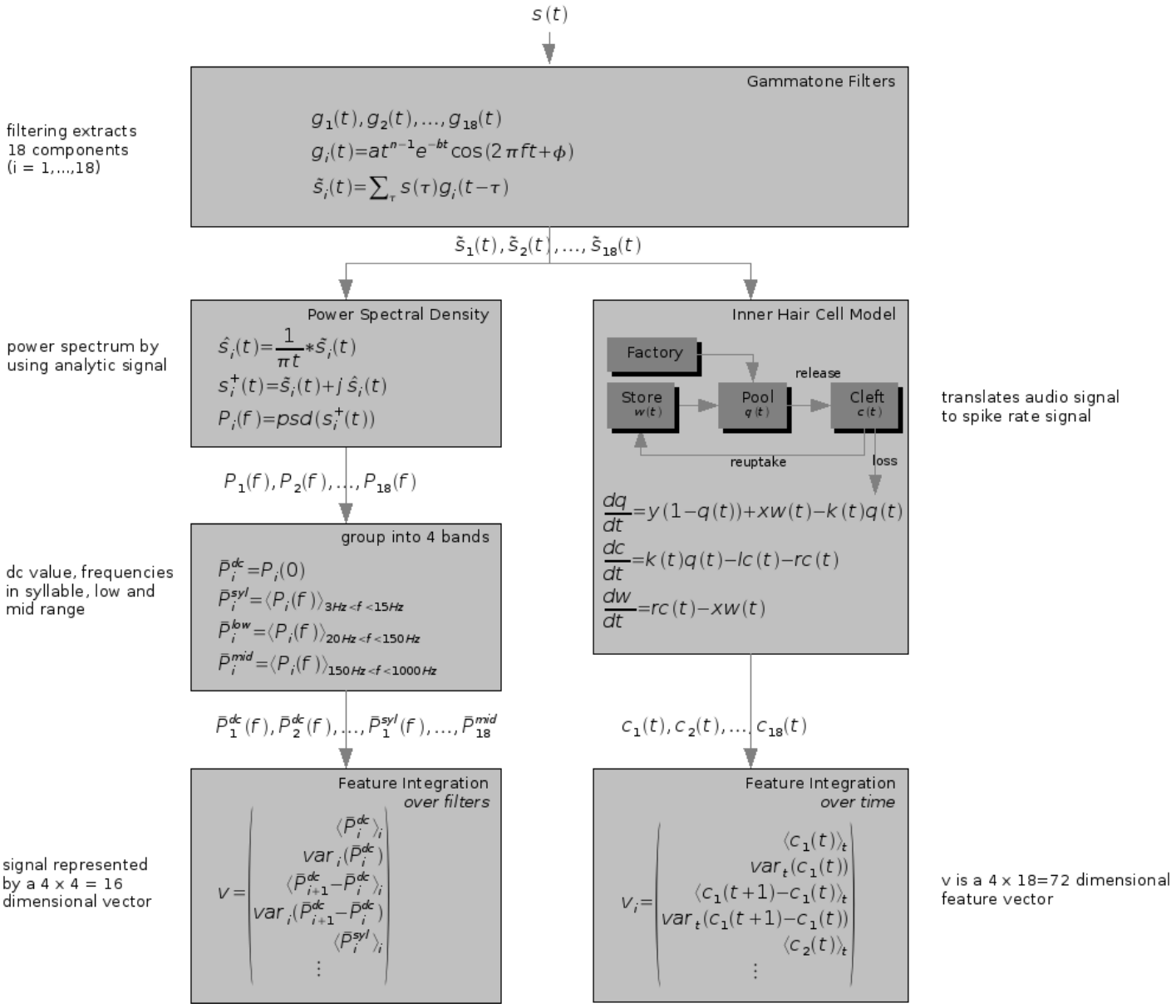

Figure 1: Steps of preprocessing. The left side shows the representation that emphasises the spectral features of the signal. The right one shows the steps using the time domain analysis with the inner hair cell model. In the SVM experiments the whole sound example was used as signal $s(t)$ in opposite to the HMM, for which the signal was windowed and the procedures were applied on each frame resulting in a series of vectors v_i^t.

But it captures in addition also features that refer to frequencies in the range of roughness or 'rhythmical' content.

The Gamma-tone filters can be combined with other representations, in order to obtain a more complete representation scheme.

2.1.2 Hilbert Transform

The first method with which we combine the gamma-tone filters is the Hilbert Transform [3]. The Hilbert transform of a signal is the convolution of the time domain signal with $\frac{1}{\pi t}$.

Combining the Hilbert transformed signal with the original signal, we obtain the analytic signal. This process deletes the negative components of the signal in the frequency domain, and doubles the amplitudes on the positive side. Furthermore the analytic signal is a base band signal. The power spectrum of the analytic signal is the modulation spectrum of the temporal envelope. In many roughness estimations, this step is the very first step of the roughness estimation procedure [11].

2.1.3 Grouping in 4 bands

However, the huge dimensionality of the power spectrum of each filter output should be reduced, in order to be able to create a feature vector for a sound. Therefore we summarise these values in four frequency bands [6] [7] by simply taking the average of the power values corresponding to these frequency bands. These frequency bands are the DC values, the frequency interval 3-15 Hz (syllable rate in speech, or rythms), 20-150 Hz (roughness), and 150-1000Hz (low pass).

2.2 Temporal content

2.2.1 Inner Hair Cell Model

To analyse the signal in time, we directed the output of the gamma-tone filters into the inner hair cell model of Meddis [12]. In this model, the firing rate of the inner hair cells, connected to the basilar membrane, is modelled. The inner hair cells fire, when a stimulus arrives. This happens when the basilar membrane is deflected at the point of a resonance frequency, where the hair cell sits. This firing is simulated by the dynamics of production and flow of transmitter substance. A certain amount of transmitter substance is released into the synaptic cleft between the hair cell and another neuron, depending on the strength of the stimulus. For each arriving stimulus, the Meddis inner hair cell model calculates these amounts iteratively. In our representation, we use the rate of transmitted part of the transmitter substance [5].

2.3 Feature Integration

In both analyseses (2.1 and 2.2) we end up with a number of serieses of values that have to be compressed into one final feature vector. We applied a feature integration scheme that calculates the mean and variance of the sequence as well as the mean and variance of the first derivative (difference of two consecutive values).

In the spectral content case these are 4 sequences (4 bands) of 18 energy values of the filters, which are compressed each to the 4 values described above. This makes a features vector of 16 values. Note, that the derivatives (deltas) were taken over the the filter outputs, i.e. in frequency domain.

In the temporal case we used the same schema in the time domain. Here we had 18 time serieses from the Inner Hair Cell model, which were compressed each to 4 mean/variance values resulting in a 72 dimensional vector.

3 Learning Methods

3.1 SVM Settings

In these experiments, we used c-support vector machines, a classification tool based on the maximisation of the margin of the classification boundary between two classes. The c-SVM has two parameters (c and g), which should be determined beforehand. In order to find the optimal values for these two parameters, we performed a grid search, where we changed these two parameters slightly to find optimal parameter settings. As grid values we use all combinations of $c = 2^{12}, 2^{13}, 2^{14}, \ldots, 2^{16}, g = 2^{-16}, 2^{-15}, 2^{-14}, \ldots, 2^{-4}$.

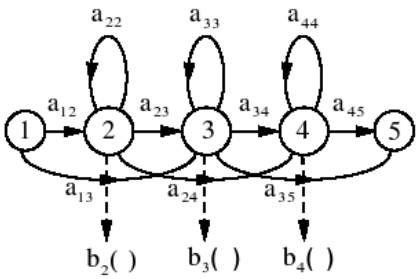

Figure 2: Hidden Markov Model Architecture

3.2 HMM Settings

A Hidden Markov model with a feed forward architecture was used, as it is common in speech recognition. The model assumes that the sound is constructed of segments of steady states in time. How many segments appear is not known. Thus, HMMs with different state numbers were used in the evaluation. Each state models the sound as a single multivariate normal distribution with parameters $\mathbf{\Sigma}$ and mean μ, which are estimated while training for each segment. The covariance matrix $\mathbf{\Sigma}$ was approximated by its diagonal to keep the number of (scalar) parameters small.

The transition matrix allows only state changes to the next or to next but one, as depicted in figure 2. It is a model with 3 emitting states (Gaussian $b_i(\Sigma, \mu)$) and transition matrix $a_{(ij)}$.

A grid search was performed to find the best architecture. The number of states from 2 to 10 and the window size of the preprocessing step was varied: 512, 102 and 2048 samples. Smaller windows increase the time resolution and more states increase the number of steady state segments.

4 Data Sets

For the experiments, recordings of footsteps are selected from the sound collection "Sound Ideas" [14].

The footsteps are a subset of the Foley collection. It consists of recordings of different kinds of shoes (heels, boots, barefoot, sneakers, leather) on various grounds (concrete, wood, dirt, metal, snow, sand). We picked the following movement modes: walking, running, jogging from male and female subjects. The recordings are cut, when necessary, such that exactly one step is contained in one sample.

For the experiments on concrete and wood floor (table 1, 2 we used 44 files per class and in the last 6-class experiment we used 40 files per class (table 3.

We performed experiments on different sole types, namely barefoot, sneakers, leather, heels and boots. These sole types were classified on two different floor types, namely on concrete and wood floor. Furthermore, we performed an experiment to classify six floor types ignoring the sole types. These floor types were snow, sand, metal, dirt, wood, concrete.

The labels of these data-sets are mostly psycho-acoustically validated. We did not perform detailed psycho-acoustical experiments, but we checked the sounds by listening to them by ourselves. We discarded sounds whose class cannot be identified while listening to them.

5 Experiments and Results

In order to evaluate the results of our experiments we used leave-one-out cross validation, where all sounds but one are incorporated into the training set. Then the trained algorithms are tested on the remaining single sample that previously had been excluded from the training set. This procedure is repeated for all possible partitions into training/test sets.

We performed multi-class experiments for footstep sounds, where we used five different sole types on concrete floor, performing five separate binary classification runs. In each run, one class is classified against the rest, which consisted of the other 4 classes. Therefore the rest-class constituted a heterogeneous mixture of sounds. We took the average of these five binary classifications. Table 1 shows not only the overall average of the separate classification experiments, but also the results of the separate binary classification experiments.

The percentages are 100% - BER (balanced errors rate), thats weights the two class errors equally.

	SVM		HMM	
	GT Hil	GT Med	GT Hil	GT Med
Barefoot	97.7%	99.7%	85.2%	94.0%
Sneakers	90.1%	83.0%	73.2%	74.6%
Leather	90.1%	92.6%	90.0%	92.0%
Heels	95.9%	97.1%	88.0%	89.4%
Boots	94.3%	92.1%	89.9%	84.6%
Average	**93.6**%	92.9%	85.3%	86.9%

Table 1: Experiment 1: Classification of the sole types on concrete floor

The same sole types hitting a wooden floor were classified in experiment 2. Following the same criteria as explained before. The results for this experiment are shown in Table 2.

	SVM		HMM	
	GT Hil	GT Med	GT Hil	GT Med
Barefoot	95.6%	95.9%	88.0%	84.3%
Sneakers	87.5%	75.9%	69.8%	67.2%
Leather	98.5%	94.8%	91.6%	89.8%
Heels	98.5%	100%	98.8%	99.4%
Boots	96.2%	92.7%	88.1%	92.7%
Average	**95.3**%	91.9%	87.3%	86.7%

Table 2: Experiment 2: Classification of the sole types on wood floor

The third experiment we classified different floor types: snow, sand, metal, dirt, wood, concrete.

	SVM		HMM	
	GT Hil	GT Med	GT Hil	GT Med
Snow	90.0%	99.5%	94.0%	89.9%
Sand	86.4%	91.0%	88.2%	92.8%
Metal	86.2%	98.8%	89.2%	92.9%
Dirt	77.7%	90.0%	77.5%	83.1%
Wood	98.3%	100.0%	89.4%	94.8%
Concrete	91.5%	98.0%	92.5%	89.3%
Average	88.4%	**96.2**%	88.5%	90.5%

Table 3: Experiment 3: Classification accuracy of the floor types

Except in one case SVM performed better than HMM. For shoe type, on average the power specturm based method (left in Figure 1) performed better than the Meddis based right method. However for the floor types the Meddis based method clearly performs better.

6 Discussion

From the theoretical view point, support vector machines try to optimise the classification boundary between two classes, by maximising the margin between the two classes. In order to do this, the method tries to find the optimal input vectors, called support vectors, which maximise the margin. On the other hand, using a multi-variate Gassian for approximation of the observation density, the hidden Markov models try to predict the state transition probabilities and the mean-variance pair for each dimension in the input parameter space. Hence, the search space of the hidden Markov models is much bigger than the support vector machines. Therefore the hidden Markov models need much more input samples to improve the prediction quality. Considering this fact, the results, which we obtained with the hidden Markov models are satisfactory.

Conclusions from these experiments for classification capabilities in a real world scenario have to be drawn with care. In this work we have listed one class against the rest balanced prediction rates[2]. For a more suggestive error measure of multi-class classification, a simple measure would be the sum of the precision: the normalized diagonal of the confusion matrix. All sounds used here stem from the same database, implying that they may have been recorded under similar conditions (same microphone, same person, same ground, one particular pair of shoes). So the classification accuracy would have to be investigated for sounds outside this data base.

7 Conclusion

We have investigated the potential of two gamma-tone based representations, and two classification methods, which have been used in combination with these representations for classifying impact like everyday sounds.

As perprocessing we used energy and spikerate from an auditory model calculated based on the output of a bank of gamma-tone filters. These two representation methods were tested with support vector machines and hidden Markov models to find out the best representation and classifier pair for this corpus. For simple classification tasks they classified the sounds perfectly.

The comparison of these experiments showed that, in the case of the step sounds, gammatone based representation methods performed well for the classification of the sounds no matter which classification method has been used. However, support vector machines performed in general better than hidden Markov models.

Further experiments are needed. For example, the hidden Markov models should be applied to sounds that have been analyzed by a filterbank with biologically plausible lower frequency that is higher than the lowest audible frequency. In addition, the preprocessing in the hidden Markov model should be consistent with the preprocessing for the support vector machine. Also it would be informative to compare the gammatone filter with MFCC coefficients with the same lowest frequency and the same number of filters. Experiments in [3] point in the direction that the Gammatone filter bank yields significantly better results than the MFCCs. Experiments should be extended to capture the classification scheme of everyday sounds introduced by Gaver [15], guided by the physical properties of the interaction between objects that produce the sound.

Acknowledgements

Thanks to the European CLOSED project (FP6-NEST-PATH "measuring the impossible" project no. 29085) for funding. We would like to thank Mathieu Pellerin, Guillaume Lemaitre, and Olivier Houix from IRCAM for their support. We thank Perfecto Herrera from MTG for helpful comments.

[2] 1−balanced error rate

References

[1] Alexander Ekimov and James M. Sabatier, "Vibration and sound signatures of human footsteps in buildings," *J. Acoust. Soc. Am.*, vol. 120, no. 2, pp. 762–68, 2006.

[2] D. Olynyk and T. D. Northwood, "Subjective judgments of footstep-noise transmission through floors," *J. Acoust. Soc. Am.*, vol. 38, pp. 1035–9, 1965.

[3] E. Martinez, K. Adiloglu, R. Annies, H. Purwins, and K. Obermayer, "Classification of everyday sounds using perceptual representation," in *Proceedings of the Conference on Interaction with Sound*. Fraunhofer Institute for Digital Media Techology IDMT, 2007, vol. II, pp. 90–95.

[4] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *International Symposium on Music Information Retrieval*, 2000.

[5] C. Spevak and R. Polfreman, "Analysing auditory representations for sound classification with self-organizing neural networks," in *Proceedings of the International Conference on Digital Audio Effects*, Verona, 2000.

[6] J. Breebaart and M. McKinney, "Features for audio classification," in *Proceedings of the Philips Symposium of Intelligent Algorithms*, Eindoven, 2002.

[7] J. Breebaart and M. McKinney, "Features for audio and music classification," in *Proceedings of the International Conference on Music Information Retrieval*, Baltimore, 2003.

[8] R.D. Patterson and J. Holdsworth, "A functional model of neural activity patterns and auditory images," *Advances in Speech, Hearing and Language Processing*, vol. 3, pp. 547–563, 1996.

[9] B.R. Glasberg and B.C.J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103–138, 1990.

[10] M. Slaney, *A matlab toolbox for auditory modeling work*, Interval Research Corporation, 1998.

[11] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, Springer-Verlag, Berlin Heidelberg NewYork London, 22 edition, 1990.

[12] R. Meddis, "Simulation of mechanical to neural transduction in the auditory receptor," *Journal of the Acoustical Society of America*, vol. 79-3, pp. 702–711, 1986.

[13] J.J. Burred and A. Lerch, "A hierarchical approach to automatic musical genre classification," in *Proceedings of the International Conference on Digital Audio Effects*, London, 2003.

[14] "Sound ideas sound database, http://www.sound-ideas.com," .

[15] W. W. Gaver, "How do we hear in the world? explorations in ecological acoustics," *Ecological Psychology*, vol. 5, pp. 285–313, 1993.