Perceptual Representations for Classification of Everyday Sounds

Elena Martinez Hernandez[1,2], Kamil Adiloglu[1], Robert Annies[1],
Hendrik Purwins[1,2], Klaus Obermayer[1]
,[1] Neural Information Processing Group, Berlin University of Technology,
{kamil,robokopp,oby}@cs.tu-berlin.de
[2] Music Technology Group, Pompeu Fabra University,
{emartinez,hpurwins}@iua.upf.edu

Abstract. In the recognition and classification of sounds, extracting perceptually and biologically relevant features yields much better results than the standard low-level methods (e.g zero-crossings, roll-off, centroid, energy, etc.). Gamma-tone filters are biologically relevant, as they simulate the motion of the basilar membrane. The representation techniques that we propose in this paper make use of the gamma-tone filters, combined with the Hilbert transform or hair cell models, to represent everyday sounds. Different combinations of these methods have been evaluated and compared in perceptual classification tasks to classify everyday sounds like doors and footsteps by using support vector classification. After calculating the features a feature integration technique is applied, in order to reduce the high dimensionality of the features. The everyday sounds are obtained from the commercial sound database "Sound Ideas". However, perceptual labels assigned by human listeners are considered rather than the labels delivered by the actual sound source. These perceptual classification tasks are performed to classify the everyday sounds according to their function, like classifying the door sounds as "opening" and "closing" doors. In this paper, among the gamma-tone-based representation techniques, other spectral and psycho-acoustical representation techniques are also evaluated. The experiments show that the gamma-tone-based representation techniques are superior for perceptual classification tasks of everyday sounds. The gamma-tone filters combined with a inner hair cell model and with the Hilbert transform yield the most accurate results in classifying everyday sounds.

1 Introduction

Unlike music and speech, the audio analysis and recognition of everyday sounds have not yet received so much attention in the literature. Nonetheless, everyday sounds play a significant role in communication, localisation and interaction. In this paper, we focus on the ability of machine learning algorithms to identify and classify sounds by example to gain insight what aspects of – mostly complex – sounds lead to a certain categorisation by humans.

Everyday sounds in urban environments are emitted from machines, human interaction with mechanical devices, and natural phenoma. In a natural environment, sounds get modulated by the acoustical properties of the environment itself and get mixed with different sources of sounds. This shows the complexity of everyday sounds. However, for analysis purpose, it is necessary to reduce the complexity of the representation to meaningful features.

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. Several classification studies have been made to distinguish different kinds of everyday sounds from other types of sounds such as music, speech, etc. However, in our approach we classify only everyday sounds based on the function they fulfil, their material or shape, or the objects they interact with. Obviously, this task, classifying everyday sounds is much more difficult then distinguishing everyday sounds from other sound classes, which have totally different characteristics.

We would like to thank Mathieu Pellerin, Guillaume Lemaitre, and Olivier Houix from IRCAM for their support.

This work is supported by the European CLOSED project (FP6-NEST-PATH "measuring the impossible" project no. 29085).

2 Representation

Low level spectral features, like zero-crossings, centroid, roll-off etc. of a given sound give useful information about the sound. However this information is not sufficient to understand sound perception. In order to be able to understand how perception works, psychoacoustical facts should be considered, and psychoa-

coustical features should be utilised to define efficient representation schemes.

Mel Frequency Cepstrum Coefficients (MFCC's) [7] are well established representation scheme, which dominate applications in speech recognition and music processing. They have rarely been applied to environmental sounds. MFCC's are short-term spectral based features, based on the Mel scale. The Mel scale is a mapping between the actual frequency values and the perceived pitch. This mapping is linear for the low-frequency values. As the frequency increases, the mapping becomes logarithmic. Generally the first 13 MFCC features are used as descriptors to represent a given sound in different tasks, because as it has already been shown, these first coefficients concentrate most of the signal energy [17].

However, there are other psychoacoustically or biologically motivated methods, which take the critical bands into account, where inputs whose frequency difference is smaller than the critical bandwidth cause the so-called beats. Another methods aim to simulate the cochlea in the inner-ear. In our representation method, we introduce the gamma-tone filters, which are considered to be biologically motivated as well.

2.1 Gamma-tone Filterbank

A gamma-tone auditory filterbank [11, 12] incorporates two insights into auditory physiology: 1) the higher frequency resolution for low frequencies, 2) the higher temporal resolution for high frequencies. With increasing centre frequency the spacing of the gamma-tone filters increases and the length of the filter decreases. Mimicking the basilar membrane, spacing and bandwidth of the filter is based on the equivalent rectangular bandwidth (ERB) [4]. Roughly, the centre frequencies are spaced linearly for low frequencies and logarithmically for high frequencies. ERBs are similar to the Bark or the Mel scale. Due to its properties, the ERB scale is a highly biologically plausible representation.

As a pre-processing of the everyday sounds, we use the gamma-tone filter implementation in Malcolm Slaney's Auditory Toolbox [13, 14]. Starting with the lowest centre frequency of 3 Hz, we use 18 gamma-tone filters in total. Therefore, for each given sound, we obtain 18 filter responses from the gamma-tone filter bank.

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

2.2 Hilbert Transform

The first method, which we can combine the gamma-tone filters with is the Hilbert Transform. The Hilbert Transform is mainly used to explain the relationships between the real and imaginary parts of a signal. The Hilbert transform of a signal is nothing but 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.

In our representation scheme, the Hilbert Transform is applied to each gamma-tone filter response. After applying the gamma-tone filterbank the Hilbert transform is calculated for each filter response. Then the power spectral density is calculated for each of the Hilbert transformed filter responses [15]. The power spectral density is the Fourier Transform of the autocorrelation of the signal. The periodogram method is generally used to calculate the power spectral density. In the standard case, the power per unit frequency is calculated, where the results have the unit $\frac{power}{frequency}$. However, we calculate the mean-squared spectrum. The mean-squared spectrum is calculated for each frequency value depending on the sampling rate. Therefore, the unit of the mean-squared spectrum values are *power*. Figure 1 shows an example mean-squared spectrum gamma-tone filter. In this figure, a closing door sound has been taken. After applying gamma-tone filters onto the sound, we obtained 18 filter responses, one for each filter in the filterbank. Then a single filter response was taken, in order to apply the Hilbert Transform onto the filter response. As the last step, the mean-squared spectrum was calculated for the Hilbert transformed filter response.

After these steps, we obtain the mean-squared spectrum for each Hilbert transformed filter response which can be considered as a matrix. However, we should reduce the dimensionality of this matrix, in order to be able to use them as a representation of the sound. Therefore we sum up these values within each of four groups of adjacent centre frequencies [1]. We take the average of the calculated values for each group. These groups are the DC values, the frequency interval 3-15 Hz, 20-150 Hz, and the rest. The interval 3-15 Hz emphasises the speech syllabic rates. The interval 20-150 Hz is the perceptual roughness [18]. After this step, we obtain four values for each filter response.

2.3 Inner Hair Cell Model

Another method, which can be combined with the gamma-tone filters is the inner hair cell model of Med-

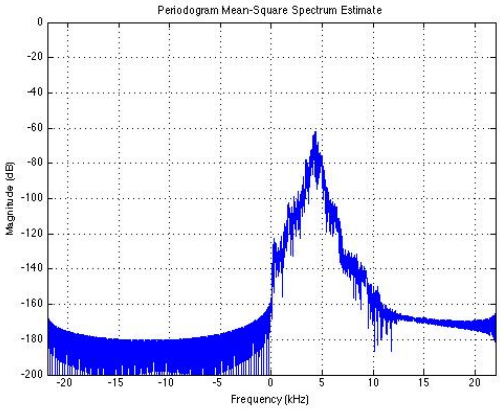

Figure 1: A single gamma-tone filter output transformed by Hilbert transform. The mean-squared spectrum of the signal is shown.

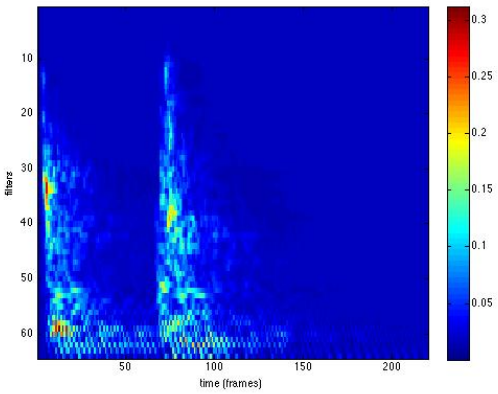

Figure 2: An example representation of gamma-tone filters combined with the inner hair cell model is shown. The sound of a closing door is analysed.

dis [8] [9] [10]. In this model, the firing rate of the inner hair cells, connected to the basilar membrane, is modelled. Hence, the gamma-tone filters and the inner hair cell model complement each other [16]. The inner hair cells fire, when a stimulus arrives and the basilar membrane is deflected at a 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. Figure 2 shows a closing door sound represented by gamma-tone filters combined with the inner hair cell model.

3 Feature Integration for SVM

3.1 Mean, Variance, Derivatives

The gamma-tone filter bank yields a multi-channel response for a single sound. Besides, the filter bank does not decrease the length of a given sound. Summing up the filter outputs reduces the dimensionality of the representation from 18 to 4 bands. The details of this method are given in the corresponding representation section. However, we compress the outputs of the bands across the entire length of the sound into a single representation vector. In order to integrate the responses of individual filters, we take the mean and variance of each of the values in the four bands. In order to be able to track the change in those values between different filters, we also calculate the mean and the variance of the first derivative of these filter responses [2]. After these calculations we obtain a 16 dimensional feature vectors for each sound.

3.2 SVM Settings

In these experiments, we used support vector machines [5], which are known to be one of the best classification tools, based on the maximisation of the margin of the classification boundary between two classes, c-SVM, which is implemented in the libsvm software library [3]. The c-SVM has two parameters (c and g), which should be determined beforehand. In order to find the optimal parameter 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 $c = 2^8, 2^9, 2^{10}, \ldots, 2^{12}, g = 2^{-16}, 2^{-15}, 2^{-14}, \ldots, 2^{-4}$. For the final experiment in Table 5 we use a wider step size for the g parameter: $g = 2^{-16}, 2^{-14}, 2^{-12}, \ldots, 2^{-4}$ In the results, we will present the best results, which have been obtained by the optimal values of these two parameters.

4 Data Sets

Recordings are taken from the sound collection "Sound Ideas" [6]. For the experiments, recordings of footsteps and opening/closing doors are selected. The door sounds have a complex temporal structure whereas the steps are short and only consist of a few onsets. In both cases a mixture of temporal patterns and spectral properties can reveal information about material

or function of the sound.

The doors dataset includes a series of recordings of wooden doors being closed or opened. A single sample contains either an opening or a closing door sound, not both one after another. 74 closing and 38 opening doors in total are used as input for the experiments.

The footstep dataset consists of recordings of different kind of shoes (heels, boots, barefoot) on various grounds (concrete, gravel, wood, dirt). They are taken from CD 16 and 17 from [6]. The recordings are cut, when necessary, such that exactly one step in contained in one sample. The limit of steps taken from the same recording is 10 to avoid overfitting effects. This leads to a dataset with 125 footsteps per class.

The datasets that we use in our experiments are listed out in the following:

- doors
  - opening - closing
- footsteps
  - high heels - boots
  - high heels - non high heels
  - barefoot - sneakers
  - on wood - on dirt

Besides performing binary classification experiments on these data sets, we evaluate our representation schemes on a multi-class classification experiment as well. In this experiment, we use five different types of footstep sounds (barefoot, sneakers, leather, heels, boots), all on concrete or marble.

The labels of these datasets are all psycoacoustically validated. We did not perform detailed psycoacoustical experiments, but we checked the sounds by listening them by ourselves. We discard sounds whose class cannot be identified while listening to them.

5 Experiments and Results

In order to evaluate the results of our experiments in a reliable way we use cross validation to calculate the average accuracy. In particular, we use the leave-one-out cross validation method. In leave-one-out cross validation experiments, all sounds but one are put into the training set. Then the trained algorithm is tested on the remaining single sample, that had been previously excluded from the training set. This procedure is repeated for all possible partitions into training/test set. The total accuracy of the system is the average accuracy of all tests.

The experiments that we perform are all binary classification experiments. The door sounds are classified as opening and closing sounds. Table 1 shows the results of SVM classification using four different representations: 1) gamma-tone filters combined with the Hilbert transform (GT Hil), 2) gamma-tone filters combined with the inner hair cell model (GT Med), 3) MFCC's, and 4) MFCC's combined with the low-level spectral features (SLL), which contain zero-crossings, roll-off and centroid.

GT Hil	GT Med	MFCC	SLL
84.0%	63.2%	60.4%	67.1%

Table 1: Classification of opening/closing door sounds by SVMs using various representations (cf. text).

The footstep sounds were classified based on the sole types (high-heels vs. boots, high-heels vs. not high-heels), and the floor (concrete vs. gravel, wood vs. dirt). Furthermore, we also designed a multi-class experiment based on the sole types as well. In the latter, we use five different sole type classes (barefoot, sneakers, leathers, heels, combat boots) on concrete floor.

In Table 2 the results of the binary classification experiment heels vs. combat boots are shown. All representation methods work almost perfect for this classification task. Only the MFCC's do not yield 100% accuracy for the task heels vs. combat boots, but 98.3% accuracy is still good. Based on these results, this task can be considered as the conceptual proof that our representation methods work at least as well as traditional methods. In order to make this task a little bit more complicated, so that we can observe differences between the accuracy rates of different representations, we perform another experiments, where we classify heels vs. non-heels. The results of this second experiment are shown in Table 3. Here, the gamma-tone based methods outperform the MFCC based methods.

GT Hil	GT Med	MFCC	SLL
100.0%	**100.0**%	98.3%	**100.0**%

Table 2: Classification results of the footstep sounds heels vs. boots are shown.

GT Hil	GT Med	MFCC	SLL
91.6%	**100.0**%	80.9%	89.2%

Table 3: Classification results of the footstep sounds heels vs. not heels are shown.

We also perform binary classification experiments by

using different sole types on two different floor types, namely wood floor vs. dirt, leaves, sand and gravel. The second class of sounds contain four different floor types. However, all these floor types are perceived almost equal to the human listener. Therefore, creating a single class out of these sounds does not any harm to the experiment from a psychoacoustical point of view. Table 4 shows the classification results of this experiment. Again the accuracy is 100% for all different representation methods. Hence, these results prove that the representation methods we propose work well for general binary classification tasks.

GT Hil	GT Med	MFCC	SLL
100.0%	**100.0**%	**100.0**%	**100.0**%

Table 4: Classification results of the footstep sounds wood floor vs. dirt are shown.

Beside these binary classification tasks, we also perform a multi-class experiment. In this experiment, we use five different sole types, and classified each of those classes against the others. Each class consists of 40 instances. We perform five separate binary classification runs. In each, one class is classified against the rest, the other four classes. As a result, we obtain five binary classification results. In order to calculate the overall accuracy of the experiment, we take the average of these five separate classifications. Table 5 shows not only the overall average of the separate classification experiments, but also the results of the separate binary classification experiments themselves. In contrast to Table 3, the classification heels vs. non-heels in Table 5 uses a different set of samples. In the latter only steps on concrete are used.

	GT Hil	GT Med	MFCC	SLL
Barefoot	94.4%	**100.0**%	94.4%	94.4%
Sneakers	**94.4**%	83.3%	76.6%	82.1%
Leather	86.3%	**100.0**%	**100.0**%	**100.0**%
Heels	85.7%	92.8%	98.3%	**100.0**%
Boots	85.0%	**85.7**%	78.7%	79.3%
Average	92.2%	**94.2**%	88.8%	89.3%

Table 5: Classification accuracy of sole types in step sounds. Altogether, the gamma-tone based representations outperform the MFCC based methods.

6 Conclusion

We have investigated the potential of four representations for classifying everyday sounds, in particular opening/closing doors and the sole and floor material in footsteps. In each experiment, the best method performs with at least 84% accuracy, in several instances even with 100%.

In contrast to low-level descriptors (such as zero-crossing rate, spectral centroid, roll-off), sound can be pre-processed by models physiologically inspired by the basilar membrane. MFCC's and gamma-tone filter banks are prominent examples, both modelling the decreasing frequency resolution for high frequencies. In addition, gamma-tone filterbanks also model the low temporal resolution for low frequency signals and their impulse response closely resembles physiological measurements in the basilar membrane.

The experiments show that in general more physiologically relevant models, gamma-tone based representations, outperform other pre-processing methods. We combined the gamma-tone filters with the Hilbert transform and with Meddis' inner hair cell model. In order to reduce the dimensionality firstly we summarised the filter responses in four different frequency bands, and then applied a feature integration method to these representations. In the end, we obtained feature vectors, which can be used to perform classification experiments with support vector machines. Beside these two representation schemes, we performed classification experiments with MFCC's, and MFCC's combined with low-level spectral features. For the simple classification tasks they classified the sounds perfectly. These simple tasks are considered to be the proof of concept in general for these representation schemes.

However, we performed more complicated experiments, in order to observe the classification accuracy of the gamma-tone based representation compared to the MFCC based methods. Comparison of these experiments showed that, in general, gamma-tone based representation methods outperformed MFCC based representation methods. Although there are several special cases, where the MFCC's performed better than the gamma-tone based methods, gamma-tone representations yielded better results.

Interestingly, on the most complex data set, the door sounds, the gamma-tone / Hilbert transform method performed significantly better then the other methods (17% better than the second best method). On the other hand, the inner hair cell combination yielded slightly better results for the footstep sounds than the Hilbert transform.

The representations used here are basically stationary and include only little information about the temporal dynamics of the sounds. The MFCC's used here include variance, discrete first derivation, and the variance of the latter. Thereby, MFCC's momentarily capture some temporal behaviour. The gamma-tone

filterbank gives better temporal resolution for higher frequencies than for lower frequencies. The Meddis hair cell model essentially emphasises on- and offsets. This feature may be the reason why a gammatone filterbank with subsequent Meddis hair cell model sometimes performs better than the Hilbert transform. Combining the Hilbert transform based representation with delta coefficients like the MFCC's or an onset detection feature like the Meddis hair cell could improve this representation. For a more complex consideration of the time course of the events a higher level analysis would be useful. Promising approaches include dynamic time warping, hidden Markov models, some sort of analysis of the rhythm (regularity, acceleration, deceleration) generated by the onsets of the signal.

References

[1] J. Breebaart and M. McKinney. Features for audio and music classification. In *International Conference on Music Information Retrieval*, 2003.

[2] J.J. Burred and A. Lerch. A hierarchical approach to automatic musical genre classification. In *International Conference on Digital Audio Effects*, 2003.

[3] C.C. Chang C.W. Su and C.J. Lin. *A Practical Guide to Support Vector Classification*. Department of Computer Science, National Taiwan University, 2007.

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

[5] S. Haykin. *Neural Networks*. Prentice Hall, London, 2 edition, 1999.

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

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

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

[9] R. Meddis. Simulation of auditory-neural transduction: Further studies. *Journal of the Acoustical Society of America*, 83-3:1056–1063, 1988.

[10] P.B. Ostergaard. Implementation details of a computation model of the inner hair-cell/auditory-nerve synapse. *Journal of the Acoustical Society of America*, 87-4:1813–1816, 1990.

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

[12] K. Robenson R.D. Patterson and J. Holdsworth. Complex sounds and auditory images. *In Proceesings of Auditory Physiology and Perception*, 9:429–446, 2004.

[13] M. Slaney. *An Eifficient Implementation of the Patterson-Holdsworth Auditory Filter Bank*. Apple Computer, 1993.

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

[15] J. O. Smith. *Spectral Audio Signal Processing*. W3K, Stanford, march 2006 draft edition, 2006.

[16] C. Spevak and R. Polfreman. Analysing auditory representations for sound classification with self-organizing neural networks. In *International Conference on Digital Audio Effects*, 2000.

[17] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transaction on speech and audio processing*, 10-5, 2002.

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