Context-based sound event recognition
Niessen, Maria Elisabeth

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2010

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
APPENDIX A: GLOSSARY

audio component
Single element in the audio signal, which is (possibly together with other components) the result of a sound event. (p. 14)

audio pattern
Configuration of grouped audio components that constitute a single sound event. (p. 4)

audio signal
Signal transmitted by a physical action recorded or heard at the location of a receiver. (p. 2)

auditory episode
Cognitive representation of a sonic environment. (p. 28)

auditory event
Cognitive representation of a sound event. (p. 28)

context
The learned associations of an event to environments and/or co-occurring events. (p. 26)

context-based
When an algorithm or model processes knowledge of the context (cf. top-down). (p. 46)

physical action
Action or process of a source that results in a sound event. (p. 28)
segregate
Select and group components in the audio signal that are likely to constitute a single sound event. (p. 4)

semantic
When an interpretation or representation is based on the experienced properties of something and its relation to other things. It is similar to the meaning that can be attributed by people. However, it can be given by systems as well, whereas meaning can only be attributed by people. (p. 26)

signal-driven
When an algorithm or model processes the (audio) signal (cf. bottom-up). (p. 8)

sound
Sound that is not attributed to a specific event. (p. 2)

sound event
Sound that is the result of a single physical action. (p. 7)

sound source
Source involved in the physical action that produces a sound event. (p. 7)
APPENDIX B: AUDIO PROCESSING

The audio processing used in this thesis, which is developed and described by Krijnders (2010), is shortly summarized in this appendix, based on the description in Krijnders et al. (2010). The first step consists of the conversion of the audio signal in a time-frequency representation, the cochleogram. In the cochleogram, tones and pulses are selected based on their local properties, which can be connected in time or frequency to form signal components. Tonal signal components can be combined into harmonic complexes when they comply with certain properties. Finally, broadband events are defined and extracted as a time delimited energy increase compared to the background noise.

B.1 COCHLEOGRAM

The audio time signal is processed using a gammachirp filter bank (Irino and Patterson, 1997). The response of each gammatone filter $g_t$ is calculated as

$$g_t(t) = at^{N-1}\exp(-2\pi b B(f_c) t)\exp(j(2\pi f_c t + c \log_{10} t))$$

(B.1)

where $f_c$ is the center frequency of the channel, $N$ the order of the gammatone ($N = 4$) and constants $a = 1$, $b = 0.71$, and $c = -3.7$. A logarithmic frequency distribution is used for 100 channels between 67 and 4000 Hertz (Hz). The bandwidth of each filter is given by (Moore and Glasberg, 1996):

$$B(f_c) = 24.7 + 0.108 f_c.$$  

(B.2)

The filter output is squared and leaky-integrated with a segment-dependent time constant ($\tau_s = 2/f_c$). The resulting energy representation is down-sampled to 200 Hz, resulting in a frame size of 5 milliseconds. The energy is compressed logarithmically and expressed in decibels (dB). We call this representation a cochleogram (see for example Figure 2.1).
B.2 TONE FIT AND PULSE FIT

To extract tones and pulses from the cochleogram, we apply channel dependent matched filters that respond to ideal tones and pulses. For each channel an ideal tone is generated and processed using the filter bank. In Figure B.1, panel (a) shows the time-frequency representation of the tone, and panel (b) the energy of the tone in one time frame. The width of the response in frequency at a threshold below the energy maximum is calculated, depicted in panel (c). This threshold is set to twice the standard deviation of the logarithmic energy of white noise in the channel ($2\sigma_n$). This standard deviation is independent of the power spectral density of the noise in the logarithmic energy domain. The width of the response is the filter parameter for the tone fit (TF). For the pulse fit (PF) the response width in time of a pulse is taken.

When the filters are applied in a cochleogram, the energies at the widths below ($sb_1$) and above ($sb_2$) a time-frequency point are averaged. The difference between the energy at the point for that channel and the average forms the filter output (Figure B.2). The application of the filters to the cochleogram results in two representations that reflect to what extend the direct environment of each point of the cochleogram resembles a tone or a pulse. These representations are thresholded to create a binary mask. This threshold is set to twice the standard deviation of
the TF or PF when applied to white noise. Areas that are too small to be either valid tones or valid pulses are discarded. This pruning, in combination with the mask threshold, limits the number of spurious areas that are caused by broadband signals, while allowing tonal or pulse-like signals. Within the remaining areas the energy maxima of the cochleogram are strung together horizontally to form tonal, or vertically to form pulse-like signal components (see Figure 5.2).

If possible, the tonal signal components are combined into harmonic complexes (HCs). Harmonic complex formation starts by selecting concurrent signal components that have a harmonic relation. These hypotheses generate new hypotheses at fundamental frequencies in the range between 300 and 1200 Hz by shifting harmonic positions of the signal component. These hypotheses are extended with more and more signal components. The process ends by selecting the hypotheses that comply best to a well-formed HC by maximizing score $S$:

$$S = n_{sc} + b_{f_0} + n_h - \sum_{sc} \text{rms}_{sc} - \sum_{sc} \Delta f_{sc}$$

(B.3)

where $n_{sc}$ is the number of signal components in the group, $b_{f_0}$ is a boolean for the existence of a signal component at the fundamental frequency, $n_h$ is the number of sequential harmonics in the group, $\text{rms}_{sc}$ are the root mean square values of the difference of a signal component and the fundamental frequency after the mean frequency difference is removed, and $\Delta f_{sc}$ is the mean difference between
the fundamental frequency and the frequency of the signal component divided by its harmonic number.

For each harmonic complex we calculate nine features: The duration, score $S$ (equation B.3), the ratio of these two, and the number of signal components indicate the strength of a harmonic complex. The mean energy and standard deviation under the signal components, the spectral tilt of the signal components, and the mean and standard deviation of the fundamental frequency are copied from Van Hengel and Andringa (2007) and Zajdel et al. (2007) to discriminate between similar harmonic sounds, such as speech and laughing.

### B.3 Broadband events

Broadband events are defined as slow broadband changes in the signal that have to satisfy the following criteria: The change in signal must last at least 2 seconds, and 30% of the frequency channels must be more than 6 dB above the long-term background. The long-term background is calculated per channel as the energy value that is exceeded more than 95% of a time interval. This level of 95% assumes that each channel is dominated by background noise at least 5% of the time interval with a temporal scope depending on the data set—the experiments described in chapter 5 use the length of the recordings, typically between one and three minutes. The energy must exceed the background by three standard deviations of white noise in that channel.

The broadband events are described with a feature vector of 20 features. The first 15 features are three properties calculated in five frequency bands. Every frequency band contains 20 channels. The 5 remaining features are the first five cepstral coefficients that describe the spectral envelope. The three properties for the five bands are only computed for the 10% most energetic time frames per event. The first property is the correlation between points in time separated by half a second. This correlation is typically high for slowly changing events and low for fast changing events. The second property is the distance between the frequency band and the average energy, in terms of standard deviations of white noise. This property is level independent and reflects the energy distribution over the bands. The third property is the average foreground-to-background ratio for each band, which reflects the total energy per band compared to the background.
APPENDIX C: REVERBERATION FEATURES

The seven features that indicate the reverberation level measure one of three types of fluctuation: variation in the energy of the harmonic tracks, variation in the salience of the harmonic tracks, and variation in the frequency of the harmonic tracks. The calculation of the features is described in the following sections. We apply the following notation: $h$ is the harmonic track, extracted from a cochleogram (see appendix B), $E_h(t)$ is energy development of the harmonic track in decibels (dB), and $f_h(t)$ is the frequency development of the harmonic track in Hertz (Hz).

C.1 ENERGY VARIATION

**Peak rate (PR)** The number of peaks (energy maxima) of at least 1 dB in the energy of the harmonic track, $E_h$, normalized for the length of the track.

**Var $E_h$** Variation difference between the energy of the harmonic track and its moving average (normalized for the length of the track), calculated as

$$\text{Var } E_h = \sum_t \left| \frac{dt}{dt} E_h(t) \right| - \sum_t \left| \frac{dt}{dt} E_{\text{avg}}(t) \right|,$$

where $dt E_h(t)$ is the differential of the harmonic track energy $E_h(t)$, and $E_{\text{avg}}(t)$ is its moving average:

$$E_{\text{avg}}(t) = \frac{1}{k} \sum_{i=t}^{t+k-1} E_h(i), \text{ for } t = 1 : n - k + 1,$$

where $E_h$ has a total length of $n$ frames (1 frame is 5 milliseconds), and the applied window size is $k = 7$ frames.

C.2 HARMONIC ENERGY SALIENCE

**$\Delta E_h(f)$** The energy slope is calculated as the mean difference in energy between a harmonic track and a reference track $E_{\text{ref}}$ at a higher frequency and at a lower
frequency side of the harmonic track:

$$\Delta E_h(f) = \overline{E_h} - E_{\text{ref}},$$

where $E_{\text{ref}}$ is determined by the difference between the fundamental frequency $f_0$ and a track at $k = 1.2$ or $0.8$ of $f_0$ (the fundamental frequency is determined through the harmonic complex, see appendix B):

$$E_{\text{ref}}(f, t) = E(f_h(t) \pm |k f_0(t) - f_0(t)|) \cdot t, \text{ for } t = 1 : n,$$

where $n$ is the length of the harmonic track and the fundamental frequency.

**Var $\Delta f_h$** Variation difference between the frequency width of the harmonic track and an ideal sinusoid (normalized for the length of the track), calculated as

$$\text{Var } \Delta f_h = \sum_t \left| dt \Delta f_h(t) \right| - \sum_t \left| dt \Delta f_s(t) \right|,$$

where $dt \Delta f_h(t)$ is the differential of the width of an ideal sinusoid at the same frequency as the harmonic track at a particular time, and $\Delta f_h(t)$ is the width in frequency channels of the harmonic track at time $t$, determined by the tone fit (TF), a frequency-dependent filter based on the ideal tone response of the cochlea model:

$$\Delta f_h(t) = \frac{\theta_{\text{TF}} - \text{TF}(s_{\text{down}} + 1, t)}{\text{TF}(s_{\text{down}} + 1, t) - \text{TF}(s_{\text{down}}, t)} - \frac{\theta_{\text{TF}} - \text{TF}(s_{\text{up}} - 1, t)}{\text{TF}(s_{\text{up}}, t) - \text{TF}(s_{\text{up}} - 1, t)},$$

where $\theta_{\text{TF}} = 1.3$ is the threshold of the tone fit, $\text{TF}(s_{\text{down}}, t)$ is the response of the TF filter at time $t$ in channel $s_{\text{down}}$, the channel at the high frequency end of the harmonic region, and $s_{\text{up}}$ is the channel at the low frequency end of the harmonic region. This harmonic region is determined through a tone mask, a binary mask created by thresholding the TF filter response of the cochleogram at twice the standard deviation of the TF filter when applied to white noise (see appendix B).

---

1 The channel numbers start at the high frequency side of the cochleogram and end at the low frequency side.
**Mean $\Delta f_h$**  Mean difference between the frequency width of the harmonic track and an ideal sinusoid, calculated as

$$\text{Mean } \Delta f_h = \overline{(\Delta f_s - \Delta f_h)},$$

where $\Delta f_s(t)$ and $\Delta f_h(t)$ are calculated as above.

### C.3 Harmonic Frequency Salience

**Var $f_h/\text{MA}$**  Variation difference between the frequency of the harmonic track and its moving average (normalized for the length of the track), calculated as

$$\text{Var } f_h = \sum_t \left| dt f_h(t) \right| - \sum_t \left| dt \overline{f}_h(t) \right|,$$

where $dt f_h(t)$ is the differential of the harmonic track frequency $f_h(t)$, and $\overline{f}_h(t)$ is its the moving average:

$$\overline{f}_h(t) = \frac{1}{k} \sum_{i=t}^{t+k-1} f_h(i), \text{ for } t = 1 : n - k + 1,$$

where $f_h$ has a total length of $n$ frames (1 frame is 5 milliseconds), and the applied window size is $k = 7$ frames.

**Var $f_h/P$**  Variation difference between the frequency of the harmonic track and its polynomial approximation (normalized for the length of the track), calculated as

$$\text{Var } f_h = \sum_t \left| dt f_h(t) \right| - \sum_t \left| dt P(f_h(t)) \right|,$$

where $dt f_h(t)$ is the differential of the harmonic track frequency $f_h(t)$, and $P(f_h(t))$ is its *piecewise cubic Hermite interpolating polynomial* (Fritsch and Carlson, 1980), calculated with five equally spaced time-frequency points on the track, that is, the start, the end, and three points in between.\(^1\)

---

\(^1\) This function is provided as *pchip* in Matlab (© 1984-2004 The MathWorks, Inc.).
C.4 Predictive strength of features

A two-layer feed-forward backpropagation neural network (NN) was trained with the features on the data set described in section 2.4.2, of which 2595 samples were used for training, and 1297 samples for validating and testing, the results of which are depicted in the following figure: