Audio Classification and Localization for Incongruent Event Detection

Jörg-Hendrik Bach, Hendrik Kayser, and Jörn Anemüller

Carl von Ossietzky Universität Oldenburg,
Medical Physics,
26111 Oldenburg, Germany
{j.bach,hendrik.kayser,joern.anemueller}@uni-oldenburg.de
http://medi.uni-oldenburg.de

Abstract. A method is presented that detects unexpected acoustic events, i.e., occurrence of acoustic objects that do not belong to any of the learned classes but nevertheless appear to constitute meaningful acoustic events. Building on the framework of Weinshall et al. [7], general and specific acoustic classifiers are implemented and combined for detection of events in which they respond in an incongruous way, indicating an unexpected event. Subsequent identification of events is performed by estimation of source direction, for which a novel classification-based approach is outlined. Performance, evaluated in dependence of signal-to-noise ratio (SNR) and type of unexpected event, indicates decent performance at SNRs better than 5 dB.

Keywords: theory of incongruence, auditory scene analysis

1 Introduction

Human detection of acoustic objects and events is remarkably efficient even when sounds of interest occur within an uninteresting floor of background activity. Further, we are also very good at detecting acoustic sounds even when they belong to a class we have not encountered before. In such a situation, two questions need to be answered simultaneously: Does the present acoustic scene contain a significant event at all? And if so: Does the event belong to any of the event classes know from previous experience? If the latter is not the case, the event is termed "unexpected", and commonly subsequent actions for its further identification and description ensue, such as localizing the direction in which the event occurred, increasing attention towards this direction, etc.

The present contribution outlines a system for performing the unexpected event detection task in the auditory modality. It builds on the general framework based on levels of ("general" and "specific") classifiers and detection of incongruences thereof, put forth in [7]. Implementation of such a system necessitates the specification of several processing modules, each of which is presented in the present work: Signal pre-processing is outlined in sections 2.1 and 2.2, yielding salient feature sets for classification and localization of acoustic events.

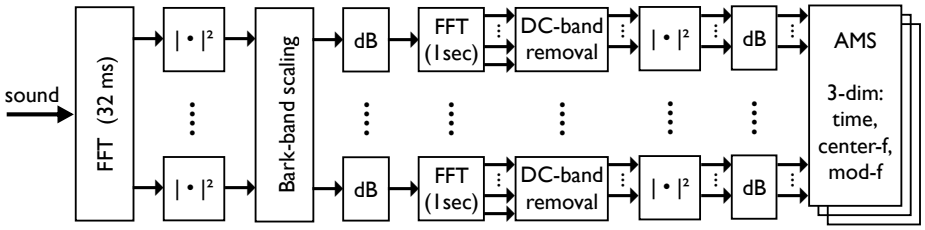


Fig. 1. AMS feature extraction.

The incongruous event detection framework is summarized in section 2.3 and the resulting audio classifier architecture presented. Section 3 highlights results of detection and localization of events.

2 Methods

2.1 Object Detection

Feature Extraction The features used in this task are Amplitude Modulation Spectrograms (AMS), which are motivated by the importance of temporal modulations for acoustic object recognition. This importance has been confirmed in numerous psychophysical, physiological and applied studies (e.g. [6]). AMS represent a decomposition of the signal along the dimensions of acoustic frequency, modulation frequency and time, and are computed by a (modulation) spectral decomposition of sub-band spectral power time-courses in overlapping temporal windows. The processing stages of the AMS computation are as follows (see Figure 1). The signal decomposition with respect to acoustic frequency is computed by a short-term fast Fourier transformation (FFT with 32 ms Hann window, 4 ms shift, FFT length 256 samples, sampling rate 8 kHz.), followed by squared magnitude computation, summation into rectangular, non-overlapping Bark bands and logarithmic amplitude compression. Within each spectral band, the modulation spectrum is obtained by applying another FFT (1000 ms Hann window, 500 ms shift, FFT length 250 samples) to the temporal trajectories of subband log-energy. Finally, an envelope extraction and a further logarithmic compression are applied. The resolution of 17 Bark acoustic frequency bands and 29 modulation frequency bands (2 Hz to 30 Hz) results in a signal representation with 493 feature-values per time step, at a temporal resolution of 1000 ms. Robustness of the AMS features with respect to constant and slowly-varying time-domain convolutive factors (channel noise) is enhanced by construction of the feature extraction scheme: The first Fourier transform step approximately converts the convolution to a multiplicative term in each (acoustic) frequency band. The subsequent logscale compression converts it into an additive term that is captured in the DC (0 Hz) modulation spectral band after the second FFT (depending on parameter choice, additional low modulation-frequency may be influenced by smearing of DC components). Deliberately discarding the affected

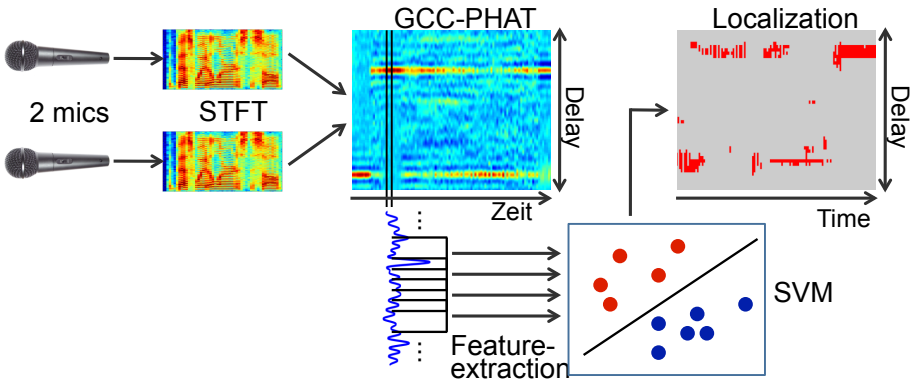


Fig. 2. Schematic of the classification approach for sound source localization.

modulation bands therefore results in AMS features that are approximately invariant to a time-domain signal convolution with short impulse responses such as microphone transfer functions and early reverberation effects.

Classification The object detection task is mapped onto a discrimination task as follows: instead of using using a generative approach to model the presence/absence of a target object, a number of known objects is classified against one another using a 1-vs-all discriminative approach. Support Vector Machines [3] are employed as a back-end. For each object, a binary classifier is trained to discriminate between "target object in background" and "different object in background".

2.2 Localization

The classification approach for sound source localization (sketched in Figure 2) employs features extracted from the generalized cross correlation (GCC) function [5] between two audio input signals. A sound source is localized by estimating the azimuthal angle of the direction of arrival (DOA) relative to the sensor plane defined by the two front-microphones of the AWEAR platform which is an mobile audio-visual recording device developed within the DIRAC project.
Support-vector-machines (SVM) [3] are used to classifiy the presence or absence of a source at each angle. This approach enables the simultaneous localization of more than one sound source in each time-frame.

The GCC is an extension of the cross power spectral density function, which is given by the Fourier transform of the cross correlation. Given two signals $x_1(n)$ and $x_2(n)$, it is defined as:

$$G(n) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H_1(\omega) H_2^*(\omega) \cdot X_1(\omega) X_2^*(\omega) e^{j\omega n} \, \mathrm{d}\omega, \tag{1}$$

where $X_1(\omega)$ and $X_2(\omega)$ are the Fourier transforms of the respective signals and the term $H_1(\omega)H_2^*(\omega)$ denotes a general frequency weighting.

In the present work PHAse Transform (PHAT) weighting [5] has been used, which normalizes the amplitudes of the input signals to unity in each frequency band, $H_1(\omega)H_2^*(\omega) = \frac{1}{|X_1(\omega)X_2^*(\omega)|}$:

$$G_{PHAT}(n) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{X_1(\omega)X_2^*(\omega)}{|X_1(\omega)X_2^*(\omega)|} e^{j\omega n}\,\mathrm{d}\omega, \tag{2}$$

such that only the phase difference between the input signals is preserved.

The inter-microphone distance in the AWEAR setup (45 cm) corresponds to a maximum delay of 1.32 ms ($\widehat{=} \pm 90°$) in each direction. The window length and the length of the Fourier transform used to compute the GCC in the spectral domain are chosen such that a 257-dimensional GCC-PHAT vector is produced for every time frame of 5.33 ms length.

From the GCC data, angular features are extracted for classification in the following way: A rectangular window is slid over the GCC-vector, subdividing the data of a single time frame into 61 feature-vectors. This covers the field of view homogeneously in terms of time delay with a resolution of $\frac{1}{3}$ ms. The mapping from the time delay τ to the angle of incidence θ is non-linear: $\theta = \arcsin(\tau \cdot \frac{c}{d})$, where c denotes the speed of sound and d the distance between the sensors. This results in a non-homogenous angular resolution in the DOA-angle-space, with higher resolution near the centre and lower resolution towards the edges of the field of view.

2.3 Incongruous Event Detection

Framework We use the concept of incongruence proposed in [7] to detect unexpected events in acoustic scenes. In a nutshell, an incongruence is defined as a disagreement between models of different levels of abstraction. An event is considered incongruous if the classifier for the more abstract (general) concept is confident about its classification, whereas the less abstract (specific) classifier is not. This allows the detection of events that are unknown to the combination of classifiers while not necessarily novel to each individual classifier.

Implementation for Audio Signals In previous work, we focussed on detecting known acoustic objects in unknown background sounds [2]. Here we aim at detecting unknown ("incongruent") objects in known background sounds using general and specific object classifiers.

The general object detector is based on a RASTA-PLP feature extraction [4]. The PLPs were extracted using 17 Bark bands, 25 ms windows with 10 ms shift and the original log-RASTA filter. For each time frame, a positive vote for the presence of an object ("blob") is collected if the sum over all coefficients is

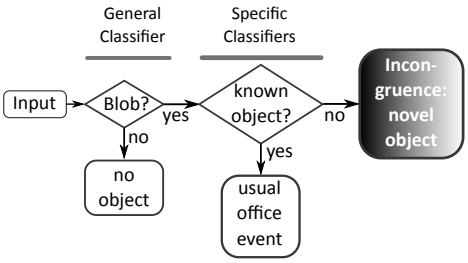


Fig. 3. Flow chart of the novelty detection based on incongruence between models for general and specific acoustic objects.

above a certain threshold θ. If at least 5% of the votes within a 1 s window are positive, that window is classified as containing an acoustic blob. The detector is evaluated by varying the threshold θ above which an input frame is classified as 'object present'. The test was performed using the background sound for the 'no object' class and all known office objects as 'object present'. Hit rates and false alarms derived from this test are summarized in receiver operating characteristics (ROC). False positives refer to background wrongly classified as blob.

The specific detectors use amplitude modulation spectrograms (AMS, see Section 2.1) as input features. Using these features, support vector machines (SVM, [3]) are trained as models for the specific acoustic objects. The training is performed using a 1-vs-all approach. The performance of the specific models has been evaluated as follows: each of the four office objects is defined as *new* once and left out of the training set. This new object is detected by running the data through the classifiers for all *known* objects. Because of this procedure, separate models are trained for each defined train set. For example, when using speech as novel event, the door detector is trained as "door vs. {keyboard,telephone}".

3 Results

3.1 Data

Data for Classification The data for all classes (including the office background noise) except speech has been recorded in a typical office at the University of Oldenburg. Separate recording sessions of approximately 15 min each have been used for train and test data. The office background noise is dominated by an air conditioning ventilation system and comparatively stationary. Speech data is provided by the TIMIT database. The sound objects have been mixed into the background at SNRs from +20 dB to −20 dB using a long-term (over the whole signal) and broad band SNR computation.

SNR in dB	20	15	10	5	0	-5	-10	-15	-20
Accuracy in %	97.4	96.5	94.5	90.6	86.7	80.3	73.4	64.5	56.9

Table 1. Accuracy of the blob detector taken at the EER point (equal false alarm and miss rates).

Data for Localization To produce a suitably large set of training data for the localization model, sound sources were simulated by using mono-channel speech recordings from the TIMIT speech corpus and generating a second channel as a delayed version of the same data, thus introducing directional information into the data. The simulated DOA angle ranged from $-80°$ (left) to $+80°$ (right) in steps of $10°$, for each direction 10 seconds of speech were used. A spherically symmetric diffuse pink noise field was generated and superposed on the speech signals with varying long-term SNR ranging from -20 dB to +20 dB in steps of 5 dB.

3.2 Detection of Incongruous Events[1]

The general model (acoustic blob detector) has been tested on the test set of the office data. All four objects as well as the background sounds have been used, and the full test set has been balanced to ensure equal amounts of data for 'object present' and 'background only'. The results are summarized in Table 1.

The results of the specific classifiers for detection of speech as incongruent event are shown in Figure 3.2, left panel. These are combined with the result of the general detector to an overall detection score for new objects as follows: the ROC performance of the specific models is scaled by the hit rate of the general detector, reflecting the fact that in the full system, the specific models are fed only those data points specified as blobs by the general detector (see Figure 3). This is equivalent to rescaling the ROC axes from $[0, 1]$ to $[0, \text{blob hit rate}]$. The accuracy at the EER as a function of the SNR is given in Figure 3.2, right panel.

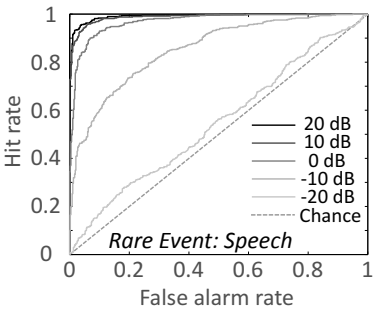
The detection of incongruous events works well ($> 75\%$ detection rate on average) down to 0 dB and approximates chance level at -10 dB. At SNRs below -10 dB, the hit rate of the blob detection is too low, preventing the rescaled ROC of the overall performance to cross the EER line, therefore no EER values can be given.

In summary, the results indicate that the framework used is suitable for novelty detection based on incongruence of different models.

3.3 Localization of Events

The evaluation showed a high robustness of the approach. The performance is stable under missmatch of training and test SNRs and independent of the direction from which a sound source impinges. Results for a single sound source

[1] This part of the results has been submitted to Interspeech 2010 [1].



SNR in dB	20	15	10	5	0	-5	-10
Speech	95.0	92.6	91.1	87.4	81.3	70.5	55.3
Keyboard	92.9	90.8	89.2	84.5	79.5	68.9	55.8
Telephone	89.2	88.4	83.4	80.2	74.2	66.1	55.3
Door	89.5	87.6	81.3	75.2	67.4	64.5	54.2

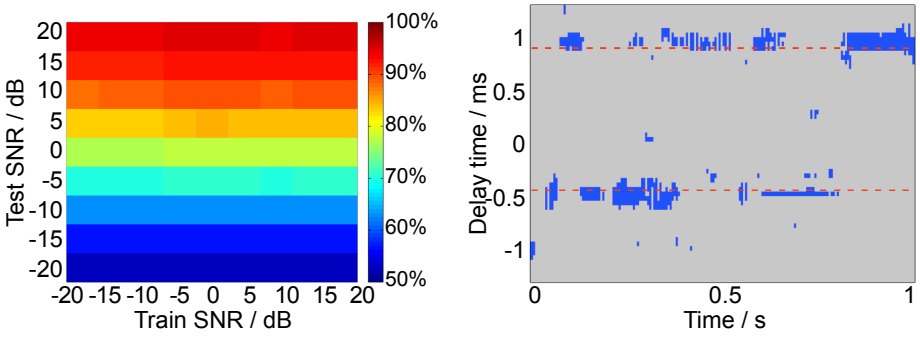
Fig. 4. Left panel: ROC performance of the specific models at different SNRs, testing with speech as the incongruent object. Right panel: Accuracy of the detection of incongruous events taken at the EER point. Below −10 dB, the EER could not be determined.

are shown in Figure 5(a).
A classification example for two speech sources with interference by a noise field at an SNR of 5 dB is shown in Figure 5(b). The theoretical position of the two speech sources at −20° and +45° are indicated by the dotted lines in the right panel. To suppress the "salt and pepper" noise contained in the results of the classification, median filtering over 3 adjacent angles and 7 time frames was applied.

4 Conclusion

We have presented a system for detection and identification of unexpected events in the auditory modality, building on the theory of incongruous event detection [7] and implementing it using classifiers optimized for acoustic signals. Classification built on the hierarchical classification scheme with general and specific classifiers incorporating previously developed acoustic classification methods. Identification of events subsequent to their detection has been incorporated by a novel localization algorithm that performs classification-based estimation of source-direction.

Results indicate that the system can successfully detect and localize incongruous events when signal-to-noise ratio (SNR) is better than 5 dB, in which case equal-error-rate detection performance exceeds between 74% and 81%, depending on the type of incongruous event. At beneficial SNR (20 dB), performance exceeds 89% to 95%. The data also demonstrate that the given numbers reflect performance of the combined scheme, i.e., general (blob) classifier, specific (object) classifiers and incongruence detection between them. Hence, further work will address all three parts to further optimize performance.



(a) (b)

Fig. 5. (a) Performance of the localization algorithm for different and missmatched train and test SNRs. The percentage of correct decisions (source / no source) is shown for a single sound source. (b) Results of the classification for two speech sources at $-20°$ and $+45°$ in a isotropic noise field at an SNR of 5 dB. The dotted lines in the right plot denote the theoretical positions of the sources.

Acknowledgement

This work was supported by the EC project FP6-IST-027787 DIRAC (Detection and Identification of Rare Audiovisual Cues).

References

1. Bach, J.H., Kollmeier, B., Anemüller, J.: Detecting novel objects in acoustic scenes through classifier incongruence. In: Submitted to Interspeech 2010 (2010)
2. Bach, J.H., Kollmeier, B., Anemüller, J.: Modulation-based detection of speech in real background noise: Generalization to novel background classes. In: IEEE (ed.) Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 41–45 (2010)
3. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001), software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
4. Hermansky, H., Morgan, N.: Rasta processing of speech. Speech and Audio Processing, IEEE Transactions on 2(4), 578 –589 (oct 1994)
5. Knapp, C., Carter, G.: The generalized correlation method for estimation of time delay. IEEE Transactions on Acoustics, Speech and Signal Processing 24(4), 320–327 (August 1976)
6. Tchorz, J., Kollmeier, B.: SNR estimation based on amplitude modulation analysis with applications to noise suppression. IEEE Transactions on Speech and Audio Processing 11(3), 184–192 (2003)
7. Weinshall, D., Hermansky, H., Zweig, A., Luo, J., Jimison, H., Ohl, F., Pavel, M.: Beyond novelty detection: Incongruent events, when general and specific classifiers disagree. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems 21, pp. 1745–1752 (2009)