



Detecting novel objects in acoustic scenes through classifier incongruence

Jörg-Hendrik Bach, Jörn Anemüller

Medical Physics, University of Oldenburg, Germany

j.bach@uni-oldenburg.de, joern.anemuller@uni-oldenburg.de

Abstract

In this study, a new generic framework for the detection and interpretation of disagreement (“incongruence”) between different classifiers [1] is applied to the problem of detecting novel acoustic objects in an office environment. Using a general model that detects generic acoustic objects (standing out from a stationary background) and specific models tuned to particular sounds expected in the office, a novel object is detected as an incongruence between the models: the general model detects it as a generic object, but the specific models can not identify it as any of the known office-related sources.

The detectors are realized using amplitude modulation spectrogram and RASTA-PLP features with support vector machine classification. Data considered are speech and non-speech sounds embedded in real office background at signal-to-noise ratios (SNR) from +20 dB to -20 dB. Our approach yields approximately 90% hit rate for novel events at 20 dB SNR, 75% at 0 dB and reaches chance level below -10 dB.

Index Terms: sound classification, acoustic objects, event detection, novelty detection, modulation spectrogram

1. Introduction

The classification of everyday sounds is a broad field of research that has contributed to a large number of applications such as acoustic event detectors, diarization systems, voice activity detection, to name but a few. One important issue with all classification approaches is how to deal with open sets, i.e. with data classes that have not been presented during the training stage. This problem is usually referred to as “Novelty Detection”, for which a number of solutions based on statistical approaches have been presented and successfully applied to real-world problems such as handwritten character recognition or process control. For a review, see [2, 3]. The basic idea behind novelty detection is to model the statistical distributions of the known classes and — using a distance measure such as the Mahalanobis distance — to determine based on a threshold criterion if the data is to be declared “novel” (or equivalently “unknown”).

In previous work [4], the robustness of audio classifiers when faced with known objects in unknown backgrounds has been tackled. Building on that, the present work focusses on detecting novel acoustic objects embedded in a known background.

This task deviates from an overall soundscape recognition task [5] since we are not only interested in classifying the overall scenario, but in the individual sources comprising the scene. Identifying separate acoustic objects and sources other than speech has been employed particularly as a tool in video segmentation and classification [6, 7], where the audio stream provides semantic information that cannot be deduced from the video image alone. Furthermore, the analysis of a particular

acoustic scene (such as is done in the present work) has been tackled by tracking activation patterns for different classes, e.g. in train station scenes [8]. A particular application to office scenes can be found in [9], where microphones and several other sensors were used to track the activity of workers in an office, judging for example how much time is spent working on the computer.

The literature on novelty detection in nonspeech audio is not abundant, most importantly the area of music information retrieval has found an application for novelty detection in song similarity and genre classification. For example, [10] uses the angle between the spectra of different frames as a simple similarity measure. Correlating with a symmetric (in time) and an antisymmetric kernel, the self-similarity within a frame and the similarity between future and past frames is compared, and their difference used as a measure of novelty. By tuning the time constants of the features and the kernels, the thus detected novelties in music can be caused by a change of note (short-term) or a change of theme (long-term). In [11], statistical modelling has been employed to identify songs that do not belong to a known genre. Basically, each genre is modelled by a Gaussian Mixture Model (GMM) over mel-frequency cepstral coefficients of the corresponding songs, and the novel songs are classified by these GMMs. Putting a threshold on the likelihood, the ratio of detected novel songs over wrongly rejected songs can be adapted. The reported results reach approximately 85% hit rate for novelty at almost 50% false alarms, and only 20–50% recall for false alarms below 20%.

In this paper, we propose a different approach: using the more general concept of incongruence, we demonstrate that novelty detection can be tackled by looking for incongruous events. The underlying framework has been proposed in [1]. 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. The approach differs from classic novelty detection (which is closely related to outlier detection) and represents a novel view on detecting new variants of known concepts. In this case, individual classifiers do not detect a novelty, but the overall scene (composition of classes) does not make sense to the combined system, hence the name “incongruence” instead of “novelty”.

In this work, the incongruence detection framework is applied to the detection of unknown acoustic sources (objects) in office scenarios. The general model is provided by an acoustic “blob” detector whose task is to detect acoustic activity popping out of the background. The underlying assumption is that a true background noise is stationary in nature, while significant change in spectral activity can be attributed to the presence of an

10.21437/Interspeech.2010-607

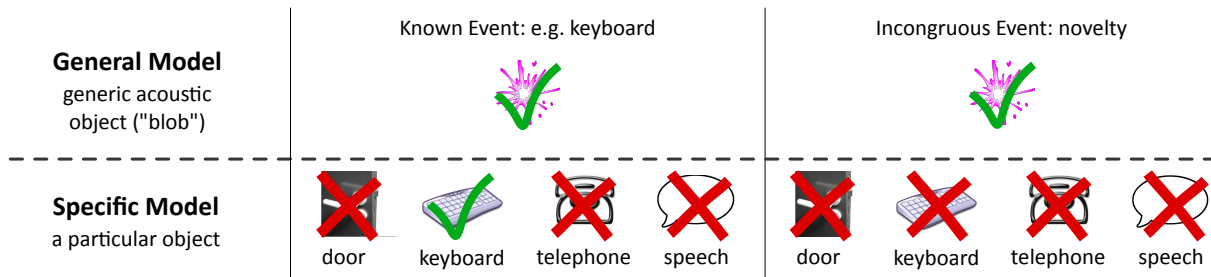


Figure 1: Diagram of the interaction between general and specific models to detect novel events. The green ticks and red crosses signify whether or not the detector — specified by the superposed icon — has a positive detection. Known events (left example) trigger both the blob detector and the appropriate detector for the object, whereas novel objects (right example) trigger the blob detector, but none of the detectors for known objects.

interesting source. At this level of generality, strongly fluctuating backgrounds fall into the category “object” and would have to be dismissed as background by some additional intermediate classifier.

The specific models on the other hand are trained to discriminate individual office-related objects including telephone, speech, doors opening/closing, and keyboard typing. When a new object is presented, an incongruence between these two models is to be expected: the acoustic blob should show high confidence, while all of the known objects have low confidence (see Figure 1).

2. Methods

All classifiers described in the following have been evaluated on the frame basis of 1 s frames.

2.1. General acoustic “blob” detector

The general object detector is based on a RASTA-PLP feature extraction [12]. 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 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.

2.2. Specific object classifiers

The specific detectors use amplitude modulation spectrograms (AMS) as input features. The AMS extracts the temporal modulation of the signal by applying an STFT in 17 Bark-scaled frequency bands. 1 s windows are used to obtain a 1 Hz resolution in the modulation frequency space. This results in a 493-dimensional feature space (17 frequency bands \times 29 modulation frequencies from 2 Hz to 30 Hz). This particular set of parameters has proven successful both for speech and non-speech sound detection [13, 14]. Using these features, support vector machines (SVM, [15]) 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 evalu-

ated 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. This implicates that 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}”.

2.3. Incongruence Detection

The evaluation whether an object is considered incongruous is based on the margin distance of the SVM models for the specific classifiers of known objects. An input sample is considered novel if the (signed) margin distance is smaller than a threshold γ for *all specific classifiers*. By varying γ , ROC curves are obtained that display the trade-off between correct novelty detection and false alarm rate on known events.

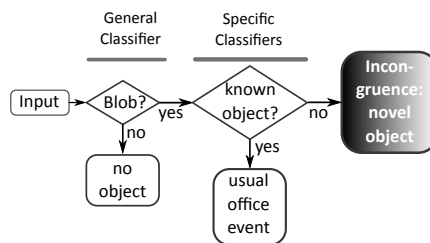


Figure 2: Flow chart of the novelty detection based on incongruence between general and specific models.

The combination of general blob detection and specific object identification (Figure 2) determines the performance of the overall novelty detection: if the blob detector does not detect any object at all, it cannot be classified as novel (or as known, for that matter). This results in an overall ROC-type performance whose maximal hit rate is given by the hit rate of the blob detector, while the shape of the curve is given by the ROC of the specific classifiers. The accuracy at the equal error rate (equal hit and false alarm rates, EER) is used as a figure of merit for the novelty detection. The novel event detection is performed on the four available office objects in a leave-one-out fashion: one event is declared as novel and left out of the training set; the models are trained on the remaining data, and the test is done using the novel object. This is repeated using each of the objects as novel once.

2.4. Data

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 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. The amount of data is identical for all classes.

3. Results and Discussion

3.1. General model

The 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 depicted in Figure 3. The equal error

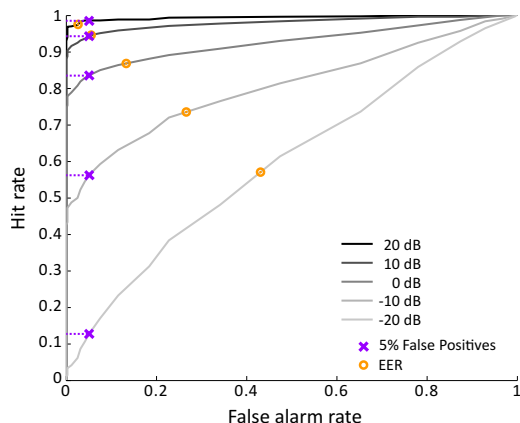


Figure 3: ROC performance of the blob detector (general model) at different SNRs.

rate can be used as an estimate of the optimal operating point (assuming equal prior probabilities for both classes) along the ROC. We choose, however, to work at a fixed false alarm rate of 5% to ensure that most of the background noise is correctly rejected for the novelty detection. The corresponding hit rate is marked because it defines the upper limit for the overall novelty detection (see Section 2.3). Obviously the blob detector works well between 20 dB and 0 dB. At -10 dB, the limit for meaningful application seems to be reached, and at -20 dB it operates essentially at chance level. Given the simplicity of the detector, this is an acceptable working range for the detection of significant events.

3.2. Specific models

The results of the specific models at different SNRs are shown in Figure 4. The objects speech and keyboard typing are easiest to detect as novel events. This can be attributed to the fact that these two types of sound show characteristic modulations in the low modulation bands (1...30 Hz) and are best captured by the AMS feature extraction. Since the two objects have a certain overlap in modulation space, confusions are expected (speech

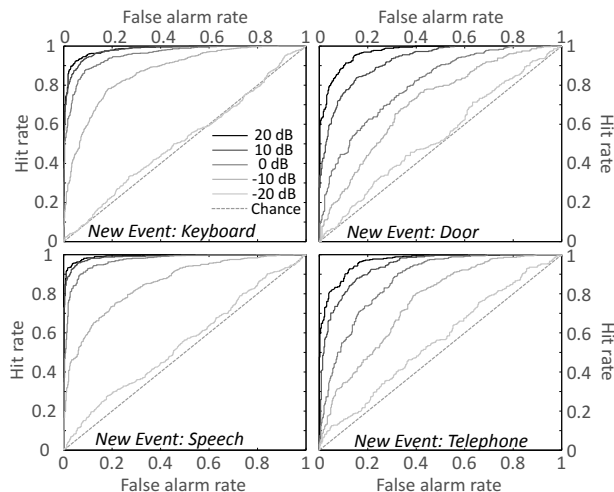


Figure 4: ROC performance of the object detectors (specific models) at different SNRs. The hit and false alarm rate when classifying an object as “new” are displayed for 4 different novel objects. The legend is identical in all 4 subplots and only shown once for clarity.

classified as keyboard and vice versa). When using one of the two objects as novel (speech, say), these confusions may limit the performance because the speech would simply be classified as keyboard typing and not recognized as unknown. This is not the case, indicating that the speech in this case is located at a greater distance from the separating hyperplane of the SVM than the keyboard data (that was trained on). This indicates that the amount of confusions between these similar classes could be reduced by heuristically biasing the decision margin or choosing a different classifier.

The door and telephone sounds appear to be more difficult to detect as novel events: particularly the sounds of opening and closing a door were not sufficiently characterized by their modulation content in the low modulation frequencies. This causes greater fluctuations in the modulation representation because the feature space occupied by those classes is more noisy than for sounds which are perfectly captured by the representation. Thus the clustering of these classes is less compact which results in greater difficulty classifying them.

In summary, the specific models work well at non-negative SNRs and reach the limit of applicability at negative SNRs. At -20 dB, chance level is reached.

3.3. Overall performance of incongruence detection

The results of the general and specific classifiers are combined 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 2). This is equivalent to rescaling the ROC axes from [0, 1] to [0, blob hit rate]. The accuracy at the EER as a function of the SNR is shown in Figure 5.

The novelty detection 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

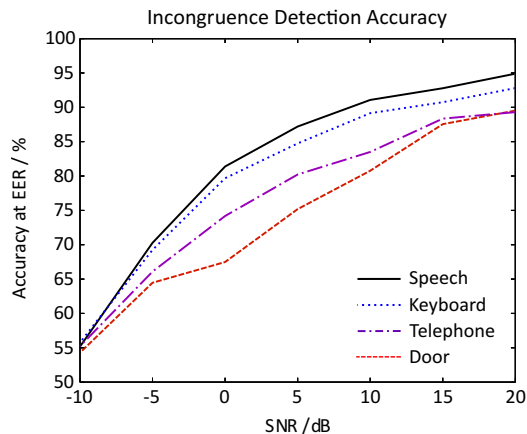


Figure 5: Accuracy of the novelty detection. One curve per type of novel event (see legend). The accuracy is taken at the EER point (equal false alarm and miss rates). Below -10 dB, the EER could not be determined.

can be given.

In summary, the results indicate that the framework used is suitable for novelty detection based on incongruence of different models. Since two models of different abstraction are used, the overall performance is influenced differently by the two sets of models. In a real application where the two models work consecutively, the specific object detectors are only triggered if the blob detector is positive. On the one hand, the hit rate (recall) of the blob detector provides an upper limit to the overall performance. The specific detectors on the other hand define the profile of the ROC performance. This means that a perfect blob detector does not alter the ROC, while perfect specific models, although technically able to correctly detect every single unknown event, do not necessarily result in perfect novelty detection.

4. Conclusions

We demonstrated that the generic framework of incongruence detection covers the detection of novel acoustic events, producing very good results for SNRs between 0 dB and 20 dB. In contrast to outlier detection, which is prone to confuse noise with actual outliers, the framework of general and specific classifiers ensures that mere noise is not identified as an interesting class, because the general level is required to detect a meaningful input that fits its descriptor.

Incongruent event detection relies on distinguishable object clusters in feature space. The amplitude modulation based features with high resolution in low frequency modulation bands are well-suited for the detection of the objects considered here. However for burst-like sounds (e.g. door open/close sounds) the use of other feature sets such as spectral features and onset detectors might prove beneficial.

The simple application of a confidence threshold is sufficient for the detection of “unknown” classes. Effectively, this means that the approach of using a combination of general and specific models (as opposed to multi-class probability density estimation) does not require more complex techniques than threshold-based solutions for novelty detection. Novel objects could be reliably detected over a wide range of SNRs. This indicates that the results obtained here may be transferable to other, more complex scenarios, which will be the subject of fu-

ture work.

5. Acknowledgements

The authors thank Hendrik Kayser and Bernd T. Meyer for proof reading the manuscript. This work has been supported by the EC integrated project DIRAC (Detection and Identification of Rare Audiovisual Cues) in the 6th FWP under IST-027787.

6. References

- [1] 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: *Advances in Neural Information Processing Systems (NIPS) 2008*, 1745-1752.
- [2] Markou, M., Singh, S., “Novelty detection: a review - part 1: statistical approaches”, *Sig. Proc.* 83, 2481–2497, 2003.
- [3] Markou, M., Singh, S., “Novelty detection: a review - part 2: neural network based approaches”, *Sig. Proc.* 83, 2499–2521, 2003.
- [4] Bach, J.-H., Kollmeier, B., Anemüller, J., “Modulation-based detection of speech in real background noise: generalization to novel background classes”, *Proc. ICASSP 2010*, 41–44.
- [5] Aucouturier, J.J., Defreville, B., Pachet, F., “The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music”, *J. Acoust. Soc. Am.* 122(2), 881–891, 2007.
- [6] Bugalho, M., Portelo, J., Trancoso, I., Pellegrini, T., Abad, A., “Detecting audio events for semantic video search”, *Proc. Interspeech 2009*, 1151–1154.
- [7] Rouvier, M., Matrouf, D., Linarès, G., “Factor Analysis for Audio-based Video Genre Classification”, *Proc. Interspeech 2009*, 1155–1158.
- [8] Niessen, M., van Maanen, L., Andringa, T.C., “Disambiguating sound through context”, *J. Semantic Comp.* 2(3), 327–341, 2008.
- [9] Oliver, N., Garg, A., Horvitz, E., “Layered representations for learning and inferring office activity from multiple sensory channels”, *Computer Vision and Image Understanding* 96(2), 163–180, 2004.
- [10] Foote, J., “Automatic audio segmentation using a measure of audio novelty”, *Proc. IEEE Conf. Multimedia and Expo 1999*, 452–455.
- [11] Flexer, A., Pampalk, E., Widmer, G., “Novelty detection based on spectral similarity of songs”, *Proc. ISMIR 2005*, 260–263.
- [12] Hermansky, H., and Morgan, N., “Rasta Processing of Speech”, *IEEE Trans. Sp. Aud. Proc.* 2, 578–589, 1994.
- [13] Anemüller, J., Schmidt, D., Bach, J.-H., “Detection of Speech Embedded in Real Acoustic Background Based on Amplitude Modulation Spectrogram Features”, *Proc. Interspeech 2008*, 2582–2585.
- [14] L. Jie, B. Caputo, A. Zweig, J.-H. Bach and J. Anemüller: “Object Category Detection Using Audio-visual Cues”, *Proc. ICVS, Santorini, Greece*, 2008
- [15] 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>.