Environmental Science Processes & Impacts

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

rsc.li/process-impacts

Environmental impact statement

Around the world, a large number of people are exposed to high levels of noise pollution, having an adverse effect on human health and well-being. Accurately assessing the noise exposure in urban environments requires a large-scale and dense measurement network to capture the spatial and temporal variations. In this context, low-cost microphones are often used to reduce costs. Compared to type-approved sound level meters, these microphones have higher failure rates and produce lower quality data. Therefore, a strong measurement quality control is of great importance. In this work a multi-criteria measurement quality assessment model is presented for validating noise measurements. It was shown that the proposed approach successfully detects microphone breakdowns and anomalies typically observed in long-term, extensive measurements.

Multi-criteria anomaly detection in urban noise sensor networks

Samuel Dauwe[*a], **Damiano Oldoni**[b], **Bernard De Baets**[c], **Timothy Van Renterghem**[b], **Dick Botteldooren**[b], **and Bart Dhoedt**[a]

The growing concern of citizens about the quality of their living environment and the emergence of low-cost microphones and data acquisition systems triggered the deployment of numerous noise monitoring networks spread over large geographical areas. Due to the local character of noise pollution in an urban environment, a dense measurement network is needed in order to accurately assess the spatial and temporal variations. The use of consumer grade microphones in this context appears to be very cost-efficient compared to the use of measurement microphones. However, the lower reliability of these sensing units requires a strong quality control of the measured data. To automatically validate sensor (microphone) data, prior to their use in further processing, a multi-criteria measurement quality assessment model for detecting anomalies such as microphone breakdowns, drifts and critical outliers was developed. Each of the criteria results in a quality score between 0 and 1. An ordered weighted average (OWA) operator combines these individual scores into a global quality score. The model is validated on datasets acquired from a real-world, extensive noise monitoring network consisting of more than 50 microphones. Over a period of more than a year, the proposed approach successfully detected several microphone faults and anomalies.

1 Introduction

Over the years, the influence of environmental stressors such as noise on people's health has been the subject of growing concern[1–4]. In order to identify the sources of environmental disturbances and to inform decision makers, long-term accurate monitoring is needed. Most often, type-approved sound level meters are used for acquiring high-quality observations. However, due to the high cost of these units, deploying an extensive measurement network becomes very expensive.

Nowadays, low-cost embedded computer systems and microphones are widely available and enable the deployment of large-scale noise monitoring networks, gathering observations over long time periods at a reasonable price. For instance, Van Renterghem et al.[5] demonstrated the use of consumer-grade microphones, resulting in only small level differences compared to high-quality microphones. However, the potentially lower reliability of these cheap sensors poses some real challenges to the maintenance and error detection.

Another interesting evolution is that low-cost sensing units also open up new opportunities for participatory measurement initiatives. Concerned citizens can play an active role in the data collection phase, gathering data with high spatio-temporal resolution that can be used in discussions with local policy makers. However, non-expert users unfamiliar with the appropriate measurement conditions can introduce errors in the gathered sensor data. A second problem is that participants interested in a specific outcome of the study could potentially tamper with the measurement conditions.

In view of these problems, strong quality control of measurement data is necessary. However, detection of sensor failures and anomalous data is a difficult task that has been addressed in many studies; for a detailed overview we refer to[6]. In the literature, short duration events that significantly deviate from the normal sensor readings are regarded as outliers[7]. Detection methods for such abnormal measurements often use rule-based approaches or statistical techniques. For instance, Sharma et al.[8] demonstrated the use of heuristic rules for detecting and identifying faults that have been observed in several real-world sensor network deployments.

A frequently used method for detecting sensor failures is the comparison with measurements from redundant sensors. Bychkovskiy et al.[9] uses a dense sensor deployment in which multiple sensors are installed to measure the same physical variable, assuming that neighbouring sensors should have similar readings. However, due to the highly local nature of noise exposure in urban environments, such an approach could only be applicable in an extremely dense network of sensors or when two identical sensor are embedded in one sensor box. Such extreme physical redundancy involves extra hardware cost which makes this approach not widely feasible.

Another approach to decide about the occurrence of faults is using a fault detection and isolation (FDI) model[10]. This approach is based on comparing available measurement vectors of the monitored system (y) with their corresponding predicted vectors ($\widehat{y}$) obtained using a system model. Nonetheless, the problem with this approach is that a detailed description of a system that takes into account the effects of faults requires multiple models. Furthermore, the system behavior under normal operation could be estimated at different temporal scales. At the second and sub-second scale, noise levels are

[a] *Department of Information Technology (INTEC) - IBCN, Ghent University - iMinds, Gaston Crommenlaan 8 bus 201, B-9050 Gent, Belgium. E-mail: samuel.dauwe@intec.ugent.be*
[b] *Department of Information Technology (INTEC) - Acoustics Research Group, Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium*
[c] *KERMIT, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure links 653, B-9000 Gent, Belgium*

quite predictable since they relate to the same event (e.g. a car pass-by, a conversation). At the 15-minute average time scale one often, but not always, detects diurnal patterns that reoccur every day over and over again and thus also at this time scale predictability is high. However, microphone properties usually degrade over a long period of time, leading to inaccurate measurements. Predicting the magnitude of microphone drift over a long time period is typically a challenge because the degradation highly depends on the actual usage and the conditions the microphone is exposed to.

Different machine-learning techniques have been used for anomaly detection in sensor networks: Bayesian networks[11], rule-based systems, nearest-neighbour-based techniques[12,13] to name just a few. For a general and detailed overview, we refer to Chandola et al.[6] and Zhang et al.[14].

Finally, insert voltage calibration and Charge-Injection Calibration (CIC)[15,16] are specific techniques for monitoring operation conditions of microphones that have been developed for high-end unmanned systems. These methods are based on the principle of injecting a constant charge into the microphone capacitance and preamplifier input circuit. For a given calibration signal input, a change in the preamplifier output indicates a malfunctioning of the microphone. However, the costs of such control systems are high and therefore only suited for high-end measurement chains. Moreover, our goal is the detection of anomalies in a broader sense, breakdowns or malfunctioning of acoustic measurement setups are just a subset of all possible anomalies that could occur.

In this paper we present a multi-criteria model for sensor data validation and anomaly detection through the evaluation of the measurement quality. The paper is structured as follows. Section 2 gives an overview of the proposed quality assessment model for anomaly detection and each of the criteria the model is composed of. Section 3 illustrates the impact of microphone accuracy on quality. Section 4 discusses an evaluation of the quality assessment model based on acoustic sensor datasets from a real-world measurement network. In Section 5, a discussion on the proposed approach finally concludes the paper.

2 Anomaly detection

2.1 Anomalies: a practical definition

According to the Oxford dictionary, an anomaly is defined as *something that deviates from what is standard, normal, or expected*. This general definition needs to be contextualized and applied to the specific field of interest. In noise sensor networks, four categories of anomalies can be identified.

- *Abrupt fault or failure:* indicates a serious breakdown of a system component or function that leads to a significantly deviating behaviour of the whole measurement chain. Examples include an unplugged microphone, or a broken microphone due to direct contact of the electronics with water.

- *Incipient fault:* represents a small and often slowly developing continuous fault of which the effects on the system are in the beginning almost unnoticeable. Sensor drift is a typical example of an incipient (soft) fault.

- *Intermittent fault:* temporary wrong measurements, for example due to transient harsh environmental conditions (e.g. extremely high or low temperature, heavy rain, etc.) in which the sensor is operating.

- *Unexpected and rare sound event:* contrary to faults, this source of anomaly is not related to sensor malfunctioning. It can happen that atypical sound events occur in the proximity of the microphone, thus altering long term noise exposure evaluation indicators such as L_{day}, $L_{evening}$, L_{night} or L_{den}. Whether this event is included in the final environmental assessment depends on the context and is an operator decision. Malicious alteration of the acoustic environment around the unattended sensors to voluntarily modify the noise measurement is another example of this type of anomaly.

2.2 Multi-criteria quality assessment

In view of detecting anomalies in the four categories as described in Section 2.1, we propose a multi-criteria approach consisting of four quality models. Each criteria results in a quality score between 0 and 1 and is aggregated into a global score, by means of an ordered weighted average operator (OWA). The concepts behind each quality model are briefly presented in Fig. 1 and will be further discussed in the remainder of this section.

Each quality model has several parameters that have to be selected carefully. In order to determine the parameters for each quality model, measurement microphones (IEC type WS2F microphones according to IEC 61094-4[17]) were placed close to a few of the consumer grade microphones. We assume that the measurement microphone is stable, correctly calibrated and fit-for-purpose. Comparison of the measurements from both types of microphones enables the tuning of the parameters related to sensor faults more easily. Characteristics of the used microphones can be found in Table 1.

In view of the computational load on the system, assessing quality on very short time intervals (i.e. 1s) is not appropriate. As a compromise we aim at obtaining a quality assessment every minute.

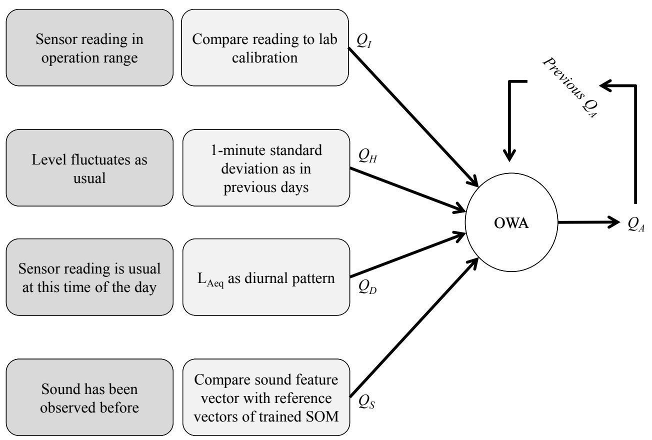

Fig. 1 Overview of the multi-criteria measurement quality assessment concepts. The dark grey boxes represent the concepts behind each quality model, the light grey boxes the implementation of such concepts and the arrows directing to the circle their related quality scores, whereas the OWA (ordered weighted average) operator in the circle combines them into a global quality score (Q_A).

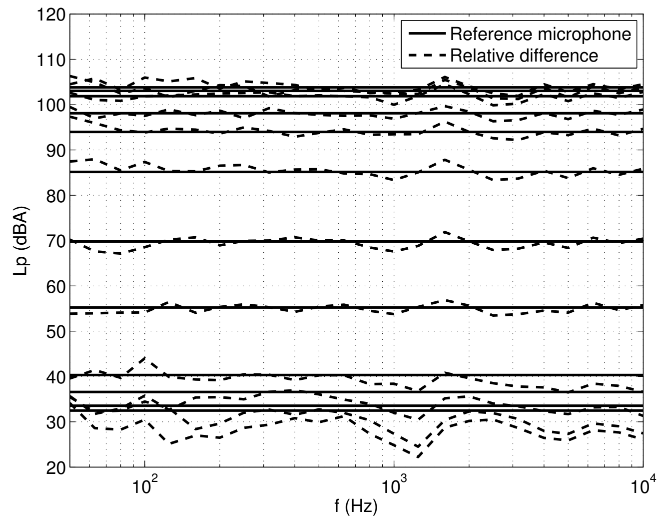

Fig. 2 Illustration of the relative difference between the reference microphone and one of the test microphones. The dashed curves show the difference $L_{P,ref.mic.} - L_{P,test.mic.}$ increased by the level of the reference microphone $L_{P,ref.mic.}$. The peak between 1000 Hz and 2000 Hz is due to the scattering of the measurement device, causing local minima and maxima near the microphone for all levels. Note that the different levels of the reference microphone are flattened out.

2.3 Intrinsic quality (Q_I)

A noise monitoring network can contain a large variety of microphone types, each having unique characteristics in terms of frequency range, noise floor, operational temperature and humidity range. Measurement microphones are expected to produce measurements that are more reliable than low-cost consumer-grade microphones.

In order to take into account the quality of the microphone characteristics, each test microphone has been placed in an anechoic room alongside a measurement microphone for determining the difference in frequency response. The measurement setup was as follows: the reference microphone with preamplifier was placed in the anechoic room connected to the National Instruments PXIe-chassis with NI-4498 DAQ card and control interface, placed outside the anechoic room. The test microphone was connected to the PC Engines Alix single-board computer[18] inside the anechoic room. Twelve different pink noise levels were emitted by a loudspeaker and for each 1/3 octave band i, the offset ($\Delta L_i = L_{i,ref.mic.} - L_{i,testmic.}$) between the reference microphone and the test microphone was measured. An illustration of the results obtained from such a measurement process for one test microphone is depicted in Fig. 2. Especially for lower noise levels there is large offset, this is due to the difference in noise floor. More details about the tested microphone can be found in Table 1.

For defining the intrinsic quality value Q_I, the expected offset ΔL_i for the i-th 1/3 octave band A-weighted sound pressure level L_i of the measurement is determined. This is done by linearly interpolating between the two closest noise levels that were measured in the anechoic room.

The Q_I score is expressed as a non-dimensional value in the range [0,1] with 1 meaning a response exactly the same as the response of a reference microphone. The formula for calculating the intrinsic quality score Q_I is:

$$Q_I(t) = e^{\frac{-\Delta(t)^2}{\sigma}} \tag{1}$$

where σ represents the acceptable deviation and $\Delta(t)$ is given as follows:

$$\Delta(t) = 10 \cdot \log_{10}\left(\frac{\sum_i 10^{\frac{L_i(t) - \Delta L_i(t) + \Delta L_{cal}}{10}}}{\sum_i 10^{\frac{L_i(t)}{10}}}\right) \tag{2}$$

Since all microphones are calibrated at 1000Hz, 94dB at the time they are deployed, the offset at this frequency and level, ΔL_{cal}, is used as reference.

2.4 Heuristic quality (Q_H)

The heuristic quality is based on the idea that short-term noise level fluctuations should be comparable at the same time of the corresponding day of the week. A drastic change in signal variability over a specific period can be considered an efficient heuristic for quality degradation or deliberate manipulation of the observed level.

Table 1 Product details, measured noise floor and prize of the test microphone and the reference microphone

Model	Type	Microphone sensitivity	Frequency range	Noise floor at 1 kHz (measured)	Cost (including pre-amplifier circuit)
Knowles FG-23329-P07*	Electret	-53 dB	100 Hz to 10 kHz	35 dB	33 €
Bruel and Kjaer type 4189 (microphone capsule)**	Electret	-26 dB	6.3 Hz to 20 kHz	13 dB	2000 €

*Test microphone
**Reference microphone

In this model, a quality score is calculated through comparing the standard deviation within a specific time period, typically one minute, at a specific time of the day, to the reference standard deviation on the energetic averages measured during the last four weeks at the same time of the day. In order to determine the integration period of the energetic averages on which the reference standard deviation has to be calculated, several periods were selected. The different integration periods were evaluated on a dataset of one day obtained for the different type of locations found in our measurement network. An overview and description of the evaluated locations is given in Table 2. The integration period was chosen based on the resulting heuristic quality score Q_H as illustrated in Table 3. Note that in this test, we used a window to select the reference levels with the size of the integration period that was moved every one minute. The standard deviation on the levels in the window was used in the actual comparison.

From the resulting quality scores we conclude that a reference standard deviation based on 1-minute energetic averages $L_{Aeq,1min}$, results in the highest heuristic quality scores. The comparison of the current minute and reference standard deviation, expressed as their ratio, is then transformed into a quality score via a Gaussian-like function centered at one.

The formula for calculating the heuristic quality score Q_H is given as follows:

$$Q_H(t) = \begin{cases} e^{\frac{-\ln(x(t))^2}{\tau_1}} & \text{if } 0 < x(t) < 1 \\ e^{\frac{-(x(t)-1)^2}{\tau_2}} & \text{if } x(t) \geq 1 \end{cases} \quad (3)$$

where $x(t)$ denotes the ratio between the standard deviation within the 1-minute period and the reference standard deviation at the same time of the day and the week. The parameters τ_1 and τ_2 represent the upper and lower thresholds determining the steepness of the Gaussian-like function. Two different functions have been chosen for x greater or less than 1 due to the fact that x has a lower limit, zero, but no upper limit by definition.

2.5 Diurnal pattern quality (Q_D)

In the context of road noise and day level assessment, Romeu et al.[19] analyzed several studies concluding that a measurement duration between 10 minutes and 1 hour can be considered representative of the noise climate at a given location. It could be expected that an integration time of this order of magnitude could be used to characterise the typical diurnal pattern of noise levels.

The temporal integration period of the diurnal pattern that results in both stable description and a usable variability in all the possible locations is expected to be the most suitable choice for the diurnal pattern quality model. Therefore, we determined for each type of location (as described in Table 2) the Pearson's correlation coefficient between the 1-minute aggregated instantaneous sound pressure levels $L_{Aeq,1min}$ and the diurnal pattern for the same time of the day. A graphical representation of this process is depicted in Fig. 3.

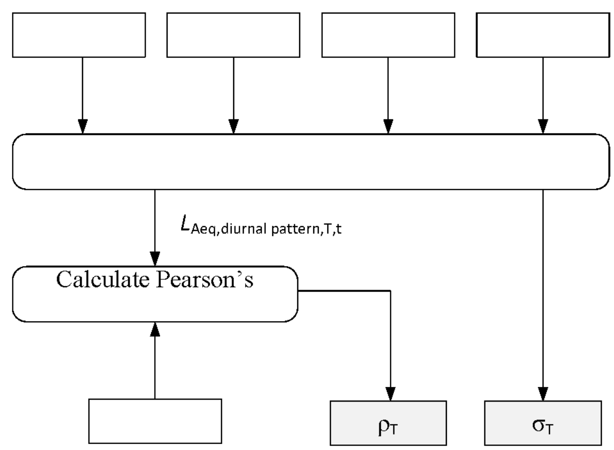

Fig. 3 Overview of the process for determining the Pearson's correlation coefficient ρ_T between the one-minute aggregated instantaneous sound pressure levels $L_{Aeq,1min,t}$ and the corresponding vector of diurnal pattern levels $L_{Aeq,diurnalpattern,T,t}$ for each minute t of one day. T denotes the temporal integration period of the diurnal pattern. To evaluate the stability of the diurnal pattern, we also calculated the standard deviation σ_T.

In order to assess the significance of the relationship between the two variables, we also determined the levels of significance (t-values). From the test results as presented in Table 4, we conclude that the 1-minute temporal integration pe-

Table 2 Overview of the different type of locations in our measurement network. All measurement points are location in Belgium.

Type of location	Description
Railway	Backyard of the first dwelling 27 meters near the railway.
Noisy façade	Street side of a building in the city center of Ghent.
Quiet façade	Internal, quiet side of a building in the city center of Ghent.
Park	Windowsill of the first floor of the castle in the Bouckenborgh park in Merksem, Antwerp.
Wind turbine	Measurement location 10 meters near the base of a wind turbine in Leuze.

Table 3 Percentages of the heuristic quality (Q_H) higher than a given threshold for good data ($Q_H > 0.5$). The quality score is calculated for different integration periods of the energetic averages. Percentages were determined based on a dataset of one day obtained from different typical locations

	$Q_H > 0.5$ (%)				
Integration period (s)	Railway	Quiet façade	Noisy façade	Park	Wind turbine
60	74.94	89.5	90.06	79.56	71.3
300	73.26	89.2	89.79	78.61	70.89
900	71.12	88.45	89.52	78.01	70.26
1800	70.96	85.12	88.96	77.23	69.89
3600	70.12	82.23	87.88	76.12	68.47

riod for the diurnal pattern results in a strong and significant relationship with the instantaneous sound pressure levels for all locations. However, standard deviation on the diurnal pattern is high when using a short integration period. Therefore, we have chosen a 15-minute temporal integration period for the diurnal pattern to account for a strong correlation and low standard deviation, thus stable pattern.

The formula for calculating the diurnal pattern quality score Q_D is given as follows:

$$Q_D(t) = e^{\frac{-(X_i(t)-\mu(t))^2}{\kappa \cdot \sigma^2_{D(t)}}} \tag{4}$$

where $X_i(t)$ represents the current 1-minute A-weighted equivalent sound pressure level $L_{Aeq,1min}$ value, $\mu(t)$ represents the 1-minute aggregated diurnal-pattern value. Parameter κ corresponds to a pre-determined constant. Based on our experience with the five test sites as described in Table 2 we chosen the value of parameter $\kappa = 2$ to optimise the correctly identified faults and decrease the erroneous detections. $\sigma_{D(t)}$ denotes the standard deviation of the energetic averages $L_{Aeq,15min}$, calculated over all days of the week or weekend during the last four weeks at the same time of the day. The latter quantity indicates how stable the diurnal pattern is at this measurement location.

2.6 SOM quality (Q_S)

This quality model is based on the idea that sounds that were never encountered before could indicate a potential sensor malfunction or an alteration of the typical acoustic environment around the microphone. This model can be seen as part of a wider computational model aiming to model the human auditory perception. More details can be found in Oldoni et al. [20].

The first stage aims to extract acoustical features inspired by human peripheral auditory processing. The model starts from the 1/3-octave band spectrum of the sound pressure level with a temporal resolution of 1/8 s. Energetic masking is simulated by a cochleagram, by means of the Zwicker loudness model [21,22]. To simulate the human auditory system, the absolute intensity and the spectro-temporal variations form the basis of the calculated acoustical features. Based on existing models for auditory saliency [23,24], a high-dimensional feature vector is calculated at each timestep by means of a centre-surround mechanism mimicking the receptive fields in the auditory cortex.

The quality score is determined in two major steps. In a first, offline phase, a neural network called Self-Organizing Map (SOM) is trained based on the temporal correlation of the occurrence of sound features. More details about the learning algorithm can be found in Oldoni et al. [20]. In a second, online phase, a sound feature vector aiming to describe the incoming sound is extracted from the 1/3-octave band spectra at each timestep. This is followed by calculating the distance between this sound feature vector and all reference vectors of SOM units. The minimum distance is then found and its related SOM unit is typically called the Best-Matching Unit (BMU). If the distance to the BMU is very high (higher than a fixed threshold), then the incoming sound can be defined as an outlier, not typical of the given acoustic environment. On the other hand, if the distance to the BMU is low, it means that the incoming sound is quite typical for that particular location. Outlier sound can be caused by microphone malfunction

Table 4 Overview of the correlation between the instantaneous one minute aggregated sound pressure levels ($L_{Aeq,1min}$) and the diurnal pattern for the same time of the day. Evaluation was done for different temporal integration periods, using a dataset of one month. The * symbol indicates the significance of the correlation coefficient. The table also contains the standard deviation on the diurnal pattern, showing a drop when the diurnal pattern integration period increases

	Correlation coefficient					*Standard deviation (dBA)*				
Diurnal pattern integration period (s)	Quiet façade	Noisy façade	Park	Railway	Wind turbine	Quiet façade	Noisy façade	Park	Railway	Wind turbine
60	0.501**	0.616**	0.158**	0.166**	0.187**	6.799	7.838	5.482	11.756	5.014
300	0.514**	0.481**	0.154*	0.157*	0.225*	7.099	5.417	5.480	7.320	3.241
900	0.541**	0.499**	0.165	0.108	0.213*	3.473	4.114	5.460	3.311	2.582
1800	0.565*	0.517*	0.174	0.109	0.286*	3.389	3.742	5.461	2.476	2.251
3600	0.483*	0.515*	0.176	0.098	0.336*	3.217	3.698	5.451	1.599	2.047

**correlation coefficient is very significant
*correlation coefficient is significant

or can indicate tampering with the device. Therefore a low quality score is assigned.

The average distance to the BMU, d_{BMU}, over a minute period is transformed into a quality score Q_S via the following function:

$$Q_S(t) = e^{\frac{-d_{BMU}(t)}{th_{mid}}} \tag{5}$$

where th_{mid} denotes the distance to the BMU for which the quality score equals to 0.5.

2.7 Aggregated quality (Q_A)

In order to get a final "quality of the measurement", i.e. a single scalar Q_A, the partial quality evaluations from each quality model have to be merged by means of an *aggregation operator* or *aggregator*. The ordered weighted averaging (OWA) operator[25] is a very flexible and tunable aggregator which performs a kind of aggregation. Many notable operators such as the minimum, maximum, arithmetic average and median are members of this class. In a more formal way, the OWA operators form a parameterized class of averaging-type aggregation operators. Details are given in the remainder of this section.

2.7.1 Inputs The input of the aggregator is the vector $Q(t)$ containing the 1-minute based quality scores of the models used for quality calculation (see Sections 2.3-2.6) and the previous output of the same OWA, $Q_A(t-1)$:

$$Q(t) = (Q_I(t), Q_H(t), Q_D(t), Q_S(t), Q_A(t-1)) \tag{6}$$

The OWA aggregator is adaptable, i.e. it can work with a variable number of quality models. This is a very important feature especially in case one or more quality models are temporarily not working or new models are added.

2.7.2 Implementation details The weights of the OWA operator are calculated as follows:

$$w_j = F\left(\frac{j}{n}\right) - F\left(\frac{j-1}{n}\right), \tag{7}$$

where $F(r) = F_\alpha(r) = r^\alpha$ and j is the index of the ordered vector $X(t) = (Q_1(t), Q_2(t), \ldots, Q_n(t))$, from the largest value to the smallest one. Vector $X(t)$ will be equal to $Q(t)$ if and only if $Q_I(t) \geq Q_H(t) \geq \cdots \geq Q_A(t-1)$. The parameter α is called the quantifier of the OWA operator and it is a positive real number which defines the orness of the average. In particular, it results in the typical arithmetical mean if $\alpha = 1$. The weights should not be calculated each time, but only in case the dimension of the quality vector changes, i.e. if a model has been added or for some reason has ceased to work. Finally the aggregated quality can be calculated as a weighted average:

$$Q_A(t) = \sum_{j=1}^{n} w_j \cdot X_j(t) \tag{8}$$

The adjective ordered comes from the fact that the quality scores $Q_j(t)$ are ordered. $Q_A(t)$ is a number between 0 and 1, where a low $Q_A(t)$ score means low quality of the corresponding sensor measurement, whereas scores near to one indicate good and trustful sensor readings.

3 Influence of microphone type on measurement quality

3.1 Quality versus level difference

In order to demonstrate the impact of sensor type on the measurement quality, a measurement microphone was placed near a lower quality microphone in a quiet urban area for comparison. Both devices have unique characteristics in terms of sensitivity and noise floor. Fig. 4 illustrates the measured noise level difference (ΔL) between the two microphones. The differences are especially noticeable in the case of low noise levels, for instance during the night, when $L_{Aeq,test.mic.}$ is constantly greater than $L_{Aeq,ref.mic.}$.

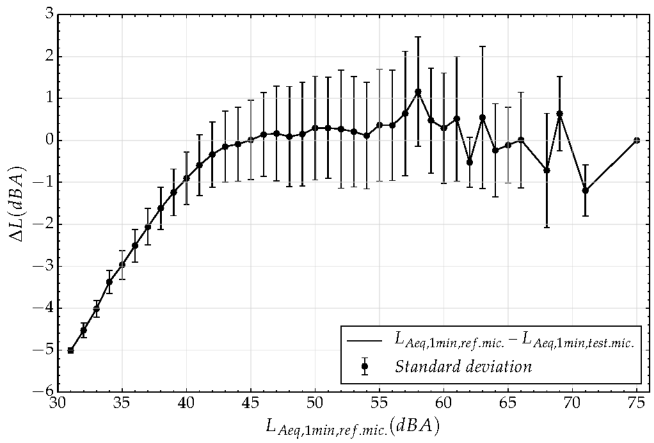

Fig. 4 Illustration of influence of microphone type on the measured noise levels. These results have been obtained using a dataset of one month.

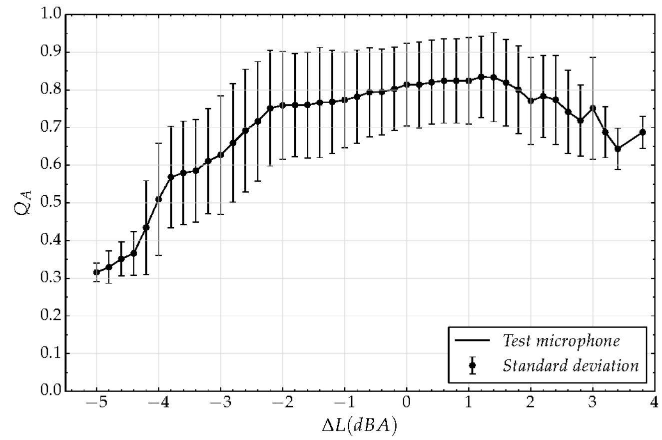

Fig. 5 Aggregated quality Q_A of a typical consumer-grade tested microphone vs. noise-level difference (ΔL) between a measurement microphone and such microphone. These results have been obtained using a dataset of one month.

Fig. 5 illustrates the aggregated quality Q_A for the test microphone as a function of the level difference. Due to the higher noise floor of the test microphone, there is a drop in quality when low noise levels occur, e.g. during night. A quality drop also occurs for high noise levels. This might be explained by the lower intrinsic quality Q_I, due to saturation of the test microphone. For high noise levels, we also noticed lower SOM quality Q_S scores.

3.2 Quality model comparison

In order to give a more detailed overview of how each quality model contributes to the global quality, a plot was made for each indicator based on the same dataset as used in Section 3.1. The distribution of the quality scores is presented in Fig. 6. Q_H, Q_D and Q_S result in a high-quality score for both the reference microphone and the tested microphone. The model designed to represent the difference in microphone quality is intrinsic indicator Q_I, resulting in significantly lower quality scores in case of the test microphone compared to the Q_H, Q_D and Q_S models.

The reference microphone was used as a standard during the response measurements in the anechoic room as described in Section 2.3. As such it should have an intrinsic quality score (Q_I) that equals to 1. However, previous work[5] showed that there are small differences between measurement microphones deployed in the field and therefore we used a value of 0.95.

This comparison also shows that the quality assessment is high, yet not perfect for the measurement microphone. Unidentified, outlier sound events reduce the average value of the Q_S quality indicator but there are few very low values. The Q_H indicator results in a small differentiation between both microphones for short fraction of the observation time. Temporal fluctuation during the night with unresolved low level intervals could cause this.

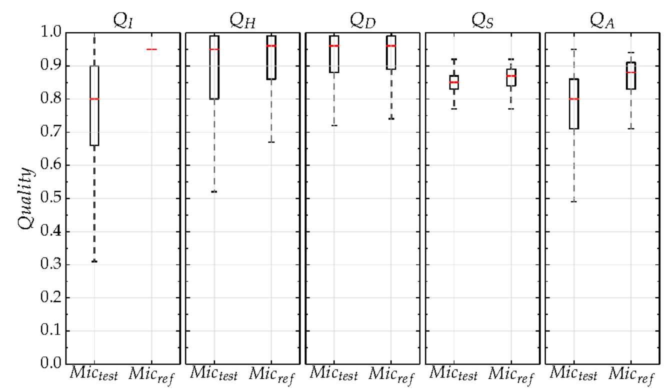

Fig. 6 Distribution of all quality models for the test microphone Mic_{test} (on the left of each box plot) and the measurement microphone Mic_{ref} (on the right of each box plot) using a dataset of two weeks. From left to right: Q_I, Q_H, Q_D, Q_S and Q_A. The intrinsic quality of the measurement microphone is set to the fixed value 0.95.

4 Detection of significant quality loss

In this section we focus on a set of microphone faults and anomalies that have been observed in a real noise-measurement network. In particular one or more examples for each of the four anomaly categories described in Section 2.1 will be shown and examined.

4.1 Noise monitoring network

For the evaluation of our multi-criteria anomaly detection approach, we used datasets acquired from an extensive noise monitoring network consisting of more than 50 nodes with microphones deployed in major cities in Belgium and the Netherlands. For more information about the used microphones we refer to Table 1. The PC Engines ALIX 3d3 is used to pre-process the raw audio sampled at 48 kHz from the on-board Realtek ALC203 AC'97 audio codec. The processing step includes the calculation of 1/3-octave band spectra with temporal resolution of 1/8s. These measurements are periodically transmitted to a central server for storage and further processing. Over a period of more than a year, our approach successfully detected several sensor faults and anomalies, as described in the remainder of this section.

4.2 Sensor breakdown

The most basic error in a measurement network is a sensor breakdown, which is illustrated in Fig. 7. This example shows how a period of normal operation is abruptly stopped due to a hardware malfunctioning, resulting in a signal loss. The quality scores of all models are depicted in Fig. 8. Due to the low electronic noise level which has not been encountered during microphone testing, Q_I is very low after the abrupt breakdown. Also Q_H and Q_D are able to detect the error due to the high difference with the diurnal pattern and the reference period. The Q_S score is not significantly lower, because the Q_S model is trained on the noise floor that was reached during the night. As a consequence, when the microphones malfunctioned, it was recognised as a typical situation for the acoustic environment of that specific microphone.

This type of sensor breakdown could obviously be detected from the noise levels directly. However, this example shows that a more holistic approach also allows detecting this trivial situation.

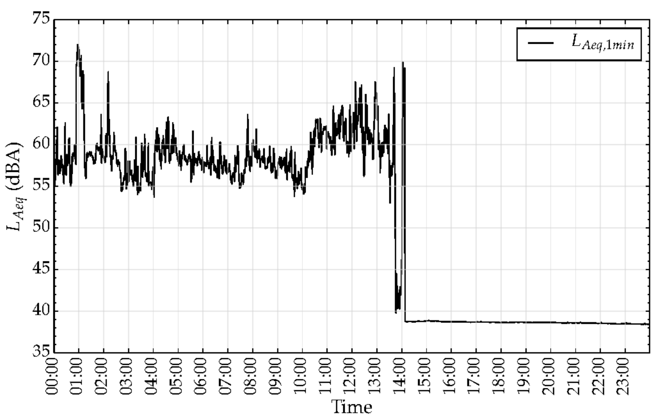

Fig. 7 $L_{Aeq,1min}$ graph illustrating a microphone breakdown, resulting in a signal loss.

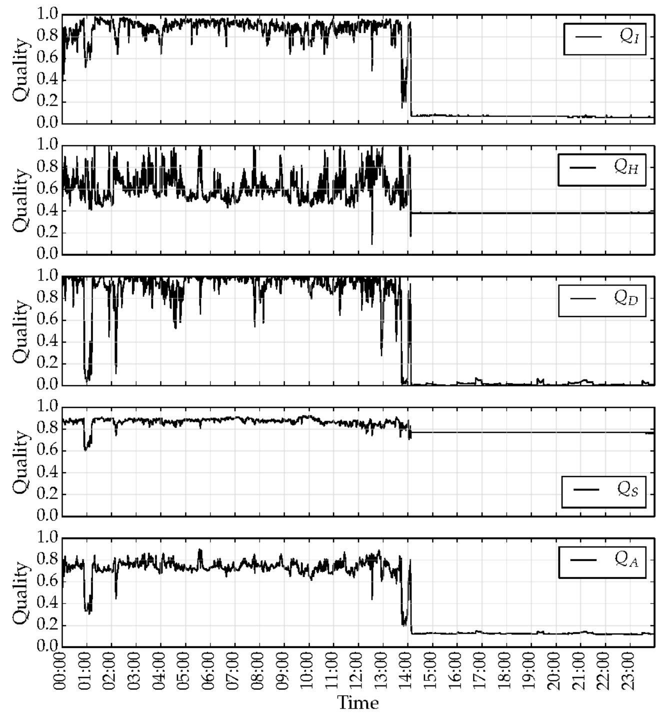

Fig. 8 From top to bottom: Q_I, Q_H, Q_D, Q_S and the final aggregated quality score Q_A. The latter is shown to drop to very low scores compared to the quality calculated in normal operating conditions: the microphone breakdown is thus clearly detected.

4.3 Incipient failure

In order to demonstrate this type of fault, we have chosen an event that occurred in our noise measurement network. Over a period of a few hours, the windshield became detached from the microphone. As illustrated in Fig. 9, during the detachment the $L_{Aeq,1min}$ increased anomalously, afterwards the $L_{Aeq,1min}$ decreased to almost the same sound level as before the detachment, but with a slightly increased noise floor due to the wind-related noise. The Q_D model results in high quality scores during the error period as shown in Fig. 10. This can be explained by the location of the measurement station, which is in a quiet area in close proximity to a wind turbine. As a consequence, no clear diurnal pattern can be observed. For the same reason, Q_H was unable to detect the error. Only the Q_S model was able to detect the error due to the differences in the acoustical cues between the normal operating conditions and the unshielded microphone.

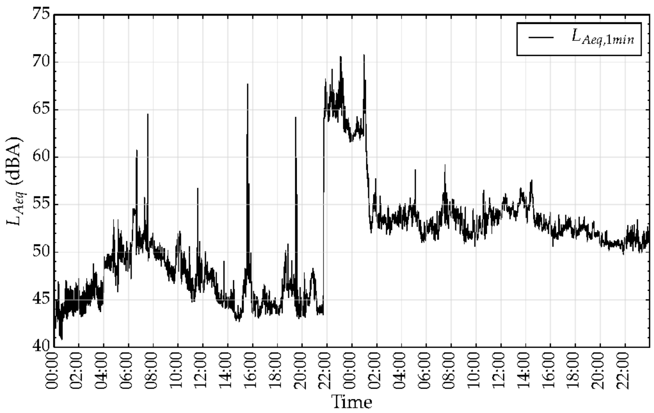

Fig. 9 $L_{Aeq,1min}$ graph illustrating an incipient failure in which the windshield became detached from the microphone, resulting in an increased noise floor after the detachment.

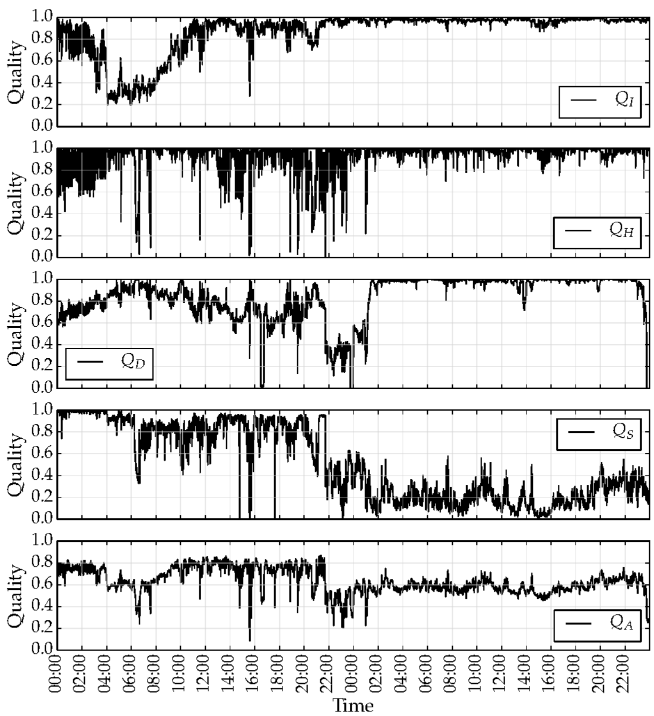

Fig. 10 From top to bottom: Q_I, Q_H, Q_D, Q_S and the final aggregated quality score Q_A. The only model which seems to detect the detachment of the windshield is Q_S, which causes the aggregated quality Q_A to decrease moderately.

4.4 Short term variations in the acoustic environment

Sometimes the anomaly is not related to the measuring system - no failure occurs - but concerns a change in the acoustic environment where the microphone is placed. Such changes have to be considered anomalous because long-term noise exposure indicators can be affected. In this example, atypical high noise levels were measured during working hours as shown in Fig. 11.

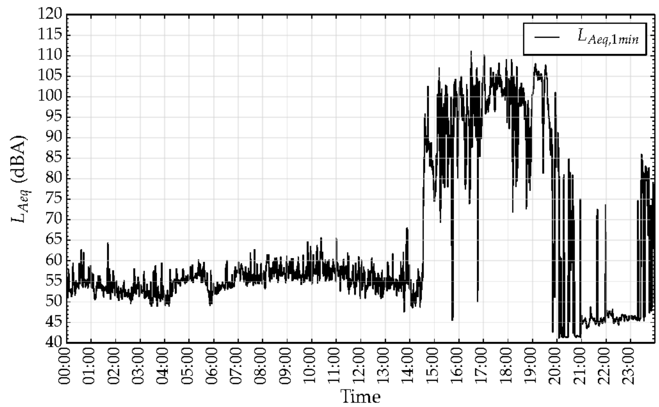

Fig. 11 $L_{Aeq,1min}$ graph illustrating short-term variations in the acoustic environment. Such an anomaly is thus not related to microphone failures.

Due to the high difference with the diurnal pattern and the reference period, the Q_D and Q_H scores are very low during the error period. Moreover, because this type of sound is uncommon for the location, Q_S is able to detect the error as well. The aggregated quality score Q_A is thus very low and the anomaly is detected as shown in Fig. 12.

Note that an operator could still decide to include these measurements if a microphone error can be excluded or the event is considered as part of the normal sound environment. In this example, its origin could be determined: an inhabitant of the house near the measurement location confirmed that construction works near the microphone occurred.

5 Conclusions and discussion

In this paper, we presented a multi-criteria measurement quality-assessment approach for noise-sensor networks. Various strategies were presented for detecting malfunctioning sensors and anomalous data. Data obtained from an extensive noise-monitoring network was used to validate the proposed detector. It was shown that the approach successfully detects sensor breakdowns and anomalies typically observed in long-term measurements.

A major strength of the proposed approach is that no software modification or extra hardware is needed in the sensor network, resulting in a cost-efficient solution. Furthermore, the proposed model is very flexible. The aggregation still works if one or more quality models are not available or if models are added dynamically.

Each quality model has several parameters that should be

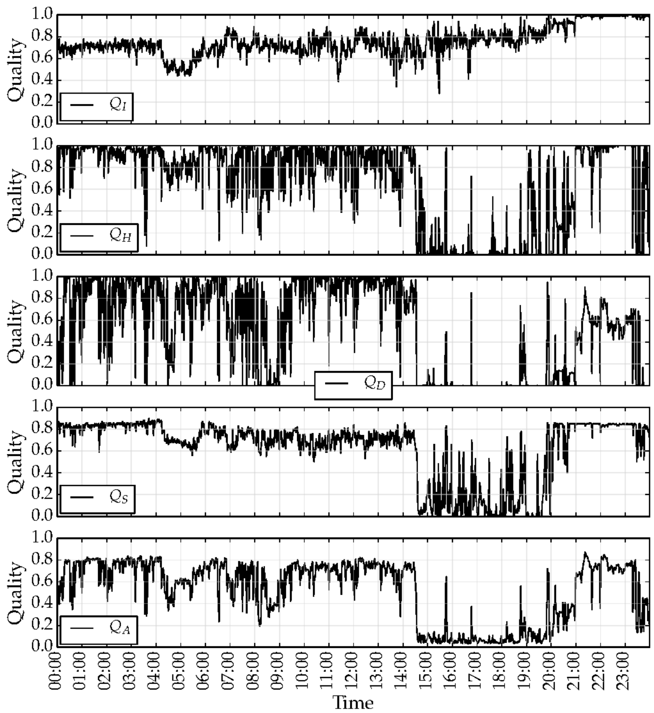

Fig. 12 From top to bottom: Q_I, Q_H, Q_D, Q_S and the final aggregated quality score Q_A. The latter is shown to drop significantly due to the atypical high noise levels, resulting in the successful detection of the anomaly.

chosen carefully. Tuning these parameters over time and according to type of location may be needed. Fully automating the parameter tuning is a challenge, due to the variety of acoustic environment and number of possible sound events. As a consequence a human intervention will still be needed. An expert in the domain can more reliably assess whether the anomaly is related to the measurement system or to changes in the acoustic environment. Such a person should also have the opportunity and knowledge to tune the model parameters to decrease the overall number of false alarms when necessary.

6 Acknowledgements

This research is part of the IDEA (Intelligent, Distributed Environmental Assessment) project, a 4-year strategic basic research project, financially supported by the IWT-Vlaanderen (Flemish Agency for Innovation by Science and Technology).

References

1 W. Babisch, B. Beule, M. Schust, N. Kersten, and H. Ising. Traffic noise and risk of myocardial infarction. *Epidemiology*, 16(1):33–40, 2005.

2 S. A. Stansfeld and M. P. Matheson. Noise pollution: non-auditory effects on health. *British Medical Bulletin*, 68(1):243–257, 2003.

3 A. S. Niskar, S. M. Kieszak, A. E. Holmes, E. Esteban, C. Rubin, and D. J. Brody. Estimated prevalence of noise-induced hearing threshold shifts among children 6 to 19 years of age: the third national health and nutrition examination survey, 1988–1994, united states. *Pediatrics*, 108 (1):40–43, 2001.

4 W. Passchier-Vermeer and W. F. Passchier. Noise exposure and public health. *Environmental health perspectives*, 108(Suppl 1):123, 2000.

5 T. Van Renterghem, P. Thomas, F. Dominguez, S. Dauwe, A. Touhafi, B. Dhoedt, and D. Botteldooren. On the ability of consumer electronics microphones for environmental noise monitoring. *J. Environ. Monit.*, 13: 544–552, 2011. doi: 10.1039/C0EM00532K.

6 V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, July 2009. ISSN 0360-0300. doi: 10.1145/1541880.1541882.

7 K. Ni, N. Ramanathan, M. N. H. Chehade, L. Balzano, S. Nair, S. Zahedi, E. Kohler, G. Pottie, M. Hansen, and M. Srivastava. Sensor network data fault types. *ACM Trans. Sen. Netw.*, 5(3):25:1–25:29, June 2009. ISSN 1550-4859. doi: 10.1145/1525856.1525863.

8 A. B. Sharma, L. Golubchik, and R. Govindan. Sensor faults: Detection methods and prevalence in real-world datasets. *ACM Trans. Sen. Netw.*, 6(3):23:1–23:39, June 2010. ISSN 1550-4859. doi: 10.1145/1754414. 1754419.

9 V. Bychkovskiy, S. Megerian, D. Estrin, and M. Potkonjak. A collaborative approach to in-place sensor calibration. In *In Proceedings of the Second International Workshop on Information Processing in Sensor Networks (IPSN*, pages 301–316, 2003.

10 A. Jabbari, R. Jedermann, and W. Lang. Application of computational intelligence for sensor fault detection and isolation. *International Journal of Computer, Information, and Systems Science, and Engineering*, 1:142–147, 2007.

11 D. Janakiram, V. Adi Mallikarjuna Reddy, and A. Phani Kumar. Outlier detection in wireless sensor networks using bayesian belief networks. In *Communication System Software and Middleware, 2006. Comsware 2006. First International Conference on*, pages 1–6. IEEE, 2006.

12 S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos. Online outlier detection in sensor data using non-parametric models. In *Proceedings of the 32nd international conference on Very large data bases*, pages 187–198. VLDB Endowment, 2006.

13 K. Zhang, S. Shi, H. Gao, and J. Li. Unsupervised outlier detection in sensor networks using aggregation tree. In *Advanced Data Mining and Applications*, pages 158–169. Springer, 2007.

14 Y. Zhang, N. Meratnia, and P. Havinga. Outlier detection techniques for wireless sensor networks: A survey. *Communications Surveys & Tutorials, IEEE*, 12(2):159–170, 2010.

15 P. Henning, K. Ginn, and H. Larsen. A new technique for monitoring condenser microphones. *Sound and Vibration*, 30(12):22–25, DEC 1996. ISSN 0038-1810.

16 G. J. L. H. Ginn, KB. Practical applications of charge injection calibration. In Newman, M, editor, *Proceedings of the 15th international congress on acoustics*, pages 61–64. IUPAP, Acoust Soc Norway, SINTEF, 1995. ISBN 82-595-8995-8. 15th International Congress on Acoustics (ICA 95), Trondheim Norway, Norway, JUN 26-30, 1995.

17 IEC 61094-4:1995, measurement microphones–part 4: Specifications for working standard microphones.

18 PC Engines. PC Engines ALIX 3d3 Product File, 2014. URL `http://www.pcengines.ch/alix3d3.htm`. Online; accessed 15 July 2014.

19 J. Romeu, M. Genescà, T. Pàmies, and S. Jiménez. Street categorization for the estimation of day levels using short-term measurements. *Applied Acoustics*, 72(8):569–577, 2011.

20 D. Oldoni, B. De Coensel, M. Boes, T. Van Renterghem, S. Dauwe, B. De Baets, and D. Botteldooren. Soundscape analysis by means of a neural network-based acoustic summary. 2011.

21 H. Fastl and E. Zwicker. *Psychoacoustics: Facts and models*, volume 22. Springer-Verlag New York Incorporated, 2007.

22 E. Zwicker, H. Fastl, and C. Dallmayr. Basic-program for calculating the loudness of sounds from their 1/3-oct band spectra according to iso 532 b. *Acustica*, 55(63):6, 1984.

23 C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis. Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology*, 15(21):1943–1947, 2005.

24 O. Kalinli and S. Narayanan. Prominence detection using auditory attention cues and task-dependent high level information. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(5):1009–1024, 2009.

25 R. R. Yager. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *Systems, Man and Cybernetics, IEEE Transactions on*, 18(1):183–190, 1988.