Computational Modelling for Soundscape Analysis Inspired by Human Auditory Perception and its Application in Monitoring Networks

Damiano Oldoni

Promotoren: prof. dr. ir. D. Botteldooren, prof. dr. B. De Baets
Proefschrift ingediend tot het behalen van de graad van
Doctor in de Ingenieurswetenschappen: Toegepaste Natuurkunde

Vakgroep Informatietechnologie
Voorzitter: prof. dr. ir. D. De Zutter
Faculteit Ingenieurswetenschappen en Architectuur
Academiejaar 2014 - 2015

UNIVERSITEIT GENT

ISBN 978-90-8578-758-7
NUR 962
Wettelijk depot: D/2015/10.500/2

Computational Modelling for Soundscape Analysis Inspired by Human Auditory Perception and its Application in Monitoring Networks

Damiano Oldoni

Dissertation submitted to obtain the academic degree of
Doctor of Engineering Physics

Supervisor:

prof. dr. ir. Dick Botteldooren
Acoustics group
Department of Information Technology
Faculty of Engineering and Architecture
Ghent University
St.-Pietersnieuwstraat 41
B-9000 Ghent, Belgium
http://acoustics.intec.ugent.be

Supervisor:

prof. dr. Bernard De Baets
KERMIT
Department of Mathematical Modelling, Statistics and Bioinformatics
Faculty of Bioscience Engineering
Ghent University
Coupure Links 653
B-9000 Ghent, Belgium
http://www.kermit.ugent.be

Members of the examining board:

prof. dr. ir. Hendrik Van Landeghem (chairman)	Ghent University, Belgium
prof. dr. ir. Timothy Van Renterghem (secretary)	Ghent University, Belgium
prof. dr. ir. Dick Botteldooren (supervisor)	Ghent University, Belgium
prof. dr. Bernard De Baets (supervisor)	Ghent University, Belgium
prof. dr. Anna Preis	Adam Mickiewicz University, Poland
prof. dr. Frank Witlox	Ghent University, Belgium
prof. dr. ir. Bert De Coensel	Ghent University, Belgium
dr. ir. Michaël Rademaker	Ghent University, Belgium

Science, what a great art!

Anonymous

Acknowledgments

Acknowledgements are at the very beginning of this book, but the last one to be written. The reason is simple: it is tremendously arduous to look at own past, at these almost 6 years spent at Ghent University, and to write about feelings and people. In these pages you will not find any name in order to not write tedious long lists nobody will remember. If you are feeling touched by one of my acknowledgements, then you will see your name in it.

My first thanks go to my supervisors: their knowledge inspired me, their research method guided me, their composure and professionalism under stress surprised me; being appreciated as researcher and individual reassured me and their support during the most decisive and troublesome moments of my PhD was an invaluable gift. There is no research without a research group, professionals from all over the world with the same passion for discovering the depths of knowledge. I am grateful for their commitment and their help: I will never forget it. Finally, there is no work environment without a coffee corner and its "population". The uncountable Italian coffee meetings and the (sometimes very odd) conversations which took place there helped me to enjoy long years at Technicum, definitely not the most human-friendly building. Thanks! I will miss such atmosphere.

Research absorbed a lot of my energies and I am not regretful for that. However, there would be no PhD without making a step out of the work's doors to live a full life. For this reason, I would like to thank my dearest friends here in Belgium and all around the world. We shared ups and downs and each one gave me his/her own special friendship. Things change continuously: new friendships arose, others have been reshaped. Life brought some of them far away making difficult to meet each other. I myself moved to Belgium after all. Friendships are like bottles of wine: you collect them all around the world, the right ones become excellent by time (and distance), others don't survive a winter.

Among all encounters, there is a very particular one which is changing my life in a so unexpected and wonderful way. Thanks for every moment I spent with you, my love. Your advices, your uncertainties, your support, your spontaneity: if this PhD arrived to the end is also thanks to you. You helped me to not overestimate failures and successes, you taught me in first person the

importance of expressing own emotions and understanding the emotions of the others. Maybe your love is imperfect, fragile, tormented, but it's so real and human I cannot ask for anything better. And your big, very big family: what an unexpected, precious gift! I felt me immediately accepted as I am: such trust is not a self-evident truth, thanks for that.

My last, but greatest thanks go to my family, in Italy. If I am now at the point to write such acknowledgements is thanks to you. No words could describe all support and disinterested love you gave me, continuously, during all these years. You have been close to me, at any moment, as nobody else. I will be always grateful and I will never forget your love, wherever I will be, forever.

Damiano Oldoni
Ghent, December 2014

Contents

Samenvatting

Laten we beginnen met een experiment. Antwoord eerst op deze vraag: “Welke geluiden hoort u in de directe omgeving van uw woning?” Maak een lijst en hou deze bij. Wandel de volgende keer wat langzamer door uw buurt en concentreer u op de geluiden die u hoort. Neem nu de lijst die u eerst maakte en vul hem met andere geluiden die u opmerkte. Wat u in dit experiment doet, lijkt triviaal, maar het is de beste manier om bewust te worden van de geluidsomgeving waarin u dagelijks vertoeft. De invloed van deze geluidsomgeving op onze stemming en zelfs onze gezondheid is in de afgelopen decennia grondig onderzocht. Haar rol in de stedelijke planning is toegenomen en kan als even relevant worden beschouwd als andere factoren, zoals visuele esthetiek, veiligheid en mobiliteit.

De geluidsomgeving behoort tot de externe wereld, buiten de luisteraar. Het luisteren naar de geluidsomgeving verwijst naar een perceptuele daad, actief uitgevoerd door de luisteraar. Bij beantwoorden van de eerste vraag uit het experiment komt bovendien retrospectieve beoordeling van wat is gehoord, dus een herinnering aan geluiden, of vanuit neurologisch perspectief, een herinnering aan geluidsgeïnduceerde indrukken, of vanuit nog een ander perspectief, de activiteit van sporen in de hersenschors die externe geluidsstimuli coderen. Daarom is de geluidsomgeving slechts de helft van het verhaal, de andere helft begint bij onze oren, het perifere auditieve systeem, waarvan de signalen in de hersenen worden verwerkt. Om rekening te houden met deze perspectief-verschuiving, van geluidsomgeving naar luisteraar, heeft Schafer in 1969 de term “soundscape” (geluidslandschap) gesuggereerd. De term is ondertussen vastgelegd in een ISO standard: “The soundscape is the acoustic environment as it is perceived and understood by the individual or by society”.

Het is interessant op te merken dat het soundscape concept is ontstaan uit hedendaagse muziekbewegingen. Muzikanten uit de jaren 60 vroegen zich wat muziek eigenlijk is. Emblematisch is het antwoord van John Cage dat, op een bepaalde manier, de basis van het soundscape concept omvat: “Music is sounds, sounds around us whether we’re in or out of concert halls”.

Tijdens het uitvoeren van een wandeling in uw buurt met aandacht voor het omgevingsgeluid, doet u iets gelijkaardigs als de zogenaamde soundscape wandelingen die Schafer en zijn voormalige studenten deden en die omgevings-geluidsdeskundigen tegenwoordig soms doen om informatie over de geluids-

omgeving op bepaalde locaties te verzamelen. De wetenschappelijke wereld begon het werk van Schafer en het belang van de soundscape benadering slechts na enkele decennia te overwegen toen de beperkingen van indicatoren als het equivalent geluidsniveau — een gemiddelde energiedosis —duidelijk werden voor het voorspellen van geluidhinder en de kwaliteit van het geluidsklimaat in de publieke ruimte. Meer gedetailleerde fysische indicatoren werden en worden geïntroduceerd, maar "ergernis" of een "positieve gevoel" zijn een "state of mind" van de luisteraar en kunnen dus niet helemaal correct begrepen worden door enkel de fysieke aspecten van de geluidsomgeving te beschouwen. Daarom is grondig inzicht in soundscape onmogelijk zonder het modelleren van de menselijke auditieve waarneming. Vanuit dit standpunt, beoogt dit proefschrift een algemeen en flexibel computationeel model te vinden dat de geluidsomgeving zo nauwkeurig als een menselijke luisteraar kan analyseren: een kleine stap naar het zogenaamde mens-gebaseerde machinaal luisteren.

Een belangrijke randvoorwaarde in dit proefschrift betreft de computationele efficiëntie en de brede inzetbaarheid van het voorgestelde model. In het bijzonder voor de intrinsieke karakteristieken van het menselijke perifere auditieve systeem zijn gedetailleerde modellen ontwikkeld die leiden tot specifieke eigenschappen zoals gammatone filtering, maar deze zijn niet erg bruikbaar in gedistribueerde stedelijke geluidsmonitoring systemen. Deze doctoraatsthesis heeft dan ook tot doel een computationeel instrument te verschaffen dat (deels) op klassieke meetapparatuur zou kunnen draaien. Dus, om een evenwicht te vinden tussen nauwkeurigheid, biologische plausibiliteit en computationele efficiëntie, wordt ab initio vastgelegd dat het model moet starten van 1/3-octaafband spectrum analyse met een tijdsresolutie van 0.125 s.

Het voorgestelde computationeel model omvat drie stadia: verwerking bij het perifeer auditief systeem, mappen van eigenschappen op basis van gezamenlijk voorkomen via een zelforganiserende kaart en modelleren van aandacht en object creatie door middel van een specifiek artificieel neuraal netwerk. In dit werk wordt in het bijzonder de nadruk gelegd op een nieuwe strategie om de kaart te trainen die continue selectief leren genoemd wordt en die rekening houdt met karakteristieken van leren van geluid bij mensen. Perceptie en retrospectieve evaluatie van geluidslandschappen door mensen hangt immers niet enkel af van de frequentie waarmee de verschillende geluiden voorkomen. Terugkerend naar de hypothetische vraag die in de openingsparagraaf gesteld werd; bevatte uw lijst woorden als stilte of achtergrondgeluid? Waarschijnlijk niet en dit ondanks het feit dat — indien u in een residentiële buurt woont —niet specifiek achtergrondgeluid veel frequenter voorkomt dan duidelijk identificeerbare opvallende geluidsgebeurtenissen. Men heeft inderdaad vastgesteld dat enkel de geluiden

die een waarnemer bewust waarneemt, bijdragen tot het vormen van een mentaal beeld van de geluidsomgeving, de soundscape en uiteindelijk de beoordeling van de kwaliteit ervan. Niet enkel het aanwezige geluid bepaalt welke geluiden waargenomen worden; geluiden die uw aandacht kunnen trekken in bepaalde omstandigheden kunnen volledig genegeerd worden op andere momenten en in een andere context. In het ontwikkelde computationeel model resulteert eenzelfde geluid inderdaad eveneens in een andere respons afhankelijk van de omgeving waarvoor de zelforganiserende kaart getraind is.

Auditieve aandacht speelt naast leren eveneens een centrale rol in het menselijk auditief systeem. Het laat toe enkel relevante en noodzakelijke informatie door te geven aan het werkgeheugen. Daarom is het noodzakelijk om auditieve aandacht te modelleren in elk systeem met de ambitie het menselijke auditieve waarneming te imiteren. Aandacht wordt typisch gemodelleerd als een contributie van twee mechanismen: “bottom-up” en “top-down”, ook wel inwaarts en uitwaarts georiënteerde aandacht genoemd. Bottom-up aandacht wordt bepaald door de opvallendheid van het geluid: zeldzame en opvallende fysische karakteristieken of instinctief biologisch belang. Bottom-up aandacht onderzoekt de geluidsomgeving naar veranderingen in intensiteit, frequentieïnhoud, of ruimtelijke locatie. De lijst die u stelde zullen normaal gezien vooral geluiden bevatten die sterk opvallen in de geluidsomgeving. Top-down aandacht wordt daarentegen gedreven door de taak die de luisteraar op dat moment uitvoert en kent cognitieve resources toe aan de geluiden die belangrijk zijn voor het volbrengen van deze taak of op zijn minst aan deze taak gerelateerd zijn. Bijvoorbeeld, tijdens het koken zal men meer aandacht hebben voor het geluid van de pruttelende gerechten dan voor het geluid van auto’s die voor de gevel passeren. Bottom-up en top-down mechanismes zijn voortdurend in competitie om aandacht op de juiste geluiden te richten en sensorieel overladen te vermijden. Deze competitie is hiërarchisch gestructureerd en komt op verschillende niveaus van abstractie voor: van lage niveaus waar competitie tussen neurale representaties van basis kenmerken van het geluid worden geanalyseerd tot op het hoogste niveau waar competitie tussen auditieve stromen en objecten optreedt. Op het hoogste niveau komt daar nog de competitie of versterking door interactie tussen de verschillende zintuigen bij. Een belangrijk mechanisme dat het aandachtsproces beïnvloedt is inhibitie van terugkeer. Dit mechanisme verhindert dat aandacht gericht blijft op een steeds weerkerend geluid en veroorzaakt een natuurlijk afzoeken van de auditieve scène. Top-down, taakgerelateerde aandacht kan dit proces wijzigen waardoor het congitieve systeem gefocusseerd kan blijven op een bepaalde stimulus.

Alle bovenvermelde kennis wordt verwerkt in een computationeel aandachts-

model dat gebaseerd is op een artificieel neuraal netwerk. Dit model wordt gesuperposeerd op de zelforganiserende kaart. Elke cel in de zelforganiserende kaart wordt gekoppeld aan een neuron van het artificieel neuraal netwerk dat aandacht modelleert. De langetermijn plasticiteit van de auditieve cortex wordt vertaald naar een traag lerende zelforganiserende kaart terwijl de snellere aandachtsmechanismen door het neurale netwerk worden gemodelleerd. Zoals vermeld is selectieve aandacht ook belangrijk bij het selecteren (en soms vormen) van auditieve objecten. In dit werk wordt een alternatief model voorgesteld voor het vormen van auditieve objecten: koppelen van de zelforganiserende kaart aan een oscillerend neuraal netwerk dat de dynamische oscillator koppeling tussen sensorische cortex neuronen geëxciteerd door een auditieve stimulus modelleert.

In dit werk worden eveneens verschillende toepassingen van de ontwikkelde modellen gepresenteerd. De eerste toepassing betreft het automatisch selecteren en opnemen van een verzameling van typische geluiden voor een buurt, "akoestische samenvatting" genoemd. Er wordt aangetoond dat deze compacte verzameling van geluiden het geluidsklimaat van een locatie karakteriseert volgens mensen die in deze buurt wonen. Met andere woorden, het theoretisch model beantwoordt de vraag "welke geluiden hoor je in de directe omgeving van je woning?". Deze soundscape analyse kan gebruikt worden als startpunt van de ontwikkeling van een geluidslandscahp dat de perceptie van de geluidsomgeving door de gebruikers van een ruimte verbetert. Dit verbeterd ontwerp wordt typisch gerealiseerd door aangename en gewenste geluiden toe te voegen of te accentueren zodat ongewenste geluiden gemaskeerd worden of op zijn minst niet langer in de focus van de aandacht liggen. Er wordt aangetoond dat de voorgestelde modellen gebruikt kunnen worden om de effecten van mogelijke interventies te beoordelen. Hierdoor is het niet langer nodig gebruik te maken van dure luisterpanels om de effectiviteit van maatregelen te beoordelen. Betekenis hechten aan de waargenomen geluiden — het proces dat toelaat de geluidservaring met woorden te beschrijven —wordt in het voorgestelde computationele model niet expliciet ingebouwd. Desalniettemin gaat een tweede toepassingsvoorbeeld hier verder op in door een support vector machine te koppelen aan de zelforganiserende kaart. Met deze methode kunnen bijvoorbeeld de geluiden in de akoestische samenvatting van een label voorzien worden. Tot slot bespreekt dit werk een aantal andere, complementaire toepassingen: herkennen van specifieke geluidsgebeurtenissen en detecteren van atypische geluiden als deel van een multicriteria aanpak voor anomalie en fout detectie in stedelijke sensornetwerken.

Een bespreking van de belangrijkste bevindingen en mogelijke verdere ontwikkelingen besluiten deze scriptie. Verschillende aspecten van menselijke

auditieve perceptie werden niet beschouwd of slechts in een zeer vereenvoudigde vorm meegenomen. Bijvoorbeeld, binaurale effecten werden buiten beschouwing gelaten, gesuperviseerde training of reïnforced leren kwamen niet aan bod. Het complexe cognitieve proces dat geluiden identificeert en er betekenis aan toekent, kon niet in de rekenmodellen opgenomen worden; de support vector machine aanpak lijkt te weinig performant in vergelijking tot de menselijke competenties. Dit zijn maar enkele voorbeelden die moeten aantonen dat het pad naar een computationeel haalbaar en volledig omvattende mens-achtige machinale luisteraar nog steeds een uitdaging vormt. Hopelijk heeft dit werk dit pad een stap korter gemaakt.

Summary

Let us start with an experiment. First, answer this question: "Which sounds do you hear in the direct surroundings of your home?" Make a list and take it with you. Then, next time you are near home, try to walk slower and focus on the sounds you hear. In case you want to add some sounds, take the list out and complete it. This little experiment might seem somewhat trivial. However, it is the best way to become self-conscious about the sonic environment you experience every day. Its influence on our mood and even our health has been very well studied in the last decades and its role in the urban planning process has grown a lot, so that it can now be considered as relevant as, for example, visual aesthetics, safety and mobility.

The sonic environment is related to the external world, outside of the listener's head. However, listening to the sonic environment refers to a perceptual act, actively performed by the listener. Additionally, answering the initial question involves a retrospective assessment of what has been heard, thus involving a recollection of sounds, or, from another perspective, a recollection of sound-induced impressions, or, from yet another perspective, memory traces in the brain cortex encoding external sound stimuli. The sonic environment is therefore only half of the story. The other half starts from our ears, the peripheral auditory system, whose signals are further processed in the brain. In order to account for this shift in perspective, from the sonic environment to the listener, the term soundscape has been suggested in the 1969 by Schafer and defined in a ISO standard: "The soundscape is the acoustic environment as it is perceived and understood by the individual or by society".

It is interesting to notice that the concept of soundscape arose from trends in contemporary music. Musicians of the 1960s wondered what music actually is. Emblematic is the answer given by John Cage, in some way containing the seed of the soundscape approach: "Music is sounds, sounds around us whether we're in or out of concert halls".

Walking in the surroundings of your home, paying attention to the sounds you hear, you are doing something very similar to the so-called soundwalks that Schafer and his former students performed and that environmental acousticians are sometimes doing nowadays to collect information about soundscapes in given locations. In fact, the scientific community started to consider the work of

Schafer and the importance of the soundscape approach only a few decades later, when indicators such as the energy equivalent sound pressure level — an averaged energy level —began to show limitations in the prediction of noise annoyance and the quality of the sound climate in public areas. Other indicators, with the same goal, were and are introduced. But what is certain is that "annoyance" or a "positive feeling" are a "state of mind" and cannot be completely understood just by considering the physical aspects of the sonic environment. Therefore, analyzing the soundscape is impossible without modelling human auditory perception. This statement is the starting point of this dissertation, which aims to find a general and flexible computational model able to analyse the sonic environment as accurately as a human listener would: a small step towards the so-called human-like machine listening.

An important precondition of this dissertation refers to the computational efficiency of the proposed model. In particular, several detailed methods have been conceived for the intrinsic properties of the human peripheral auditory system like gammatone filtering, but they cannot be easily used in distributed urban monitoring systems. Therefore, this dissertation aims to provide a computational tool that could ultimately (if only partly) run on classic measurement equipment. Thus, in order to find a balance between accuracy, biological plausibility and computational efficiency, the presented model is based on a 1/3-octave band spectrum analysis at 0.125 s.

The proposed computation model is composed of three stages: peripheral auditory processing, mapping of acoustical features based on co-occurrence by means of a self-organizing map and modelling auditory attention and auditory object creation by means of specific neural networks. In particular, this work has focussed on a novel strategy to train the map, called continuous selective learning, which accounts for aspects of human auditory learning. In fact, human perception and retrospective assessment of soundscapes do not depend exclusively on the rate of occurrence of sounds that are heard. Returning to our experiment: would expressions as silence, background noise or quietness be mentioned in your list? Probably not, although — in case you live in a residential area —not specific background sound occurs more often than identifiable and noticeable sound events. In fact, only the sounds that the listener consciously notices will contribute to the creation of a mental image of the sonic environment, the listener's soundscape, and ultimately will shape his/her perception of its quality. Not only the physical sound features determine which sounds are perceived; sounds that could attract our attention in certain circumstances would be totally ignored at other moments or in other contexts. In the developed computational model, the same sound input would result

in different model outputs depending on the sound environment where the self-organized map was trained.

Together with learning, auditory attention (inter)plays a central role in human auditory perception. In fact, it allows us to select the information needed to be passed on to the working memory. For this reason, modelling the auditory attention is of the utmost importance in every system aiming to imitate human auditory perception. Typically, attention is modeled as the contribution of two kinds of mechanisms, respectively called bottom-up and top-down mechanisms. Bottom-up attention is related to the conspicuity of the sound: rare or novel physical features or instinctive biological importance. Bottom-up mechanisms perform a novelty detection task, thus monitoring the acoustic environment for changes in intensity, frequency or spatial location of the sound stimuli. Typically, the list of sounds you would come up contains mostly these conspicuous sounds. At the other hand, top-down attention is driven by the task performed at that moment by the listener and focus cognitive resources on sounds that are important for accomplishing such task or are at least task-related. For example, if you are cooking, you would most likely pay attention on the simmering food instead of sound coming from cars passing by in front of your house. Bottom-up and top-down mechanisms interact continuously in a competitive selection process in order to direct attention on the right sounds and avoid sensory overload. This competition is also hierarchically structured and acts at different levels of abstraction: from low levels, where competition among neural representations of basic sound features occurs, up to the highest cognitive levels where competition among different auditory streams occurs. The last level would be the competition among the information from the different senses, eventually arising in cross-sensory reinforcements. An important mechanism modulating attentional process is the so-called inhibition-of-return. It prevents attention to be continuously focused on the same auditory stream, thus naturally generating an attentional scan of the auditory scene. Top-down, task-related attention mechanisms can alter this process in order to keep the working memory focused on the desired task.

This empirical knowledge is transferred into a computational attention model based on an artificial neural network that takes into account both bottom-up and top-down mechanisms and includes inhibition-of-return as well. Such a model can be seen as a second layer, superimposed on the self-organizing map. Each unit of the self-organizing map is coupled to a unit in the neural network modelling attention. The long-term plasticity of the auditory cortex is modeled by a slow learning self-organizing map, while the faster attention mechanisms by the neural network. As mentioned before, attention is important in selecting

(and in some cases forming) auditory objects. This dissertation also proposes an alternative model of auditory object formation by coupling the self-organizing map with an oscillatory neural network, which models the dynamic oscillatory correlation of sensory cortex neurons excited by an auditory stimulus.

This dissertation also presents several applications of the theoretical model. The first one is the automated selection and recording of a collection of typical sounds at a given location, called "acoustic summary". It is shown that this is a comprehensive set of sounds characterizing the specific location as judged by people living in the surroundings. In other words, the theoretical model presented here answers that very first question "Which sounds do you hear in the direct surroundings of your home?". Such a soundscape analysis can be used as the starting point for the design of a soundscape improving the auditory scene perception experienced by the users of a space. This improved soundscape is typically achieved by adding or accentuating pleasant and desirable sounds in order to mask or, at least, shift the listener's attention from undesired sounds. It is shown that the proposed model can be used to evaluate the perceptual effects of possible interventions, thus removing the need of expensive listening panels to assess their effectiveness. Attaching a meaning to sounds— the process that lets you describe with words the experienced sounds —is not tackled by the computational model proposed here. However, a second application is presented which aims to find a simple solution to this issue by using a support vector machine linked to a trained self-organizing map. This method can be used, for example, to label the sound excerpts composing an acoustic summary. Finally, the dissertation also discusses some other complementary applications of the model: the recognition of particular known sounds, i.e. specific sound event recognition, and the detection of atypical sounds as part of a multi-criteria approach for anomaly and failure detection in urban sensor networks.

The dissertation is concluded by discussion of the main findings and possible improvements. In fact, several aspects of the human auditory perception have been either simplified or not considered in the proposed model. For example, no binaural effects were taken into account, and no supervised and reinforcement learning strategies were modeled in the training of the self-organizing map. Moreover, the complexity of the cognitive act of identifying a sound and attaching a meaning to it could not be included in the theoretical model and the coupling of the model with a support vector machine appears to be too simplistic when compared to human proficiency. These are just a few examples showing that the path towards a computationally feasible and totally comprehensive human-like machine listening remains challenging. Hopefully, this work made it a step shorter.

List of Abbreviations

ANN	Artificial Neural Network
BMU	Best-Matching Unit
CASA	Computational Auditory Scene Analysis
CSL	Continuous Selective Learning
CWT	Continuous Wavelet Transform
ERP	Event-Related Potentials
FN	False Negative
FP	False Positive
FPR	False Positive Ratio
IDEA	Intelligent Distributed Environmental Assessment
INTEC	Information Technology Department
IOR	Inhibition of Return
ISO	International Organization for Standardization
IEEE	Institute of Electrical and Electronics Engineers
LEGION	Locally Excitatory Globally Inhibitory Oscillator Network
MAS	Multi-Agent System
MFCCs	Mel-Frequency Cepstral Coefficients
O&M	Observations and Measurements
OGC	Open Geospatial Consortium
OWA	Ordered Weighted Average
SNR	Signal-to-Noise Ratio
SOM	Self-Organizing Map
STPL	Short-Term Partial Loudness
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TPR	True Positive Ratio
UGent	Ghent University

List of Symbols

Commonly used noise indices

L_{Aeq}	A-weighted Energy Equivalent Sound Pressure Level [dB(A)]
L_{day}	Day Level [dB(A)]
L_{evening}	Evening Level [dB(A)]
L_{night}	Night Level [dB(A)]
L_{den}	Day-Evening-Night Level [dB(A)]

Symbols related to the sound feature extraction model

$s(f,t)$	Simplified Cochleagram
f	Spectral Value [Bark]
$g(f,t)$	Set of 2D Gaussian and Difference-of-Gaussian Filters
r_i	Convolution of the Cochleagram with the $i-$th Filter
$\vec{e}$	Standard Basis Vector in Euclidean Space

Symbols related to SOM and its training

$\vec{r}$	Sound Feature Vector
M_x, M_y	x and y Dimensions of the SOM Lattice
M	Total Number of SOM Units
$\vec{\mathbf{m}}_i$	SOM Unit Position in the $2D$ Lattice
c	SOM Best-Matching Unit Index
$\vec{m}_i$	Reference Vector of SOM Unit i
h	Neighbourhood Function
α	Learning Rate
α_0	Initial Learning Rate
σ	Width of the 2D Neighbourhood Function
σ_0	Initial Width of the 2D Neighbourhood Function
N	Number of Samples Used for the Training
T_{up}	Activation Threshold

T_{down}	Deactivation Threshold
s	Measure of Overall Auditory Saliency
E	Average Quantization Error

Symbols related to LEGION and SOM-LEGION coupling

x_i	Excitatory Unit of the $i-$th LEGION Oscillator
y_i	Inhibitory Unit of the $i-$th LEGION Oscillator
I_i	External Stimulation of the $i-$th LEGION Oscillator
I	Positive Constant
H	Heaviside Function
p_i	Lateral Potential of the $i-$th LEGION Oscillator
ϕ	Threshold
h	Neighbourhood Function
S_i	Overall Coupling Contribution from Near Oscillators
ρ	Source of Gaussian Noise
γ	Regulating Parameter
ϵ	Regulating Parameter
β	Regulating Parameter
T	Permanent Connection Weights
T_{max}	Maximal Permanent Connection Weight
M_h	$h-$order Simple Moving Average of the Inverse of the Distance of the BMU
λ	Relative Threshold
t_L	Internal Time or LEGION Time
W	Dynamic Connection Weights
W_T	Total Dynamic Connection Weight
z	Global Inhibitor
W_z	Global Inhibitor Weight
θ	Global Inhibitor Threshold
u_i	Variable Measuring whether the Oscillator i is Externally Stimulated or not
η	Constant
ν	Constant
ω	Parameter
δ	Similarity of two Neighbouring SOM Units
ϕ	Scaling Factor

Symbols related to the Artificial Neural Network

IOR	Inhibition-of-return
E	Excitation
EL	Local Excitation
IG	Global Inhibitor
A	Activation

Symbols related to the fuzzy excitation model

S	Similarity Measure
E	Excitation of the SOM Units
d	Distance of the Sound Feature Vector to a SOM Unit
$\tilde{d}$	Distance of the Sound Feature Vector to the BMU Unit
k	Positive Constant
t_i	Duration of the $i-$th Target Sound
E_i	Excitation Map Related to the $i-$th Target Sound
$\tilde{E_h}$	Excitation Prototypical Map Related to the $h-$th Cluster
$\bar{t_h}$	Prototypical Sound Event Duration Related to the $h-$th Cluster
T	Dynamic Threshold

Symbols introduced in Section 3.5

Q_S	Quality Score Calculated by SOM-based Quality Model
d_{BMU}	Distance to the BMU
T_{mid}	Distance to the BMU for which the Quality Score Equals to 0.5
Q_I, Q_H, Q_D	Other Quality Scores
Q_A	Aggregated Quality Score
$\mathbf{W}$	Quality Vector Composed of the 1-minute based Quality Scores
$\tilde{\mathbf{Q}}$	Ordered Quality Vector
α	Quantifier of the OWA Operator

Symbols introduced in Appendix C

c_i	Value Combining Saliency and Frequency of Occurrence of the $i-$th SOM Unit
o_i	Number of Time Steps the $i-$th SOM Unit is the BMU
β_{occ}, β_{sal}	Positive Weighting Coefficients
Y	Number of Played Sounds
a	Vector Containing the Coefficients of the Regression Model
a_1, a_2	Coefficients of the Regression Model
b	Constant Term of the Regression
x	Two-Dimensional Categorical Variable Coding for the Different Acoustic Summary Criterion
H_0	Null Hypothesis
R^2	Adjusted $R-$Square
F	F Statistic
p	p Value
FP	Number of False Positives
TP	Number of True Positives
FN	Number of False Negatives
TN	Number of True Negatives
std	Standard Deviation

List of Publications

Articles in international journals

- A. Bockstael, L. Dekoninck, A. Can, D. Oldoni, B. De Coensel, and D. Botteldooren, Reduction of wind turbine noise annoyance: an operational approach, *Acta Acustica united with Acustica*, Vol. 98, no. 3, pp. 392-401, 2012

- D. Oldoni, B. De Coensel, M. Boes, M. Rademaker, B. De Baets, T. Van Renterghem, and D. Botteldooren, A computational model of auditory attention for use in soundscape research, *Journal of the Aocustical Society of America*, Vol. 134, no. 1, pp. 852-861, 2013

- D. Oldoni, B. De Coensel, A. Bockstael, M. Boes, B. De Baets, and D. Botteldooren, The acoustic summary as a tool for representing urban soundscapes, *Landscape and Urban Planning*, (in revision)

Abstracts in international journals

- D. Botteldooren, D. Oldoni, and B. De Coensel, Acoustic summary as a tool for soundscape analysis and design, *Journal of the Aocustical Society of America*, Vol. 128, no. 4, p. 2370, 2010. Presented at *The 2nd Pan-American/Iberian Meeting on Acoustics*, Cancún, Mexico, Nov. 2010

Articles in peer-reviewed conference proceedings

- D. Oldoni, B. De Coensel, M. Rademaker, T. Van Renterghem, B. De Baets, and D. Botteldooren, Context-dependent environmental sound monitoring using SOM coupled with LEGION, *Proceedings of the IEEE International Joint Conference on Neural Networks*, Barcelona, Spain, Jul. 2010

- M. Boes, D. Oldoni, B. De Coensel, and D. Botteldooren, Attention-driven auditory stream segregation using a SOM coupled with an excitatory-

inhibitory ANN, *Proceedings of the IEEE International Joint Conference on Neural Networks*, Brisbane, Australia, Jun. 2012

- M. Boes, D. Oldoni, B. De Coensel, and D. Botteldooren, A biologically inspired recurrent neural network for sound source recognition incorporating auditory attention, *Proceedings of the IEEE International Joint Conference on Neural Networks*, Dallas, USA, Aug. 2013

Articles in other conference proceedings

- D. Botteldooren, D. Oldoni, and B. De Coensel, Human-mimicking environmental sound measurement, *MINET Conference : Measurement, Sensation and Cognition: Measuring the Impossible*, London, UK, Nov. 2009

- D. Oldoni, B. De Coensel, M. Rademaker, T. Van Renterghem, B. De Baets, and D. Botteldooren, Imitating human auditory processing for urban soundscape measurement, *Proceedings of the Institute of Acoustics & Belgium Acoustical Society*, Ghent, Belgium, Apr. 2010

- T. Van Renterghem, D. Oldoni, and D. Botteldooren, Sound propagation in a park over the year, *Proceedings of the 39th International Congress and Exposition on Noise Control Engineering (Inter-Noise 2010)*, Lisbon, Portugal, Jun. 2010

- D. Oldoni, B. De Coensel, M. Rademaker, T. Van Renterghem, D. Botteldooren, and B. De Baets, Computational soundscape analysis based on a human-like auditory processing model, *Proceedings of the EAA EuroRegio 2010 congress*, Ljubljana, Slovenia, Sept. 2010

- A. Bockstael, L. Dekoninck, B. De Coensel, D. Oldoni, A. Can, and D. Botteldooren, Wind turbine noise: annoyance and alternative exposure indicators, *Proceedings of Forum Acusticum 2011*, Aalborg, Denmark, Jun. 2011

- D. Botteldooren, B. De Coensel, D. Oldoni, M. Boes, and P. Lercher, Biologically inspired modeling of environmental noise perception, *Proceedings of the Institute of Acoustics*, from *10th International Congress on Noise as a Public Health Problem*, London, UK, Jul. 2011

- 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, *Proceedings of the 40th International Congress and Exposition on Noise Control Engineering (Inter-Noise 2011)*, Osaka, Japan, Sept. 2011

- M. Boes, B. De Coensel, D. Oldoni, and D. Botteldooren, A biologically inspired model adding binaural aspects to soundscape analysis, *Proceedings of the 40th International Congress and Exposition on Noise Control Engineering (Inter-Noise 2011)*, Osaka, Japan, Sept. 2011

- D. Botteldooren, B. De Coensel, D. Oldoni, T. Van Renterghem, and S. Dauwe, Sound monitoring networks new style, *Acoustics 2011: Proceedings of the Annual Conference of the Australian Acoustical Society*, Queensland, Australia, Nov. 2011

- D. Botteldooren, M. Boes, D. Oldoni, and B. De Coensel, The role of paying attention to sounds in soundscape perception, *Proceedings of the Acoustics 2012 Hong Kong Conference*, Hong Kong, China, May 2012

- D. Oldoni, B. De Coensel, M. Boes, T. Van Renterghem, and D. Botteldooren, A computational auditory attention model for urban soundscape design, *Proceedings of the 41st International Congress and Exposition on Noise Control Engineering (Inter-Noise 2012)*, New York, NY, USA, Aug. 2012

- X. Valero, D. Oldoni, D. Botteldooren, and F. AlÃas, Support vector machines and self-organizing maps for the recognition of sound events in urban soundscapes, *Proceedings of the 41st International Congress and Exposition on Noise Control Engineering (Inter-Noise 2012)*, New York, NY, USA, Aug. 2012

- X. Valero, D. Oldoni, D. Botteldooren, and F. AlÃas, Sound event recognition in urban soundscapes with self-organizing maps and support vector machines, *Soundscape of European Cities and Landscapes*, Merano, Italy, Mar., 2013

- B. De Coensel, M. Boes, D. Oldoni, and D. Botteldooren, Characterizing the soundscape of tranquil urban spaces, *Proceedings of Meetings on Acoustics*, from *International Congress on Acoustics*, Montreal, Canada, Jun., 2013

- D. Botteldooren, T. Van Renterghem, D. Oldoni, S. Dauwe, L. Dekoninck, P. Thomas, W. Weigang, M. Boes, B. De Coensel, B. De Baets, and B. Dhoedt, The internet of sound observatories, *Proceedings of Meetings*

on Acoustics, from *International Congress on Acoustics*, Montreal, Canada, Jun., 2013

Computational Modelling for Soundscape Analysis Inspired by Human Auditory Perception and its Application in Monitoring Networks

Chapter 1

From human auditory perception to machine listening

In this introductory chapter the basic principles of human auditory perception are laid before. In this context, the soundscape approach is introduced. Important aspects of human auditory perception such as auditory scene analysis, auditory attention and learning are discussed. Finally, machine listening is introduced as an attempt to synthesize the knowledge on human auditory perception in computational models and particular attention is given to the novelty of the approach presented in this work.

1.1 Perception, sensation and computation

Human beings have always been concerned with understanding how the external world is represented in consciousness, the so-called *perceptual problem*, and how the act of perception can affect their emotions. An immediate extension, called the *correspondence problem*, is how accurate the perception represents the external reality considered the fact that we perceive the external world via intrinsic limited senses. It should thus not surprise that perception has been a primary concern in many philosophical theories. The philosophical school linked to Plato considered the sensory inputs as an inaccurate copy of the external world, eventually corrected by the Reason, thus providing us a perfect representation of it. Such view was several centuries afterwards extended by the German philosopher Immanuel Kant: the intellect creates the perceived phenomena, based on whatever inadequate information is provided by the

sensory systems, by means of the innate primary concept of time and space and several innate categories defining object quantities, qualities and the relations among such objects. Therefore, studying perception means studying an aspect of cognition and reason, noticeably reducing the importance of studying the sensory systems. This school of thought has been criticized since the time of Plato by the philosophical tradition called *sensism* or *sensualism*. Aristippo, a contemporary of Plato, argued that the senses are enough accurate to represent the external world and therefore there is no need to correct them by the platonic Reason; before him the presocratic Protagoras said that "Man is nothing but a bundle of sensations". This view was rephrased by Hobbes in the 17th century and later by Locke. In their works the importance of the sensory inputs is the key for understanding the mind. As Hobbes wrote:

> there is no conception in a man's mind which hath not at first, totally or by parts, been begotten upon the organs of sense (1).

Locke went further conceiving the mind as a *tabula rasa*, with no innate categories or rules, a kind of white paper written by sensory processes. In this view studying the perception means studying the sensation and the afferent sensory systems.

The modern and scientific approach to sensory perception seems the result of a third thought of school influenced by both the two extreme opposing positions. Aristotle was the first who tried at compromise followed by St. Thomas Aquinas. In more recent times, the compromise led to the separation of *sensation* and *perception*, as first introduced by Thomas Reid (2). Such distinction is still nowadays broadly used and it is commonly accepted in psychology. The term sensation refers to the stimulation of a sensory receptor which converts energy into a nerve signal reaching the brain. In case of audition, the energy contained in the sound pressure waves is transferred by the eardrum and the middle ear to the cochlea where the hair cells transform it into a nerve signal reaching the brain by means of the auditory nerve. Perception is an intellectual process referring to the selection, organization and interpretation of the sensory inputs. It is worth noting that Reid, as his predecessor Berkeley, was very influenced in his work by Newton's inductive ascent reasoning in his attempt of ascending towards more unifying and general laws (3). The scientifically-based study of sensation and perception is however due to the work of Hermann von Helmholtz, considered one of the pioneers of perceptual psychology. He was indeed one of the first to use his insights on nerve physiology and ophthalmology to formulate empirical theories of depth perception, colour and motion perception and to infer from them a more general perceptual theory. In his works about theory of vision (4) he introduced the concept of *unconscious conclusion* or *unconscious*

inference to describe the reflex-like mechanism by which perception is formed. The inputs coming from the sensory systems, the sensation, are pre-rationally worked out by humans to form a percept. However, such unconscious process is not regulated by absolute categories as in Plato or Kant's philosophy: individual (as cultural and sociological) differences can lead to different percepts. The Helmholtz's formulation of unconscious inference implies also that perception has an important developmental and behavioural aspect: the influence of different past experiences can eventually end up in different percepts of the same sensory inputs.

Helmholtz's great contribution to perception is not limited to vision: his book "On the Sensations of Tone as a Physiological Basis for the Theory of Music" (5) in the 1862, together with the contemporary work of Fechner (6), is a milestone of the not yet born disciplines called psychoacoustics and musical psychology. In the twentieth century the progress in fields as neurology, audiology and psychology provided decisive insights on how humans perceive sounds, focussing on both sensation and perception and the relation between them. Such progress was also possible thanks to the birth of computer science in the 1930s and the following tremendous technological development. But the increase of the calculus performance resulted also in an increase in theoretical research on possible computational models simulating biological mechanisms, (auditory) perception and sensation included. Insights in psychoacoustics started thus to be included in theoretical computational models: the history of A-weighting is likely the first tangible example. Fletcher and Munson (7) introduced in the early 1930s the psychoacoustic concept of loudness defined as —"the magnitude of auditory sensations" and determined experimentally the first equal-loudness contours (the so-called Fletcher-Munson curves), thus characterizing the non-linear relationship in the frequency domain between sound pressure level and loudness. Three years later, in 1936, these findings could already be introduced by means of the A-weighting curve (the 40 phon Fletcher-Munson curve) in the first standards for sound level meters (8). Sometimes biological mechanisms inspire computational models which go further the primary scope of modelling them. Artificial neural networks (ANNs), for example, have been conceived since the very beginning, in the 1940s and 1950s, as computational counterparts of biological neurons and still nowadays are fundamental in modelling human brain. However, ANNs showed very soon their potential in solving a wide variety of tasks of machine learning. Therefore, ANNs have been often used in many sectors where their biological plausibility was not relevant. The transfer of knowledge about human auditory perception into computational models will be taken into account in Section 1.6.

1.2 Reductionism and the soundscape approach

Understanding the physiological and psychological mechanisms underlying the human auditory perception is the final goal of psychoacoustics, already mentioned in the previous section. This branch of science uses typically a reductionist approach for modelling the many aspects of human hearing. The equal-loudness contours mentioned before, for example, used pure tones, energetic masking was accurately studied by means of artificial sounds, such as sequences of tones or broadband noises (9), or using speech (10) and the experiments were always performed in laboratory conditions.

The same reductionist approach is overall used in environmental acoustics too. Traffic noise, for example, is typically assessed as the sum of the noise produced by all vehicles, typically modelled as different one-dimensional noise sources: engine, tires, exhaust and aerodynamic noise. Many parameters are used, as type of vehicle, speed, type of road surface and driving behaviour. Therefore, empirical equations have been developed within international project studies like RoTraNoMo, Harmonoise and Nord2000. Reductionism is also applied in studying sound propagation and noise control engineering, resulting in international standards (ISO 9613-1 (11) and ISO 9613-2 (12)). Even classic noise annoyance assessment and noise abatement policies tend to evaluate the quality of a sonic environment based on overall A-weighted noise level, L_{den}, thus trying to reduce the intrinsic complexity of the studied issue. However, this approach has received several critics (13; 14) as many other factors contribute to the quality of a sonic environment: spectral (15) and temporal structure (16) are essential descriptors and many other indicators will be added likely in the future. Researchers started thus to wonder whether such reductionist approach will ever be sufficient: a paradigm shift from noise annoyance to sound quality occurred (17) and a more holistic and qualitative approach for assessing the quality of the sonic environment started to be used (18; 19). Such approach, called the *soundscape approach* or *soundscaping* (20) is however not new.

The word and concept of soundscape became popular at first in the music world thanks to the composer and environmentalist R. Murray Schafer and his books: *The New Soundscape* (21) and *The Tuning of the World* (22), respectively from 1969 and 1977. However, no exact definition of soundscape is given in such books, probably supposing the analogy with landscape would be sufficiently clear to the reader. The first definition seems to be the following one from the *Handbook for Acoustic Ecology* (23):

> An environment of sound (or sonic environment) with emphasis on the way it is perceived and understood by the individual, or by a

> society.

It seemed the community of acousticians needed musicians such as Schafer to make this shift of perspective. In the 60s, in fact, musicians started to reflect about what is music. A very emblematic answer has been given by avant-garde composer John Cage and was reported by Schafer (21):

> Music is sounds, sounds around us whether we're in or out of concert halls.

Such a definition can in some way be considered as the seed the soundscape approach was born from. Musicians started going out music halls to listen to the sounds produced by nature, man or machines: the soundwalk method was thus born and became a common practice in the context of the Schafer's *World Soundscape Project* (22), the first project focusing on soundscape. Thanks to it, soundscape interest started to spread out of the musicians community and drummed up the interest of acousticians, who added their expertise for finding a link among physical sound features and the soundscape description. Therefore, soundscaping combines the physical registration of relevant acoustical parameters with the evaluation of perceptual effects via specific interviews and questionnaires involving community members who live in the location under study (24). Soundscaping is often deployed by means of soundwalks in urban outdoor spaces while conducting sound measurements and perceptual interviews, or by means of even longer term strategies mainly based on mobile sound measurements and extensive community involvement, in particular from public workers such as local police officers (20).

Soundscaping arose the consciousness of the scientific community on the fact that not every urban sound is noise and that the strive to reduce noise annoyance is not exactly equivalent to noise abatement. The soundscape researcher would eventually attempt to preserve and eventually accentuate *soundmarks*, i.e. sounds that are unique to an area, as defined by Schafer (22). Soundscape is therefore referred as an essential aspect of urban planning, at the same level of importance as visual aesthetics. Nowadays, soundscape research is a very active interdisciplinary research field involving acousticians, psychologists, musicians, linguists, architects and urban planners, just to mention the main professional groups.

1.3 Analysing the auditory scene

Humans have a great proficiency in disentangling mixtures of incoming sounds into coherent perceptual representations of objects (called auditory streams),

usually related to individual sound sources, based on a combination of auditory and visual cues. Understanding how humans perform this task is crucial especially within the soundscape approach. In a simplifying manner, this process of auditory scene analysis is often regarded as a two-stage analysis-synthesis process (25). In the first stage (segmentation), the acoustic signal is decomposed into a collection of time-frequency segments. In the second stage (grouping), segments that are likely to have arisen from the same environmental source are combined into auditory streams. Traditionally, it has been assumed that the perceptual mechanisms behind this process are largely pre-attentive: only after auditory streams are formed, they can become an object of attention (26; 27). Although this view is appealing because of its conceptual simplicity, recent findings suggest that attention also plays a role in the formation of auditory streams (28; 29). Overall, it can be stated that the process of auditory scene analysis draws on low-level principles for segmentation and grouping, but is fine-tuned by selective attention (30) which will be investigated in the next section.

1.4 Auditory attention

Attention plays an important role in audition. Auditory attention in fact allows us to focus our mental resources on specific aspects of the acoustic environment, while ignoring all other aspects (31), thus avoiding cognitive overload. As formerly mentioned, auditory attention plays a role even in the auditory stream formation and not only in auditory stream selection.

Central in most theories on attention — visual as well as auditory —is the interplay of bottom-up (saliency-based, depending on the characteristics of the stimulus) and top-down (voluntary, depending on the state of the listener) mechanisms in a competitive selection process (30; 32). The bottom-up mechanism selectively enhances responses to sounds that are conspicuous, for example because they have rare or novel physical features, or are of instinctive biological importance. This is accomplished in the sensory cortex by a novelty detection system that continuously monitors the acoustic environment for changes in frequency, intensity, duration or spatial location of stimuli (33; 34). In contrast, the top-down mechanism focuses processing resources on the auditory information that is most relevant for the current goal-directed behaviour of the listener. The selection of information for entry into working memory is found to be a competitive and hierarchically structured process (32). Selective attention is thus typically compared to a stagelight (35), sequentially "illuminating" different elements of the auditory scene. To do this, an important process

inhibition-of-return (36; 37) (IOR) occurs. Originally investigated in the vision domain (38), IOR prevents attention from permanently focusing on a particular element of the (auditory or visual) scene. In Posner's first experiments it was found out that measured reaction time for detecting visual objects in previously cued locations was longer when compared to locations not previously cued (38). IOR can to a certain extent be inhibited by voluntary selective attention, which may prohibit involuntary switching of attention to task-irrelevant distractor sounds (39).

1.5 Auditory learning

The importance of auditory attention per se would not be sufficient to correctly perceive and identify sounds. Learning, specifically auditory learning, is crucial and essential for surviving, but the mechanisms underlying it are still not clearly understood. Desired or undesired familiar sounds are in fact more easily detected (40) than unknown sounds. Sensitivity to particular acoustical features of a sound are learned in early childhood, but new sounds can be learned at all ages (41). Once sounds become familiar, they are identified more easily. It must be noted that learning effects are not limited to high-level associative memory. Several neurophysiological studies have reported on the capacity for holding memory traces (enduring neural records) in the primary auditory cortex (see Weinberger (42) for an extensive review). In particular, the number of neurons of the representational area of a sound is tuned by its importance (43) and the bigger the area, the stronger the memory effects (44). Neurophysiological correlates of cognitive processes such as selective attention (45; 46), expectancy (47), concept formation (48) and crossmodality effects (49) have been found in the primary auditory cortex, suggesting that due to neuronal plasticity, the primary auditory cortex is not merely an acoustic analyser, but an adaptive auditory problem solver (42). Another important property of the auditory cortex is tonotopy: neurons next to each other are typically excited by similar stimuli. Tonotopic maps have been observed in the auditory cortex of animal species such as cats (50) and monkeys (51; 52). The human cortex also contains several topologically ordered regions (53; 54; 55), similar to regions observed in the macaque monkey brain (55). In order to develop a human-mimicking model for machine listening, continuous learning and tonotopic mapping of the auditory cortex have to be taken into account. As explained in the next chapter, such aspects will involve a key role in the model.

1.6 Machine listening: accuracy or complexity? The human-like listening approach

After the publication in 1990 of the groundbreaking book of Albert S. Bregman *Auditory Scene Analysis* (25), many researchers started wondering whether it would be possible to transfer such knowledge about human auditory perception into computational models. In a very short time the multidisciplinary branch called *computational auditory scene analysis*, CASA, was born (see Wang and Brown (56) for an overview of computational auditory scene analysis models).

However, most of the researchers in this new research field started to focus very soon on speech and how extracting as clean as possible foreground sound signal from background noise. Reproducing the cocktail-party effect [1] by computational means became thus the key research topic.

The model presented in this work aims to mimic human evaluation of the sonic environment not trying to extract sounds that are as clean as technically possible, but trying to analyse the scene as accurately as a human listener would. This model is therefore flexible: no restrictions about neither the type of heard sounds nor sound processing applications are considered. Counterexamples can be voice and speech recognition techniques: they are restricted to a certain time-frequency domain and aim to solve specific issues as word target recognition and voice recognition. The presented model, on the contrary, aims to cover all aspects of machine listening without losing the ultimate goal of this research, i.e. human-like machine listening.

Most of the CASA models are based on very detailed and sophisticated sound analysis techniques such as Gammatone filters and Mel-frequency cepstral coefficients (MFCCs). Gammatone filtering is in fact the most computationally feasible solution for simulating the cochlear response, while MFCCs are widely used in speech recognition techniques. The Gammatone filtering technique has, however, the disadvantage that off-the-shelf sound measurement equipment cannot be used as a front-end, which decreases its applicability in environmental acoustic sensor networks. The alternative to simply record and transmit the sound at all microphones continuously is also infeasible due to data storage and transmission bandwidth limitations. The MFCCs have been very successful for single-source, non-reverberant speech recognition (58). However, they suffer from high sensitivity to noise and reverberation, much like other techniques such as the continuous wavelet transform (CWT). Good performances for specific

[1] Phenomenon of focusing auditory attention on a specific auditory stream thus filtering out all other auditory stimuli. The phenomenon was first defined as the cocktail-party problem by Colin Cherry (57) in 1953: "how do we recognize what one person is saying when others are speaking at the same time?" (57).

environmental sound recognition tasks could be achieved provided that the recording contains single and clean sources (59), clearly an unrealistic assumption for actual environmental sounds (60). Their use in long-term environmental monitoring networks seems therefore disadvantageous.

Thus, in order to find a balance between accuracy, biological plausibility and computational feasibility, the presented model is based on 1/3-octave band spectrum analysis at 0.125 s. The choice of such time resolution can be justified by noting that a wide range of outdoor environmental sounds have a relatively slowly varying temporal envelope (61; 62; 63) and that attention mechanisms work on the same time scale as suggested by measured event-related potentials (ERP) (64).

Chapter 2
Theoretical model

The empirical knowledge on human auditory processing (auditory scene analysis, masking, detection of sounds, learning, and auditory selective attention) summarized in the previous chapter, is worked out into a human-mimicking computational model of auditory processing as described in the following sections. The input of the model is the sound signal recorded by a microphone (only monaural inputs are considered). The model output depends on the application the model is used for. Computational efficiency is advantageous for long-term deployment, high feasibility and possible extension to a multi-node sensor network: simplified computational auditory processing models are thus preferred. The proposed computation model is composed of three stages: (a) peripheral auditory processing, (b) mapping of acoustical features based on co-occurrence and (c) modelling auditory attention and grouping. The following sections will give an overview of each of these stages.

2.1 Peripheral Auditory Processing

The sound signal measured by the microphone is collected by the off-the-shelf sound measurement equipment and the 1/3-octave band spectrum from 20 Hz to 20 kHz is calculated with a temporal resolution of $1/8\,\mathrm{s}$[1]. Such sound representation is not as detailed as the one obtained by using a gammatone filterbank, but has the important advantage to be computationally light and supported by all off-the-shelf sound measurement equipment, which can be used as a front-end, thus increasing the applicability of the model. Subsequently, a simplified *cochleagram* is calculated using the Zwicker loudness model (9; 65), which accounts for energetic masking in analogy to the initial processing by

[1] In Appendix A a temporal resolution of 1 s is used.

the cochlea and basilar membrane. The complete audible frequency range is considered (0 to 24 Bark) with a spectral resolution of 0.5 Bark, resulting in 48 spectral values at each time step.

The auditory system segregates sounds and triggers the bottom-up auditory attention based on individual auditory features such as spectral or temporal irregularities (25; 66; 67; 68; 69). These sound features have been used for constructing computational auditory saliency models (34; 70; 71). Based on them, measures for intensity, spectral and temporal modulation, also called contrast, are calculated using a centre-surround mechanism (72), thus taking into account the receptive fields in the primary auditory cortex (67; 72; 73; 74). This is done by convolving at each time step the cochleagram with several 2D gaussian and difference-of-gaussian filters encoding respectively intensity and spectro-temporal gradients at 16 different scales: 4 for intensity and 6 for both spectral and temporal contrast, similarly to (34) [2]. A cross section of these filters is shown in Figure 2.1. Two-dimensional difference-of-gaussian filters are a very good approximation of the more computationally demanding laplacian of gaussians, the multi-dimensional generalization of the mexican hat wavelets. By using a set of these functions it is possible to detect changes in time and frequency at various scales, thus measuring the spectral and temporal contrast of the input sound.

The resulting vector $\vec{r}(t)$, called *sound feature vector* or simply *feature vector*, is thus composed of $16 \times 48 = 768$ values. It encodes the informative content about the sonic environment at a given time step and it will be henceforth used for the following steps of the models.

2.2 Sound feature grouping based on co-occurrence and topological mapping

The feature vector provides the informative content about the sonic environment at a given time step. However, it would be useless if such information would not be coupled with a model of the continuous learning effects typical of humans. It has been shown that sensitivity to particular acoustical features of a sound are learned early childhood and that new sounds can be learned at all ages (41). Moreover, the primary auditory cortex can hold memory traces (see (42) for a review) and the number of neurons of the representational area of a sound is tuned by its importance (43) and the size of such area is

[2] In (34) 4 instead of 6 filters are used for modelling both spectral and temporal modulation. However, the cochleagram is calculated with very higher temporal resolution.

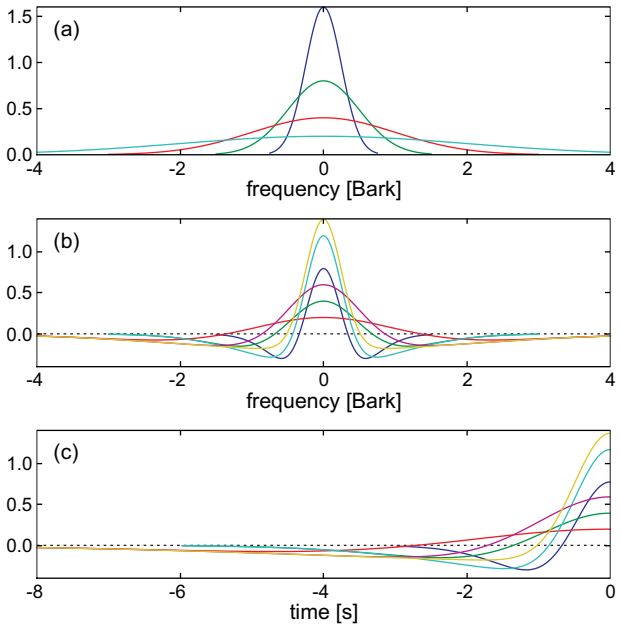


Figure 2.1: Cross section of the receptive filters that are used to calculate (a) intensity, (b) spectral contrast and (c) temporal contrast. For the latter, causality is preserved by only convolving with the past.

proportional to the strength of the memory effects (44). In order to model these properties of human learning, the use of a neural network-based approach by using a *self-organizing map* (SOM), also called *Kohonen map* (75), coupled with the continuous selective learning discussed in Section 2.3 is proposed in this work. A reason to use a SOM neural architecture in this model comes to the fact that several topographic maps have been observed in visual and auditory cortex (50; 51; 76; 77) and the SOM has been originally conceived as an abstract mathematical model of topographic mapping. However, the SOM original incremental algorithm described in this section trains the SOM purely on the rate of occurrence of heard sounds although human perception and retrospective assessment of soundscapes does not depend exclusively on it. For this reason a new learning technique called continuous selective learning is proposed in this work and it will be discussed in Section 2.3. By means of an extensive training on sound feature vectors at the microphone location, the SOM would eventually learn which features co-occur, thus allowing to categorize the typical sounds composing the specific sonic environment. The training of the

SOM should usually last for a long period, that can vary from few days to some weeks, depending on the variety of the sonic environment at the given location. Moreover, the SOM, typically described as an unsupervised learning-based method for clustering and high-dimensional data visualization (78), seems a very suitable model to manage the high-dimensionality of the sound feature vectors. The presented model is intrinsically context dependent: typical sounds at a given location can fit or not fit in the sonic environment of other locations; such aspect has analogies in human cognitive processes studied by means of detection task experiments: familiarity with the sound to be detected makes the detection easier (79). In the rest of this section, a formal description of the SOM architecture and the original incremental algorithm used for the first training of the SOM are presented.

The SOM used in this model is a 2D network of M equally-spaced *units* in a regular hexagonal M_x by M_y lattice and will be denoted as $\vec{\mathbf{m}}_i = (\mathbf{m}_x, \mathbf{m}_y) \in \mathbb{R}^2$. Each unit i has an associated *reference vector* $\vec{m}_i$ in the high-dimensional sound feature space. The initial position of the reference vectors is calculated by means of principal component analysis on an input data subset, resulting in vectors lying in the hyperspace spanned by the eigenvectors corresponding to the two principal components (78). After initialization, their coordinates are modified during the first training phase which is based on the Original Incremental SOM Algorithm (75) wherein the following three steps are repeated at each time step. First, the sound feature vector $\vec{r}(t)$ calculated at time step t is selected. Then the best-matching unit (BMU) $\vec{m}_{c(t)}(t)$ is calculated. The BMU is the unit corresponding to the closest reference vector:

$$c(t) = \arg\min_i \|\vec{r}(t) - \vec{m}_i(t)\| \ . \tag{2.1}$$

Third, the reference vector corresponding to the BMU and, to a lesser extent, those of the neighbouring units in the 2D lattice are moved closer to the input high-dimensional data point:

$$\vec{m}_i(t+1) = \vec{m}_i(t) + h_{c(t),i}\left(\vec{r}(t) - \vec{m}_i(t)\right) \ , \tag{2.2}$$

where h is the neighbourhood function. In our model we opted for a Gaussian function of the distance between $c(t)$, i.e. the BMU at time step t, and the unit i:

$$h_{c(t),i} = \alpha(t) \exp\left(-\frac{\left\|\vec{\mathbf{m}}_i - \vec{\mathbf{m}}_{c(t)}\right\|^2}{2\sigma^2(t)}\right) \ , \tag{2.3}$$

where the learning rate $0 < \alpha(t) < 1$ and the width of the 2D neighbourhood

$\sigma(t)$ are two time-step dependent parameters and both are strictly monotonically decreasing functions:

$$\alpha(t) = \alpha_0 \frac{C}{C+t}, \quad C = \frac{N}{\sqrt{10}}, \tag{2.4}$$

$$\sigma(t) = 1 + (\sigma_0 - 1)\,(\frac{N-t}{N}), \tag{2.5}$$

where $\alpha_0 = \alpha(t=0)$ is the initial learning rate and N is the number of samples used for the training. Note that $h_{c,i} = \alpha\,(t)$ for the BMU. The SOM adaptation governed by the neighbourhood function is thus, due to Eq. (2.4) and (2.5), strictly decreasing in time and, due to the Gaussian function in Eq. (2.3), is also decreasing for units farther from the BMU in the lattice.

After this training, the reference vectors of the SOM units can be seen as a non-linear discrete 2D mapping of the probability density function of the sound feature vectors $\vec{r}(t)$ used for training. In particular, specific regions of the sound feature space contain more reference vectors than other only sparsely represented regions, thus preserving the high-dimensional relationships underlying the input feature vectors (75). In other words, feeding a new sound feature vector $\vec{r}(t)$ to the trained SOM, the smaller the distance to the BMU $\left\|\vec{r}\,(t) - \vec{m}_{c(t)}\right\|$, the more often similar sound feature vectors occurred during the training phase.

2.3 SOM extension: Selective Continuous Learning

Mapping the probability density function of the incoming sounds is not sufficient: human perception and retrospective assessment of soundscapes do not depend exclusively on the rate of occurrence of heard sounds. For instance, salient but less often occurring sound events would be better remembered than nonsalient ones that might occur more often but stay out of attention focus. The SOM trained with the Original Incremental SOM Algorithm is thus used as a starting point for a second much longer training phase which will be referred as *continuous selective learning* (CSL). At each time step t the sound feature vector $\vec{r}(t)$ is calculated and the BMU $\vec{m}_{c(t)}\,(t)$ is found as before. However, not all input sound feature vectors $\vec{r}(t)$ are used as inputs during the CSL: a learning phase is triggered only if the distance to the BMU is bigger than an activation threshold T_{up}:

$$\left\|\vec{r}\,(t) - \vec{m}_{c(t)}\,(t)\right\| > T_{up}\,. \tag{2.6}$$

All the input vectors calculated at the following time steps are selected as inputs for training until the distance to the BMU becomes lower than a deactivation

threshold T_{down}:

$$\left\|\vec{r}\,(t) - \vec{m}_{c(t)}\,(t)\right\| < T_{down}\,. \tag{2.7}$$

The Eqs. (2.2) - (2.5) are still valid. Furthermore, sound feature vectors occurring a few seconds before the triggered learning period are included for smoothing and causality issues. Typically two seconds are considered, corresponding to 16 feature vectors.

The advantages of the CSL have been discussed for the first time in Oldoni et al. (80). In particular, it has been shown that a SOM trained using sound feature vectors from a quiet site can match traffic sound in terms of distance to the BMU after a training with the CSL [3] using sound feature vectors from a street as input. The distance to the BMU of this SOM is comparable to the distance to the BMU of the SOM trained exclusively on traffic sound (80). Moreover, a measure of overall auditory saliency [4], $0 < s\,(t) < 1$, is used for calculating the learning rate parameter α:

$$\alpha\,(t) = \alpha_0 \cdot s\,(t)\,(0.5 + s\,(t))^2\,\frac{C}{C+t}, \quad C = \frac{N}{\sqrt{10}}, \tag{2.8}$$

where $\alpha_0 = \alpha(t=0)$ is the initial learning rate and N is the total number of samples in analogy with Equations 2.4 and 2.5 [5]. The measured saliency is used as a learning strength modulator: the learning of sound feature vectors whose related saliency values are higher than 0.5 is enhanced, while sound feature vectors with lower saliency are somewhat suppressed. The goal of using saliency in CSL is to increase the effectiveness in reducing the SOM units whose reference vectors are related to often occurring but not relevant background noise or quiet moments. The thresholds T_{up} and T_{down} are empirically chosen in such a way that less than 10% of all sound feature vectors are used as input for CSL.

Another important aspect besides computation is the way to visualize the SOM 2D grid of units after training. The high-dimensionality of the sound feature space makes the visualization via projection on a particular hyperspace not sufficiently informative. The ultimate purpose of SOM visualization techniques is to easily identify groups of neighbouring units with similar high-dimensional reference vectors by locally investigating the morphology of the map. For this reason a typical SOM visualization technique uses the U-matrix (82), a matrix

[3] in Oldoni et al. (80) the CSL is referred to as dynamic learning

[4] A detailed description of the algorithm for calculating the auditory saliency can be found in De Coensel and Botteldooren (81)

[5] The only difference is that during the SOM incremental algorithm explained in Section 2.2 all N sound feature vectors are used for training, while during the CSL N is much larger than the number of vectors effectively used as inputs.

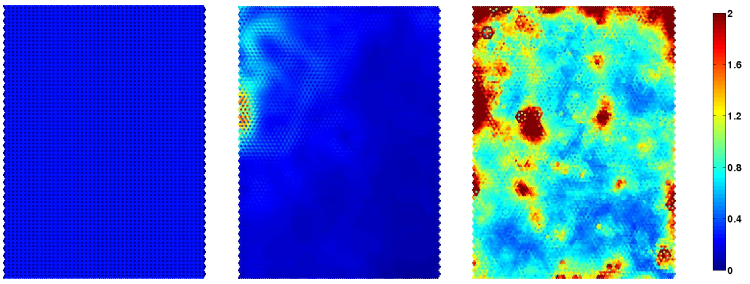


Figure 2.2: U-matrix of the SOM: (left) after initialization via principal component analysis, (centre) after training based on the original SOM incremental algorithm, (right) after training based on continuous selective learning.

containing both the distances between nearest neighbouring units and their average. The U-matrix has dimensions $[2M_x - 1, 2M_y - 1]$ and the distance is colour-coded, thus making it possible to distinguish regions of SOM units with similar reference vectors from regions showing higher variability. In Figure 2.2 the effects of the training based on the SOM incremental algorithm and subsequent CSL are shown.

The reference vectors referring to the trained SOM units can be seen as representative abstract sound prototypes encoded by their sound feature vectors. Such prototypes can be heard by recording sound samples which are the most similar to them in the sound feature space. Such set of sounds can be seen as a sound library describing the sonic environment at the measurement location. Further details on such library and an application based on it will be discussed in Section 3.2.

2.4 Auditory object formation, auditory attention and environmental sound recognition techniques

In this section several submodels based on the trained SOM are discussed. First, the attempt of modelling auditory object formation over time via an oscillating neural network will be presented. Some ideas from this model have been used also to model the role of selective auditory attention. Finally, in the last two sections a Support Vector Machine and a fuzzy excitation model approach are coupled to SOM for identifying sounds.

2.4.1 Grouping and auditory object formation by means of an oscillatory neural network

In the 1980s insights into the oscillatory correlation properties of neurons in the sensory cortex resulted in an increase of theoretical research on possible computational models of the corresponding biological mechanisms. The first model constructed by von der Malsburg and Schneider was later extended by Wang in his *shifting synchronization theory* by means of a Locally Excitatory Globally Inhibitory Oscillator Network (LEGION) (83). In general, each input pattern is represented by the synchronization of a group of LEGION's oscillators. Contrarily, desynchronization among different groups of oscillators representing different input patterns occurs. Thanks to its ability of segmenting multiple input patterns, LEGION was applied in 2D image scene segmentation (84; 85; 86). Similarly, LEGION was used to solve CASA issues as auditory stream segregation (87; 88), speech detection (89) and has been used in a computational framework modelling auditory selective attention (90). In all these LEGION based computational framework the auditory signal is in some way transformed in a 2D representation which represents the input for LEGION. In such representation, where time is explicitly or implicitly included, the issue of auditory segregation is thus very similar to image scene segmentation.

In this work, the similarity of a sound feature vector and a specific SOM unit is, from a neural oscillatory point of view, a measure of the external stimulation that a LEGION oscillator receives. As stated in Appendix A, a measure of similarity can be defined by means of a decreasing function of the distance to the BMU. It should be underlined that the SOM unit and the LEGION oscillator can be conceptually considered the same formal neural unit expressing two different functionalities of ideal neurons: the long term memory formation is modelled by the SOM extensive training, while the dynamic oscillatory correlation of sensory cortex neurons excited by an auditory stimulus is modelled by LEGION. Moreover, in this SOM-LEGION coupled model, the so-called permanent weights among near oscillators are determined during the SOM training, being related to the similarity of two neighbouring SOM units: near SOM units which are very distant in the multi-dimensional sound feature space are therefore loosely connected. In case more than one group of oscillators is excited by the input feature vector, desynchronization among them naturally arises as shown in Figure 2.3. For more details about LEGION and its coupling with SOM, see Appendix A.

However, this synchronization within groups and desynchronization among groups of oscillators occurs after a transient phase. Eventually, it occurs after each change of the external stimulation, letting the oscillators to recombine

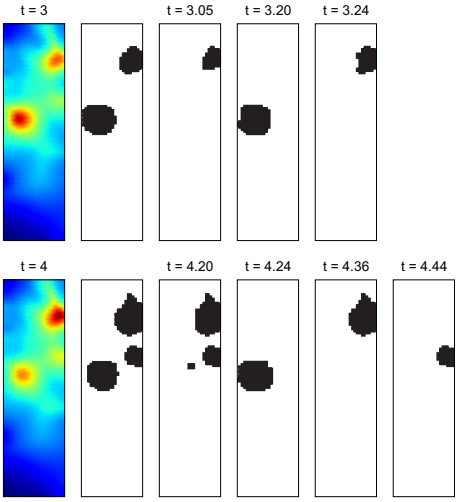


Figure 2.3: Left (2 columns): similarity (inverse of the distance) of two input samples at $t = 3\,\mathrm{s}$ (top) and $t = 4\,\mathrm{s}$ (bottom), before (1st column) and after (2nd column) binarization. Right (4 columns): some snapshots of LEGION taken at different times.

their dynamic connection weights to adapt themselves to the new external input. However, such an abrupt reset of the oscillation periodicity makes impossible any attempt to follow the streams in a dynamic auditory scene, thus limiting the useability of this approach. Such problem has not yet been solved and still limits the straightforward use of LEGION in long running applications. It is worth to mention that the LEGION applications to CASA cited before don't tackle this issue, analysing only few seconds long sound excerpts. Such applications involve also a very detailed time-frequency analysis which is not feasible in the framework of this work as already mentioned in Section 1.6. Actually, calculating the trajectories of all LEGION's oscillators is very computationally demanding as well, especially in this work where a high number of oscillators are used. In order to speed up the computational process the singular limit method (91) is extensively used as mentioned in Appendix A. However, such approximation doesn't reduce the disruptive transient on the oscillation periodicity after each change of the external stimulation.

2.4.2 Modelling Auditory Attention

To model auditory attention an excitatory-inhibitory artificial neural network (ANN), simulating the auditory cortex, is introduced. Such a network is to a certain degree similar to LEGION due to the fact that it uses the same concept of local excitation and global inhibition in order to reach a dynamic winner-take-all situation. The formal duality between the SOM unit and the LEGION oscillator is also maintained: the long term memory formation is expressed by the extensive SOM training and the excitation-competition mechanisms underlying auditory attention are modelled by ANN.

Each neuron i is excited by input sounds with feature vectors that are similar to the reference feature vector of the corresponding SOM unit. In order to simulate the importance played by conspicuous and salient sounds in bottom-up attention mechanisms, the saliency $s(t)$ of the incoming sound is used as an excitation modulator factor in a similar way it was used in Section 2.3 for modulating the auditory learning in the SOM. The IOR, $IOR_i(t)$, is also introduced as in (81) and it is implemented by increasing the inhibition of the activated neurons. In this way a scan of the sonic environment is promoted, shifting from a group of neurons to another. A leaky integrator is used to implement both $IOR_i(t)$ and excitation $E_i(t)$ for all neurons.

The top-down attention can be also included in such model as an IOR modulator by changing the time constants for neurons related to a certain group of SOM units, thus delaying or even halting the shift of attention.

As in LEGION, local excitatory and global inhibition terms are added in the model to achieve the same goal of clustering, due to the local excitation $EL_i(t)$, and competitive selection, due to the global inhibitor $IG(t)$. As for the weights among LEGION oscillators, the local excitatory term of neuron i is based on the excitation of the neighbouring neurons and is weighed based on the similarity among the reference vectors of the corresponding SOM units as explained in Section 2.4.1. In this way neighbouring neurons related to similar sounds (in the sound feature space) are strongly connected, while neurons encoding dissimilar sounds are weakly connected. As in LEGION, a global inhibition term $IG(t)$ is added to all neurons when the activation term is higher than a certain threshold, thus enhancing competitive selection among groups of excited neurons. The activation of a neuron, $A_i(t)$, can be then written as the sum of all these terms:

$$A_i(t) = \max\{0, E_i(t) + EL_i(t) + IOR_i(t) + IG(t)\}, \tag{2.9}$$

where the maximum guarantees non negative activation values.

2.4.3 Classifying sounds via Support Vector Machines

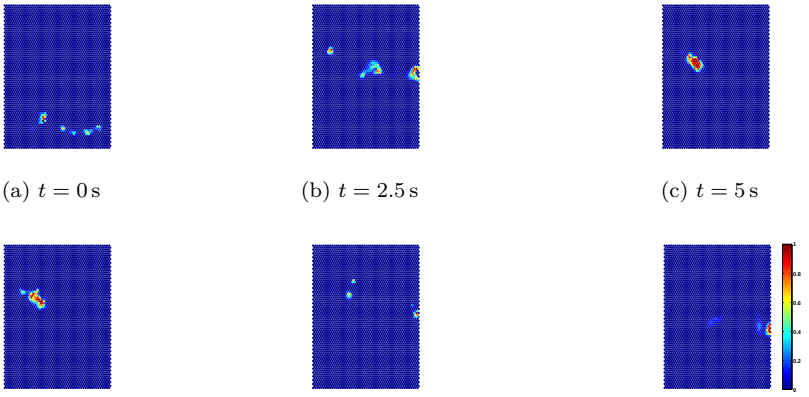
One of the goals of machine listening is creating automated systems able to classify sound events occurring in a certain environment. In the context of outdoor environments, an approach based on a Support Vector Machine (SVM) coupled with SOM is considered here.

The sound library in Section 2.3 representing the sound prototypes encoded by SOM units can be listened to one by one and manually labeled by an expert listener. However, it requires a lot of time and thus this approach is unfeasible for being used at a large scale. A supervised learning method based on a SVM can be used to automatically label the SOM units. The SVM is first trained using reference vectors of SOM units inheriting the labels given by an expert listener to the corresponding sound samples forming the sound library; after the training the SVM can be used for labelling sounds of other SOMs from other time periods or other locations. The more the sonic environments are similar, the more such approach is effective.

The SVM can also be used for a basic environmental sound recognition tool: once the SOM units are coupled with labels, the incoming sound can be at each time step identified by using the label of the corresponding BMU, thus forming a model for continuous environmental sound recognition. A drawback of the SVM method is that only generic labels or categories of sounds can be used in order to maintain the training set related to each label statistically significant. Even so, a non-negligible number of units cannot be covered by labels due to the high specificity of the represented sounds. A way to solve this would be to include them in even broader sound categories at the expense of loosing further accuracy. In Section 3.3 an application of the use of a SVM is shown and its performance in labelling a sound library is discussed.

2.4.4 Identifying sounds via a fuzzy excitation model

Identifying specific target sounds is a common subject in CASA models and in many environmental sound monitoring applications. Typically, they have been deployed in order to identify a specific target sound and cannot be easily adapted for identifying other type of sounds. Although the sound feature extraction and the use of SOM have not been originally conceived for this goal, a model for identifying environmental target sounds is here proposed. It has the advantage of being very general and usable for every sort of sounds present in the sonic environment under study. It combines SOM, supervised learning and a similarity measure, $S(t)$, with $0 < S(t) < 1$, as in De Coensel et al. (92). After training a SOM at a given location as in Sections 2.2 and 2.3, a set of target



(a) $t = 0\,\mathrm{s}$ (b) $t = 2.5\,\mathrm{s}$ (c) $t = 5\,\mathrm{s}$

(d) $t = 7.5\,\mathrm{s}$ (e) $t = 10\,\mathrm{s}$ (f) $t = 12.5\,\mathrm{s}$

Figure 2.4: Some snapshots of SOM units' excitation taken at different times during a target sound event, i.e. the passage of a train.

sounds are manually selected from the sonic environment, the corresponding sound feature vectors are extracted and the excitation of the SOM units is calculated via a Gaussian transformation similarly as in Sections 2.4.1 and 2.4.2:

$$E_i(t,j) = e^{-\frac{d_i(t,j)-\tilde{d}_i(t)}{k\tilde{d}_i(t)}}, \tag{2.10}$$

where i is the index over the set of target sounds, j the index over the SOM units, $d(t,j)$ is the distance of the sound feature vector at time t to the $j-$ SOM unit, $\tilde{d}(t)$ is the distance to the BMU at time t and k is a positive constant defining the width of the Gaussian function. In particular, $E_i(t,j) \leq 1$ and $E_i(t,j) = 1$ only for the BMU. The idea is to create a prototype of the excitation of the target sound using the set of known target sound samples and measuring the similarity $S(t)$ between it and the excitation due to the incoming sound: the higher $S(t)$, the more similar the incoming sound to the prototype. However, the duration of the target sound samples is typically variable: the easiest solution would be averaging over time and over samples, but it cannot be applied in case of not steady target sounds where excitation of different regions of SOM at different moments for each target sound occurs. An example is given in Figure 2.4 where the excitation of the SOM units is shown at different time steps of a target sound, in this case the passage of a train.

In this example, changing the type of train, its speed and the number of

railcars could modify the excitation pattern among the SOM units. In this work, this issue is solved by summing for each target sound i the excitation of the SOM units over the target sound duration t_i:

$$E_i(j) = \sum_{t_i} E_i(t, j), \tag{2.11}$$

and then clustering the resulting excitation maps $E_i(j)$. The number of clusters should not be chosen a priori, depending on the variability among the target sounds. In the application shown in 3.4 a $k-$means clustering is performed and the optimal number of clusters, N, is found by using the Davies-Bouldin index (93). In this way not one, but multiple, although typically a very limited number, excitation prototypical maps $\tilde{E_h}(j)$ are created. Each cluster inherits a prototypical sound event duration, $\bar{t_h}$, by averaging the durations of all the sound samples belonging to it. By means of time windows as long as the prototypical sound event durations, as many excitation maps of the incoming sound as the number of prototypes are calculated:

$$(\sum_{t}^{t+\bar{t_h}} E(t, j), \quad h = 1, ..., N \tag{2.12}$$

Finally, similarity measures $S_h(t)$ between the incoming sound and the prototypical excitation maps $\tilde{E_h}(j)$ are thus computed and coupled to corresponding dynamic thresholds $T_h(t)$, $0 < T_h(t) < 1$:

$$T_h(t) = \begin{cases} 1, & \text{if } S_h(t) > T_h(t-1) \\ e^{-(\ln T_{h_0})\frac{t-t_{h_0}}{L}}, & \text{if } t - t_{h_0} < L \text{ and } S_h(t) \leq T_h(t-1) \\ T_{h_0}, & \text{if } t > t_{h_0} + L \text{ and } S_h(t) \leq T_h(t-1) \end{cases} \tag{2.13}$$

where the threshold $0 < T_{h_0} < 1$ is a fixed real number, t_{h_0} is the most recent time that $T_h(t) = 1$ and L is a broad time window centred on a sound target event. An example of the functioning of the dynamic threshold is shown in Figure 2.5. A target sound is considered being detected if $T_h(t) = 1$ at least for one of the N dynamic thresholds during a broad time window centred on the target sound event. Vice versa, it would be considered not detected if $T_h(t) \neq 1$, for all thresholds during the same broad window. A study of the performance of this model follows in Section 3.4 where a case study is discussed. As already mentioned, the use of SOM, as presented in Sections 2.2- 2.3, has not been conceived for detecting specific target sounds. In fact, the proposed method would fail in detecting out of context sounds or very rare sounds, due to the fact

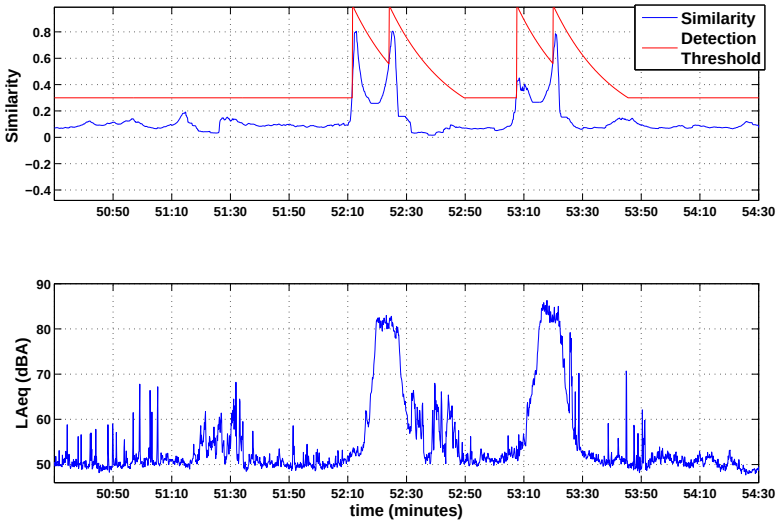


Figure 2.5: Example of the functioning of the dynamic threshold. Above: similarity $S(t)$ and dynamic threshold $T(t)$. Below: sound pressure level. The two peaks are related to two target sound events.

that they would never or almost never occur during the CSL-based training. These outliers could be better detected considering the fact that the distance to the BMU would be very high. More details on such outlier detection method can be found in Section 3.5.

Chapter 3
Applications

In this chapter, several applications of the theoretical model are presented. First an application in the context of urban soundscape analysis is shown: a collection of typical sounds at a given location, called acoustic summary, can be automatically extracted. It is shown that this is a comprehensive set of sounds characterizing the specific location as judged by people living in the surroundings. An example of soundscape design is also provided assessing the perceptual effect of introducing additional sounds. The evaluation of the SVM-based method to automatically label sound samples is presented next. This method is very useful, for example, to label the sound excerpts composing an acoustic summary. The third section presents an application of the fuzzy excitation model for identifying specific target sounds. Finally, it is shown how the theoretical model can be used for anomaly detection in urban noise sensors as part of a multi-criteria approach. In addition, the theory underlying the aggregation of multiple indicators is discussed.

3.1 Introduction

The theory presented in the previous chapter will be deployed here in several applications. Although machine listening cannot completely reproduce or mimic human listening, several aspects of it will be investigated here. In particular, the following topics will be considered:

- selecting typical sounds (acoustic summary tool, Section 3.2),
- attaching meaning to them (automatic sound labeling via SVM, Section 3.3),

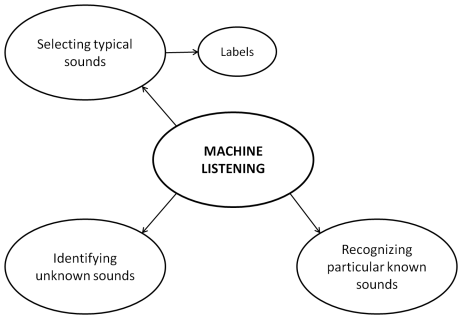


Figure 3.1: Aspects of machine listening considered in this work.

- recognizing particular known sounds (specific sound event recognition, Section 3.4)
- detecting atypical sounds (anomaly and failure detection, Section 3.5).

3.2 Soundscape mapping and design

Liveability of the urban environment has always been a primary issue for architects, landscape and urban planners. In particular, there is an increasing awareness of the fact that the sonic environment forms an essential component of the urban environment that requires as careful planning as the landscape (94; 95). Therefore, the current challenge for acousticians is to develop specific tools in order to efficiently characterize and represent the soundscape of a specific location and, based on that, design possible future soundscapes in order to evaluate the potential of planned urban interventions.

3.2.1 Soundscape mapping: the acoustic summary tool

Once the learning presented in Section 2.3 has ended, the reference vectors associated to the trained SOM units can be seen as representative abstract sound prototypes encoded by their sound feature vectors. Once a SOM is trained, it can be used for constructing a library of sounds, whereby sound samples that are most similar in the sound feature space to the sound prototypes within the SOM are recorded. The first step in constructing the acoustic summary is calculating feature vectors for the sound observed at each time step as explained in Section 2.1. The BMU is then selected, and the distance between its reference vector and the current sound feature vector is calculated. Based on this distance,

sound recording is triggered if the selected SOM unit has not been the BMU before (meaning that the encountered sound has not occurred before during the sound sample retrieval phase), or if the distance to the BMU is smaller than any earlier distance for this BMU (meaning that a better matching sound sample is encountered). These steps have to be taken with low latency due to the limited audio recording buffer of typical measurement stations. Sound samples are recorded from 3 seconds before to 2 seconds after the recording trigger. It turns out that for typical urban soundscapes, the bulk of the SOM units is represented by an audio sample after a few days of sound sample retrieval. This large set of sounds can be seen as a sound library describing the soundscape at the measurement location.

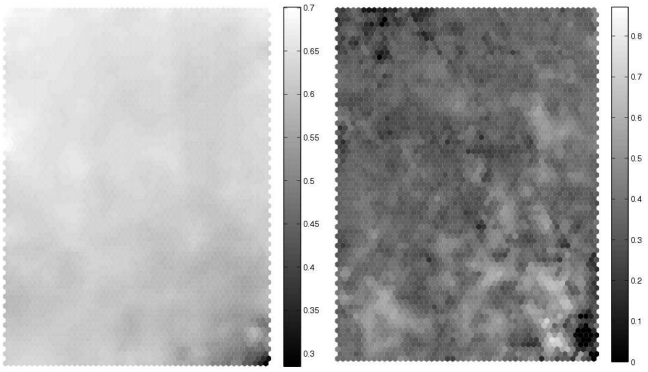
Due to the high number of sounds, such sound library could be way too big and unpractical for easily exploring the given soundscape by listening. Therefore, a selection is needed: three ranking criteria are presented that can be used to select a subset of the most representative sounds for the given soundscape; this subset is called the *acoustic summary*. The acoustic summaries used in the experiment described in Appendix C are composed of 32 sounds.

The first proposed ranking criterion is based on saliency: the higher the saliency, the more likely the sound sample will be representative and the higher its ranking. As explained in Section 2.3, a measured overall saliency value can be calculated at each time step from the sound feature vector. The SOM reference vectors lie in the sound feature space, therefore saliency values can be calculated for each of them, resulting in a saliency overlay on the SOM as shown in Figure 3.2.

It could happen that only similar sounds encoded by very close SOM units would be selected if a SOM region is the most salient one. In order to avoid this, a constraint on the distance among the 2D SOM units could be introduced. For example, in the construction of the acoustic summaries presented in Appendix C no contiguous SOM units could be selected.

A second criterion is based on how often each of the SOM units was selected as the BMU during a given time interval, typically one day or more, resulting in a frequency of occurrence overlay on the SOM as shown in Figure 3.2. However, the frequency of occurrence of sounds is not likely to be a sufficient criterion to represent the sounds that will be noticed and remembered, as already mentioned in Section 2.3. For this reason, a third method is proposed which combines saliency and frequency of occurrence of each SOM unit.

Also for these two other selection criteria the same constraint on the 2D distance among the SOM units can be used. In Appendix C no contiguous SOM units could be selected.



(a) (b)

Figure 3.2: Map of (left) saliency and (right) frequency of occurrence of the reference vectors linked to the units of a SOM trained at a location in Ghent, Belgium. The occurrence map is calculated based on one full working day, 691200 samples, and shows the relative logarithmic occurrence as BMU of each SOM unit.

A listening test involving local experts and described in Appendix C has been performed to evaluate the ability of the model to produce acoustic summaries representative of the soundscape at a number of urban locations. In a first experiment, local experts could almost always identify their own living environment from a set of 3 locations after listening to the acoustic summaries constructed by means of all the three selection criteria. However, as resulted from a second experiment, the acoustic summary criterion combining saliency and frequency of occurrence of the sound events generally produces the most accurate acoustic summary as shown in Figure 3.3 where the answers of all participants for all selection criteria are given. Three acoustic summaries, all coming from the location where the participant lives, but either formed by the saliency, the frequency of occurrence, or the mixed criterion were presented. The participants were asked to rank the presented fragments based on perceived accuracy in representing the surroundings of the participant's own home, from 1 for the most appropriate one up to 3 for the least appropriate one. The saliency-based criterion produces good acoustic summaries as well but risks to outweigh highly informative and salient sounds especially in urban residential areas. In addition, participants judged the acoustic summaries based on frequency of occurrence alone to be the least representative due to the prevalence of quiet sound fragments, which are much less informative of the given soundscape even

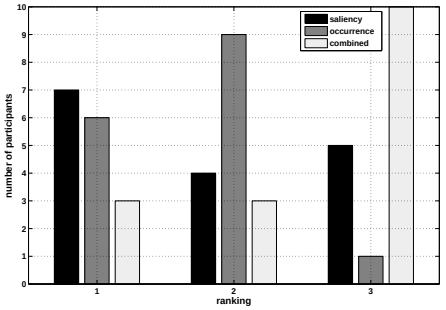


Figure 3.3: Ranking given by all participants to three acoustic summaries related to their own surroundings. The three acoustic summaries were selected by means of three different criteria: saliency (black), frequency of occurrence (grey) and their combined measure (white).

though they occur very often in residential areas. In the third experiment, each participant was asked to construct his/her own collection of sounds that represented the direct surroundings of its home, by selecting sounds from a set of 64 sounds. Half of the sounds the participant could choose from came from his home location, the other half was from two other randomly chosen locations (more details about the choice of the locations is given in Appendix C). All participants except one score better than a random guess. Moreover, the link from the results of the first and the third experiment demonstrates that the representativeness of an acoustic summary is a direct consequence of the quality of each sound composing it. Nevertheless, the number of false negative and false positive cannot be in general neglected: the sound samples composing an acoustic summary can, most of the time, be associated to more than one location, if the sound samples are considered separately from the others. Therefore, results of this experiment confirm the validity of using an acoustic summary for representing or evoking a soundscape. Finally, the test demonstrated that typically only a few sounds are needed to represent the soundscape of an urban area, confirming the choice of 32 sounds for each location as a sufficient number of sounds. A more detailed discussion of the experiments' results can be found in Appendix C.

3.2.2 Designing new soundscapes

The assessment of the soundscape at a given location as described in Section 3.2.1 is the first step of any soundscape designing technique. In this section, the

next step is considered, i.e. designing new soundscape scenarios by assessing the perceptual effects of introducing additional sounds. The introduction of green areas in the urban environment or changing the end use of a street (e.g. converting a street into a pedestrian area) are typical examples of interventions that would eventually lead to a soundscape alteration and the task of the urban (soundscape) planner is to evaluate beforehand the benefits of such interventions in function of cost-effectiveness analysis. The soundscape design case study presented here is based on the formation of a sound library of the modified soundscape and on the use of the auditory attention model explained in Section 2.4.2.

A fixed sound measurement station was installed in the city of Ghent, next to an urban road, carrying about 3000 vehicles/day during a typical work day. The sonic environment at the chosen location mainly consists of a mixture of road traffic noise due to private and public transport, and noise from pedestrians due to the proximity of several shops and one educational institution. A SOM was trained during a period of 3 weeks using the CSL as explained in Section 2.3.

The aim of the case study is to assess the perceptual effects of introducing a green urban area, thus attracting songbirds at the microphone location, a measure that is often proposed to increase the pleasantness of a soundscape (96). In case no green area can be inserted, the use of audio islands by means of (camouflaged) loudspeakers can be planned (97; 98). For this, a 1-h sound recording was performed during a work day not included in the period used for training. The L_{Aeq}during this 1-h period was 68.2 dB(A). Subsequently, a series of 30 artificial one-hour sonic environments were created by mixing the original recording with an increasing number of bird sounds at random instances in time. For this, a series of bird vocalisations without background noise, with a duration of up to a few seconds, were used, for which the peak level was adjusted to match the peak level of the few bird sounds present in the original recording thus creating a realistic level for the individual chirps. The 1-h $L_{\mathrm{Aeq}}$of the added bird sound ranged from 46.3 dB(A), representing very few sporadic twitters, to 75.8 dB(A), representing a quasi-continuous bird chorus, resulting in a signal-to-noise ratio (SNR) from -21.9 dB to 7.6 dB.

The 30 artificial sound mixtures were used in random order for a second CSL phase. This phase is particularly important to let the SOM get acquainted with the new bird sounds mixed with the actual sonic environment. A sound library was then created and the SOM units whose sound samples contained bird vocalizations were marked by an expert listener. They are mainly grouped in two separate regions as shown in Figure 3.4: the first region could be

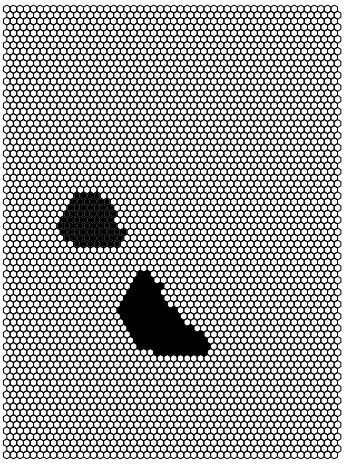


Figure 3.4: The two regions of the SOM related to individual bird chirps (above left) and bird chorus (below right).

associated to individual bird twitters while the second region is devoted to bird chorus. The presence of multiple SOM regions for bird sounds is due to the fact that individual chirping sounds and bird choruses are automatically located in different regions in the multidimensional sound feature space and are therefore represented by different groups of SOM reference vectors.

SOM units belonging to these two regions are more frequently the BMU as the SNR of bird sound increases as shown in Figure 3.5. In particular, the frequency of occurrence of the SOM units of the first region increases monotonically until a peak is reached at a SNR equal to -2 dB and then decreases to zero: as SNR increases the bird twitters are more and more frequent so that bird chorus fragments start to form. Therefore, the frequency of occurrence of SOM units belonging to the second region increases significantly and becomes dominant at a SNR equal to 0 dB. Taking into account the auditory attention mechanisms explained in Section 2.4.2 and considering the same two regions, the percentage of time that human attention is focused on bird sound can be estimated and it is shown in Figure 3.6. In particular, it can be seen that for lower SNR such percentage is slightly higher than the frequency of occurrence shown in Figure 3.5 due to the high saliency of bird sounds in comparison with the background. In contrast, bird sounds are continuously present for high SNR and IOR will cause attention to shift temporarily away from it. As a result the fraction of the time that the sound is expected to attract attention is lower than the fraction of the time that the sound is present. These results are in

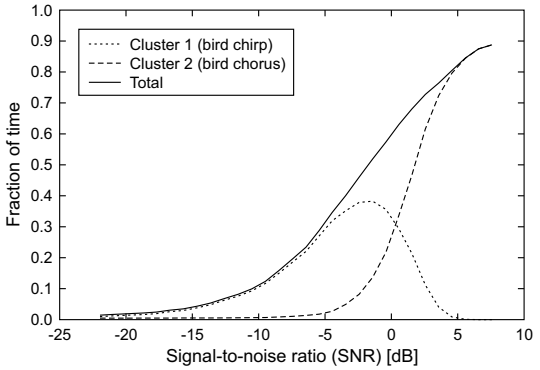


Figure 3.5: Evolution of the fraction of time the BMU is located in region 1 (bird chirp, dotted line), region 2 (bird chorus, dashed line) and their sum (total, continuous line) as a function of SNR between background and foreground. For each sound scenario, one hour (3600 1 s testing samples) has been used.

accordance with empirical results found by De Coensel et al. (99) assessing that already at a SNR of -10 dB bird sounds can increase the pleasantness of the soundscape significantly. For more details about this case study, see Appendix B.

3.3 Attaching meaning to sounds: automated labeling

In this section a case study is presented, based on the model explained in Section 2.4.3.

Two sound libraries related to two trained SOMs are considered here. The two SOMs have been trained using CSL on sound feature vectors calculated from continuous data collected from the same location but during two different periods, three weeks in October and three weeks in November 2011 respectively. Both sound libraries contained sound samples recorded the day after the end of the training. In particular, they were composed of 2369 and 2892 samples respectively, i.e. 68% and 83% of the total 3500 SOM nodes. An expert listener (a researcher specialized in environmental acoustics) listened to the 5 s long samples composing the sound libraries and observed that the most common sound events could be assigned to the following classes: *bird*, *chatting people*, *car*, *truck*, *motorbike/scooter*, *tram* and *background noise/hum*. The same person selected the sounds belonging to these classes. Two sets corresponding to the two sound libraries were then created: the first one was composed of 1046 sound

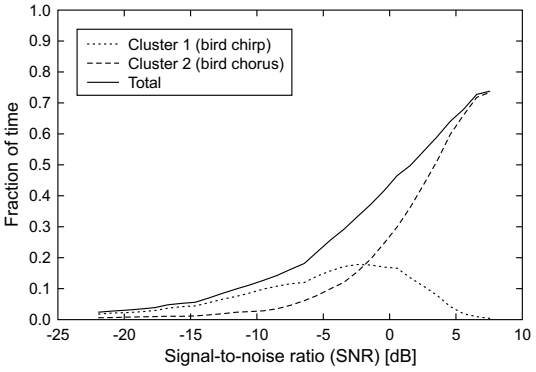
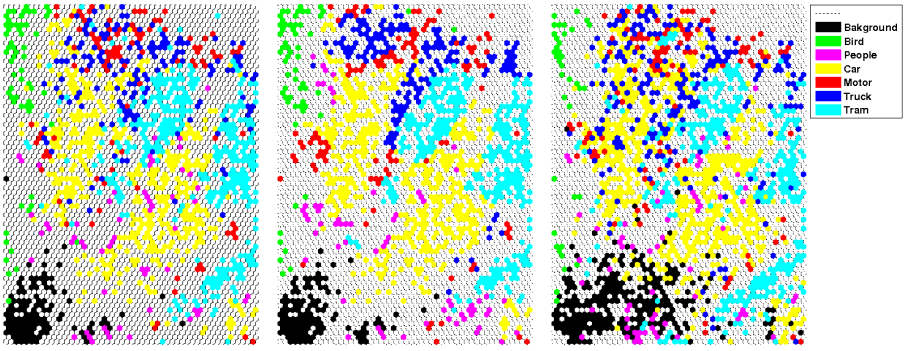


Figure 3.6: Evolution of the fraction of time the auditory attention is located in region 1 (bird chirp, dotted line), region 2 (bird chorus, dashed line) and their sum (total, continuous line) as a function of SNR between background and foreground. For each sound scenario, one hour (3600 1 s testing samples) has been used.

fragments, while the second set was composed of 1206 sound fragments, i.e. 44% and 42% of the total number of samples composing the sound libraries. The not included sound samples were either considered by the expert listener as a mixture of sounds from different classes or could not being classified in any of the 7 classes. The reference vectors of the SOM nodes related to the first set were used to train a SVM as explained in Section 2.4.3, while the second library referring to the second SOM was used for testing. The distribution of the classes over the test sound library as given by the expert listener is shown in Figure 3.7a. It can be noticed that some SOM nodes are either not represented in the sound library or they are not labeled: some less frequent sound classes were not considered (church bells, different kinds of alarming sounds as horns, etc.), neither mixtures of co-occurring sounds. As shown in Figure 3.7b, the proposed SVM automated method provides a SOM labeling quite similar to the one given by the expert (917 matching labeled sounds, 76% of the 1206 sounds labeled by the expert listener) suggesting that the proposed SVM labeling method is able to reproduce with a certain degree of accuracy the human SOM labeling when sufficient data are available. In order to confirm such findings, a comparison test was performed: a second listener was asked to listen to the second (test) sound library and to classify the sound samples using the classes already used by the expert listener. In Figure 3.7c some differences may be found when compared to the sound library labeled by the expert listener. The labels belonging to road vehicle categories (car,

truck and motorbike/scooter) seem to be slightly more mixed in the case of the second listener. Also his/her perception of background noise is different, which is reflected in the bigger cluster of sounds assigned to that class. Summing up all the classes, the second listener gave a higher number of labels than the expert (1543 and 1206, respectively). All these results confirm a natural human variability in distinguishing and tagging sounds. Moreover, the labeling deviation between these two human listeners is slightly larger than the deviation between the expert human listener and the SVM classification based on the SOM reference vectors, making it an interesting solution for automating urban sound labeling.



(a) (b) (c)

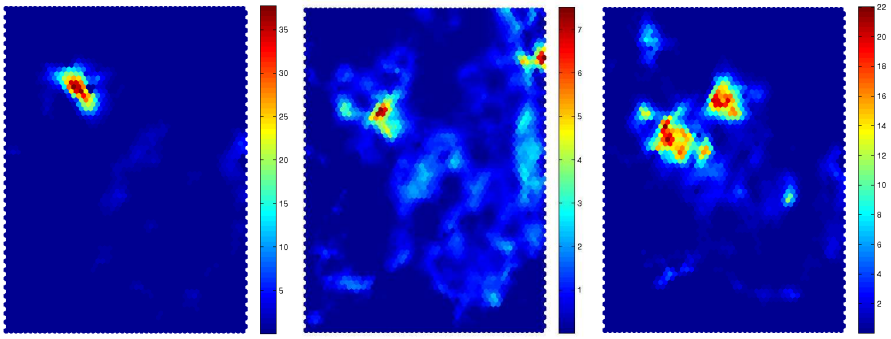
Figure 3.7: Distribution of the 7 classes over the sounds of the test sound library as provided by (a) an expert listener, (b) the SVM trained on the reference vectors of another SOM used for training, (c) a second listener. The not coloured SOM units are either not represented in the sound library or refer to some less frequent sound classes which were not considered (church bells, different kinds of alarming sounds as horns, etc.) or refer to mixtures of co-occurring sounds.

3.4 Measure of similarity for specific sound recognition

In many applications the recognition of a specific type of sound, typically called *target sound* is needed. In this section a case study of the model proposed in Section 2.4.4 is presented and its performance discussed. A microphone was placed at 20 m from railways in the suburbs of the Belgian city of Ghent. A SOM was trained based on the feature vectors calculated from continuous data collected during October and November 2012. A few months later two separate

recording sessions were performed on two different days for a total of 3.5 hours of recordings. Thirty-seven train passages were detected by listening and are considered as the target sounds of this case study. The duration of the train passages is highly variable, from 7 seconds up to 24 seconds, depending on type of train, speed and number of railcars. The target sounds were divided randomly in two equally populated groups (18 train passages in the first group and 19 train passages in the second) for a 2-fold cross-validation test, i.e. the two groups are both used for training and testing.

The excitation pattern for every train passage in the training group is calculated and averaged over time as explained in Section 2.4.4, resulting in $N_t r$ excitation maps, E_i, $i = 1, .., N_t r$, where $N_t r$ is the number of train passages in the training set, $N_t r = 18$ or $N_t r = 19$, for the two cross-validation tests. Next, a $k-$means clustering is performed in order to define the excitation prototypical maps. The optimal number of cluster, three, was first found by using the Davies-Bouldin index (93). The corresponding excitation prototypes, $\tilde{E}_1$, $\tilde{E}_2$ and $\tilde{E}_3$ are shown in Figure 3.8. The average train passage durations related to



(a) (b) (c)

Figure 3.8: The three excitation prototypical maps for train sound detection after clustering.

the three prototypes $(\bar{t_1}, \bar{t_2}, \bar{t_3})$, were used as time windows for measuring the sum of the excitation of the SOM units related to the incoming sound:

$$(\sum_{t}^{t+\bar{t_1}} E(t,x,y), \sum_{t}^{t+\bar{t_2}} E(t,x,y), \sum_{t}^{t+\bar{t_3}} E(t,x,y)), \tag{3.1}$$

A train passage belonging to the testing group is considered being detected (*true positive*) if at least one of the dynamic thresholds: $T_i(t) = 1$, at least once

during a broad window centred around the train passages. Vice versa, it would be considered not detected (*false negative*) if $T_i(t) \neq 1$, for all thresholds in the same broad window. An erroneous detection (*false positive*) occurs if at least one of the dynamic thresholds $T_i(t) = 1$ for t not belonging to time windows related to train passages. The performance of this model can be studied by means of the following statistical measures: *precision* Pr, *recall* Re and their combination called F_1-*score* F_1, defined as follows:

$$\begin{aligned} Pr &= \tfrac{TP}{TP+FP} \\ Re &= \tfrac{TP}{TP+FN} \\ F_1 &= \tfrac{2Pr \cdot Re}{Pr+Re} \end{aligned} \tag{3.2}$$

where TP, FP and FN are respectively the number of true positives, false positives and false negatives. In the left graph of Figure 3.9 such measures are plotted in function of the threshold parameter T_0: $Re = 1$, i.e. $FN = 0$, for $T_0 \leq 0.12$, $Pr = 1$, i.e. $FP = 0$, for $T_0 \geq 0.21$ and $F_1 \approx 0.9$ for $0.1 < T_0 < 0.2$. In particular, for $T_0 = 0.12$ all trains from the testing group are correctly detected and three false positives occur. Very similar results are obtained interchanging the roles played by training and testing sets as shown in the right graph of Figure 3.9.

The same model has been also tested using as input the 1/3-octave band spectrum at 0.125 s instead of the excitation of the SOM nodes. Slightly better results are obtained in this study case. However, it is probably due to its peculiarities: very loud train passage events, low background noise and the almost total absence of road noise sources. In fact, this simplified approach results in many false positives if tested in a location situated in an urban area.

3.5 Contribution to a multi-criteria approach of measurement anomaly detection

Nowadays, it is possible to deploy a distributed noise sensor network due to the technological development of low-cost consumer grade microphones. However, a strong and as much as possible automated quality control of the measured data is necessary in order to handle the lower reliability of such microphones. In sound sensor networks, four type anomalies can be identified:

- *abrubt fault* or *failure*: breakdowns leading to a significantly deviating

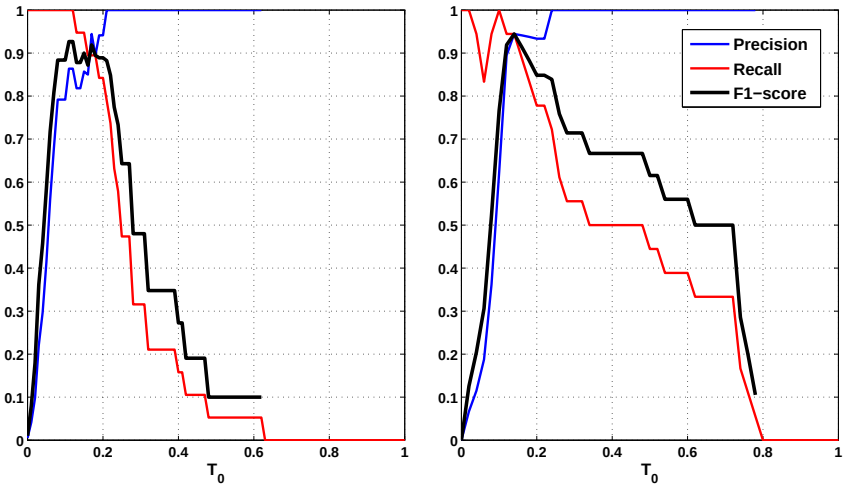


Figure 3.9: Precision, recall and F_1-score of the 2-fold cross-validation test as functions of the threshold parameter T_0.

behaviour of the whole measurement chain

- *incipient fault*: small and often slowly developing continuous fault such as for example sensor drift

- *temporary fault*: temporary wrong measurements, for example due to transient harsh environmental conditions

- *unexpected and rare sound events*: sound events considered atypical compared to the sonic environment experienced where the microphone is placed, thus altering long term noise exposure evaluation indicators such as L_{day}, L_{evening}, L_{night}or L_{den}. Malicious alteration of the acoustic environment around the unattended microphones to voluntarily modify the noise measurement is another example of this type of anomaly.

In order to detect these anomalies, a multi-criteria approach consisting of four quality models has been proposed. Each criterion results in a quality score between 0 and 1 and is aggregated into a global score, by means of an ordered weighted average (OWA) operator (100). The quality control system

was developed as a team effort and has been published by Dauwe et al. (101)[1]. In this section the personal contribution to the multi-criteria approach will be discussed, i.e. the quality model based on SOM and the calculation of the global score by means of the OWA aggregator.

After an extensive training on the sonic environment in the proximity of the microphone, a SOM can be used to detect anomalies in particular of the last two categories. In fact, after the CSL is performed, the distance of the BMU to the incoming sound feature vector can be seen as a measure of how typical the sound is within the sonic environment at the measurement location. So, if the distance to the BMU is very high (higher than a fixed threshold), then the incoming sound can be defined as an outlier for the given acoustic environment, and a low quality score is assigned. The average distance to the BMU over a minute period, $d_{\mathrm{BMU}}(t)$, is transformed into a quality score Q_S via the following function:

$$Q_S(t) = e^{-\ln 2 \frac{d_{\mathrm{BMU}}(t)}{T_{mid}}} \tag{3.3}$$

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

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) operators form a very flexible and tunable parametrized class of averaging type aggregation operators. Many notable operators such as the minimum, maximum, arithmetic average and median are members of this class.

The input of the aggregator is the vector $\mathbf{Q}(t)$ composed of the 1 minute based quality scores of the anomaly detection models and the previous output of the same aggregator, $Q_A(t-1)$:

$$\mathbf{Q}(t) = [Q_I(t), Q_H(t), Q_D(t), Q_S(t), Q_A(t-1)], \tag{3.4}$$

where $Q_I(t)$, $Q_H(t)$ and $Q_D(t)$ are the partial quality scores calculated by three other models described by Dauwe et al. (101)which will be not discussed in this work. A noticeable feature of the OWA aggregator class is its versatility to work with a variable number of quality scores, a situation which can occur in case one or more quality models are temporarily not working or new models are added.

[1]It will be reported also in the PhD dissertation of Samuel Dauwe.

The weight vector $\mathbf{W} = [w_1, w_2, \cdots, w_n]$ containing the weights of the OWA is calculated as follows:

$$w_j = F(\frac{j}{n}) - F(\frac{j-1}{n}), \quad j = 1, ..., n, \tag{3.5}$$

where $F(r) = F_\alpha(r) = r^\alpha$, n and j are respectively the dimension and an index of the ordered vector [2]:

$$\tilde{\mathbf{Q}}(t) = [Q_1(t), Q_2(t), \cdots, Q_n(t)], \quad j = 1, ..., n, \tag{3.6}$$

from the largest value to the smallest one. Note that $\tilde{\mathbf{Q}}(t) = \mathbf{Q}(t)$ 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 affects the orness of the OWA defined as:

$$orness(W) = \frac{1}{n-1} \sum_{j=0}^{n-1} j\, w_{n-j}. \tag{3.7}$$

The weight vector should be calculated only in case the dimension of the quality vector changes, i.e. if a quality model has been added or for some reason has ceased to work. Finally the aggregated quality can be calculated as a weighted average:

$$Q_A(t) = \mathbf{W} \cdot \mathbf{Q}(t) = \sum_{j=1}^{n} w_j \cdot Q_j(t). \tag{3.8}$$

The adjective ordered comes from the fact that the quality scores $Q_j(t)$ are ordered. Following the definitions in Equations 3.5 and 3.6, the OWA operator results in the arithmetic mean if $\alpha = 1$ ($orness(W) = 0.5$), the maximum if $\alpha = 0$ ($orness(W) = 0$) and the minimum if $\alpha \to \infty$ ($orness(W) = \frac{1}{n-1}$). The aggregated quality $Q_A(t)$ 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.

In the application discussed here $\alpha = 2$, so that $\mathbf{W}$ results to be:

$$\mathbf{W} = [0.04, 0.12, 0.2, 0.28, 0.36], \quad orness(W) = 0.7, \tag{3.9}$$

In this way Q_A is very sensitive to Q_5, the smallest quality indicator. This choice is motivated considering the fact that each criterion has been developed for detecting a different kind of anomaly: high orness is therefore important to

[2] In this specific case: $n = 5$.

detect every anomaly. In Figure 3.11 the quality scores of all four methods and the aggregated quality for a period of two entire days are shown. The windshield of the microphone was first attacked by birds around 10 pm of the first day and eventually became detached after a few hours in the night. As illustrated in Fig 3.10, during the detachment the $L_{\mathrm{Aeq,1min}}$ increases anomalously, afterwards it decreases to almost the same sound level as before the detachment, but with an increased noise floor due to the wind-related noise. This example was chosen because the other quality criteria failed to detect the anomaly. For an overview of all possible anomalies and other examples, please refer to (101).

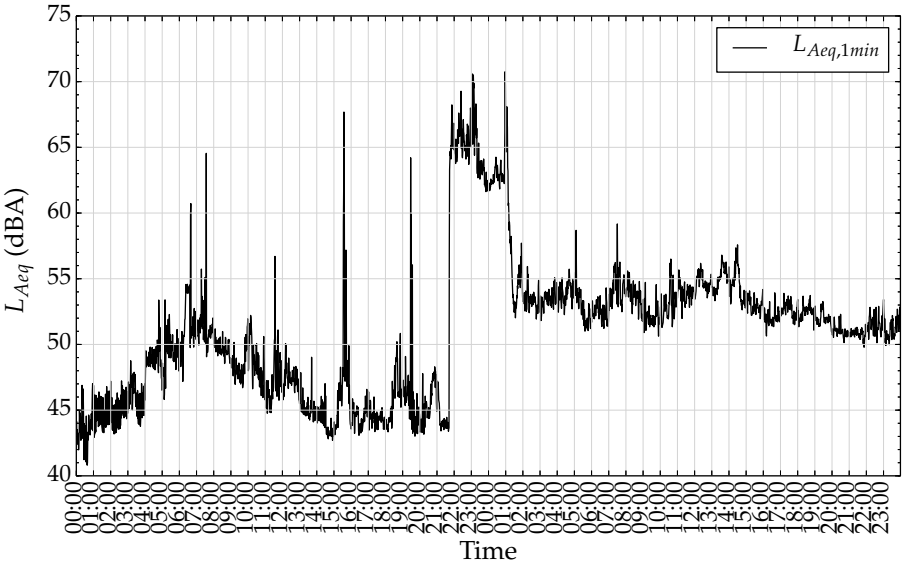


Figure 3.10: $L_{\mathrm{Aeq,1min}}$ graph illustrating a failure in which the windshield became detached from the microphone after being attacked by birds, resulting in anomalous peaks followed by a slightly increased noise floor after the detachment.

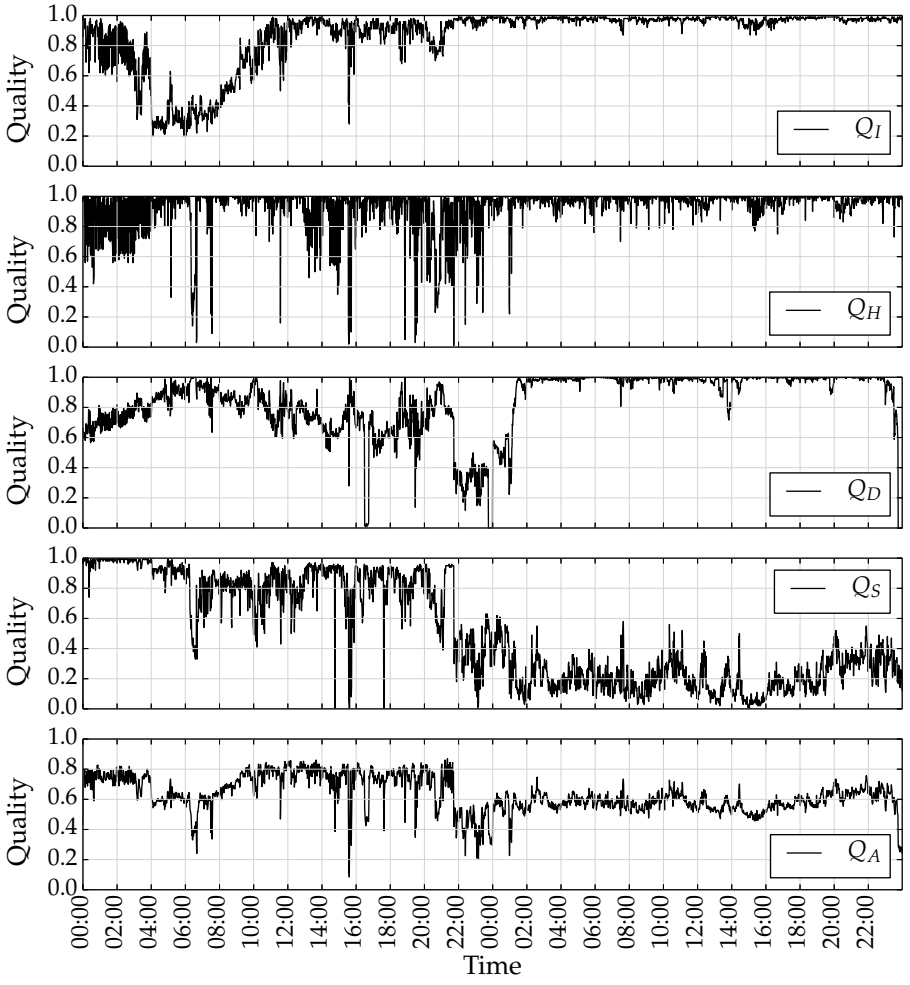


Figure 3.11: From top to bottom: Q_I, Q_H, Q_D (quality scores not considered in this work), Q_S and the final aggregated quality score Q_A. Only Q_S seems to detect the detachment of the windshield and the aggregated quality Q_A decrease thus moderately.

Chapter 4
Conclusions and future work

This chapter presents several conclusions regarding the model and its applications. Special attention is given to future work, needed to improve the main aspects of the model in order to increase the performance and the biologial plausibility in analyzing the auditory scene and deploy it in extended long-time running sound sensor networks.

The model presented in this work consists of several submodels:

1. processing the incoming sound signal in a broad set of sound features mimicking the peripheral auditory processing and assessing its auditory saliency;

2. topographically mapping the sound features based on co-occurrence and a learning technique — continuous selective learning, CSL —conceived within this work for working with long time series and therefore tuning a SOM to the typical heard sounds, thus aiming to replicate the effects of the underlying mechanisms of human learning;

3. auditory object formation via LEGION replicating the oscillatory correlation properties of neurons in the sensory cortex;

4. modelling the main aspects of human auditory attention and competitive selection making a distinction between bottom-up and top-down attention;

5. attaching meaning to sounds via automated labeling by means of a SVM coupled to the SOM trained at a given location;

6. identifying specific target sounds by means of a fuzzy excitation model using the trained SOM and a similarity measure.

Moreover, several important applications based on (part of) this model are presented in Chapter 3. In the remainder of this conclusive chapter, the key aspects and findings will be presented with a view to the unsolved issues to be tackled in future research.

The first step of the computational model for peripheral auditory processing presented in Section 2.1 was based on 1/3-octave band spectrum from 20 Hz to 20 kHz of monaural sound signal performed at a temporal resolution of 1/8 s. This approach presents surely its limitations as regards biological plausibility, but it is needed for maintaining a high level of computational feasibility being supported by all off-the-shelf sound measurement equipment in use nowadays. The cochleagram performed using the Zwicker loudness model accounted for energetic masking, which is a key aspect of peripheral human auditory processing. The receptive fields in the auditory cortex were mimicked by means of multiscale centre-surround mechanisms aiming to encode intensity, spectral and temporal contrast of the incoming sound. Again, biological plausibility was limited by computational feasibility: the human auditory system is particularly proficient in recognizing tonality within the incoming sound signal, an aspect linked to speech recognition by evolutionary reasons. The aim of the model presented in this work is limited to analyse outdoor environmental scenes, therefore speech analysis is not considered. However, the use of simple tonal features could be useful for a better detection and characterization of human voices and bird vocalizations. Another interesting idea for future work would be the use of sound features based on dynamic ripples instead of Gaussian and difference-of-Gaussian filters. Dynamic ripples combine one spectral modulation rate with one temporal modulation rate. They are therefore ideal as primary signals in several neurological researches investigating the auditory stimuli processing in the (primary) auditory cortex (45; 102; 103; 104; 105). In particular, they seem promising in capturing the interactions between spectral and temporal responses in the human cortex (106).

The sound feature vector formed by the set of sound features encoding intensity, spectral and temporal contrast of the incoming sound was used as input for a neural architecture called SOM which was introduced in Section 2.2 and was first inspired by topographic mapping of various regions of the sensory cortex and its plasticity during learning. The use of sound features based on dynamic ripples coupled with SOM should improve the biological plausibility of the model. Selective tuning to combined spectro-temporal modulations occurs in the human primary and secondary auditory cortex as demonstrated by means of dynamic ripples (103), which were also used in studies on receptive fields in animal auditory cortex (45; 107; 108). In this work continuous selective

learning, a specific learning technique, was introduced which aims to mimic the tendency of auditory system to enhance the learning of salient although not so often occurring sounds. In particular, a measurement of the auditory saliency was used as learning modulator.

Auditory object formation was proposed coupling each SOM unit with a neuron of a neural oscillatory network called LEGION, thus expressing two functionalities of ideal neurons: long term memory formation based on SOM training and the dynamic oscillatory correlation of sensory cortex neurons excited by an auditory stimulus schematized by LEGION. However, the transient phase occurring after each change of the external stimulation disrupts the oscillation periodicity making thus impossible any attempt to follow the auditory object(s) and limiting the use of this kind of networks in cognitive computational models. More appropriate computational oscillatory models are thus needed.

However, some concepts of LEGION dynamics such as local excitation and global inhibition were used in an ANN conceived to model auditory attention: IOR mechanisms and bottom-up attention were modelled as well as the way to introduce in the model the top-down focused attention. ANN can produce auditory objects without the help of any supervised learning technique. However, an issue not tackled by this research is the link among such auditory objects in the neural network and the auditory objects as perceived by humans. This issue is due to the fact that only unsupervised learning was considered in this research, while it is clear that supervised and reinforcement learning strategies should be considered as well. A second aspect not taken into account in this work is the process of attaching meaning to auditory objects, therefore identifying sounds. It has been shown that the meanings attributed to sounds act as a determinant for soundscape quality evaluations (109; 110), and therefore identification of sounds is an important factor in the context of soundscape design; sounds that are not identified are expected to influence overall soundscape appraisal to a lesser degree. The model presented by Boes et al. (111) could help to solve these issues by introducing a concept layer and a human-based labeled sound library used for the supervised learning of the ANN. In the same work it is also shown how the role of the "teacher" is important at an early stage but decreases as the training runs. However, attaching meaning to well-defined sound events by means of labels as a human listener would do is not trivial: the influence of inter-individual differences and linguistic issues have to be solved (110). Inter-individual differences are very often related to the context in which the sound event occurs, an aspect of utmost importance. For example, the sounds of the shoreline and of a distant highway are very similar: only the context could solve the ambiguity. Or, from a different perspective, the

same sounds could be described differently depending on the familiarity of the listener with them. Just using the same example, a man living in a sea town would describe those two sounds as being recorded at the shoreline, while people living in developed urban areas would labeled them as coming from a distant highway. A human listener would typically describe a known sound event by either its source or by the action generating it or even using both together: the labels "tram", "braking" and "tram braking" are just an example of these three different ways to attach a meaning to the heard sound and actually none of these three labels can be considered more correct than the others, although the third label seems the more precise. In the model combining the trained SOM with a SVM presented in Sections 2.4.3 and 3.3 a small set of standardized labels referring exclusively to the sound source were used by the expert listener: car, truck, motorcycle, people, bird, tram and background. The sound library, composed of 2369 sound samples, did not contain sufficient sound samples of several less frequently occurring sound sources as bells, shutting doors, etc. and so they could not be used for training the SVM on these other sound categories. Therefore, this approach and the work of Boes et al. (111) show that a much broader human labeled sound dataset is urgently needed and it would represent an important step towards human-like automated labeling. The solution proposed by the acoustics group of INTEC-UGent is to use the broad diffusion of internet and mobile apps to reach as many potential "teachers" as possible. The sounds composing the sound libraries and many other sounds are organized in a tailored database and coupled with an online game [1] where participants are asked to listen to sound excerpts and to label them in order to score. This approach assures not only the construction of a very broad dataset of labeled environmental sounds but also preserves the inter-individual variability in labeling. Moreover, such variability in labeling can be used for tuning the meaning attachment phase on the desired level of accuracy.

Although the future developments discussed here are needed and seem very promising, the current model can already produce remarkably interesting results. The acoustic summary, as explained in Section 3.2, revealed to be an interesting tool for detecting and selecting the most representative sound events as humans would do, thus characterizing the soundscape of a specific location. The validation test performed by people living in the surroundings of the microphone locations demonstrates the goodness of the automatically constructed acoustic summaries. In particular, the selection of sounds based on both saliency and frequency of occurrence seems to be the most simple and valid strategy for representing a soundscape. Moreover, it has been shown that

[1] `www.noiseplay.org`

the model can be used for designing future soundscape scenarios too. The SOM can be trained on artificial mixtures of sounds which simulate the sonic environment after an urban planning intervention. Exploring the sound library, specific regions of the SOM related to the new sounds can be found and the influence in the perceived soundscape can be estimated by means of the auditory attention modelling ANN.

The SOM architecture combined with the CSL can also be a starting point for specific applications in environmental acoustics such as target sound recognition (see Sections 2.4.4 and 3.4) and anomaly detection (see Sections 3.5 and 3.4). The distance to the SOM units of a set of manually selected target sounds can be used to create a prototypical excitation map, a sort of distinctive imprint. The excitation due to the incoming sound can therefore be compared to such map and detection occurs by measuring its similarity with the excitation map. Successfully tested on train passage detection in Section 3.4, this technique is worth to be further tested on different target sounds. The SOM can also be integrated in a multi-criteria approach for anomaly detection. The distance of the BMU is a non-linear measure of how typical the sound is within the sonic environment at the measurement location. Therefore, the incoming sound can be considered not typical of the given acoustic environment when the distance to the BMU is high. A low quality score is therefore assigned thus making the detection of unexpected and rare sound events possible. Not only, frequent low quality scores could also indicate malfunctions, failures or deliberate tampering.

The model here presented has been partly developed in the context of the IDEA project [2] (Intelligent Distributed Environmental Assessment), focussing on traffic related environmental stressors such as air pollutants and noise. In such context, an important goal was the development of an extensive distributed measurement network maintaining the necessary level of computational feasibility of long-running data-driven applications. A well structured database model based on the Open Geospatial Consortium (OGC) observations and measurements standard (O&M) was needed in order to define a domain-independent, conceptual model for representing standardized spatiotemporal data. In particular, the 1/3-octave band spectrum calculated by all sensor nodes was continuously stored in such database, which was queried in order to retrieve the necessary data to run the computational model presented in this work. Moreover, in order to construct and maintain a well-ordered sound library in real-time as defined in Section 2.3 a specific database has been developed. The next challenge would be to implement such model in a flexible network architecture making use of multi-agent systems (MASs) in order to integrate

[2] `www.idea-project.be`

the model here presented in a computing environment capable of managing and optimizing several applications autonomously (112). In this way the applications presented in Section 3 could run simultaneously for several microphones enhancing sensibly the computational efficiency.

Appendix A

Context-dependent environmental sound monitoring using SOM coupled with LEGION

Damiano Oldoni, Bert De Coensel, Michaël Rademaker, Timothy Van Renterghem, Bernard De Baets, Dick Botteldooren

Published in *Proceedings of IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* Barcelona, Spain, pp. 1413-1320, 2010.

Environmental sound measurement networks are increasingly applied for monitoring noise pollution in an urban context. Intelligent measurement nodes offer the opportunity to perform advanced analysis of environmental sound, but trade-offs between cost and functionality still have to be made. When using a tiered architecture, local nodes with limited computing capabilities can be used to detect sound events of potential interest, which are then further analysed by more powerful nodes. This paper presents a human-mimicking model for detecting rare and conspicuous sound events. Features encoding spectro-temporal irregularities are extracted from the sound, and a Self-Organizing Map (SOM) is used to identify co-occurring features, which most likely belong to a single sound object. Extensive training allows this map to be tuned to the typical sounds that are heard at the microphone location. A Locally Excitatory Globally Inhibitory Oscillator Network (LEGION) is used to group units of the SOM in order to construct distinct sound objects.

A.1 Introduction

Advances in the design of low-cost computing devices and sensors, together with an increase in bandwidth and covering power of low-cost wireless networks, are forming a technological push for the use of wireless sensor networks (113). Acoustical sensor networks in particular provide a wide range of applications, such as audio surveillance for public security (114; 115), habitat monitoring (116; 117) or environmental noise pollution monitoring (118; 119). Information retrieved from the latter could be used to assess potential noise annoyance or sleep disturbance, to validate noise maps or even to locally steer activities, e.g. via intelligent traffic systems.

Although the hardware, storage capacity and communication bandwidth needed for building environmental sound measurement networks is increasingly becoming cheaper, trade-offs between cost and functionality still have to be made. For example, it is infeasible to perform advanced sound source recognition using small, cost- and energy-efficient nodes, while it is also infeasible to simply record and transmit the sound at all microphones continuously, due to data storage and transmission bandwidth limitations. A solution for this problem is to use a tiered architecture (see e.g. (117)), in which the spatial resolution of the network is exploited by using cheap local nodes with limited computing capabilities, which select and transmit sound fragments of possible interest to be processed by more powerful nodes (usually centrally located).

One of the most basic techniques for sound event detection is thresholding: when the instantaneous sound pressure level exceeds a predefined threshold, the occurrence of a sound event is assumed, and the node starts recording for a given period of time. In case of adaptive thresholding, the threshold is relative to the background level, which can vary in time slowly (120). More recently, a number of techniques for selecting salient parts of the auditory scene have been proposed, inspired by the neural mechanisms that guide human attention (34; 70; 71). However, a major disadvantage of current techniques is that no distinction is made between frequently occurring and thus expected sound events, and rare events. Moreover, the kind of expected sound events depends on the context of the microphone. For example, the sound of birds singing can be expected near a microphone situated inside an urban park, while the sound of cars passing by is expected in a busy street.

The ideal node in an environmental sound measurement network for monitoring noise pollution should, in a computationally efficient way, be able to learn and discern the sounds frequently occuring at the location of the microphone, thus distinguishing between common and rare or conspicuous sound events. In

this paper, we show how this goal could be achieved using a simple biologically inspired technique.

Features encoding spectro-temporal irregularities are extracted from standard 1/3-octave band levels, which can be measured with off-the-shelf sound level meters. Subsequently, sound events are discerned using a combination of two types of neural networks: a Self-Organizing Map (SOM) (75) that allows—after extensive training—to identify co-occurring sound features and a Locally Excitatory Globally Inhibitory Oscillator Network (LEGION) (84) for grouping and segregation of corresponding sound fragments. The combination of both neural networks models two essential features of the brain: the SOM mimics the plasticity (during the learning phase) and complex morphology of the network of neurons forming the auditory cortex, while the LEGION approximates the dynamic oscillations between connected neurons.

In Section A.2 we provide a description of the coupled SOM-LEGION network, starting from the sound extraction, and the specific solutions adopted. The model was applied in different real scenarios: the results and some discussions are provided in section A.3. Finally, a section with conclusions follows in A.4.

A.2 Methodology

A.2.1 Sound feature extraction

In a first stage, a *feature vector* is extracted, at regular time intervals, from the sound signal measured by the node microphone. Instead of calculating a detailed time-frequency representation of the raw sound wave, the model starts from the 1/3-octave band spectrum, calculated with a temporal resolution of 1 s. This procedure has the main advantage that off-the-shelf sound measurement equipment can be used as a front-end, which reduces the computational load on the measurement node. The choice of time resolution can be justified by noting that the sounds of main importance for environmental noise pollution monitoring (cars, trains, aircraft, fans etc.) have a relatively slow varying temporal envelope (61; 121). A simplified cochleagram $s(f, t)$ is then calculated using the Zwicker loudness model (9), which accounts for energetic masking. The complete hearable frequency range is considered (0 to 24 Bark) with a spectral resolution of 0.5 Bark, resulting in 48 spectral values at frequencies $f_j = \frac{1}{2} j$ Bark, for each timestep.

The mechanism for extracting the feature vector, which characterizes the amount of novelty in the sound signal, is inspired by the way the human

auditory system biases its attention toward particularly conspicuous events. The auditory system is, next to absolute intensity, also sensitive to spectro-temporal irregularities. Based on existing models for auditory saliency (34; 70; 71), the proposed model calculates measures for intensity, spectral and temporal modulation using a center-surround mechanism, which mimicks the receptive fields in the auditory cortex. In particular, multi-scale features are calculated in parallel by convolving the cochleagram with various 2D gaussian and difference-of-gaussian filters $g_i(f,t)$. The former encode intensity, while the latter subtract between a "center" fine scale and a "surround" coarser scale, and encode the spectral and temporal gradient of the cochleagram at 16 scales (4 for intensity, 6 for spectral contrast and 6 for temporal contrast):

$$r_i(f,t) = (s * g_i)(f,t) \tag{A.1}$$

with $i = 1, 2, \ldots, 16$. Fig. A.1 shows a section of the filters along the time or frequency axis. Finally, a feature vector $\vec{r}(t)$ is constructed at each timestep, consisting of $16 \times 48 = 768$ values:

$$\vec{r}(t) = \sum_{i=1}^{16} \sum_{j=1}^{48} r_i(f_j,t)\vec{e}_{48(i-1)+j} \tag{A.2}$$

with $\{\vec{e}_k : 1 \leq k \leq 768\}$ the standard basis for the 768-dimensional Euclidean space.

A.2.2 Feature co-occurrence analysis: Self-organizing map

The self-organizing map (SOM), an abstract mathematical model of topographic mapping from the (visual) sensors to the cerebral cortex (77), is most often described as an unsupervised technique for the visualization of high-dimensional data (78). It does so using typically a 2D network of *units* or *nodes*. Their representation in the high-dimensional space is provided through *reference vectors*. After initialization, their coordinates are modified during the training process wherein the following (vastly simplified) steps are repeated until a stopping criterion is met:

1. Feed an input high-dimensional data point to the SOM.
2. Determine the best-matching unit (BMU) i.e. the unit corresponding to the closest reference vector.
3. Move the reference vector corresponding to the BMU and, to a lesser

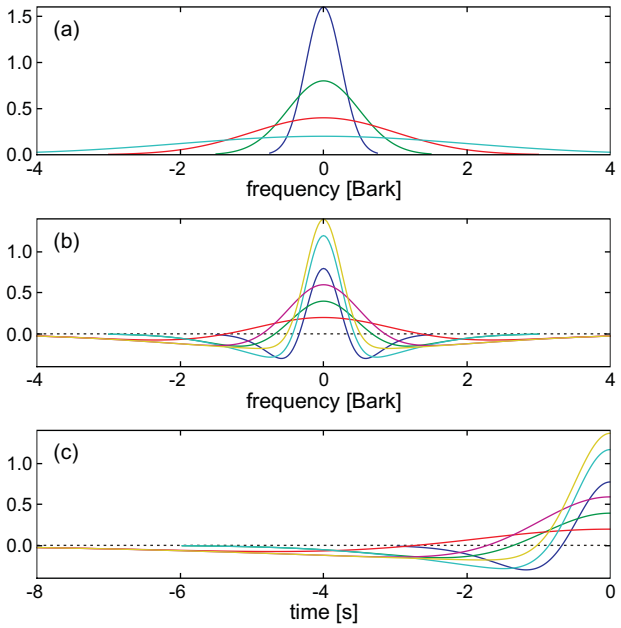


Figure A.1: Cross section of the receptive filters that are used to calculate (a) intensity, (b) spectral contrast and (c) temporal contrast. For the latter, causality is preserved by only convolving with the past.

extent, those of the neighbouring units in the 2D grid, closer to the input high-dimensional data point.

In practice, the training process and the resulting SOM are strongly influenced by a number of parameters, such as the size of the SOM, the type of initialization of the units, the strength of learning and the type of neighbourhood considered in the third step, as well as the evolution over time of the learning parameters.

Nevertheless, after training it is clear that the frequency distribution of the input data in the high-dimensional space will be approximated by the reference vectors of SOM units, possibly leading to dense high-dimensional clusters interspaced by regions where the reference vectors of the SOM units are more distant. This emerging order is the basis for the effective visualization in the SOM. Consequently, the SOM can also be considered to perform a kind of abstraction, compressing information while preserving the most important high-dimensional relationships (75). A trained SOM could then be understood as a nonlinear 2D projection of the probability density function of the high-dimensional input data. An intuititive quantification of the SOM quality is then

the average high-dimensional distance of a set of data points to their respective BMUs.

Now, we provide a brief, more formal description of the SOM technique, based on the description in (78). More formally, we consider an n-dimensional input space $\mathbb{R}^n$, in our application the 768-dimensional space of raw sound features. The SOM units are represented by the reference vectors $\vec{m}_i \in \mathbb{R}^n$, with index i identifying the unit. The M units in the 2D network are aligned on a regular M_x by M_y grid and are represented as $\vec{\mathbf{m}}_i = (\mathbf{m}_x, \mathbf{m}_y) \in \mathbb{R}^2$. As the vectors $\vec{m}_i$ are adapted during training, we will write $\vec{m}_i(t)$ to denote the vector at time-step t during training, and use $\vec{m}_i$ only when training is complete. Input data is represented as $\vec{r} \in \mathbb{R}^n$, and at time-step t, the sample $\vec{r}(t)$ is processed by the SOM. The BMU at time-step t is then found by considering

$$c(t) = \arg\min_i \|\vec{r}(t) - \vec{m}_i(t)\| \ . \tag{A.3}$$

Thus, at time step t, $\vec{m}_{c(t)}(t)$ denotes the BMU for the input sample $\vec{r}(t)$. Adapting the BMU or, indeed, any unit, is then performed as follows:

$$\vec{m}_i(t+1) = \vec{m}_i(t) + h_{c(t),i}\left(\vec{r}(t) - \vec{m}_i(t)\right) \ , \tag{A.4}$$

where h, the neighbourhood function, performs a non-linear smoothing selection on the discrete 2D neighbourhood structure. Often used is a Gaussian function of the distance between the BMU at time step t, $c(t)$, and the generic unit i:

$$h_{c(t),i} = \alpha(t) \exp\left(-\frac{\left\|\vec{\mathbf{m}}_i - \vec{\mathbf{m}}_{c(t)}\right\|^2}{2\sigma^2(t)}\right) \ . \tag{A.5}$$

The time-step dependent parameters governing the behaviour of this type of neighbourhood function are the learning rate $0 < \alpha(t) < 1$ and the width of the 2D neighbourhood $\sigma(t)$. Both are monotonically decreasing in t:

$$\alpha(t) = \alpha_0 \frac{C}{C+t}, \quad C = \frac{N}{100}, \tag{A.6}$$

$$\sigma(t) = 1 + (\sigma_0 - 1)\left(\frac{N-t}{N}\right) \tag{A.7}$$

where N is the number of samples. Observe that $h_{c,i} = \alpha(t)$ only for the BMU, and is strictly decreasing for units farther from it in the 2D grid. Thus, for a constant similarity, the BMU is adapted to a stronger extent than any neighbouring units.

A final point concerns the visualisation of the 2D grid after training. Due to the high dimensionality of the raw feature space, the visualization of the trained map via projection on particular planes is rarely informative—one could argue that if such an approach would lead to satisfactory results, there was less need to apply the SOM algorithm in the first place. Rather, in order to easily identify regions with similar high-dimensional representations, it will be more informative to display how close in the high-dimensional space a unit in the map is to its neighbouring units. In fact, a typical way to visualize the morphology of the map uses the so-called U-matrix (82), which is a matrix of dimensions $[2M_x - 1, 2M_y - 1]$ containing both the distances between the nearest neighbours and their average. Color-coding the units on the map on the basis of their average distance to their nearest neighbours allows distinguishing regions where 2D neighbouring reference vectors are similar, from regions of high variability. We provide an example in Section A.3.

When training is complete, the SOM quality can be assessed on the basis of two concepts. The first is the average distance between each input vector from a set of test samples and its BMU, the so-called *average quantization error* E. It is computed as follows for a set of test samples $\vec{r}(1), \ldots, \vec{r}(N)$:

$$E = \frac{\sum_{t=1}^{N} \left\| \vec{r}(t) - \vec{m}_{c(t)} \right\|}{N}, \tag{A.8}$$

with $\vec{m}_{c(t)}$ now denoting the BMU for test sample $\vec{r}(t)$.

The second concept is the *topographic error*, the proportion of test samples for which the BMU and the next-best-matching unit are not neighbours. A low topographic error can be considered to be indicative of a focused SOM, clustering the units around the dense regions in $\mathbb{R}^n$.

These concepts complement each other quite well: if all the units are widely spaced in the $\mathbb{R}^n$ space formed by the eigenvectors, the SOM is very likely to obtain a quite low average quantization error, while the topographic error is likely to be large. If, in contrast, the units are packed too tightly around the densest regions in $\mathbb{R}^n$, the average quantization error can be expected to be high, while the topographical error is expected to be low.

To reduce both the average quantization error and the topographic power of the map, it is usually sufficient to reduce the initial width of the neighbourhood function σ_0 and/or the learning rate α_0, making the map less flexible, while simultaneously increasing the number of training runs to compensate for the slower learning (75).

A hexagonal lattice was used in this paper, allowing a 2D grid of equal-spaced units while maximizing for any value of $\sigma(t)$ the number of neighbours in the grid.

The unit reference vectors were initialized by the linear initialization function, resulting in a regular array of vectorial values that lie on the subspace spanned by the eigenvectors corresponding to the two largest principal components of input data used during the training (78). In our application, the high-dimensional space is composed of the raw sound features, meaning each unit corresponds to an abstract prototype of a sound. The goal is thus to group similar (in the raw feature sense) sound fragments in the SOM. Sound feature values that often arise together, and are thus often part of the same sound fragment, are then expected to have the same BMU, or to even cluster close together in the SOM. In order to allow this behaviour to arise, a proper choice of features is crucial, as well as proper values for the parameters governing the SOM construction, training and resulting performance.

The training phase has to take into account a very large number of input data: in our case 86400 samples (the number of seconds in one day) were used. Afterwards, the trained SOM is ready to receive new data samples and localize the BMU.

As we will now show, a natural link between SOM and LEGION then arises: the similarity of a raw feature vector and a specific SOM unit is, from a neural oscillatory point of view, a measure of the external stimulation that a LEGION oscillator receives. Conceptually the SOM unit and the LEGION oscillator can be considered the same formal neural unit. In fact, the two neural networks are the expression of two different functionalities of ideal neurons: the long term memory formation is modeled by the SOM extensive training while the dynamic oscillatory correlation of sensory cortex neurons excited by an auditory stimulus is schematized by LEGION. The LEGION network model and the details of the SOM-LEGION coupling are developed in the next section.

A.2.3 Segregation: LEGION

Increased insight in the oscillatory correlation properties of the neurons in the sensory cortex during the 1980s resulted in an increase in theoretical research on possible computational models of the corresponding biological mechanisms. One of the first models thus constructed, by von der Malsburg and Schneider (122), was later extensively developed in the auditory context by Wang (87; 88) using a so-called *shifting synchronization theory*, based on oscillatory correlation, where neuronal oscillators representing the neuronal counterpart of specific sound features are used.

In that context, each sound object was represented by synchronization of a group of oscillators corresponding to the relative sound features. Contrarily, desynchronization among different groups of oscillators meant that the sound is

the sum of different auditory streams.

The Wang model is based on a particular network architecture referred to as LEGION (84): it is generally composed of a 2D grid of oscillators, in our coupled SOM-LEGION architecture corresponding to the units in the SOM. The dynamics of the i-th oscillator is the combined activity of an excitatory unit x_i and an inhibitory unit y_i:

$$\dot{x_i} = 3x_i - x_i^3 + 2 - y_i + I_i H\left(p_i - \varphi\right) + S_i + \rho\,, \tag{A.9}$$

and

$$\dot{y_i} = \epsilon\left(\gamma\left(1 + \tanh\left(x_i/\beta\right)\right) - y_i\right)\,, \tag{A.10}$$

where I_i is the external stimulation, H is the Heaviside function, p_i is the so-called lateral potential, φ is a threshold, S_i is the overall coupling contribution due to the near oscillators of the network and $\rho < 0$ is a source of Gaussian noise. There are three regulating parameters: γ, ϵ and β, where the last two are small positive constants.

The external stimulation I_i in (A.9), together with the permanent connection weights T_{ik} (explained later), entails the core of the SOM-LEGION coupling. I_i depends on the distance between the input raw feature vectors $\vec{r}(t)$ and the i-th unit of the trained SOM, closely related to the quantization error in the SOM. It is computed as

$$I_i\left(t\right) = IH\left[\left\|\vec{r}\left(t\right) - \vec{m}_i\right\|^{-1} - \lambda M_h\left(t\right)\right], \tag{A.11}$$

where I is a positive constant, H the Heaviside function, $\left\|\vec{r}\left(t\right) - \vec{m}_i\right\|^{-1}$ is a measure of the similarity of the input vector to the i-th unit, $M_h\left(t\right)$ is the h-order simple moving average of the inverse of the distance of the BMU and $0 < \lambda < 1$ is a relative threshold. Because of the use of H, this formulation of the external stimulation can be referred to as a binarization: oscillators similar enough to the raw feature vector are stimulated, while those too far away are not.

It must be clear by now that all variables in (A.9)-(A.10) are dimensionless. It holds true for the variable of integration which is naturally referred as time and that here we call internal time or LEGION time and indicated as t_L; at the contrary in (A.11) the real time is involved. The simplest way to match them is to fix a certain LEGION time interval τ_L and impose the equality $\tau_L = 1\,\mathrm{s}$ thus avoiding the confusion between two different time scales. Returning to

(A.9)-(A.10) it means that:

$$\dot{x_i} = \frac{x_i}{dt_{_L}} = \frac{1}{\tau_{_L}} \frac{x_i}{dt}, \quad dt = \tau_{_L}\, dt_{_L}, \tag{A.12}$$

and the same holds for y_i.

If I_i is positive and $H = 1$, the i-th oscillator produces a near-steady stable orbit between a so-called silent phase (left branch of the $\dot{x}$-nullcline cubic function in (A.9)) and an active phase (right branch). The passage between them occurs at a faster time scale compared to motion within each phase, thus resulting in a sort of jumping. Finally, the parameter γ in (A.10) influences the relative time spent in each phase.

The coupling term S_i is typically composed of two terms:

$$S_i = \sum_{k \in N(i)} W_{ik} H\left(x_k - \theta_x\right) - W_z H\left(z - \theta_{xz}\right) , \tag{A.13}$$

with the first term taking into account the phase of the oscillators in the neighbourhood, $N\left(i\right)$, through the use of *dynamic* connection weights (explained later) and the second term referring to the activity of a global inhibitor z weighted by W_z. If at least one oscillator is in the active phase, $z \to 1$ at a slow time scale whereas $z \to 0$ if all oscillators are in the silent phase, thus allowing the activation of new oscillators (for more details on the form of z and the threshold θ_{xz}, see ((87))).

Terman and Wang (83) formulated a procedure called *dynamic normalization*, significantly speeding up the synchronization within each oscillator block. It involves the dynamic connection weights, which can be assessed from the external stimulation, and the so-called *permanent connection weights*:

$$\dot{u_i} = \eta \left(1 - u_i\right) I_i - \nu u_i \, , \tag{A.14}$$

$$\dot{W}_{ik} = W_T T_{ik}\, u_i u_k - W_{ik} \sum_{j \in Ni} T_{ij}\, u_i u_j - \omega \nu W_{ik} \, , \tag{A.15}$$

where the variable u measures whether the oscillator i is stimulated, the constants $\eta \gg \nu$ are chosen so that u_i tends to 1 quickly if the oscillator i is stimulated, while it relaxes slowly to 0 when it doesn't receive any external stimulation. In (A.15), W_T is the so-called total dynamic connection weight and the last term, not explicitly dependent on u, is here for the first time introduced as a dissipating term weighted by the parameter ω: this term does not affect appreciably the normalization if $\omega\nu \ll 1$. When using this procedure, all the oscillators belonging to the same externally excited group receive the same

amount of coupling from their neighbours, irrespective of whether they are completely surrounded by externally stimulated oscillators or not, being one of oscillators at the border of the group. The T_{ik} are called *permanent* connection weights and, contrarily to the dynamic weights W_{ik}, are fixed between two neighbouring oscillators, being the expression of the hardwired connections in the network. In the SOM-LEGION coupled model, these permanent weights are determined during training, being related to the similarity of two neighbouring units, $\delta_{ik} = \|\vec{m}_i - \vec{m}_k\|^{-1}$:

$$T_{ik} = T_{max}\left[1 + \phi\left(\frac{\delta_{ik} - \delta_{min}}{\delta_{max} - \delta_{min}} - 1\right)\right], \tag{A.16}$$

where the constant T_{max} is the maximal permanent connection weight and $\phi < 1$ is a scaling factor in order to have $(1-\phi)\,T_{max} \leq T_{ik} \leq T_{max}$. Thus, the more similar two units of the SOM are, the higher the coupling between the two corresponding oscillators is.

The study of the dynamics of our LEGION network implies solving hundreds of coupled differential equations, rendering impossible any attempt to process in real-time the massive amount of data acquired by a sound measurement network. To speed up the computational process the *singular limit method* developed by Linsay and Wang (91) is extensively used. This method, in the form of an algorithm, allows skipping most of the computation by considering the fact that the oscillatory system feels the effect of oscillator changes only when oscillators jump up or down: only at those moments the lateral potential and global inhibitor values can change. Thus, the only information needed to know the dynamics of the entire system is the branch occupied by each oscillator and the time at which a jump occurs (for more details on the method, see (91)).

The lateral potential, as implemented in (91), is not suited for dynamic external stimulation $I(t)$. In this paper a different and simpler approach was used: at the end of each cycle of the algorithm the active oscillators that do not have at least 1 of 6 neighbours active are forcedly inhibited by moving them to the left branch.

A.3 Results

In our work we have focused on two different sound scenarios: a typical urban sound environment defined by a mixture of light and heavy traffic noise, labelled as T, and a park, with typical natural sounds and only marginally affected by human presence, labelled P. Two fixed measurement stations, one for each scenario, recorded standard 1/3-octave band levels calculated with a time

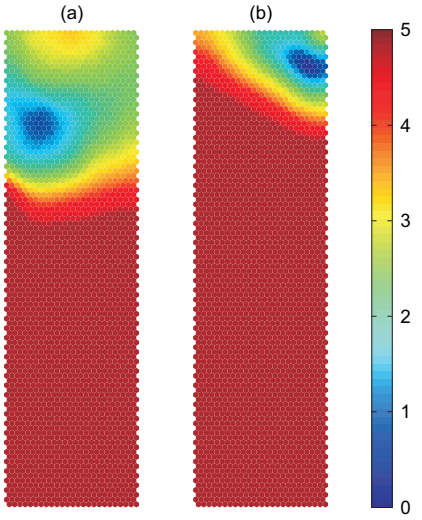


Figure A.2: Distance of the raw feature vector related to a typical sample from T and two maps trained at the same location, but with different initial parameters: (a) $\alpha_0 = 0.6$, $\sigma_0 = 50$, high flexibility; (b) $\alpha_0 = 0.03$, $\sigma_0 = 10$, low flexibility. Training length: 86400 samples.

resolution of 1 s. Different values for some SOM parameters were tested in order to improve the ability of SOM to identify co-occurring sound features. The dimensions of the 2D grid seemed to be not critical above lower limit values. In this paper they were fixed to $M_x = 25$ and $M_y = 100$. The most critical parameters were found to be the length of training runs, the initial value of the learning rate, α_0, and the width of the 2D neighbourhood σ_0. To evaluate the quality of the SOM training, some sound excerpts were recorded at the same scenarios but not used during the training phase. An example is provided in Fig. A.2, wherein the distance between the raw feature vector related to a quiet moment at T and the units of two maps trained at T but with different flexibility are plotted. The less flexible map, which is the one trained with smaller α_0 and σ_0, displays a better focusation and is thus preferable.

Training maps in fixed scenarios result in a strong sound-context dependency. Thus, all of the units of a map trained in P are very dissimilar to raw feature vector corresponding to a typical sample from T, as can be seen in Fig. A.3(d). Obviously, the units in such a map display a better matching for a quiet natural sound sample, as shown in Fig. A.3(b). In contrast, the map trained in T shows good focusation and a low quantization error for both the samples Fig. A.3(a) and (c), as even in a road traffic environment, silent periods are present (e.g.

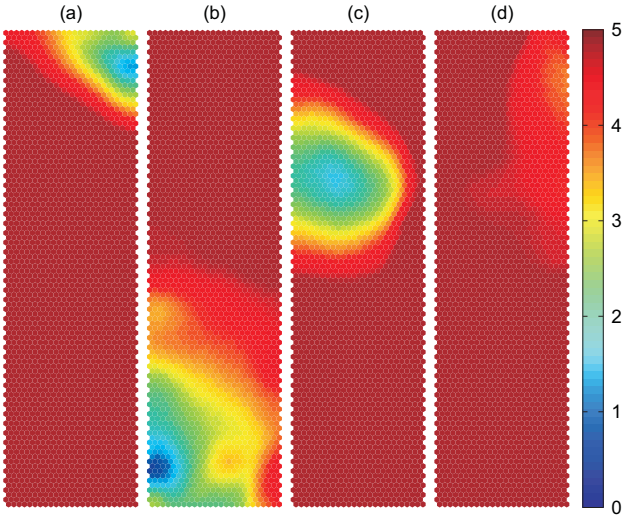


Figure A.3: Distance of the raw feature vector related to a typical sample from P and the units in the SOM trained in (a) T, (b) P. Distance of the raw feature vector related to a typical sample from T and the units in the SOM trained in (c) T, (d) P.

during the nocturnal part of the recording used for training). This context dependency allows an intuitive way to distinguish between common and rarely occurring (or even new) sound events, possibly triggering an alert or a more detailed analysis of the sound events: by re-training the map with that specific input, a later occurence of the same sound will no longer trigger an alert. The context dependency can also be exploited in a different way: feeding a sample to a number of SOMs, each of which was trained on a different context, and comparing the focusations and the quantization errors, can yield information about which context the sample most likely belongs to.

The context dependency can be reduced by training the SOM with excerpts coming from various scenarios. There is an interesting parallel between this situation and the human brain, which is exposed to a lot of different sound contexts during life. To approximate this multi-context learning, a series of 51 sound excerpts of 15 minutes where recorded at various locations in and around the city of Ghent, including traffic-free shopping streets, street canyons with low and high traffic intensity, residential areas, open squares, urban parks and quiet areas at the edge of the city. The new sound samples replaced partly of the night time samples of each scenario, T and P respectively, thus creating two more heterogeneous scenarios called HT and HP. Two new SOMs were trained,

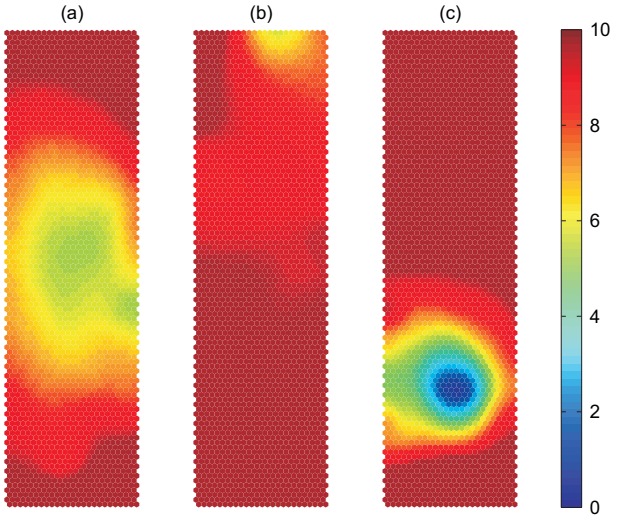


Figure A.4: Distance between the raw feature vector of a sample from a crowded street and the units in the SOM trained in (a) T, (b) P, (c) HT. Training length: 86400 samples, $\alpha_0 = 0.03$, $\sigma_0 = 10$.

one in HT and the other in HP. The units of the old SOMs cover the new sounds only poorly, having been trained exclusively with inputs coming from their specific scenario, T or P. In contrast, the new SOMs are very versatile and can match practically all types of inputs corresponding to the wide range of scenarios they have been trained on. In Fig. A.4 this aspect is visualized by taking into account a 1 s sound fragment from a crowded shopping street, where talking passers-by can be heard. Moreover, the new SOMs still show a low quantization error for samples from T or P, as shown in Fig. A.5.

The U-matrix of the SOM trained in HT is shown in Fig. A.6, revealing how the SOM is composed of regions where neighbouring units are very similar and regions where the opposite holds true. This is common if the SOM has been trained on the basis of a very diverse set of sounds (e.g., coming from very different contexts).

As explained at the end of Section A.2.2, the LEGION oscillators and the SOM units are two different functional representations of the same neural units. In particular, the units best matching the input can be interpreted as externally excited neuronal oscillators, in accordance with (A.11). LEGION thus provides:

1. grouping of contiguous excited oscillators representing particular raw feature vectors, by means of coherent oscillations;

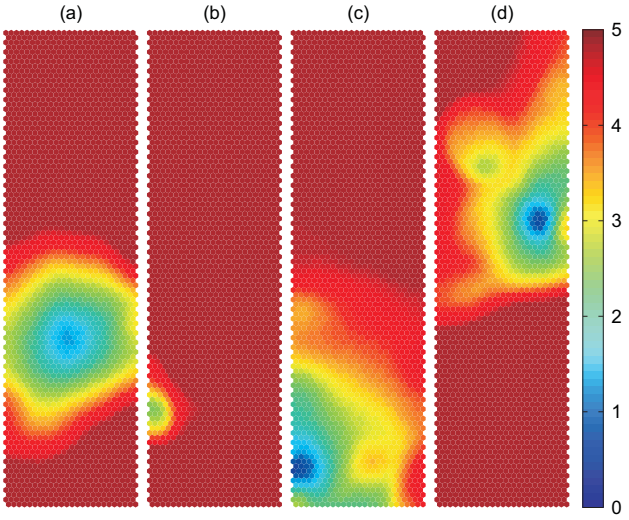


Figure A.5: Distance between the raw feature vector of a typical sample from T and the units of the SOM trained in (a) T, (b) HT. Distance between the raw feature vector of a typical sample from P and the units in the SOM trained in (c) P, (d) HP. Training length: 86400 samples, $\alpha_0 = 0.03$, $\sigma_0 = 10$.

2. segmentation of distinct groups of oscillators by introducing a phase among the groups oscillation.

For the simulation shown in Fig. A.7 the SOM trained in P was chosen. The parameters for SOM training were set as follows: $\alpha_0 = 0.03$, $\sigma_0 = 10$. The values $h = 3$ and $\lambda = 0.92$ were used for binarization in (A.11) and the external stimulation I was set respectively to 0.2 and 0 for stimulated and unstimulated oscillators. The neighbourhood was composed of the 6 nearest neighbours. The maximal value of the global inhibitor W_z was set to 1.7. The following values for the parameters regarding the dynamic connection weights W_{ik} in (A.14)–(A.15) were used: $\eta = 3.0$, $\nu = 0.1$ and $\omega = 1$. The permanent connection weights, as defined in (A.16), were calculated using $\phi = 0.5$ and $T = 1.5$. Of the parameters in (A.9)–(A.10) governing the dynamics of a single oscillator, only γ is needed if the singular limit method is adopted, and it is set to 6.5 here. Finally, for the LEGION time, the value $\tau_L = 15$ was used.

Fig. A.7 shows oscillatory dynamics of LEGION together with the similarity to the SOM units and the external stimulation $I(t)$ for a period of 2 s. It is a clear example of the ability of LEGION to segregate different groups of stimulated oscillators by letting them move to the active phase at different

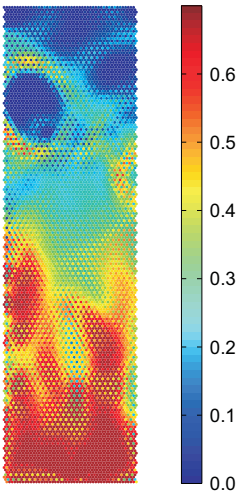


Figure A.6: U-matrix of a SOM trained in HT. Training length: 86400 samples, $\alpha_0 = 0.03$, $\sigma_0 = 10$.

times. In particular, Fig. A.7 at $t = 4.2\,\mathrm{s}$ shows the transient phase wherein the oscillators recombine their dynamic connection weights to adapt themselves to the new external input.

A.4 Conclusions

A model for context-dependent environmental sound monitoring, rigidly grounded on neurological mechanisms, was constructed in this paper. The plasticity of the human cortex, in the context of processing spectro-temporal features, was simulated by the use of a Self-Organizing Map (SOM) based on 1 s standard 1/3-octave band levels. Much as human beings do, sounds were learned within the context in which they were usually heard, resulting in a high context dependency and a high tuning of the model on the typical sounds heard in the specific scenario. In other words, how the presence or absence of a sound during training influenced the SOM, depends on the other sounds perceived during training. After training, the map could be used to assess how typical a new sound fragment is by determining its similarity to the units of the SOM.

A different manifestation of the context dependency is the number of nodes a SOM devotes to a specific type of sound (e.g. car passages, near-silence, pedestrian chatter). Correspondingly, the more heterogeneous the soundscape on which a SOM is trained, the smaller the number of nodes dedicated to each

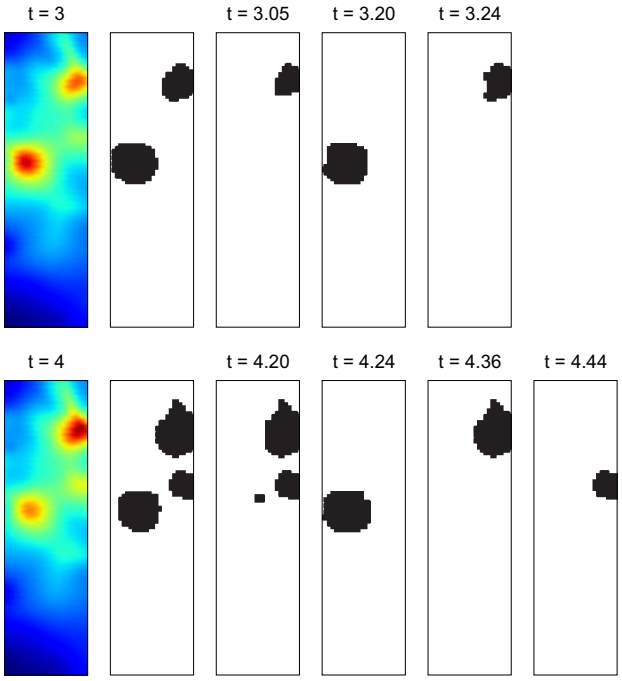


Figure A.7: Left (2 columns): similarity (inverse of the distance) of two input samples at $t = 3\,\mathrm{s}$ (top) and $t = 4\,\mathrm{s}$ (bottom), before (1st column) and after (2nd column) binarization ($\lambda = 0.92$, moving average order: $h = 3$). Right (4 columns): some snapshots of LEGION taken at different times. The samples used here are extracted from test input data recorded in scenario P.

specific type of sound.

By coupling the SOM to a Locally Excitatory Globally Inhibitory Oscillator Network (LEGION), which simulates the oscillatory correlation activity of the neuronal sensory cortex, we were able to use the coupled model for object formation and segregation tasks, where an object in our context is a group of contiguous units similar to the new sound sample.

The model could be used to distinguish between common and rare sound events in a context-specific manner. Moreover, we feel the model merits further research in order to assess its suitability for specific environmental sound recognition and segregation. In order to do so, future work will have to focus on increasing the time resolution and the sound stream formation ability by reducing the transient time in the LEGION oscillatory dynamics.

A different avenue of interest is increasing the biological plausability of the

SOM-LEGION coupling. More to the point, previously unheard sound events can result in little activation of the map, while one would prefer to have a LEGION-segregation between known and unknown components even in such a setting, especially in case of highly salient, though unknown, events. Segregation could also be performed for unknown sound events by changing the binarization threshold, perhaps by considering local maxima in activation of the SOM. In the current implementation, previously unheard components will likely be ignored due to them having a smaller activation than the known components of the sound event. Another issue regards the training phase. In this paper there is a sharp distinction between training and testing phase, which is not biologically plausible: to a certain extent, connections in the brain remain flexible, and training from external stimuli remains possible. A possible improvement of the model could be to trigger a new SOM learning phase when conspicous but unknown sound events are observed.

Appendix B

A computational model of auditory attention for use in soundscape research

Damiano Oldoni, Bert De Coensel, Michiel Boes, Michaël Rademaker, Bernard De Baets, Timothy Van Renterghem, Dick Botteldooren

Published in *The Journal of the Acoustical Society of America* vol. 134, no. 1, pp. 852–861, 2013.

Urban soundscape design involves creating outdoor spaces that are pleasing to the ear. One way to achieve this goal, is to add or accentuate sounds that are considered to be desired by most users of the space, such that they mask undesired sounds, or at least distract attention away from undesired sounds. In view of removing the need for a listening panel to assess the effectiveness of such soundscape measures, the interest for new models and techniques is growing. In this paper, a model of auditory attention to environmental sound is presented, which balances computational complexity and biological plausibility. Once the model is trained for a particular location, it classifies the sounds that are present in the soundscape and simulates how a typical listener would switch attention over time between different sounds. The model provides an acoustic summary, giving the soundscape designer a quick overview of the typical sounds at a particular location, and allows to assess the perceptual effect of introducing additional sounds.

B.1 Introduction

Sound is an integral part of the urban environment, and there is a growing awareness that acoustical aspects should be considered at the same level of importance as architecture and visual aesthetics in urban planning and the design of urban outdoor spaces (95; 123; 124). For example, it has been shown that easy access to nearby outdoor (green) spaces for public amenity, such as urban squares and parks, leads to important positive effects on stress restoration (125) and general well-being (126; 127) of urban residents. In order to create this kind of urban spaces, environments that are of high acoustic quality, it is essential that auditory aspects and knowledge on human perception of environmental sound are included during the urban planning and design process. The goal of the soundscape designer is to compose acoustic environments that are as much as possible pleasing to the ear. More in particular, this means creating spaces in which the sounds that the listener identifies as desired in that context are often heard, while undesired sounds remain mostly hidden to the human ear, or at least are not noticed by the user of the space. This approach obviously goes beyond noise abatement and the striving for silence, and as such, there is a growing need for new models and techniques for soundscape analysis and design, well grounded in human auditory perception.

In this paper, a human-mimicking computational model for soundscape analysis is presented, which combines a self-organizing map of acoustical features with a functional model of auditory attention. The model classifies the sounds that are present in the soundscape over time, and simulates how listeners would switch their attention over time between different sounds. As such, it can be used within the soundscape design process to assess the influence of soundscaping measures (e.g. adding desired or removing undesired sounds) in the field. Next to this, the model involves constructing an acoustic summary through extensive training, tuning the model to the typical sounds that are heard at a particular location. The latter could be used to quickly provide an overview of a specific soundscape for the soundscape designer.

Auditory scene analysis has already been studied extensively by computational means (see Wang and Brown (56) for an overview). The ultimate goal of most of these models is to extract clean sound samples for individual components of the auditory scene, e.g. for separating speech from background noise. The ultimate aim of the present model is to mimic human evaluation of the sonic environment. In contrast to these previous models, it does not aim at extracting sounds that are as clean as technically possible, but at analysing the scene precisely as accurate as a human listener would be capable of. However,

as the model is aimed to be integrated in equipment for long-term outdoor sound measurement, it presents a compromise between biological accuracy and computational efficiency. Furthermore, because of the huge variation between listeners, the model is aimed to be valid on a statistical basis, rather than on an individual basis. It has to be noted that the model, in its present form, does not involve the automated labeling of classified sounds; rather, it simulates how a soundscape will be analytically perceived by the listener. This work refrains from designing methodologies for identifying which sounds are desired in a given environment with a particular use (128; 129).

In Section B.2, a short overview of the literature on auditory scene analysis, attention and masking is given, summarizing the empirical foundation for the model, without going into much detail on the neurobiological basis. In Section B.3, a detailed formulation of the model is presented. In Section B.4, a case study that illustrates the use of the model as a tool in soundscape design is presented. Finally in Section B.5, conclusions and perspectives for future research follow. The work described in this paper builds upon different ideas presented in earlier works (81; 120; 130; 131; 132).

B.2 Empirical background

B.2.1 analysing the auditory scene

Outdoor acoustic environments are usually composed of a wide range of sounds that often overlap in time or frequency. Humans have a great proficieny in disentangling this mixture of incoming sounds into coherent perceptual representations of objects (called auditory streams), usually related to individual sound sources, based on a combination of auditory and visual cues. In a simplifying manner, this process of auditory scene analysis is often regarded as a two-stage analysis-synthesis process (25). In the first stage (segmentation), the acoustic signal is decomposed into a collection of time-frequency segments. In the second stage (grouping), segments that are likely to have arisen from the same environmental source are combined into auditory streams. Traditionally, it has been assumed that the perceptual mechanisms behind this process are largely pre-attentive: only after auditory streams are formed, they can become an object of attention (26; 27). Although this view is appealing because of its conceptual simplicity, recent findings suggest that attention also plays a role in the formation of auditory streams (28; 29). Overall, it can be stated that the process of auditory scene analysis draws on low-level principles for segmentation and grouping, but is fine-tuned by selective attention (30).

B.2.2 Detecting and identifying a sound

Some sounds, although present in the auditory scene, will not be detected; no matter how hard the listener tries, these sounds remain masked. Masking effects have been widely studied using artificial sounds, such as sequences of tones or broadband noises (9), or using speech (10), but basic research on auditory masking of environmental sound is lacking (133). Two types of masking are generally distinguished (134; 135): energetic and informational masking. Energetic masking concerns competing sounds (maskers) overlapping in time and frequency such that parts of one sound (the target) are rendered inaudible. Informational masking regards difficulties to detect a target sound which cannot be accounted for by interfering energy patterns at the peripheral auditory system, but are caused by auditory mechanisms at higher levels of processing. An example of the latter is the inability to separate elements of the target sound from elements of the masker sound, due to similarity between the target and the masker (136).

At this point it is useful to distinguish between detecting and identifying environmental sounds. Detecting a sound means that the listener can observe that a sound is present. Identifying a sound means that the listener can name the sound. For simple sounds such as pure tones, detecting is almost equal to identifying, but for speech and environmental sound this is not the case. It has been shown that the meanings attributed to sounds act as a determinant for soundscape quality evaluations (109; 110), and therefore identification of sounds is an important factor in the context of soundscape design; sounds that are not identified are expected to influence overall soundscape appraisal to a lesser degree.

Detectability of a particular target sound within a soundscape is expected to depend on the spectral characteristics of both the target sound and the background sound, as can be concluded from knowledge on (energetic) masking. However, one should keep in mind that both target and background sound may exhibit considerable temporal variations. For example, the use of water sounds for masking road traffic noise in urban parks has recently gained some scholarly interest (99; 133). Reducing the detectability of road traffic noise to 10 % of the time by adding water sound might therefore require water sound with an equivalent level up to 10 dB(A) above the equivalent level of road traffic noise. The model of Glasberg and Moore (137; 138) summarizes the knowledge on (partial) loudness due to energetic masking, and may be used to quantify the audibility of time-varying sounds in the presence of background sound.

Leech *et al.* (139) and Gygi and Shafiro (140) investigated the particular characteristics of a sound that allow it to be identified in familiar auditory back-

ground scenes. The signal-to-noise ratio between the sound and the background noise was found to be the most important factor in their studies. They also found that contextual congruency between the sound and the background noise plays a role, in the sense that sounds that are not readily expected within a given environment are more easily identified. They could not prove that this was due to potential similarities in acoustic features of background noise and congruent target sounds.

Identifying a sound not only involves the ear but also the brain, and both have their limitations. It can be expected that information content plays a role: the more information is embedded in a sound, the easier it will be to detect it. In the experiment by Gygi and Shafiro (140), some of the physical components of the sounds that made them more identifiable were standard deviation of the spectrum and the number of bursts or peaks. Both characteristics are related to the information content of a sound, and both make the target sound less likely to be masked completely. More generally, it can be expected that identifying a sound within a complex auditory scene also depends on how many unique features the sound has. For example, broadband noises are less likely to be identified than vocalizations that contain a rich variety of tones and tonal fluctuations.

Furthermore, familiarity of the listener with the sound to be detected makes it easier for the listener to detect it (40). This mechanism could work for desired as well as for undesired sounds. Sensitivity to particular acoustical features of a sound are learned in early childhood, but new sounds can be learned at all ages (41). Once sounds become familiar, they are identified more easily. It must be noted that learning effects are not limited to high-level associative memory. Several neurophysiological studies have reported on the capacity for holding memory traces (enduring neural records) in the primary auditory cortex (see Weinberger (42) for an extensive review). In particular, the number of neurons of the representational area of a sound is tuned by its importance (43) and the bigger the area, the stronger the memory effects (44). Neurophysiological correlates of cognitive processes such as selective attention (45; 46), expectancy (47), concept formation (48) and cross-modality effects (49) have been found in the primary auditory cortex, suggesting that due to neuronal plasticity, the primary auditory cortex is not merely an acoustic analyzer, but an adaptive auditory problem solver (42). Another important property of the auditory cortex is tonotopy: neurons next to each other are typically excited by similar stimuli. Tonotopic maps have been observed in the auditory cortex of animal species such as cats (50) and monkeys (51; 52). The human cortex also contains several topologically ordered regions (53; 54; 55),

similar to regions observed in the macaque monkey brain (55).

B.2.3 Paying attention to a sound

Although a particular sound within the acoustic environment may be hearable if one listens to it, this does not imply that one actually has to. Users of the space may not notice the sound because they are performing tasks—auditive or not—or are involved in activities that require their attention. On a longer time scale, the sounds that we consciously notice will contribute to the creation of a mental image of the acoustic environment at a location, and ultimately will shape our perception of its quality. As such, not noticing a sound can be positive if the sound is not part of the acoustic design, while it is negative if the sound is considered a unique soundmark (141) of the location.

Auditory attention allows us to focus our mental resources on specific aspects of the acoustic environment, while ignoring all other aspects (31). More in particular, the auditory attention mechanism is responsible for selecting the information that is to be processed in more detail in working memory, and thus that may be used for making decisions and taking actions (32). It is an essential mechanism in human input processing, as it avoids sensory overload. Central in most theories on attention (visual as well as auditory) is the interplay of bottom-up (saliency-based, depending on the characteristics of the stimulus) and top-down (voluntary, depending on the state of the listener) mechanisms in a competitive selection process (30; 32).

The bottom-up mechanism selectively enhances responses to sounds that are conspicuous, for example because they have rare or novel physical features, or are of instinctive biological importance. This is accomplished by a novelty detection system that continuously monitors the acoustic environment for changes in frequency, intensity, duration or spatial location of stimuli (33; 34). This pre-attentive mechanism operates rapidly and independently of the nature of the particular task that the listener may be performing. In contrast, the top-down mechanism focuses processing resources on the auditory information that is most relevant for the current goal-directed behavior of the listener. This mechanism is guided by information already held in working memory, through sensitivity control, in which the relative strengths of different information channels that compete for access to working memory are regulated (32). Examples are directing eye movement or changing the orientation of the head, or modulating the sensitivity of the neural circuits that process the information. Finally, the selection of information for entry into working memory is found to be a competitive, hierarchically structured process (142). At low hierarchical levels, competition occurs within neural representations of basic sound parameters;

at higher levels, competition occurs between different auditory streams; at the interface with working memory, competition occurs between information from the different senses. At each level, the stimulus with the highest relative strength is selected (combining bottom-up and top-down effects), in a winner-takes-all fashion. This is why selective attention is often compared to a stagelight (35), sequentially illuminating different parts of the scene for further analysis. An important factor in this process is inhibition-of-return (37; 143) (IOR), which prevents attention from permanently focusing on the most salient components of the scene, naturally generating an attentional scanpath over time. The process of voluntary selective attention involves working memory, sensitivity control and competitive selection operating in a recurrent loop (32), and may prohibit involuntary switching of attention to task-irrelevant distractor sounds (39).

It is difficult to determine whether a particular sound within the acoustic environment is noticed or not, using psychophysical experiments. Simply asking people about the sound may point their attention towards it and make them notice the sound. In laboratory conditions, biophysical measures such as event-related potentials (ERP) can be used to assess the influence of attention to sounds during the performance of various non-auditory tasks (144; 145). Such research suggests that effective orientation of attention towards particular sounds is influenced by a wide range of top-down, personal factors: the prior experience of the listener with the sound and the significance of the sound to the listener, the listener's intentions and activities, its emotional state (146; 147) or even a possible genetic component (148; 149). Next to this, the emotional cues carried by a sound also affect the degree to which it captures attention. Unpleasant sounds are known to attract human attention more than neutral sounds (150), even when the peak sound amplitudes are similar (151).

B.3 Computational framework

B.3.1 General considerations

In the following paragraphs, the knowledge on human auditory processing of environmental sounds summarized in the previous section is worked out into a computational model of auditory attention that can be used for analysing outdoor soundscapes. The model takes as input the sound signal recorded by a microphone at a particular location and has as output a measure of the potential of various soundscape components (related to sound sources) for attracting attention. In view of long-term deployment of the model in outdoor measurement equipment, and for evaluating simulated soundscape design interventions,

computational efficiency (low data communication rates, real-time operation, etc.) is advantageous. Consequently, the use of detailed auditory processing models, such as those existing for loudness (137), masking (138), stream segregation (56), auditory saliency (34) or auditory attention (152) is not feasible. Instead, simplified models for each step of the soundscape analysis process are proposed. Next to this, the proposed model only accounts for monaural sound, disregarding the influence of spatial cues on attention, and does not perform automated labeling of sounds. The proposed computational model is comprised of three stages, illustrated in Figure B.1: (a) peripheral auditory processing and the calculation of a measure of auditory saliency, (b) mapping of acoustical features based on co-occurrence and (c) modelling auditory attention. A detailed description of each of the three stages follows.

B.3.2 Peripheral auditory processing

In a first stage, a feature vector is extracted, at regular time intervals, from the sound signal measured by the microphone. Instead of calculating a detailed time-frequency representation of the raw sound wave (e.g. using a gammatone filterbank), the model starts from the 1/3-octave band spectrum (31 bands from 20 Hz to 20 kHz), calculated with a temporal resolution of 1 s. This procedure has the main advantage that off-the-shelf sound measurement equipment can be used as a front-end, which increases the applicability of the model. The limited data rate (31 values per second) makes it possible to implement the model on a large-scale measurement network and to store data for longer periods of time. Furthermore, the choice of time resolution can be justified by noting that a wide range of outdoor environmental sounds have a relatively slowly varying temporal envelope (61; 62; 63). Subsequently, a simplified cochleagram is calculated using the Zwicker loudness model (9; 65), which accounts for energetic masking. Again, the complete hearable frequency range is considered (0 to 24 Bark) with a spectral resolution of 0.5 Bark, resulting in 48 spectral values at each time step.

The mechanism for extracting the feature vector, which characterizes the strength and spectro-temporal variability in the sound signal, is inspired by the way the human auditory system biases its attention towards particularly conspicuous events. Based on existing models for auditory saliency (34; 153), the proposed model calculates measures for intensity, spectral and temporal contrast using a center-surround mechanism, which mimics the receptive fields in the auditory cortex. In particular, multiscale features are calculated in parallel by convolving the cochleagram with various 2D gaussian and difference-of-gaussian filters. The former encode intensity, while the latter encode the spectral and

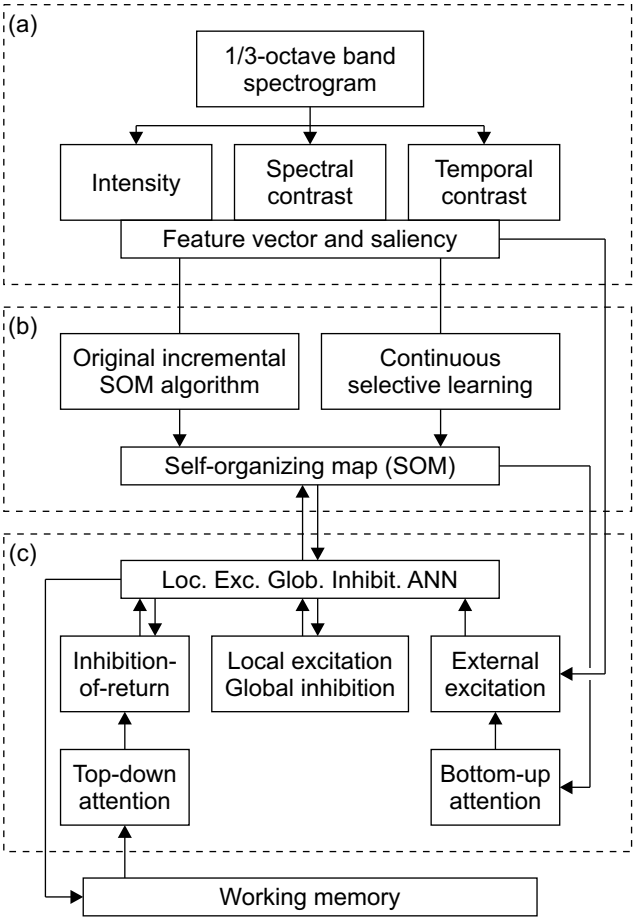


Figure B.1: Schematic overview of the proposed computational model: (a) peripheral auditory processing, (b) self-organizing map of acoustical features based on co-occurrence, and (c) auditory attention.

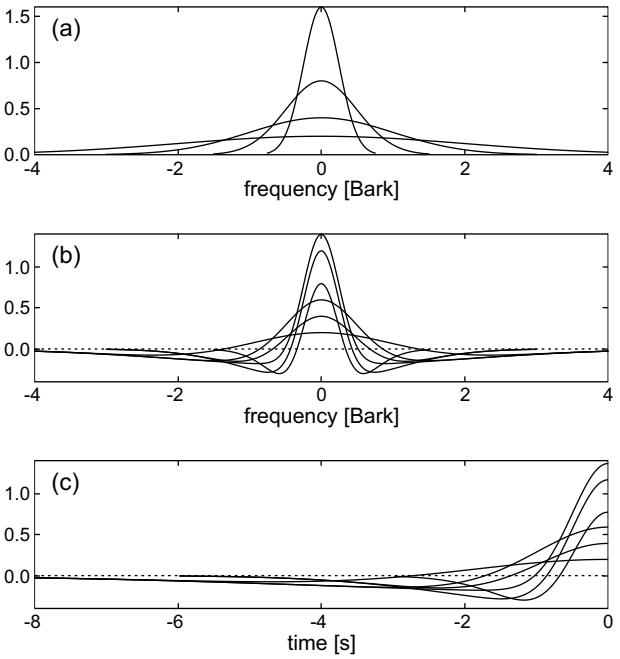


Figure B.2: Cross section of the receptive filters that are used to calculate (a) intensity, (b) spectral contrast and (c) temporal contrast. For the latter, causality is preserved by only convolving with the past.

temporal gradient of the cochleagram. In total, 16 scales (4 for intensity, 6 for spectral contrast and 6 for temporal contrast) are considered. Figure B.2 shows a section of the filters along the time or frequency axis. Using this procedure, a feature vector is constructed at each time step, consisting of $16 \times 48 = 768$ values.

Based on the feature vector, a measure for the saliency of the sound at each time step is calculated. The calculation largely follows the scheme presented by Kalinli and Narayanan (153), with the major adjustment that the effects of spectro-temporal orientation and pitch are not considered. First, rectified center-surround differences are calculated from the raw features obtained at different scales within the same modality (intensity, spectral or temporal contrast), mimicking the properties of local cortical inhibition (34). The resulting center-surround differences are then scaled to a common range, in order to eliminate the difference in dynamic range between the different modalities and scales, and normalized using an iterative nonlinear algorithm that simulates competition

between neighbouring salient locations on the tonotopic scale, promoting peaks while suppressing background noise (154). The normalized center-surround difference vectors are then combined (added) across scales within each modality, and the resulting vectors are again normalized using the same algorithm, and combined to achieve a single tonotopic vector, encoding the saliency of the sound at each time step and at each frequency channel. Finally, a single saliency score at each time step is calculated by summing all values of the saliency vector, hereby assuming that saliency combines additively across frequency channels (70). A detailed description of the algorithm can be found in De Coensel and Botteldooren (81).

B.3.3 Co-occurrence mapping of features

Biological systems learn which auditory features belong to the same auditory object based on co-occurrence. However, auditory learning as described at the end of Section B.2.2 is not a straightforward process, and is still far from being fully understood and computationally replicable. Moreover, learning and memory are not observable phenomena; they have to be inferred from behaviour (42). Nevertheless, in the computational framework here presented, an initial unsupervised learning strategy based on feature co-occurrence is used. It is implemented as a Self-Organizing Map (SOM) or Kohonen Map (155), an abstract model of topographic mapping in the sensory cortex (see Section B.2.2).

A SOM is a two-dimensional grid of units, each of which is represented in the high-dimensional feature space through a reference vector. The Original Incremental SOM Algorithm (155) to train the map consists of iterating the following two steps until some stopping criterion is met:

1. An input feature vector is provided at each time step and the unit corresponding to the closest reference vector, generally called the best-matching unit (BMU), is found.

2. The reference vector corresponding to the BMU and those of units near to the BMU are moved closer to the input feature vector.

The second step underlies the topological preservation. After training, the reference vectors of the SOM units tend to a nonlinear discrete mapping of the distribution of the input data. Some regions of the feature space will be densely mapped by the reference vectors of the SOM units, while other regions will only be sparsely represented. This way, the high-dimensional relationships underlying the input feature data are projected on a two-dimensional map (155). Once

the projection is sufficiently accurate, as quantified by the stopping criterion, training stops.

Machine learning purely based on co-occurrence does not account for the influence of several factors influencing human learning such as attention that were mentioned in Section B.2. Therefore, the basic SOM training is extended with a second training phase that accounts for saliency and novelty of the sound, thus attributing more weight to sounds that are likely to attract attention. The implemented strategy, called continuous selective learning (132), can be seen as a series of much shorter learning periods, triggered whenever the distance between the new feature vector and the BMU is higher than an activation threshold, T_1, and halted when less than a deactivation threshold, T_2, with $T_2 \leq T_1$. Moreover, in order to give more importance to salient sound events, the overall saliency as calculated in Section B.3.2 is used as a modulator of the learning strength. It is observed that after a couple of weeks of continuous selective learning, the SOM is capable of identifying—in terms of distance to the BMU—most of the sounds occurring in a specific acoustic environment. In other words, after such training, the reference vector of each SOM unit corresponds to a representative sound prototype. In order to translate the information encoded in the SOM into hearable sound samples, a sound recording session can be used, during which representative 5-second sound samples with feature vectors closest to each SOM unit are stored. We call this compilation of sounds the "acoustic summary" of the given soundscape (132). Note that the sound samples of the acoustic summary are not labeled automatically in this work. Instead, this can be performed by an expert listener (e.g. an acoustician acquainted with the soundscape of the given location), who explores the acoustic summary and identifies regions in the map corresponding to specific classes of sounds used to present the results in Section B.4.2.

B.3.4 Modelling auditory attention

In order to identify sounds that will be heard on the basis of a trained SOM, an excitatory-inhibitory artificial neural network (ANN), simulating the auditory cortex, is introduced. With each unit of the SOM, a neuron is associated, to be excited by input sounds with feature vectors that are similar to the reference feature vector of the corresponding SOM unit. In order to achieve this, first, a measure for similarity between the input feature vector and the SOM reference vectors needs to be calculated. This is done by calculating the Euclidean distance between the two vectors. Low values of this distance indicate high similarity and vice versa, but, as high excitation is desired in case of high similarity, a gaussian-like function, centered around zero, is used

to convert the Euclidean distances to excitation values, resulting in excitation values approaching 1 for highly similar, and 0 for dissimilar vectors. To take into account the fact that the excitation process is not instantaneous, a leaky integrator is used, with different time constants for increasing and decreasing excitation values.

Bottom-up attention, as explained in Section B.2.3, is a rapidly operating process and is independent of the activity the listener is involved in, facilitating the detection of conspicuous and salient sounds. This is implemented in the model by weighing neuron excitation with a saliency factor, calculated based on the reference vector of the corresponding SOM unit.

As in De Coensel and Botteldooren (81), IOR is introduced to prevent auditory attention from staying focused on one particular source, thus enabling a listener to scan his/her auditory environment. At each time step, only a certain number of neurons will finally be activated, indicating that attention is focused on these neurons. In the current model, IOR is implemented as an increasing inhibition term for these neurons, causing activation to decrease and eventually to fall back to zero. For neurons that are not activated, and thus are not a candidate to get attention, IOR decreases to zero, such that activation is made possible again. This way, IOR causes attention to be continuously shifted from one zone to another. As with excitation, a leaky integrator is used for the implementation, again with different time constants for increasing and decreasing values.

The effect of top-down or outward oriented attention is implemented as a factor modulating the IOR mechanism. By changing the IOR time constants for neurons related to certain zones of the SOM, the shifting of attention can be delayed or even halted when it is focused on neurons corresponding to one of these zones. This way, sustained attention on the sounds represented by these zones is facilitated. Modelling the cause of top-down attention itself is far beyond the scope of current computational models.

Finally, concepts of a Locally Excitatory Globally Inhibitory Oscillator Network (84) (LEGION) are used to implement clustering and competitive selection, to indicate which sound receives attention and thus is entered in working memory. In order to minimize the computational load of the model, there are no oscillators as in a LEGION, but local excitation and global inhibition terms are still used respectively for clustering and competitive selection. Local excitation is added to the input of each neuron, based on the excitation of its neighbouring neurons, weighted with precalculated connection weights that depend on the similarity of the reference vectors of the two corresponding SOM units. Neighbouring neurons which represent very similar sounds are strongly

connected, while connection is weak when the neurons represent dissimilar sounds. A preliminary unit activation can be calculated as the sum of excitation terms minus the IOR, with negative values being set to zero. Global inhibition now adds a new inhibition term to each neuron in the network, calculated based on the sum of these preliminary activations of all neurons. When this summed activation exceeds a certain preset value, global inhibition will rise, and vice versa. By subtracting this inhibition term from the preliminary activation and setting negative values to zero, the final activation is calculated. Thanks to the clustering effect of local excitation, at each time step, only one or a few clusters will have positive values for their final activation. These clusters represent the sounds that receive attention, and for which information is sent to working memory.

B.4 Case study

B.4.1 Overview

In this section, a proof of concept of the computational framework presented in Section B.3 is provided. A fixed sound measurement station was installed in the city of Ghent, next to an urban road, carrying about 3000 vehicles/day during a typical work day. The sonic environment at the chosen location mainly consists of a mixture of road traffic noise due to private and public transport, and noise from pedestrians due to the proximity of several shops and one educational institution. A standard 1/3-octave band spectrum at 1 s time intervals was measured during 3 weeks and is used to train the computational attention model (see below).

The aim of the case study was to assess the perceptual effects of attracting songbirds at the microphone location, a measure that is often proposed to increase the pleasantness of a soundscape (96). For this, a one-hour sound recording was performed during a work day (but not during the latter 3-week period used for training). The L_{Aeq} during this one-hour period was 68.2 dB(A). Subsequently, a series of 30 artificial one-hour sonic environments were created by mixing the original recording with an increasing number of bird sounds at random instances in time. For this, a series of bird vocalisations without background noise, with a duration of up to a few seconds, were used, for which the peak level was adjusted to match the peak level of the few bird sounds present in the original recording. The one-hour L_{Aeq} of the added bird sound ranged from 46.3 dB(A), representing a few sporadic vocalisations, to 75.8 dB(A), representing a quasi-continuous bird chorus, resulting in a signal-

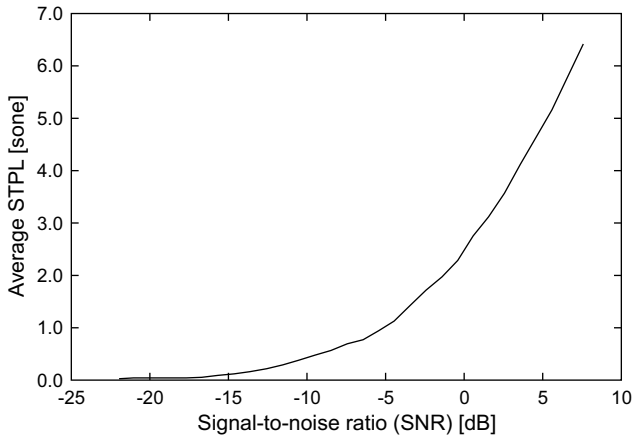


Figure B.3: The average short-term partial loudness (STPL) of the bird sound above the background noise, as a function of signal-to-noise ratio.

to-noise ratio (SNR) for bird sound versus background ranging from -21.9 dB to 7.6 dB.

B.4.2 Results

A first assessment of the effect of adding bird sound would be to check the audibility of the bird sound above the background noise. Figure B.3 shows the average short-term partial loudness (STPL) of the bird sound above the background, for the series of artificial sound mixtures, as a function of signal-to-noise ratio, as calculated with the model of Glasberg and Moore (138). The average partial loudness rises monotonically with signal-to-noise ratio, and starts to increase with a higher rate between -5 dB and 0 dB, marking the range in which the individual bird vocalisations, which can be partially energetically masked if considered separately, start to form a chorus that is audible continuously. Note that the energetic masking model by Glasberg and Moore has only limited applicability in evaluating the effect of acoustical design measures in situ, because it requires that separate recordings for foreground and background sound are available (thus only artificial sound mixtures can be used), and that, due to its computational complexity, fragments are short—for the results of Figure B.3, only the first minute of sound was used.

To demonstrate the performance of the auditory attention model presented in this work, first, acoustical feature vectors and instantaneous saliency values were calculated for the 3-week measurement period, using the algorithm presented in

Section B.3.2. Subsequently, based on this data, a SOM, composed of 50 × 75 = 3750 hexagonally placed units, was trained in three phases. During the first phase, the incremental SOM training algorithm, as presented in Section B.3.3, was applied to the features calculated during 14 hours of the first day of the measurement period. During the second phase, the selective learning algorithm was applied to the remaining 3 weeks of measurement data. During the third phase, the artificial sound mixtures containing bird vocalisations were used in random order. Training a SOM on the sounds at a specific location results in a strong sound context dependency (131). In particular, sounds not present in the training set cannot be easily classified (they will have a large distance to the BMU). Therefore, the third training phase is needed to get the SOM acquainted with the new bird sounds added to the background. From now on, we will refer exclusively to this fully trained SOM.

An acoustic summary has been created as mentioned in Section B.3.3, based on several hours of recording at the given location and the 30 artificial soundscapes. Next, SOM units related to bird sounds are marked by an expert listener, and these are shown in Figure B.4(h). They are mainly grouped into two different regions, related to individual bird chirps (region 1) and a chorus of bird song (region 2). In light of Section B.3.3, the presence of multiple SOM regions devoted to bird sounds should not be surprising: the sound of a single chirp and the sound produced by many birds in chorus result in different sound features, and thus in different regions of the map. Figure B.4(a)-(g) shows how often each of the units of the SOM become the BMU when the original sound and each of the artificial sound mixtures is presented to the model.

As expected, units inside both regions corresponding to bird sounds are more frequently the BMU as the SNR of bird sound increases. This behavior can be quantitatively evaluated by calculating the percentage of time the BMU belongs to either region 1 or region 2 as a function of the SNR. Figure B.5 shows that the percentage of the time that individual bird chirp features are dominant (BMU belonging to region 1) increases monotonically, until a peak is reached at a SNR equal to -2 dB. At that point, the percentage of the time that bird chorus features (BMU belonging to region 2) are dominant starts to increase, while the time that individual bird chirp features are dominant falls back to zero with increasing SNR, marking a quasi-continuous bird chorus present throughout the corresponding artificial sound mixtures.

Now, the same procedure is repeated, taking into account attention mechanisms. Although implemented in the general computational model (see Section B.3.4), the effect of top-down attention is not taken into account, as this would require a model for working memory, which is outside the scope of this

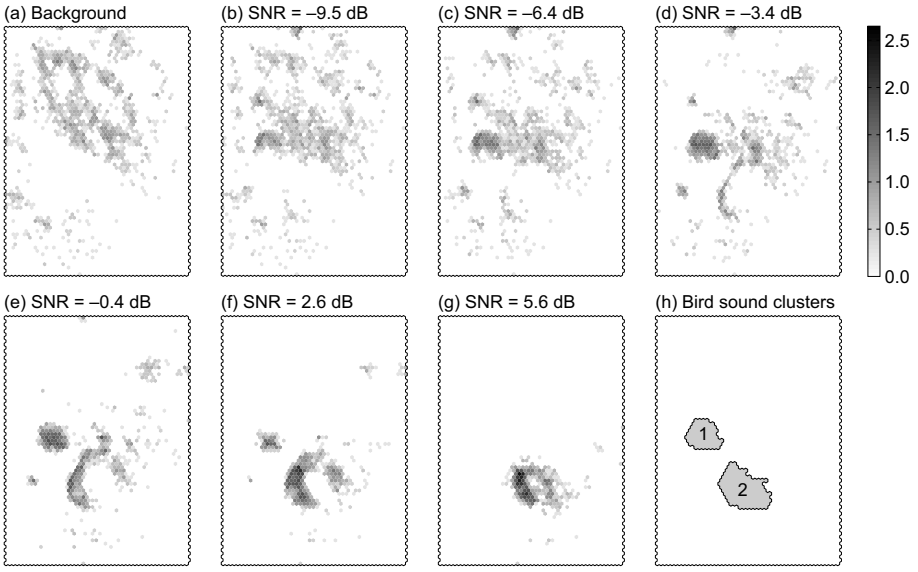


Figure B.4: Logarithmic distribution of the occurrence of the BMU among the SOM units for different scenarios: (a) background, (b)-(g) artificial soundscapes, in which bird vocalisations are progressively added to the background. For each sound scenario, one hour (3600 testing samples) has been used. (h) The two regions of the SOM related to individual bird chirps (region 1) and bird chorus (region 2).

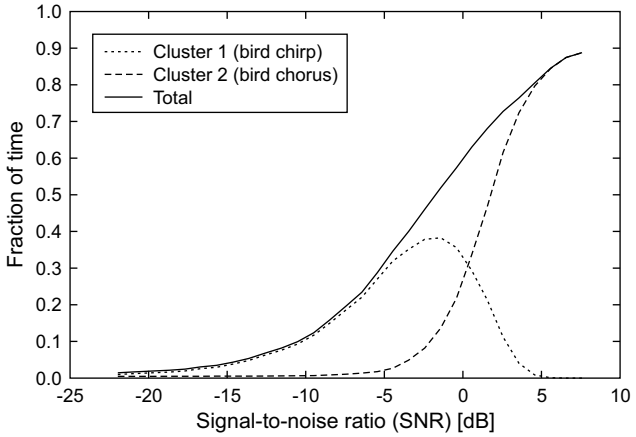


Figure B.5: Evolution of the fraction of time the BMU is located in region 1 (bird chirp, dotted line), region 2 (bird chorus, dashed line) and their sum (total, continuous line) as a function of SNR between background and foreground. For each sound scenario, one hour (3600 testing samples) has been used.

paper. Consequently, IOR time constants are the same for all neurons. The neuron with the strongest activation is now taken at each time step to represent the sound (i.e. the combination of sound features) that receives attention, and in the same way as before, a distribution of occurrence is calculated. From this distribution, the same two clusters are used to calculate the percentage of time the most strongly activated neuron is located in each of the regions, thus approximating the fraction of time that attention is focused on bird sound. The results are displayed in Figure B.6.

It can be seen that for lower SNR, the percentage of time that attention is paid to birds is slightly higher than in Figure B.5, while for higher SNR, this percentage is lower. This indeed is the expected behavior, as for lower SNR, each time bird sound is detectable, it will get attention because its saliency is higher than the background, and because inhibition-of-return will be very low. For higher SNR, bird sound will be continuously detectable, and inhibition-of-return will cause attention to shift away from it. Considering that sounds need to be audible and be paid attention to, in order to contribute to the appraisal of a soundscape, these results are also in accordance with empirical results reported by De Coensel *et al.* (99). There, it was found that, already at an SNR of -10 dB, adding (salient and intermittent) bird sound to a sonic environment dominated by road traffic noise would increase the pleasantness of the soundscape significantly, more than adding the sound of a continuously

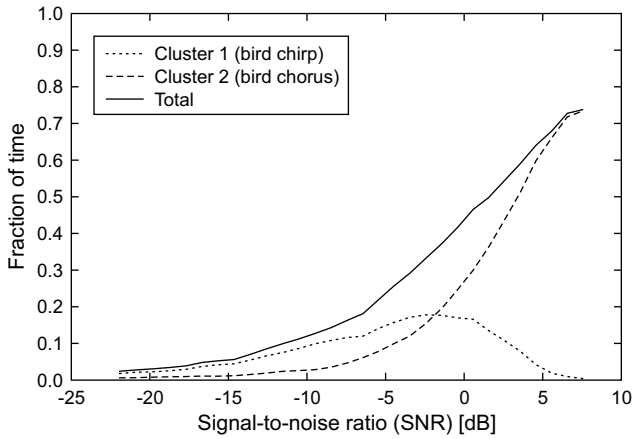


Figure B.6: Evolution of the fraction of time the auditory attention is located in region 1 (bird chirp, dotted line), region 2 (bird chorus, dashed line) and their sum (total, continuous line) as a function of SNR between background and foreground. For each sound scenario, one hour (3600 testing samples) has been used.

flowing fountain at various SNR, although the latter may be more suited to energetically mask road traffic noise (99; 133). The sounds produced by bird vocalisations and fountains are generally considered to be positive in urban and rural environments (121; 156). Consequently, in the context of soundscape planning, the presented model can be helpful to quantify the potential positive effect of introducing additional sounds in the sonic environment, e.g. through the use of audio islands (98).

B.5 Conclusions

Taking into account the mechanisms underlying human auditory perception of environmental sound is a fundamental principle in soundscape design. However, models and techniques that would assist the soundscape designer in achieving this goal are still lacking. In this work, a computational model for soundscape analysis was presented, which implements processes such as bottom-up selective attention and learning, with the goal of simulating how listeners would switch their attention over time between different sounds. The model consists of simplified implementations of several already existing submodels for auditory saliency, topographic mapping, learning, and auditory attention. It complements already existing models of attention-based auditory scene analysis (56) although

it does not provide the same level of detail. However, the novelty of this model lies in its capability to process long stretches of sound in order to accommodate for the huge variation in environmental sounds that characterize the typical urban outdoor environment. The model can be applied to construct an acoustic summary of a soundscape, i.e. a collection of the typical sounds that can be heard at a particular location, and allows to assess the influence of soundscaping measures such as adding additional sounds to distract attention away from undesired sounds. The latter use was illustrated through a case study, in which the effect of adding bird sound to an urban sonic environment was investigated, and in general, accordance with empirical results was found for this particular case. An unexpected model outcome was the emergence of two regions in the map as more and more bird sounds were entered. It was confirmed by listening to the samples that these highly activated regions corresponded to what could be labeled "bird chirps" at the one hand, and a "bird chorus" at the other.

The presented model does not take into account cross-sensory or high-level cognitive effects that lead to top-down auditory selective attention, or meaning attachment to sounds. The latter would involve accounting for the influence of inter-individual differences, and solving linguistic issues (110). Indeed, a sound can be described at two different levels, either by its source or by the action generating it, although such levels are not always clearly separated. A description of the physical properties of either the sound source or the sound itself is provided only when the listener is not able to identify the source or the activity generating the sound. In order to explain the complexity of labeling and categorizing the sounds, observe the following two examples: the sounds of a tram passing by a stopping place and the sound produced by birds. In the first example, a listener would typically label each sound based on the action that generates the sound (braking, opening the doors, warning sounds before closing the doors, accelerating), while it would be unlikely that the specific sources (brakes, engine or loudspeaker) are mentioned. In this case, activity categorization will thus be dominant. Obviously, listeners would also very likely mention the tram as a whole, referring to the sound source. In the second example, the sound produced by birds, the label "bird chirping" is generally used, thus showing again a mixture of the two levels: sound source (birds) and the activity producing such sounds (chirping). Moreover, a listener may refer to the number of birds: while one bird chirping or several birds chirping together denote the same activity, the (number of) sound sources changes. Automated labeling of the acoustic summary as compiled with the present model thus provides a challenge for future research.

Appendix C

The acoustic summary as a tool for representing urban soundscapes

Damiano Oldoni, Bert De Coensel, Annelies Bockstael, Michiel Boes, Bernard De Baets, Dick Botteldooren

Submitted to *Landscape and Urban Planning* on June 26th, 2014, in revision.

Detecting and selecting sound events is emerging as an interesting technique for characterizing and representing the soundscape of a specific location. In this article we propose a computational model for automatically constructing a so-called acoustic summary, i.e. a comprehensive collection of sounds aiming to represent the specific soundscape at a given location. Such an acoustic summary could be used by architects, soundscape designers, and urban planners to explore - by listening - the sonic environment at a certain location as it is perceived by a human listener. The model is based on a self-organizing map, a type of neural network. It starts by extracting several psychoacoustic features from the sound. A specific, extensive and unsupervised training allows this map to be tuned to the typical sounds that are likely to be heard at the microphone location. The learning algorithm takes into account some basic aspects of human perception. For example, salient events tend to be better remembered than the ones that do not stand out, even if they occur less frequently. After the training, the self-organizing map is used to form an exhaustive acoustic summary by means of automatically recording specific sound events for the microphone location. In addition to describing the proposed tool, this paper also presents a validation test with local experts in order to show the ability of the model to pick up sounds which bring out the distinctiveness and the specificity of the soundscape as a local expert would do.

C.1 Introduction

Livability of the urban environment has always been a compelling issue for urban planners. Well-being of the citizen is related to the quality of the urban environment in different ways. Person-environment mismatch at the dwelling may lead to stress and related health impacts (157) but also the quality of the public space is of utmost importance. High quality public spaces stimulate social cohesion, recreation, and physical activity (158). In particular the role of urban green areas has been extensively investigated in this respect and several studies from the last decades indicate that people's psychological restoration and well-being is enhanced by direct access to nature and restorative areas (159; 160; 161; 162; 163) by visual access to them from the dwellings (164; 165; 166) and by their perceived availability (127). The positive role played by such areas has meanly been studied from the perspective of visual diversity, naturalness and aesthetics. However, the role of the soundscape and in particular quietness and tranquillity is increasingly being stressed (127). Therefore, there is an increasing awareness of the fact that the sonic environment forms an essential component of the urban environment that requires as careful planning as the landscape (94; 95).

Classically, urban sound has been treated as a waste product to be tackled with suitable noise control policies whose most popular and visible tool has been extensive noise mapping. However, the final goal of planning and designing urban environments is not only noise abatement, but the creation of spaces with matching positive acoustic qualities (167). This approach, typically referred to as a *soundscape approach*, is getting increasing multidisciplinary attention and is the subject of several projects and studies (95; 129; 168; 169; 170). As the soundscape concept extends beyond the sonic or acoustic environment and includes the way this is perceived and understood by a typical user of the space and within a typical context, the tools at the disposal of the urban sound planner and soundscape designer should account for human auditory perception (171).

Today, physical registration of relevant acoustical parameters is commonly accepted as a first soundscape analysis step (172), followed by an evaluation of the perceptual effects by techniques such as specific interviews and questionnaires, preferably involving community members who live at the location under study (24). The combination of these two approaches is called *combined soundscape analysis* (168; 172) and it is often deployed by means of soundwalks, in which sound measurements and perceptual interviews are conducted simultaneously. In a research perspective, the results are combined in order to find quantitative relationships between physical sound indicators and perceptual attributes (173). Soundwalks are a popular methodology for understanding outdoor soundscapes (174), but they are inherently short-term and typically include only daytime. For this reason, several long-term strategies have been developed, mainly based on mobile sound measurements and community involvement, e.g. with public workers such as local police officers (172). This approach is surely more detailed and complete, but requires a considerable organizational effort and regular and constant participation, resulting in feasibility and reproducibility issues. In both short and long term approaches, a methodology for systematically selecting and recording a comprehensive collection of sound events that is representative for the sonic environment in the way that it is perceived and understood by local experts – inhabitants and visitors – would mean a significant step forward in soundscape methodology.

In this paper a neural-network-based model is proposed that automatically constructs an *acoustic summary*, i.e. a collection of representative sounds that are likely to be noticed at a particular location and together represent the soundscape at that location. The acoustic summary can provide a quick overview of a specific soundscape, thus being a useful soundscape analysis tool for the urban planner and the soundscape designer. In contrast to most of the computational auditory scene analysis (CASA) models (see (56) for an

overview) the major interest here does not lie in extracting as clean as possible sound samples for all components of the auditory scene. On the contrary, the intention is to summarize the sonic environment using only those sounds that a human observer, not particularly focusing its attention to environmental sound, would notice. For this reason, the proposed model partly takes inspiration from specific CASA techniques for extracting salient fragments of the auditory scene but it is also inspired by mechanisms underlying human attention (34; 70; 71). Moreover, CASA techniques are not context dependent. Distinguishing between frequently occurring sounds and out-of-context or rarely occurring sounds is a crucial aspect in constructing an acoustic summary. For this reason, besides a biologically inspired auditory processing model, learning is a very important aspect in the presented model. It is implemented by means of a neural network called *Self-Organizing Map* (SOM) or Kohonen Map (155) and a specifically tailored learning technique. Furthermore, the model attempts to create a compromise between biological accuracy and computational efficiency as the model is to be integrated in equipment for long-term outdoor measurement and the data processing underlying the decision whether or not to record particular sound events has to be made in real-time.

The structure of this paper is as follows: Section C.2 describes the neural-network based model to construct the acoustic summary. Section C.3 is dedicated to the results of a validation test performed by local experts in order to assess how accurately the acoustic summary is representing the soundscape in their neighborhood. Section C.4 discusses the results and future developments. Finally, in Section C.5 conclusions are presented.

C.2 Methods

C.2.1 Overview

Constructing the acoustic summary requires computational auditory scene analysis that mimics how a human observer would split the sonic environment in its relevant components. Considering the application of the model in long-term outdoor measurement stations, computational efficiency has to be considered. For this reason, existing detailed auditory processing models for loudness (137), masking (138) and auditory saliency (34) are replaced by simplified versions. The proposed model is comprised of two main stages, illustrated in Figure C.1: (I) during the learning phase a self organizing map (SOM) is tuned to the typical sounds at the given location based on the sound level and its spectrum and (II) for each class of sounds thus obtained, prototypes are recorded to form

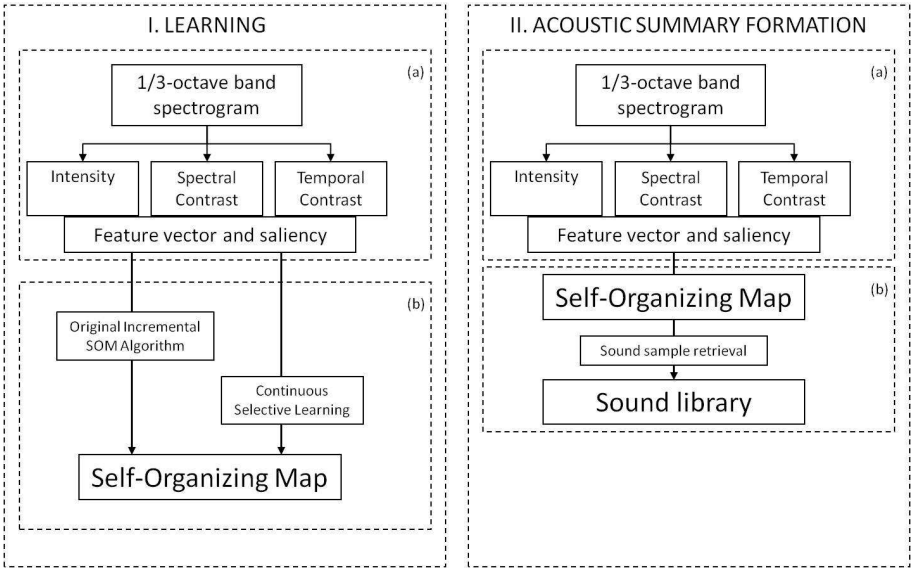


Figure C.1: Schematic overview of the proposed computational model: (I) learning and (II) acoustic summary formation. Both stages start by a simplified (I.a,II.a) peripheral auditory processing. The learning uses the output of such processing for (I.b) training a self-organizing map of acoustical features. The acoustic summary formation stage uses the trained map for (II.b) retrieving sound samples and thus forming a sound library. Finally, an acoustic summary is formed (II.c) by selecting from the library a limited number of sounds based on a ranking method.

the acoustic summary. Real-time operation is required in the second stage due to the limited sound buffer of typical outdoor measurement stations. The sound signal recorded by the microphone is first treated in a similar way as the human peripheral auditory processing (I.a-II.b) and both acoustical features and a measure of auditory saliency are calculated. The learning stage maps the acoustical features based on co-occurrence (I.b) using the incremental SOM algorithm and a training technique called *Continuous Selective Learning* (CSL) that was developed specifically for this purpose. Once the learning has ended, the trained SOM can be used for automatically triggering the recording of typical and salient sounds and thus incrementally forming a library of prototypical sounds (II.b). The acoustic summary is then formed by selecting a small number of sounds from this sound library, based on a ranking method (II.c). In this paper three different ranking methods are presented and tested during the

validation test.

C.2.2 Sound feature extraction

The sound feature extraction stage of the proposed model is highly inspired by a model for auditory attention that was developed earlier by the authors (171). A 1/3-octave band spectrum with a temporal resolution of 0.125 s is calculated starting from the raw audio signal. The relatively coarse temporal resolution was chosen considering the fact that environmental sounds usually show a slow-varying temporal envelope (61; 121). To account for energetic masking, a simplified cochleogram $s(f, t)$ is then calculated based on the Zwicker loudness model (9) covering the complete audible frequency range (0 to 24 Bark) with a spectral resolution of 0.5 Bark. The auditory system is, in addition to absolute intensity, also sensitive to spectro-temporal irregularities (25; 66; 68; 175). The proposed model therefore calculates measures for intensity, spectral and temporal modulation using a center-surround mechanism (72), based on auditory saliency models (34; 70; 71). More in detail, a convolution of the cochleogram with various 2D gaussian and difference-of-gaussian filters is performed in parallel at each time step, resulting in a set of multi-scale features called sound feature vector and consisting of 768 values. The corresponding high-dimensional vector space will be referred to as the sound feature space. More details about the sound feature extraction can be found in (171). Finally, a scalar value called overall auditory saliency is also calculated from the sound feature vector, following the algorithm developed by (81).

C.2.3 Learning

The feature vector provides extensive information about the sonic environment at a given time step. Analysis of the sonic environment should usually last for a long period ranging from a few days to several weeks, depending on the richness in sounds of the sonic environment at the given location. The crucial point is how to use such a large amount of data to construct a concise but exhaustive acoustic summary. In this paper a neural-network-based approach is proposed, which makes use of a self-organizing map. Several topographic maps have been observed in the visual and auditory cortex (50; 51; 76; 77) and the SOM has been originally conceived as an abstract mathematical model of such topographic mapping. Moreover, SOM is typically described as an unsupervised learning-based method for clustering and visualizing high-dimensional data (78), another important aspect to take into account due to the high-dimensionality of the sound feature space. In the framework of the present model, the SOM

would eventually learn which features belong to the same auditory object based on co-occurrence. Furthermore, the size of a representational area of a sound in primary auditory cortex is closely related to its importance (43) and the strength of the memory effects (44), an aspect of auditory learning very well modeled by SOM and the CSL which will be described later in this section. As mentioned in Section C.1, context dependency should be considered while selecting sounds for constructing an acoustic summary. Knowing the context can entail familiarity with the sonic environment and it has been shown that familiarity with the sound to be detected makes the detection easier (79). The extensive training on sound feature vectors at the microphone location tunes the SOM to the typical sounds composing the local sound environment and thus makes the system "familiar" with them.

The SOM used in our model is a 2D network of 3750 equal-spaced units in a regular hexagonal lattice. Each unit has an associated reference vector in the high-dimensional sound feature space. The initial position of the reference vectors is calculated by means of principal component analysis on an input data subset as in (78). After initialization, reference vector coordinates are modified during a first training phase which is based on the Original Incremental SOM Algorithm (155). In particular, at each time step, the unit with reference vector that most closely matches the current sound feature vector is selected (typically called the best-matching unit or BMU) and the reference vector of the BMU, and to a lesser extent the reference vectors of the neighbouring units in the 2D lattice, are moved closer to the input feature vector. After this initial training phase, the reference vectors of the SOM units can be seen as a non-linear discrete 2D mapping of the probability density function of the sound feature vectors used for training. In particular, some regions of the sound feature space contain more reference vectors than others, thus preserving the high-dimensional relationships underlying the input feature vectors (155). When positioning a new sound feature vector with respect to the trained SOM, the smaller the distance to the BMU, the more often similar sound feature vectors occurred during the training phase. The learning algorithm described above is purely based on frequency of occurrence and does not take into account the fact that human perception and retrospective assessment of a sonic environment also depends on the saliency of the sounds. Salient sound events would be better noticed and remembered than less salient ones (176) even if they do not occur that often. Therefore, the SOM trained with the original incremental SOM algorithm is used as a starting point for a second much longer training phase which is referred to as (continuous) selective learning. As in (171) the overall auditory saliency, a number between zero and one, is used for modulating the learning rate parameter during the

selective learning: the learning based on sound feature vectors whose related saliency values are higher than 0.5 is enhanced, while learning based on feature vectors corresponding to sounds with lower saliency is somewhat suppressed. The goal of using saliency in selective learning is to reduce the number of SOM units whose reference vectors are related to often occurring but non-relevant sounds such as the urban background hum. At each time step, the BMU is found as before. However, not all input sound feature vectors are used as inputs during the selective learning: a learning phase is triggered only if the distance to the BMU is higher than an activation threshold Tup. All subsequent input vectors are then selected as inputs for training, until the distance to the BMU drops below a deactivation threshold Tdown. Furthermore, sound feature vectors occurring a few seconds before the triggered learning period are included. In this paper, a 2 seconds pre-trigger period is used, corresponding to 16 time steps. The thresholds Tup and Tdown are chosen in such a way that less than 10% of all sound feature vectors are used as input for selective learning. In order to visualize the effects of training on the SOM reference vectors the so-called U-matrix (82) is used. This matrix shows the distances between reference vectors related to each pair of neighboring SOM units. By means of a color-coding it is thus possible to distinguish groups of SOM units with similar reference vectors and areas with high variability. The effects of the CSL on the clustering of SOM units can be seen in Figure C.2 where the U-matrix after the first training using the original incremental SOM algorithm is shown next to the U-matrix of the final SOM after the continuous selective learning phase.

C.2.4 Sound sample retrieval and selection

The reference vectors associated to the trained SOM units can be seen as representative abstract sound prototypes encoded by their sound feature vectors. Once a SOM is trained, it can be used for constructing a library of sounds, whereby sound samples that are most similar in the sound feature space to the sound prototypes within the SOM are recorded. As shown in the schematic overview in Figure 1, the first step in constructing the acoustic summary is calculating feature vectors for the sound observed at each time step as explained in Section C.2.2. The BMU is then selected, and the distance between its reference vector and the current sound feature vector is calculated. Based on this distance, sound recording is triggered if the selected SOM unit has not been the BMU before (meaning that the encountered sound has not occurred before during the sound sample retrieval phase), or if the distance to the BMU is smaller than any earlier distance for this BMU (meaning that a better matching

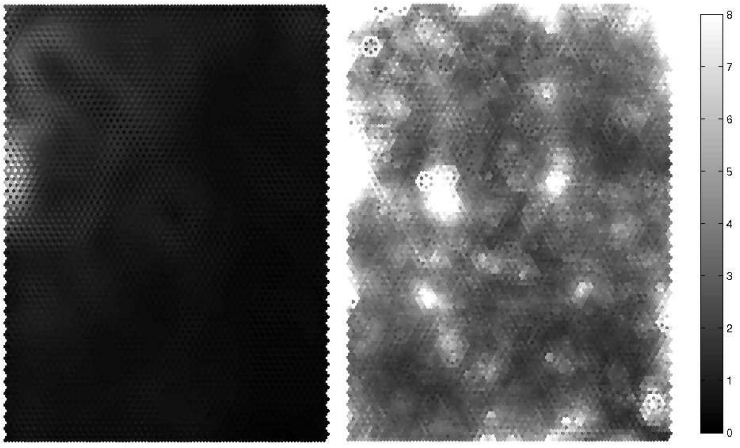


Figure C.2: U-matrix showing the distance by means of color-coding among the reference vectors of neighboring SOM units after (left) the first training session using the original Incremental SOM algorithm and (right) the continuous selective learning.

sound sample is encountered). These steps have to be taken with low latency due to the limited audio recording buffer of typical measurement station. Sound samples are recorded from 3 seconds before to 2 seconds after the recording trigger. It turns out that for typical urban soundscapes, the bulk of the SOM units is represented by an audio sample after a few days of sound sample retrieval. This set of sounds can be seen as a sound library describing the soundscape at the measurement location. The large number of audio samples that is gathered through the procedure described above is unpractical for easily exploring the given soundscape by listening. For this reason, three ranking criteria are presented, which can be used to select a subset of sounds that is most representative for the given soundscape; this subset is then called the acoustic summary. The first proposed ranking criterion is based on saliency: the higher the saliency, the more likely the sound sample will be representative and the higher its ranking. As explained in Section C.2.2, a measured overall saliency value can be calculated at each time step from the sound feature vector. The SOM reference vectors lie in the sound feature space, therefore saliency values can be calculated for each of them, resulting in a saliency overlay on the SOM. A second criterion is based on how often each of the SOM units was selected as the BMU during a given time interval, typically one day or more, resulting in a frequency of occurrence overlay on the SOM. As mentioned in Section C.2.3,

the frequency of occurrence of sounds is not likely to be a sufficient criterion to represent the sounds that will be noticed and remembered. For this reason, a third method is proposed which combines saliency and frequency of occurrence of each SOM unit

$$c_i = \beta_{occ} \cdot \frac{\log(o_i + 1)}{\log N} + \beta_{sal} \cdot s_i, \tag{C.1}$$

where o_i is the number of time steps the SOM unit i is the BMU, N is the total number of samples used for calculating the frequency of occurrence, s_i the saliency of unit i and β_{occ} and β_{sal} are two positive weighting coefficients between 0 and 1 so that $\beta_{occ} + \beta_{sal} = 1$.

The number of sounds to be selected depends on the envisaged use of the acoustic summary. In the validation test discussed in Section C.3, 32 sounds have been selected based on their ranking for each criterion. An a posteriori justification for selecting exactly this number is given in Section C.4.

C.3 Validation test

C.3.1 Overview

A validation test has been designed to check the representativeness of the automatically generated acoustic summaries for an urban soundscape. Sound recording devices at 6 locations in and around the Belgian city of Ghent, that will be referred to as Bi, Ko, Bu, Sp, Be, and Dr were used. In C.1 the day-evening-night equivalent sound level, L_{den}, and a qualitative description of the sonic environment for each location are given. Four locations Bi, Ko, Bu and Sp are situated in urbanized areas, Be is located in the very heart of the city while Dr is in the suburbs. Sound recording devices were attached to the front façade of dwellings. Sixteen people living in the surroundings of the sound recording devices placed in Bi, Ko, Bu and Sp have been contacted for participating to the test as local experts, four per location. Very few people were living in the direct surroundings of the devices placed in Be and Dr, so nobody was contacted from these two locations. The acoustic summaries from these two locations were therefore exclusively used as confounders and their quality was not assessed by the validation test. For this reason, Bi, Ko, Bu and Sp will be referred to as group 1 in the remainder of the paper, while locations Be and Dr will form group 2. For each participant in the validation test, three locations were selected at the beginning of the test. The first selected location was always the location from group 1 where the participant lived. The two other locations were randomly selected: one location was chosen among the others of group

Location	L_{den}(dB(A))	Description
Ko	71.4	Urban square in the city centre. Road traffic noise due to private and public transportation, noise from pedestrians and a music fanfare on Sunday. Microphone placed on a windowsill at the third floor.
Bi	61.3	Urban no-through street in the centre of Ghent, mainly used for parking. Limited road traffic noise due to private transportation, noise from pedestrians and children playing from a recreational area in the neighbourhood. Microphone placed on a windowsill at the 1st floor.
Sp	65.5	Urban street in a residential area. Road traffic noise due to private and public transportation. Microphone placed on a windowsill at the second floor.
Bu	73.3	Urban street along the railways. Road traffic noise due to private and public transportation, train noise. Microphone placed on a windowsill at the 3rd floor.
Be	65.2	Urban street in a restricted traffic zone in the very heart of Ghent. Limited road traffic noise due to the transit of taxi and trucks for restaurants and shop delivery, noise from pedestrians due to the presence of the most important tourist attractions of the city and very distinct bell melodies from the nearby belfry.
Dr	56.4	Quite rural place, about 500 meters from railways. Microphone placed in the backyard of a house in a countryside village.

Table C.1: L_{den}(dB(A)) and qualitative description of the sonic environment at the six locations where the acoustic summary model has been tested. All the locations are situated in the Ghent municipality, five of them in the city, one in a suburban area a few kilometers from the city center.

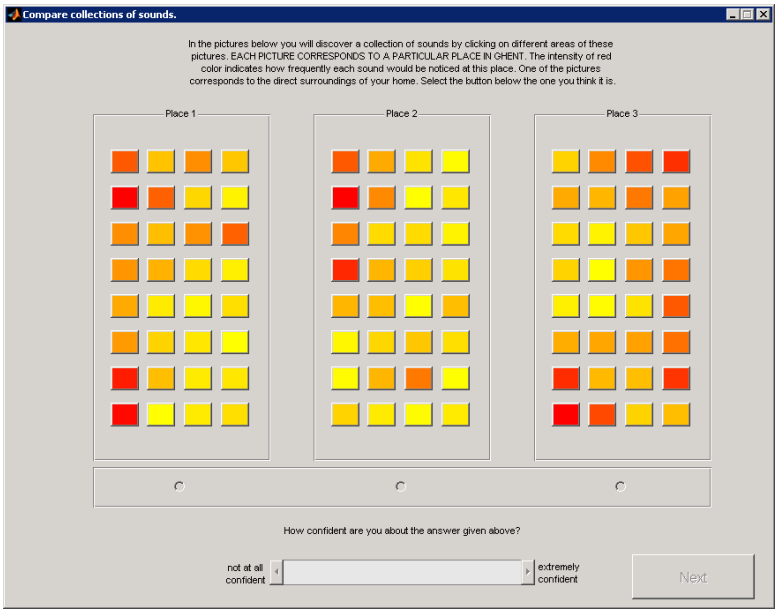


Figure C.3: Snapshot of the first experiment. In this experiment the participants were asked to perform the following task: "In the pictures below you will discover a collection of sounds by clicking on different areas of these pictures. Each picture corresponds to a particular place in Ghent. The intensity of red colour indicates how frequently each sound would be noticed at this place. One of the pictures corresponds to the direct surroundings of your home. Select the button below the one you think it is".

1, and one among the two locations of group 2. The validation test itself was composed of four consecutive experiments, followed by a small questionnaire in which comments could be formulated. The test duration was not fixed and varied among the participants from 30 minutes up to one hour. A computer with high quality sound card and a Sennheiser HD-280 PRO headphone were used for audio presentation.

C.3.2 Experiment 1

In the first experiment, the participants explored the sounds of the acoustic summaries of the three selected locations and had to select the one that they thought corresponded to the direct surroundings of their home (see Figure C.3 for a snapshot of the experiment).

This experiment was repeated three times, with acoustic summaries con-

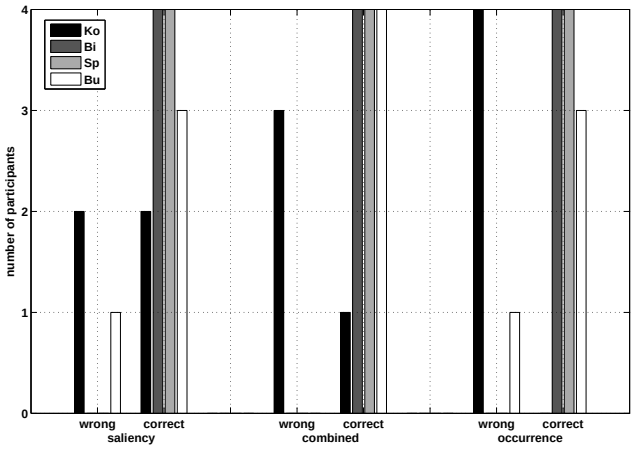


Figure C.4: Correctness of the answers given by the 16 participants from the four locations of group 1 (Ko, Bi, Sp, Bu) while being asked to select among three choices which acoustic summary represented the surroundings of their home.

structed using each of the three criteria —saliency, frequency of occurrence and combined criterion —in randomized order. Each acoustic summary was visualized as a panel of 32 buttons, each of them corresponding to a different sound sample. A colour map spanning from yellow to red was used to colour the different buttons. Depending on the three different ranking criteria, the colour encoded (1) the saliency value s_i, (2) the frequency of occurrence o_i, or (3) the combined value c_i of the corresponding SOM unit. To stress color differences, yellow was assigned to the smallest value and red to the highest value among the 32 values for s_i, o_i and c_i. Participants could listen to each of the sound samples as much as they wanted, by clicking the respective button, before selecting an acoustic summary from the three candidates shown in randomized order. In Figure C.4 the results of the first experiment are shown. In total 13 participants out of 16 correctly selected the acoustic summary that corresponded to the direct surroundings of their home for summaries constructed on the basis of saliency and the combined criterion. Only 11 participants selected the correct acoustic summary in case it was constructed on the basis of frequency of occurrence. The few errors are not equally divided among the four locations included in this test. All participants at the location Bi and Sp recognized the acoustic summaries correctly and from Bu only one error for both saliency and occurrence

criteria occurred. The acoustic summaries from Ko were hardly recognized; most errors were made for the acoustic summary formed by occurrence, followed by the combined criterion and then the saliency criterion. In general, the high and similar number of correct answers for all three ranking-selecting criteria indicates that the sound library where the sounds are selected from is composed of typical and representative sounds for the given location. To further explore possible differences between the three criteria, the number of sounds to which each participant listened before making a choice, is analysed. From the box plot in Figure C.5 it is clear that participants decided faster in case of acoustic summaries based on saliency, while on average they needed to listen to the highest number of sounds for occurrence-based acoustic summaries. These differences

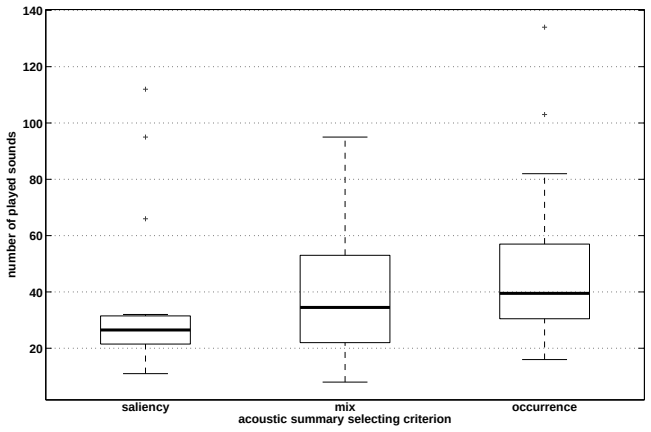


Figure C.5: Box plot of the number of sounds the participants played for deciding which acoustic summary among three choices represented the surroundings of their home. Standard whisker value, 1.5, has been used.

are statistically analysed with the linear regression model $Y = ax + b$ with Y the number of played sounds, $a = (a_1, a_2)$ the coefficients of the regression model, b the constant term of the regression and x the two-dimensional categorical variable coding for the different acoustic summary criterion, so that $x = (0, 0)$ for the acoustic summary based on saliency, while $x = (1, 0)$ and $x = (0, 1)$ for the occurrence and the combined criterion respectively. After excluding the outliers in Figure C.5, the null hypothesis H_0: $a_1 = a_2 = 0$ is rejected based on an overall $F-$test for regression: $F(2, 40) = 3.42$, $p = 0.04$. In this regard, it should be noted that, although randomized, the order in which the summaries

based on each of the three criteria were presented could in theory have influenced the number of played sounds. The order, coded as a two-dimensional categorical variable, is thus added in the previous regression model and the null hypothesis H_0: $a_1 = a_2 = b_1 = b_2 = 0$ cannot be rejected this time, being $F(4, 38) = 1.82$, $p = 0.14$. Moreover, the adjusted $R-$square, R^2, is the highest when the criterion is the only explanatory variable ($R^2 = 0.1$) and it decreases if the order of presenting the three criteria is added in the regression ($R^2 = 0.07$). The same holds true if such order is the only explanatory variable ($R^2 = 0.02$). A final proof that the number of sounds is truly influenced by the acoustic summary criterion and not by the order of presentation is given by an $F-$test comparing the two regression models. The extended regression model with the order added does not provide a significantly better fit: $F(2, 38) = 0.34$, $p = 0.72$.

C.3.3 Experiment 2

In the second experiment, three acoustic summaries, all coming from the location where the participant lives, but either formed by the saliency, the frequency of occurrence, or the mixed criterion were presented. The participants were asked to rank the presented fragments based on perceived accuracy in representing the surroundings of the participant's own home (see Figure C.6 for a snapshot of the experiment). The results of this experiment are shown in Figure C.7 where frequency of the given ranks (1, 2, or 3) is depicted per acoustic summary. The acoustic summary based on occurrence is clearly considered the least representative. The combined criterion provides an acoustic summary ranked first and second by 15 out of 16 participants.

C.3.4 Experiment 3

In the third experiment, each participant was asked to construct his/her own collection of sounds that represented the direct surroundings of its home, by selecting sounds from a set of 64 sounds (see Figure C.8 for a snapshot of the experiment). Half of the sounds the participant could choose from came from his home location, the other half was from two other randomly chosen locations: 16 sounds from the location of group 1 and 16 sounds from the location of group 2. The participants were not told about such subdivision. All sounds belonged to acoustic summaries based on the combined criterion. This inclusion/exclusion of sounds in the final sound collection can be seen as a binary classification task; therefore it makes sense to speak of true and false positives or negatives. The sounds coming from the participant's location that were rightly selected by the

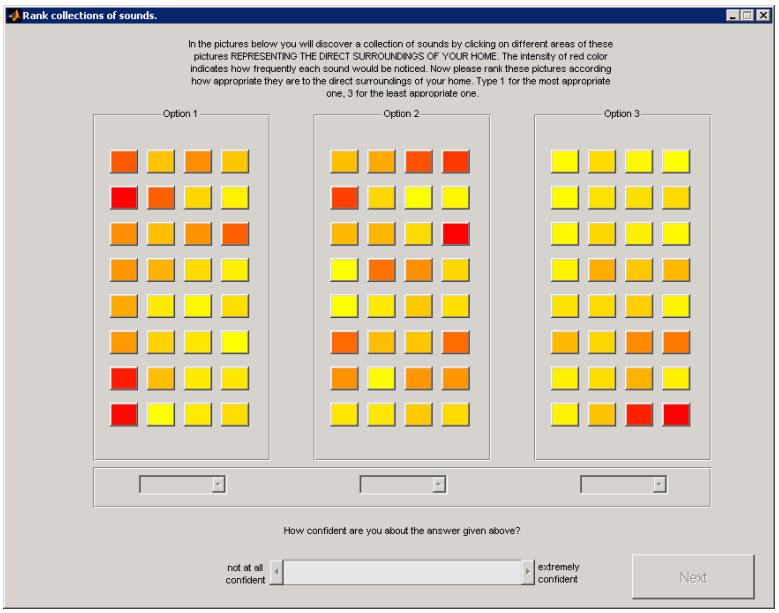


Figure C.6: Snapshot of the second experiment. In this experiment the participants were asked to perform the following task: "In the pictures below you will discover a collection of sounds by clicking on different areas of these pictures representing the direct surroundings of your home. The intensity of red colour indicates how frequently each sound would be noticed. Now please rank these pictures according how appropriate they are to the direct surroundings of your home. Type 1 for the most appropriate one, 3 for the least appropriate one."

participant are called true positives (TPs), while selected sounds recorded at other locations are called false positives (FPs). The true negatives (TNs) are the sounds from other locations correctly not selected and the false negatives (FNs) are the sounds from the surrounding of the participant's home not selected. The higher the number of TPs and TNs, the better the acoustic summary model has captured the peculiarities of the soundscape at each location. An overview of the results for all participants is shown in Figure C.9. The high variability among participants was to be expected. Nevertheless, 10 of the 16 participants scored TPs and TNs both greater than 16, with 16 being the expected result of a random guess. The False Positive Ratio (FPR) and the True Positive Ratio (TPR) are calculated and shown in Figure C.10. The FPR is defined as the ratio between the FPs and the number of sounds from other locations, i.e. 32, while the TPR is the ratio between the TPs and the number of sounds from

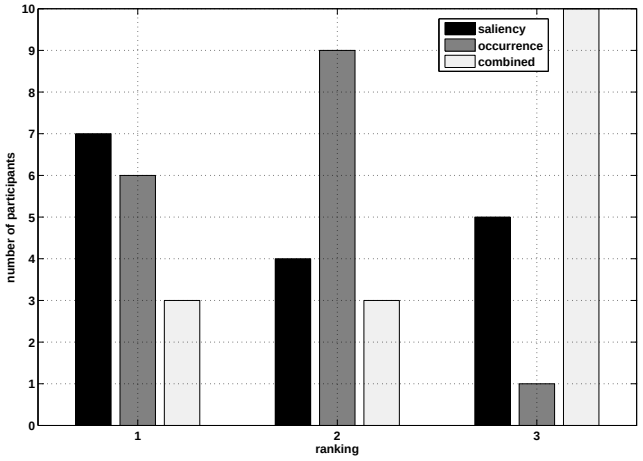


Figure C.7: Overview of the results of the second experiment: each participant was asked to rank three acoustic summaries of its own surroundings based on how representative of its location were. The three acoustic summaries, already presented in the first experiment, were selected by means of three different criteria: saliency, frequency of occurrence and their combined measure.

the participant's location, again 32. The higher the TPR and the lower the FPR are, the more convincing the acoustic summary. In Figure 7 one can see that all participants except one score better than a random guess which would give a point along the diagonal line, the so-called line of no-discrimination. Moreover, the participant called Ko_2 in Figure C.9 is very far from this line too, showing that this participant was completely misled by the proposed sounds. In fact, from Figure C.9 it can be seen that he only selected sounds from the two other locations. The results of the third experiment support the findings from the first experiment. Participants from Bi and Sp —not making any mistake in the first experiment—scored on average better than participants from Bu, who, in turn, scored better than participants in Ko, as shown in Figure C.11 where the accuracy, defined as $(TPs + TNs)/64$, is plotted. The participants from Ko in addition show the highest variability: the first and second participant respectively have the best and the worst accuracy among all participants. It can be noted that the accuracy of the participants from Ko follow the results they obtained during the first experiment: the first participant got the best score in the first experiment making only one mistake, the third participant made two mistakes out of three, while the other two participants

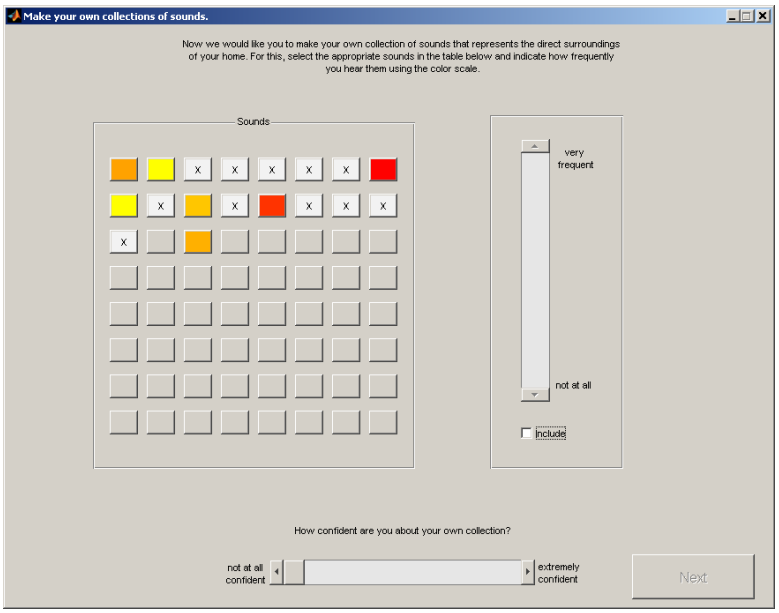


Figure C.8: Snapshot of the third experiment. In this experiment the participants were asked to perform the following task: "Now we would like you to make your own collection of sounds that represents the direct surroundings of your home. For this, select the appropriate sounds in the table below and indicate how frequently you hear them using the color scale."

could never select their own acoustic summary. It is also worthwhile checking whether accuracy was influenced by the number of played sounds in the second experiment. Participants listened exclusively to sounds coming from their own surroundings just before performing this third experiment. So it could have been possible that correct selection in the third experiment is enhanced when more sounds have been listened to in the second experiment. A $t-$test on the slope of a simple linear regression model between accuracy and number of played sounds in the second experiment does not reject the null hypothesis of unrelated variables, i.e. slope equal to zero ($T-$score$= 1.08$, $p = 0.16$). The same conclusion holds if precision, defined as $TPs/(TPs + FPs)$, instead of accuracy is considered ($T-$score$= 1.48$, $p = 0.30$). An $F-$test confirms the null hypothesis for both accuracy ($F = 2.18$, $p = 0.16$) and precision ($F = 1.13$, $p = 0.31$).

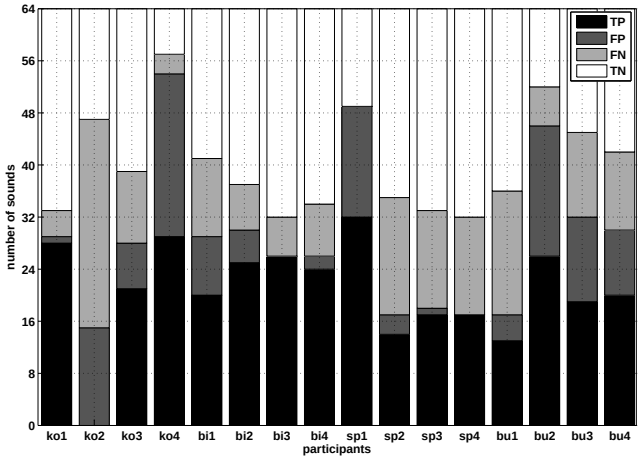


Figure C.9: Overview of the results of the third experiment where each participant was asked to make his/her own acoustic summary that represented the direct surroundings of its home. It was asked to select the appropriate sounds among 64 sounds: thirty-two formed the acoustic summary from the participant's location based on the combined criterion; the other 32 sounds were randomly selected from the acoustic summaries of the other two randomly selected locations based on the combined criterion as well. The participants are denoted by location acronym and a progressive number. In black the sounds from the participant's location correctly selected, called true positives (TP); in dark grey the sounds from a different location wrongly selected, called false positives (FP); in light grey the sounds from the participant's location not selected, called false negatives (FN); in white the sounds from other locations correctly not selected, called true negatives (TN).

C.3.5 Experiment 4

In the last experiment, participants were asked to label 20 sounds that were randomly selected from the 32 sounds composing the saliency-based acoustic summary from their dwelling location (see Figure C.12 for a snapshot of the experiment). This experiment was followed by a small questionnaire in which each participant was asked to leave free comments about the experiment (see Figure C.13). In an open question, it was asked whether there were sounds not heard in the labeling experiment that should have been included in order to better represent the surroundings of the participant's home. The comments, summarized in C.2, are important hints to better understand the obtained

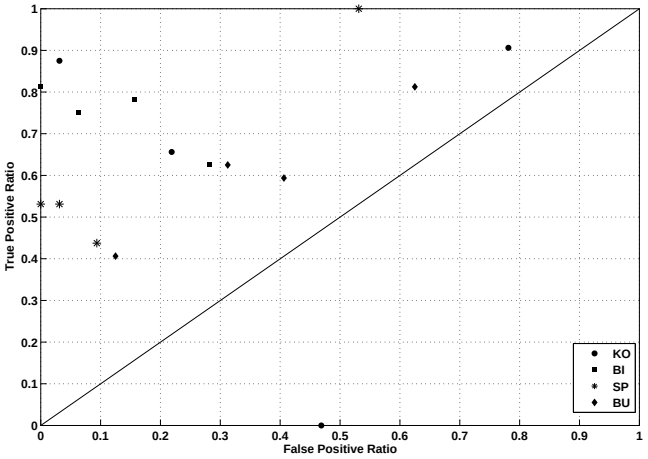


Figure C.10: Scatter plot of the True Positive Rate versus the False Positive Rate calculated based on the results shown in Figure 6. Different markers are chosen for the four locations the participants come from: circles (Ko), squares (Bi), stars (Sp) and diamonds (Bu). The line of no-discrimination is also shown: a random guess would give on average a point on such line.

results. For example, the comments written by the participants from Ko can explain their errors in the first experiment: three out of four were expecting the typical sounds of the market held each Sunday morning in their neighbourhood. Those sounds were not present in the acoustic summaries because the sound sample retrieval was not running during any Sunday, thus missing the very specific so-called soundmarks of that location (22). The same could be said about the comment of participant Ko_2: the construction works he referred to were a very recent activity starting after the sound sample retrieval stopped. In addition, the participants from Bu missed the typical sound of the elementary school located at their backside. These soundmarks were not recorded because the microphone was placed at the front façade of the house. It is worth noting that the main remarks came from the participants living in Ko and Bu, which were the only ones making errors during the first experiment.

Participants	Comment
Ko_1, Ko_3, Ko_4	It would be nice to include sounds of the music bands playing on Sunday morning and during flower market on Sunday.
Ko_2	I didn't hear noise samples of the construction works going on in the square where we live. Otherwise it was very representative. Ninety-five procent of the audio samples were traffic noises: it corresponds well to the amount of traffic we have in front of our apartment.
Bi_1, Bi_2, Bi_3	No comment or positive remarks as "good representation, typical sounds and ambience"
Bi_4	I would include some sounds from the music school at the other side of the street
Sp_1, Sp_3, Sp_4	The sounds represent our street, especially the buses.
Sp_2	More calm situations are needed.
Bu_1, Bu_2, Bu_3	I miss the sounds of the back of the house, e.g. the children playing in the playground.
Bu_4	Most of the sounds are present.

Table C.2: Main concepts expressed in the comments written by the participants after listening to and labeling 20 sounds randomly selected from the 32 sounds composing the acoustic summary based on saliency. The comments are linked to the participants who wrote them.

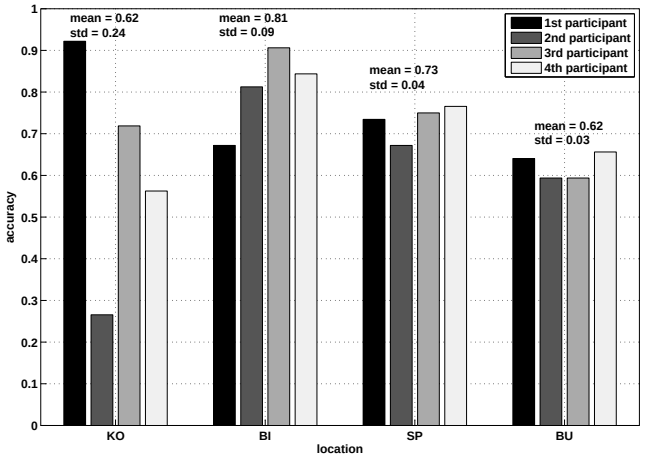


Figure C.11: Accuracy in selecting one's own acoustic summary for all participants divided per location. Mean and standard deviation are also indicated.

C.4 Discussion

The first idea emerging from this study is the importance of soundmarks in describing a soundscape: any acoustic summary which lacks soundmarks would be considered to be less representative, as occurred in Ko or, to a lesser extent, in Bu. Typically, soundmarks have a very specific temporal pattern and occurrence, thus sound sample retrieval needs to run continuously in order to include also these potentially less frequent, but highly relevant soundmarks. Together with soundmarks, spatiality also plays an important role in defining the soundscape. The present research focused on the front façade, where one would have assumed to find the majority of characteristic sounds, but it can happen that soundmarks can only be observed at the other side of the dwelling, as occurred in Bu. Participants appear to be capable of taking these spatial differences into account when judging the acoustic summaries; despite the lack of typical school sounds participants from Bu scored quite good thanks to typical sounds from the front façade. The results from the third experiment demonstrate that in general participants can identify "their" sounds better than random guessing. Moreover, there is a link between the results from the first and the third experiment, showing that the representativeness of an acoustic summary is a direct consequence of the quality of each sound composing it. Nevertheless, the number of false negative and false positive cannot be in general

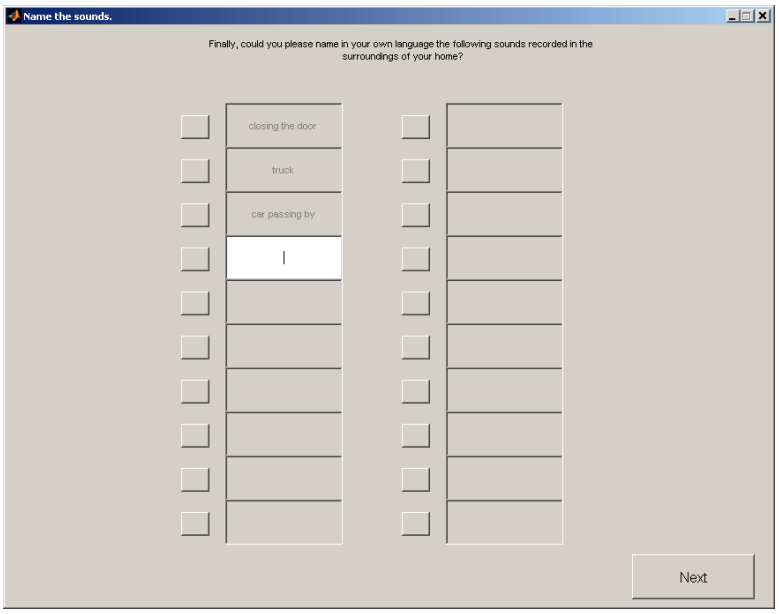


Figure C.12: Snapshot of the fourth experiment. In this experiment the participants were asked to perform the following task: "Finally, could you please name in your own language the following sounds recorded in the surroundings of your home?"

neglected: the sound samples composing an acoustic summary can, most of the time, be associated to more than one location, if the sound samples are considered separately from the others. Therefore, results of this experiment confirm the validity of using an acoustic summary for representing or evoking a soundscape. Considered as a whole, such a collection of sounds can be much more representative of the uniqueness of a sonic environment than each single sound on itself that is part of the acoustic summary. The finding that most participants were able to answer correctly for the limited number of played sounds justifies that 32 is a reasonable number of sounds for an acoustic summary to characterize a location. Thus, selecting such a limited set of sounds is as crucial as the sound sample retrieval itself: it would make no sense to continuously retrieve sound samples if the soundmarks and other typical sounds would not be selected for the acoustic summary afterwards. In this work, the number of sounds composing the acoustic summary was heuristically determined and was the same for all locations. However, the richness of a soundscape depends intrinsically on the considered location. Our model could

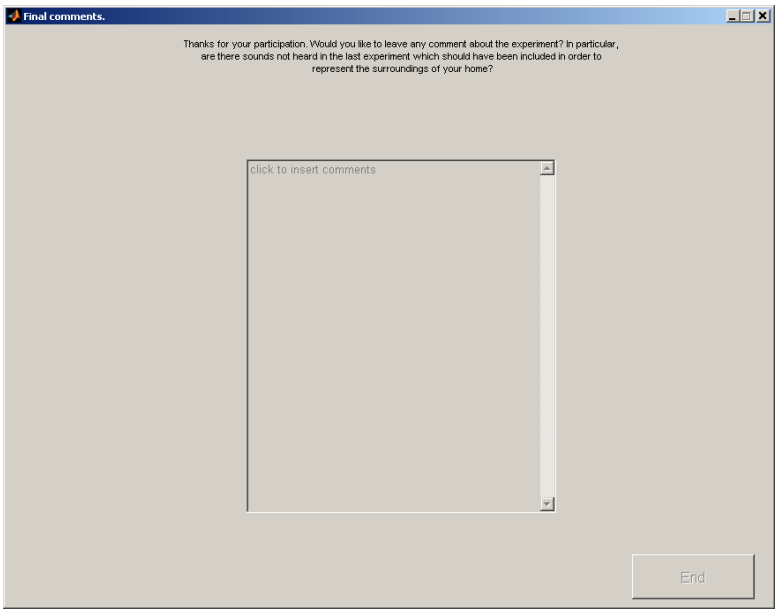


Figure C.13: Snapshot of the fourth experiment. In this experiment the participants were asked to perform the following task: "Finally, could you please name in your own language the following sounds recorded in the surroundings of your home?"

therefore be improved in future, considering acoustic summaries composed by a variable number of sounds. For example, a measure of the overall similarity among the SOM reference vectors could be used to determine the richness of the sonic environment at a given location, and consequently the number of sound samples that should be selected. The second experiment confirms that frequency of occurrence is not the best criterion for selecting the sounds composing the acoustic summary. In many locations the sounds selected based on this criterion are typically very quiet, especially in residential areas or parks, thus missing the less often occurring but much more salient sounds. Hence, saliency is a better criterion for constructing the acoustic summary, but there is still a non-negligible group of people considering it the least appropriate. Selecting only high salient sounds typically comes down to selecting loud sounds, and an excessive number of such fragments is no longer representative of the perceived soundscape in urban residential areas. Therefore, a combination of occurrence and saliency was conceived and tested. The second experiment demonstrates that such a combination is a simple and valid strategy for representing a soundscape as a

human would do. Based on these results, more advanced processing models could be tested in the future, for example, adding human-like attention mechanisms in the model as in (111). In this work, a fixed sound sample duration of 5s was used; however, every sound event has its own typical duration and it should be preserved in order to better represent the sound events composing the acoustic summary. The model presented by Boes et al. (111) could help to solve this issue.

C.5 Conclusions

This work presents a computational model for constructing a comprehensive and representative collection of sounds that are present at a given location. Such a collection, called an acoustic summary, can be a useful tool for quickly presenting and analysing the soundscape at a given location. The model consists of two stages: in a first stage, a Self-Organizing Map is tuned to the typical sounds at the given location, and, in a second stage, an acoustic summary is constructed by first collecting and then selecting specific sound samples based on the trained map. The model takes into account aspects of human auditory perception, such as bottom-up selective attention and learning. A listening test involving local experts has been performed to evaluate the ability of the model to produce acoustic summaries representative of the soundscape at a number of urban locations. The test demonstrated that the model can construct representative acoustic summaries. In particular, the model produces broad and satisfactory sound libraries from which the acoustic summary can be extracted. In general, satisfactory results are obtained from all the three tested criteria used for selecting representative audio samples from the sound library to compose the acoustic summary. However, the acoustic summary criterion combining saliency and frequency of occurrence of the sound events generally produces the best acoustic summary. The saliency-based criterion produces good acoustic summaries as well but risks to outweigh highly informative and salient sounds. In addition, participants judged the acoustic summaries based on frequency of occurrence alone to be the least representative due to the prevalence of quiet sound events, which are much less informative of the given soundscape even though they occur very often in residential areas. Finally, the test demonstrated that only a few sounds are needed to represent the soundscape of an urban area, confirming the choice of 32 sounds for each location.

Bibliography

[1] T. Hobbes, "Of man, being the first part of leviathan," in *Leviathan, or the Matter, Forme & Power of a Common-wealth Ecclesiasticall and Civill*. Yale University Press, 2010.

[2] R. Nichols, *Thomas Reid's theory of perception*. Oxford University Press, 2007.

[3] A. A. Luce and T. E. Jessop, *The Works of George Berkeley, Bishop of Cloyne*, vol. 1, 1948.

[4] H. Von Helmholtz, *Handbuch der physiologischen Optik: mit 213 in den Text eingedruckten Holzschnitten und 11 Tafeln*. Voss, 1866, vol. 9.

[5] H. L. Helmholtz, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Cambridge University Press, 2009.

[6] G. T. Fechner, *Elements of Psychophysics, 1860*. Appleton-Century-Crofts, 1948.

[7] H. Fletcher and W. A. Munson, "Loudness, its definition, measurement and calculation," *The Journal of the Acoustical Society of America*, vol. 5, no. 2, pp. 82–108, 1933.

[8] R. McCurdy, "Tentative standards for sound level meters," *Transactions of the American Institute of Electrical Engineers*, vol. 55, no. 3, pp. 260–263, 1936.

[9] E. Zwicker and H. Fastl, *Psychoacoustics. Facts and Models*, 2nd ed., ser. Springer Series in Information Sciences. Berlin, Germany: Springer-Verlag, 1999, no. 22.

[10] D. S. Brungart, B. D. Simpson, M. A. Ericson, and K. R. Scott, "Informational and energetic masking effects in the perception of multiple simultaneous talkers," *The Journal of the Acoustical Society of America*, vol. 110, no. 5, pp. 2527–2538, 2001.

[11] *ISO 9613-1:1993. Acoustics – Attenuation of sound during propagation outdoors – Part 1: Calculation of the absorption of sound by the atmosphere.* ISO, 1993.

[12] *ISO 9613-2:1996. Acoustics – Attenuation of sound during propagation outdoors – Part 2: General method of calculation.* ISO, 1993.

[13] E. Björk, "Community noise in different seasons in Kuopio, Finland," *Applied Acoustics*, vol. 42, no. 2, pp. 137–150, 1994.

[14] K. Genuit, "The use of psychoacoustic parameters combined with a-weighted spl in noise description," in *Proceedings of the 28th International Congress and Exposition on Noise Control Engineering (Inter-Noise 1999*, vol. 1999, no. 1. The Institute of Noise Control Engineering of the USA, 1999, pp. 1887–1892.

[15] M. Raimbault, C. Lavandier, and M. Bérengier, "Ambient sound assessment of urban environments: field studies in two French cities," *Applied Acoustics*, vol. 64, no. 12, pp. 1241–1256, 2003.

[16] D. Botteldooren, B. De Coensel, and T. De Muer, "The temporal structure of urban soundscapes," *Journal of Sound and Vibration*, vol. 292, no. 1, pp. 105–123, 2006.

[17] B. Schulte-Fortkamp and D. Dubois, "Recent advances in soundscape research," *Acta Acustica united with Acustica*, vol. 92, no. 6, pp. 5–8, 2006.

[18] B. Schulte-Fortkamp and W. Nitsch, "On soundscapes and their meaning regarding noise annoyance measurements," in *Proceedings of the 28th International Congress and Exposition on Noise Control Engineering (Inter-Noise 1999*, vol. 3. The Institute of Noise Control Engineering of the USA, 1999, pp. 1387–1394.

[19] B. Schulte-Fortkamp *et al.*, "The meaning of annoyance in relation to the quality of acoustic environments," *Noise and Health*, vol. 4, no. 15, p. 13, 2002.

[20] B. Schulte-Fortkamp, B. M. Brooks, and W. R. Bray, "Soundscape: an approach to rely on human perception and expertise in the post-modern community noise era," *Acoustics Today*, vol. 3, no. 1, pp. 7–15, 2008.

[21] R. Schafer, *The New Soundscape: a Handbook for the Modern Music Teacher*. BMI Canada, 1969. [Online]. Available: http://books.google.be/books?id=sJQYAQAAIAAJ

[22] ——, *The Tuning of the World: Toward a Theory of Soundscape Design*. University of Pennsylvania Press, 1977.

[23] B. Truax, *The World Soundscape Project's Handbook for Acoustic Ecology*. ARC Publications Vancouver, BC, 1978.

[24] B. M. Brooks, "Traditional measurement methods for characterizing soundscapes," *The Journal of the Acoustical Society of America*, vol. 119, no. 5, pp. 3260–3260, 2006.

[25] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, Massachusetts, USA: The MIT Press, 1994.

[26] E. S. Sussman, "Integration and segregation in auditory scene analysis," *The Journal of the Acoustical Society of America*, vol. 117, no. 3, pp. 1285–1298, 2005.

[27] E. S. Sussman, J. Horváth, I. Winkler, and M. Orr, "The role of attention in the formation of auditory streams," *Perception & Psychophysics*, vol. 69, no. 1, pp. 136–152, 2007.

[28] R. Cusack, J. Decks, G. Aikman, and R. P. Carlyon, "Effects of location, frequency region, and time course of selective attention on auditory scene analysis." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 30, no. 4, p. 643, 2004.

[29] S. A. Shamma, M. Elhilali, and C. Micheyl, "Temporal coherence and attention in auditory scene analysis," *Trends in neurosciences*, vol. 34, no. 3, pp. 114–123, 2011.

[30] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, "Auditory attention-focusing the searchlight on sound," *Current Opinion in Neurobiology*, vol. 17, no. 4, pp. 437–455, 2007.

[31] M. Elhilali, J. Xiang, S. A. Shamma, and J. Z. Simon, "Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene," *PLOS Biology*, vol. 7, no. 6, p. e1000129, 2009.

[32] E. I. Knudsen, "Fundamental components of attention," *Annual Review of Neuroscience*, vol. 30, pp. 57–78, 2007.

[33] S. Shamma, "On the role of space and time in auditory processing," *Trends in Cognitive Sciences*, vol. 5, no. 8, pp. 340–348, 2001.

[34] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: an auditory saliency map," *Current Biology*, vol. 15, no. 21, pp. 1943–1947, 2005.

[35] G. Sperling and E. Weichselgartner, "Episodic theory of the dynamics of spatial attention," *Psychological Review*, vol. 102, no. 3, p. 503, 1995.

[36] C. Spence and J. Driver, "Inhibition of return following an auditory cue: The role of central reorienting events," *Experimental Brain Research*, vol. 118, no. 3, pp. 352–360, 1998.

[37] D. J. Prime, M. S. Tata, and L. M. Ward, "Event-related potential evidence for attentional inhibition of return in audition," *NeuroReport*, vol. 14, no. 3, pp. 393–397, 2003.

[38] M. I. Posner and Y. Cohen, "Components of visual orienting," *Attention and performance X: Control of Language Processes*, vol. 32, pp. 531–556, 1984.

[39] E. Sussman, I. Winkler, and E. Schröger, "Top-down control over involuntary attention switching in the auditory modality," *Psychonomic Bulletin & Review*, vol. 10, no. 3, pp. 630–637, 2003.

[40] J. W. Lewis, W. J. Talkington, A. Puce, L. R. Engel, and C. Frum, "Cortical networks representing object categories and high-level attributes of familiar real-world action sounds," *Journal of Cognitive Neuroscience*, vol. 23, no. 8, pp. 2079–2101, 2011.

[41] L. R. Engel, C. Frum, A. Puce, N. A. Walker, and J. W. Lewis, "Different categories of living and non-living sound-sources activate distinct cortical networks," *NeuroImage*, vol. 47, no. 4, pp. 1778–1791, 2009.

[42] N. Weinberger, "Reconceptualizing the primary auditory cortex: Learning, memory and specific plasticity," *The Auditory Cortex*, pp. 465–491, 2011.

[43] R. Rutkowski and N. Weinberger, "Encoding of learned importance of sound by magnitude of representational area in primary auditory cortex," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 38, p. 13664, 2005.

[44] K. Bieszczad and N. Weinberger, "Representational gain in cortical area underlies increase of memory strength," *Proceedings of the National Academy of Sciences*, vol. 107, no. 8, p. 3793, 2010.

[45] J. Fritz, S. Shamma, M. Elhilali, D. Klein *et al.*, "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex," *Nature Neuroscience*, vol. 6, no. 11, pp. 1216–1223, 2003.

[46] J. Fritz, M. Elhilali, and S. Shamma, "Differential dynamic plasticity of a1 receptive fields during multiple spectral tasks," *The Journal of Neuroscience*, vol. 25, no. 33, p. 7623, 2005.

[47] T. Raij, L. McEvoy, J. Mäkelä, and R. Hari, "Human auditory cortex is activated by omissions of auditory stimuli," *Brain Research*, vol. 745, no. 1-2, pp. 134–143, 1997.

[48] F. Ohl, H. Scheich, W. Freeman *et al.*, "Change in pattern of ongoing cortical activity with auditory category learning," *Nature*, vol. 412, no. 6848, pp. 733–735, 2001.

[49] J. Pekkola, V. Ojanen, T. Autti, I. Jääskeläinen, R. Möttönen, A. Tarkiainen, and M. Sams, "Primary auditory cortex activation by visual speech: an fmri study at 3 T," *NeuroReport*, vol. 16, no. 2, p. 125, 2005.

[50] P. Heil, R. Rajan, and D. Irvine, "Topographic representation of tone intensity along the isofrequency axis of cat primary auditory cortex," *Hearing Research*, vol. 76, no. 1-2, pp. 188–202, 1994.

[51] A. Morel and J. Kaas, "Subdivisions and connections of auditory cortex in owl monkeys," *The Journal of Comparative Neurology*, vol. 318, no. 1, pp. 27–63, 1992.

[52] C. Petkov, C. Kayser, M. Augath, and N. Logothetis, "Functional imaging reveals numerous fields in the monkey auditory cortex," *PLOS Biology*, vol. 4, no. 7, p. e215, 2006.

[53] T. Talavage, P. Ledden, R. Benson, B. Rosen, and J. Melcher, "Frequency-dependent responses exhibited by multiple regions in human auditory cortex1," *Hearing Research*, vol. 150, no. 1-2, pp. 225–244, 2000.

[54] T. Talavage, M. Sereno, J. Melcher, P. Ledden, B. Rosen, and A. Dale, "Tonotopic organization in human auditory cortex revealed by progressions of frequency sensitivity," *Journal of Neurophysiology*, vol. 91, no. 3, pp. 1282–1296, 2004.

[55] C. Humphries, E. Liebenthal, and J. Binder, "Tonotopic organization of human auditory cortex," *Neuroimage*, vol. 50, no. 3, pp. 1202–1211, 2010.

[56] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., 2006.

[57] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[58] D. O'Shaughnessy, "Invited paper: Automatic speech recognition: History, methods and challenges," *Pattern Recognition*, vol. 41, no. 10, pp. 2965–2979, 2008.

[59] M. Cowling and R. Sitte, "Comparison of techniques for environmental sound recognition," *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2895–2907, 2003.

[60] J. D. Krijnders, M. E. Niessen, and T. C. Andringa, "Sound event recognition through expectancy-based evaluation ofsignal-driven hypotheses," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1552–1559, 2010.

[61] B. De Coensel, D. Botteldooren, and T. De Muer, "$1/f$ noise in rural and urban soundscapes," *Acta Acustica united with Acustica*, vol. 89, no. 2, pp. 287–295, 2003.

[62] D. Botteldooren, B. De Coensel, and T. De Muer, "The temporal structure of urban soundscapes," *Journal of Sound and Vibration*, vol. 292, no. 1, pp. 105–123, 2006.

[63] B. De Coensel, D. Botteldooren, K. Debacq, M. E. Nilsson, and B. Berglund, "Clustering outdoor soundscapes using fuzzy ants," in *IEEE International Joint Conference on Evolutionary Computation, 2008 (IEEE World Congress on Computational Intelligence)*. IEEE, 2008, pp. 1556–1562.

[64] G. F. Woodman, "A brief introduction to the use of event-related potentials in studies of perception and attention," *Attention, Perception, & Psychophysics*, vol. 72, no. 8, pp. 2031–2046, 2010.

[65] 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*, vol. 55, no. 1, pp. 63–67, 1984.

[66] W. A. Yost, "Auditory perception and sound source determination," *Current Directions in Psychological Science*, pp. 179–184, 1992.

[67] C. R. deCharms, D. T. Blake, and M. M. Merzenich, "Optimizing sound features for cortical neurons," *Science*, vol. 280, no. 5368, pp. 1439–1444, 1998.

[68] C. Alain, S. R. Arnott, and T. W. Picton, "Bottom–up and top–down influences on auditory scene analysis: Evidence from event-related brain potentials." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 27, no. 5, p. 1072, 2001.

[69] D. L. Woods, C. Alain, R. Diaz, D. Rhodes, and K. H. Ogawa, "Location and frequency cues in auditory selective attention." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 27, no. 1, p. 65, 2001.

[70] O. Kalinli and S. Narayanan, "A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, Antwerp, Belgium, 2007, pp. 1941–1944.

[71] V. Duangudom and D. V. Anderson, "Using auditory saliency to understand complex auditory scenes," in *Proceedings of the 15th European Signal Processing Conference (EUSIPCO 2007)*, Poznań, Poland, 2007, pp. 1206–1210.

[72] C. E. Schreiner, H. L. Read, and M. L. Sutter, "Modular organization of frequency integration in primary auditory cortex," *Annual Review of Neuroscience*, vol. 23, no. 1, pp. 501–529, 2000.

[73] A. Fishbach, Y. Yeshurun, and I. Nelken, "Neural model for physiological responses to frequency and amplitude transitions uncovers topographical order in the auditory cortex," *Journal of Neurophysiology*, vol. 90, no. 6, pp. 3663–3678, 2003.

[74] P. A. Valentine and J. J. Eggermont, "Stimulus dependence of spectro-temporal receptive fields in cat primary auditory cortex," *Hearing Research*, vol. 196, no. 1, pp. 119–133, 2004.

[75] T. Kohonen, *Self-Organizing Maps*, 3rd ed. Heidelberg, Germany: Springer-Verlag, 2001.

[76] C. Kayser, C. I. Petkov, M. Augath, and N. K. Logothetis, "Functional imaging reveals visual modulation of specific fields in auditory cortex," *The Journal of Neuroscience*, vol. 27, no. 8, pp. 1824–1835, 2007.

[77] H. Yin, *The Self-Organizing Maps: Background, Theories, Extensions and Applications*. Springer, 2008, pp. 715–762.

[78] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1-3, pp. 1–6, 1998.

[79] J. W. Lewis, W. J. Talkington, A. Puce, L. R. Engel, and C. Frum, "Cortical networks representing object categories and high-level attributes of familiar real-world action sounds," *Journal of Cognitive Neuroscience*, vol. 23, no. 8, pp. 2079–2101, 2011.

[80] D. Oldoni, B. De Coensel, M. Rademaker, T. Van Renterghem, D. Botteldooren, and B. De Baets, "Computational soundscape analysis based on a human-like auditory processing model," in *Proceedings of the EAA Euregio 2010 congress*. Slovenian Acoustical Society (SDA), 2010, pp. 1–6.

[81] B. De Coensel and D. Botteldooren, "A model of saliency-based auditory attention to environmental sound," in *20th International Congress on Acoustics (ICA-2010)*, 2010, pp. 1–8.

[82] A. Ultsch, "Self organized feature maps for monitoring and knowledge acquisition of a chemical process," in *Proceedings of the International Conference on Artificial Neural Networks*, vol. 93, Amsterdam, the Netherlands, 1993, pp. 864–867.

[83] D. Terman and D. Wang, "Global competition and local cooperation in a network of neural oscillators," *Physica D: Nonlinear Phenomena*, vol. 81, no. 1-2, pp. 148–176, 1995.

[84] D. Wang and D. Terman, "Locally excitatory globally inhibitory oscillator networks," *IEEE Transactions on Neural Networks*, vol. 6, no. 1, pp. 283–286, 1995.

[85] D. Wang, "Emergent synchrony in locally coupled neural oscillators," *IEEE Transactions on Neural Networks*, vol. 6, no. 4, pp. 941–948, 1995.

[86] D. Wang and D. Terman, "Image segmentation based on oscillatory correlation," *Neural Computation*, vol. 9, no. 4, pp. 805–836, 1997.

[87] D. L. Wang, "Auditory stream segregation based on oscillatory correlation," in *Neural Networks for Signal Processing [1994] IV. Proceedings of the 1994 IEEE Workshop*, Ermioni, Greece, 1994, pp. 624–632.

[88] D. Wang, "Primitive auditory segregation based on oscillatory correlation," *Cognitive Science*, vol. 20, no. 3, pp. 409–456, 1996.

[89] D. Wang and G. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 684–697, 1999.

[90] S. N. Wrigley and G. J. Brown, "A computational model of auditory selective attention," *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1151–1163, 2004.

[91] P. Linsay and D. Wang, "Fast numerical integration of relaxation oscillator networks based on singular limit solutions," *IEEE Transactions on Neural Networks*, vol. 9, no. 3, pp. 523–532, 1998.

[92] B. De Coensel, D. Botteldooren, K. Debacq, M. E. Nilsson, and B. Berglund, "Clustering outdoor soundscapes using fuzzy ants," in *IEEE Congress on Evolutionary Computation, 2008 (IEEE World Congress on Computational Intelligence)*. IEEE, 2008, pp. 1556–1562.

[93] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 2, pp. 224–227, 1979.

[94] J. L. Carles, I. L. Barrio, and J. V. de Lucio, "Sound influence on landscape values," *Landscape and Urban Planning*, vol. 43, no. 4, pp. 191–200, 1999.

[95] M. Zhang and J. Kang, "Towards the evaluation, description, and creation of soundscapes in urban open spaces," *Environment And Planning B: Planning And Design*, vol. 34, no. 1, p. 68, 2007.

[96] M. Raimbault and D. Dubois, "Urban soundscapes: Experiences and knowledge," *Cities*, vol. 22, no. 5, pp. 339–350, 2005.

[97] B. Hellström, M. Nilsson, P. Becker, and P. Lundén, "Acoustic design artifacts and methods for urban soundscapes," in *15th International Congress on Sound and Vibration*, 2008.

[98] B. Schulte-Fortkamp, "The tuning of noise pollution with respect to the expertise of people's mind," in *INTER-NOISE and NOISE-CON Congress*

and Conference Proceedings, vol. 2010, no. 5. Institute of Noise Control Engineering, 2010, pp. 5744–5752.

[99] B. De Coensel, S. Vanwetswinkel, and D. Botteldooren, "Effects of natural sounds on the perception of road traffic noise," *The Journal of the Acoustical Society of America*, vol. 129, no. 4, pp. EL148–EL153, 2011.

[100] R. R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decisionmaking," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 18, no. 1, pp. 183–190, 1988.

[101] S. Dauwe, D. Oldoni, B. De Baets, T. Van Renterghem, D. Botteldooren, and B. Dhoedt, "Multi-criteria anomaly detection in urban noise sensor networks," *Environmental Science: Processes & Impacts*, vol. 16, no. 10, pp. 2249–2258, 2014.

[102] D. A. Depireux, J. Z. Simon, D. J. Klein, and S. A. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *Journal of Neurophysiology*, vol. 85, no. 3, pp. 1220–1234, 2001.

[103] D. R. Langers, W. H. Backes, and P. v. Dijk, "Spectrotemporal features of the auditory cortex: the activation in response to dynamic ripples," *Neuroimage*, vol. 20, no. 1, pp. 265–275, 2003.

[104] I. Nelken, "Processing of complex stimuli and natural scenes in the auditory cortex," *Current Opinion in Neurobiology*, vol. 14, no. 4, pp. 474–480, 2004.

[105] M. Schönwiesner and R. J. Zatorre, "Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fmri," *Proceedings of the National Academy of Sciences*, vol. 106, no. 34, pp. 14 611–14 616, 2009.

[106] R. J. Zatorre and M. Schönwiesner, "Cortical speech and music processes revealed by functional neuroimaging," in *The Auditory Cortex*. Springer, 2011, pp. 657–677.

[107] J. F. Linden, R. C. Liu, M. Sahani, C. E. Schreiner, and M. M. Merzenich, "Spectrotemporal structure of receptive fields in areas ai and aaf of mouse auditory cortex," *Journal of Neurophysiology*, vol. 90, no. 4, pp. 2660–2675, 2003.

[108] D. J. Klein, J. Z. Simon, D. A. Depireux, and S. A. Shamma, "Stimulus-invariant processing and spectrotemporal reverse correlation in primary auditory cortex," *Journal of Computational Neuroscience*, vol. 20, no. 2, pp. 111–136, 2006.

[109] C. Lavandier and B. Defréville, "The contribution of sound source characteristics in the assessment of urban soundscapes," *Acta Acustica united with Acustica*, vol. 92, no. 6, pp. 912–921, 2006.

[110] D. Dubois, C. Guastavino, and M. Raimbault, "A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories," *Acta Acustica united with Acustica*, vol. 92, no. 6, pp. 865–874, 2006.

[111] M. Boes, D. Oldoni, B. De Coensel, and D. Botteldooren, "A biologically inspired recurrent neural network for sound source recognition incorporating auditory attention," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, Dallas, USA, 2013, pp. 1–8.

[112] S. Dauwe, T. Van Renterghem, D. Botteldooren, and B. Dhoedt, "Multiagent-based data fusion in environmental monitoring networks," *International Journal of Distributed Sensor Networks*, vol. 2012, 2012.

[113] H. Karl and A. Willig, *Protocols and Architectures for Wireless Sensor Networks*. Chichester, UK: John Wiley & Sons, Ltd., 2005.

[114] M. Mancas, L. Couvreur, B. Gosselin, B. Macq *et al.*, "Computational attention for event detection," in *Proceedings of ICVS Workshop on Computational Attention & Applications, WCAA*, Biedfeld, Germany, 2007.

[115] L. Couvreur, F. Bettens, J. Hancq, and M. Mancas, "Normalized auditory attention levels for automatic audio surveillance," in *International Conference on Safety and Security Engineering*, Malta, 2007.

[116] A. Mainwaring, D. Culler, J. Polastre, R. Szewczyk, and J. Anderson, "Wireless sensor networks for habitat monitoring," in *Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications*. ACM, 2002, pp. 88–97.

[117] H. Wang, D. Estrin, and L. Girod, "Preprocessing in a tiered sensor network for habitat monitoring," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 4, pp. 392–401, 1900.

[118] L. Filipponi, S. Santini, and A. Vitaletti, "Data collection in wireless sensor networks for noise pollution monitoring," in *Distributed Computing in Sensor Systems*. Springer, 2008, pp. 492–497.

[119] S. Santini, B. Ostermaier, and R. Adelmann, "On the use of sensor nodes and mobile phones for the assessment of noise pollution levels in urban environments," in *Sixth International Conference on Networked Sensing Systems (INSS), 2009*. IEEE, 2009, pp. 1–8.

[120] B. De Coensel, D. Botteldooren, T. De Muer, B. Berglund, M. E. Nilsson, and P. Lercher, "A model for the perception of environmental sound based on notice-events," *The Journal of the Acoustical Society of America*, vol. 126, no. 2, pp. 656–665, 2009.

[121] B. De Coensel and D. Botteldooren, "The quiet rural soundscape and how to characterize it," *Acta Acustica united with Acustica*, vol. 92, no. 6, pp. 887–897, 2006.

[122] C. von der Malsburg, "The correlation theory of the brain function," Max-Planck-Institute for Biophysical Chemistry, Internal Report 81-2, 1981.

[123] B. Hellström, *Noise design: Architectural Modelling and the Aesthetics of Urban Acoustic Spase, PhD Thesis.* Stockholm, Sweden: School of Architecture, Royal Institute of Technology, 2003.

[124] A. Brown and A. Muhar, "An approach to the acoustic design of outdoor space," *Journal of Environmental planning and Management*, vol. 47, no. 6, pp. 827–842, 2004.

[125] P. Grahn and U. K. Stigsdotter, "The relation between perceived sensory dimensions of urban green space and stress restoration," *Landscape and Urban Planning*, vol. 94, no. 3, pp. 264–275, 2010.

[126] E. Öhrström, A. Skånberg, H. Svensson, and A. Gidlöf-Gunnarsson, "Effects of road traffic noise and the benefit of access to quietness," *Journal of Sound and Vibration*, vol. 295, no. 1, pp. 40–59, 2006.

[127] A. Gidlöf-Gunnarsson and E. Öhrström, "Noise and well-being in urban residential environments: The potential role of perceived availability to nearby green areas," *Landscape and Urban Planning*, vol. 83, no. 2, pp. 115–126, 2007.

[128] L. Yu and J. Kang, "Factors influencing the sound preference in urban open spaces," *Applied Acoustics*, vol. 71, no. 7, pp. 622–633, 2010.

[129] A. Brown, J. Kang, and T. Gjestland, "Towards standardization in soundscape preference assessment," *Applied Acoustics*, vol. 72, no. 6, pp. 387–392, 2011.

[130] D. Botteldooren and B. De Coensel, "A model for long-term environmental sound detection," in *IEEE International Joint Conference on Neural Networks, 2008 (IEEE World Congress on Computational Intelligence)*. Hong Kong, China: IEEE, 2008, pp. 2017–2023.

[131] D. Oldoni, B. De Coensel, M. Rademaker, B. De Baets, and D. Botteldooren, "Context-dependent environmental sound monitoring using SOM coupled with LEGION," in *IEEE International Joint Conference on Neural Networks, 2010 (IEEE World Congress on Computational Intelligence)*. Barcelona, Spain: IEEE, 2010, pp. 1413–1420.

[132] 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," in *40th International Congress and Exposition on Noise Control Engineering (Inter-Noise-2011)*, vol. 5. Institute of Noise Control Engineering Japan, 2011, pp. 3988–3993.

[133] M. E. Nilsson, J. Alvarsson, M. Rådsten-Ekman, and K. Bolin, "Auditory masking of wanted and unwanted sounds in a city park," *Noise Control Engineering Journal*, vol. 58, no. 5, pp. 524–531, 2010.

[134] C. S. Watson, "Some comments on informational masking," *Acta Acustica united with Acustica*, vol. 91, no. 3, pp. 502–512, 2005.

[135] N. Durlach, "Auditory masking: Need for improved conceptual structurea)," *The Journal of the Acoustical Society of America*, vol. 120, no. 4, pp. 1787–1790, 2006.

[136] T. Y. Lee and V. M. Richards, "Evaluation of similarity effects in informational masking," *The Journal of the Acoustical Society of America*, vol. 129, no. 6, pp. EL280–EL285, 2011.

[137] B. R. Glasberg and B. C. Moore, "A model of loudness applicable to time-varying sounds," *Journal of the Audio Engineering Society*, vol. 50, no. 5, pp. 331–342, 2002.

[138] ——, "Development and evaluation of a model for predicting the audibility of time-varying sounds in the presence of background sounds," *Journal of the Audio Engineering Society*, vol. 53, no. 10, pp. 906–918, 2005.

[139] R. Leech, B. Gygi, J. Aydelott, and F. Dick, "Informational factors in identifying environmental sounds in natural auditory scenes," *The Journal of the Acoustical Society of America*, vol. 126, no. 6, pp. 3147–3155, 2009.

[140] B. Gygi and V. Shafiro, "The incongruency advantage for environmental sounds presented in natural auditory scenes." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 37, no. 2, p. 551, 2011.

[141] R. Schafer, *The soundscape: Our Sonic Environment and the Tuning of the World*. Rochester, Vermont, USA: Destiny Books, 1994.

[142] A. Baddeley, "Working memory: looking back and looking forward," *Nature Reviews Neuroscience*, vol. 4, no. 10, pp. 829–839, 2003.

[143] C. Spence and J. Driver, "Auditory and audiovisual inhibition of return," *Perception & Psychophysics*, vol. 60, no. 1, pp. 125–139, 1998.

[144] C. Escera, K. Alho, I. Winkler, and R. Näätänen, "Neural mechanisms of involuntary attention to acoustic novelty and change," *Journal of Cognitive Neuroscience*, vol. 10, no. 5, pp. 590–604, 1998.

[145] C. Escera, E. Yago, M.-J. Corral, S. Corbera, and M. I. Nunez, "Attention capture by auditory significant stimuli: semantic analysis follows attention switching," *European Journal of Neuroscience*, vol. 18, no. 8, pp. 2408–2412, 2003.

[146] J. Domínguez-Borràs, M. Garcia-Garcia, and C. Escera, "Negative emotional context enhances auditory novelty processing," *NeuroReport*, vol. 19, no. 4, pp. 503–507, 2008.

[147] J. Domínguez-Borràs, S.-A. Trautmann, P. Erhard, T. Fehr, M. Herrmann, and C. Escera, "Emotional context enhances auditory novelty processing in superior temporal gyrus," *Cerebral Cortex*, vol. 19, no. 7, pp. 1521–1529, 2009.

[148] M. Garcia-Garcia, C. Escera, I. SanMiguel, and I. Clemente, "COMT and ANKK-1 gene-gene interaction accounts for distraction effect and resetting of the gamma neural oscillations to novel sounds," *International Journal of Psychophysiology*, vol. 77, p. 231, 2010.

[149] M. Garcia-Garcia, F. Barcelo, I. Clemente, and C. Escera, "The role of dat1 gene on the rapid detection of task novelty," *Neuropsychologia*, vol. 48, no. 14, pp. 4136–4141, 2010.

[150] M. M. Bradley and P. J. Lang, "Affective reactions to acoustic stimuli," *Psychophysiology*, vol. 37, no. 02, pp. 204–215, 2000.

[151] G. Thierry and M. V. Roberts, "Event-related potential study of attention capture by affective sounds," *NeuroReport*, vol. 18, no. 3, pp. 245–248, 2007.

[152] S. Wrigley and G. Brown, "A computational model of auditory selective attention," *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1151–1163, 2004.

[153] O. Kalinli and S. Narayanan, "Prominence detection using auditory attention cues and task-dependent high level information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 1009–1024, 2009.

[154] L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 161–169, 2001.

[155] T. Kohonen, *Self-Organizing Maps*, 3rd ed., ser. Springer Series in Information Sciences, Springer, Ed. Springer, 2001, no. 30.

[156] J. Kang and M. Zhang, "Semantic differential analysis of the soundscape in urban open public spaces," *Building and Environment*, vol. 45, no. 1, pp. 150–157, 2010.

[157] R. S. Lazarus, *Emotion and Adaptation.* Oxford University Press, 1991.

[158] A. L. Bedimo-Rung, A. J. Mowen, and D. A. Cohen, "The significance of parks to physical activity and public health: a conceptual model," *American Journal of Preventive Medicine*, vol. 28, no. 2, pp. 159–168, 2005.

[159] R. S. Ulrich, "Natural versus urban scenes some psychophysiological effects," *Environment and Behavior*, vol. 13, no. 5, pp. 523–556, 1981.

[160] R. Kaplan, "The role of nature in the urban context," in *Behavior and the Natural Environment.* Springer, 1983, pp. 127–161.

[161] ——, "Nature at the doorstep: Residential satisfaction and the nearby environment," *Journal of Architectural and Planning Research*, 1985.

[162] R. S. Ulrich, R. F. Simons, B. D. Losito, E. Fiorito, M. A. Miles, and M. Zelson, "Stress recovery during exposure to natural and urban environments," *Journal of Environmental Psychology*, vol. 11, no. 3, pp. 201–230, 1991.

[163] T. Hartig, A. Böök, J. Garvill, T. Olsson, and T. Gärling, "Environmental influences on psychological restoration," *Scandinavian Journal of Psychology*, vol. 37, no. 4, pp. 378–393, 1996.

[164] R. Ulrich, "View through a window may influence recovery," *Science*, vol. 224, no. 4647, pp. 224–225, 1984.

[165] R. Kaplan, "The role of nature in the context of the workplace," *Landscape and Urban Planning*, vol. 26, no. 1, pp. 193–201, 1993.

[166] ——, "The nature of the view from home psychological benefits," *Environment and Behavior*, vol. 33, no. 4, pp. 507–542, 2001.

[167] D. Botteldooren, B. De Coensel, T. Van Renterghem, L. Dekoninck, and D. Gillis, "The urban soundscape: a different perspective," in *Duurzame Mobiliteit Vlaanderen: de Leefbare Stad*. Universiteit Gent. Instituut voor Duurzame Mobiliteit, 2008, pp. 177–204.

[168] M. Adams, T. Cox, G. Moore, B. Croxford, M. Refaee, and S. Sharples, "Sustainable soundscapes: Noise policy and the urban experience," *Urban Studies*, vol. 43, no. 13, pp. 2385–2398, 2006.

[169] B. C. Pijanowski, A. Farina, S. H. Gage, S. L. Dumyahn, and B. L. Krause, "What is soundscape ecology? an introduction and overview of an emerging new science," *Landscape Ecology*, vol. 26, no. 9, pp. 1213–1232, 2011.

[170] B. C. Pijanowski, L. J. Villanueva-Rivera, S. L. Dumyahn, A. Farina, B. L. Krause, B. M. Napoletano, S. H. Gage, and N. Pieretti, "Soundscape ecology: the science of sound in the landscape," *BioScience*, vol. 61, no. 3, pp. 203–216, 2011.

[171] D. Oldoni, B. De Coensel, M. Boes, M. Rademaker, B. De Baets, T. Van Renterghem, and D. Botteldooren, "A computational model of auditory attention for use in soundscape research," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 852–861, 2013.

[172] B. Schulte-Fortkamp, B. M. Brooks, and W. R. Bray, "Soundscape: An approach to rely on human perception and expertise in the post-modern community noise era," *Acoustics Today*, vol. 3, no. 1, pp. 7–15, 2007.

[173] B. Berglund and M. E. Nilsson, "On a tool for measuring soundscape quality in urban residential areas," *Acta Acustica united with Acustica*, vol. 92, no. 6, pp. 938–944, 2006.

[174] M. Adams, N. Bruce, W. Davies, R. Cain, P. Jennings, A. Carlyle, P. Cusack, K. Hume, and C. J. Plack, "Soundwalking as a metholdology for understanding soundscapes," *Proceedings of the Institute of Acoustics*, vol. 30, no. 2, pp. 548–554, 2008.

[175] T. Houtgast, "Frequency selectivity in amplitude-modulation detection," *The Journal of the Acoustical Society of America*, vol. 85, no. 4, pp. 1676–1680, 1989.

[176] C. Ranganath and G. Rainer, "Neural mechanisms for detecting and remembering novel events," *Nature Reviews Neuroscience*, vol. 4, no. 3, pp. 193–202, 2003.