Listening in the Wild 2015

One-day research workshop · August 28th 2015 · Queen Mary University of London

10:00 Registration, tea and coffee

10:45 Welcome (Dan Stowell)

11:00 Session 1 (Chair: Emmanouil Benetos)

* Annamaria Mesaros (Tampere University of Technology, Finland)
  Sound event detection in everyday environments
* Alison Johnston (British Trust for Ornithology)
  What proportion of birds do we detect?
  Variation in bird detectability by species, habitat and observer
* Jordi Bonada (Universitat Pompeu Fabra, Barcelona)
  Probabilistic-based synthesis of animal vocalizations

12:30 Lunch, poster session, and bells

14:00 Session 2 (Chair: Bob Sturm)

* Rob Lachlan (Queen Mary University of London)
  Analysing the evolution of complex vocal traits:
  song learning precision and syntax in chaffinches
* Alan McElligott (Queen Mary University of London)
  Mammal vocalisations: from quality to emotions

15:00 Tea and coffee

15:30 Session 3 (Chair: Dan Stowell)

* Emmanouil Benetos (Queen Mary University of London)
  Matrix factorization methods for environmental sound analysis
* Sarah Angliss (composer, roboticist and sound historian, London)
  The Bird Fancyer's Delight

16:45 Close, opportunity to continue discussions in a nearby pub/restaurant
(suggested: The Victoria, 110 Grove Road, E3 5TH)

Map

Arts Two lecture theatre:

From QMUL's East Gate, on **Mile End Road**, enter the campus then turn left. Arts Two is the third building on the left.

Map image (c) MapBox and OpenStreetMap

Wifi access:

Network: QM-Events

Passcode: taComA51ST

Listening in the Wild: speaker abstracts

Sound event detection in everyday environments

Annamaria Mesaros (Tampere University of Technology, Finland)

Our everyday environment is a complex mixture of sound sources. When describing a scene, we often explain it through labels related to the sound source and the sound production mechanism, to facilitate understanding of what is happening: dog barking, bird singing, etc. We refer to these descriptors as sound events. Detection, identification and segregation of the sound events is a natural thing for humans, but a highly challenging problem for automatic algorithms.

This presentation introduces work on sound event detection carried out at Tampere University of Technology over the last five years. Our early work on the topic concerned only the most prominent sound event at each time, to output a single sequence of events. Difficulties arise when the audio material contains a large number of sounds and many overlapping events, which is usually the case with the real world environments. The more recent approaches try to solve the problem of detecting multiple overlapping events in the audio mixtures. A promising solution is sound source separation, using the ability of non-negative matrix factorization to separate the mixture into components. Ideally, these separated components correspond to different sound sources and can be used for learning characteristics of individual sound categories. Another successful approach to dealing with multisource mixtures is deep learning. The discrimination power of the neural networks is able to overcome the complexity of the mixture and learn useful models for separate categories of sounds directly from the mixtures.

These recent developments have brought performance of sound event detection systems to a level where it it possible to develop various audio monitoring applications, for monitoring machines, people or wildlife.

What proportion of birds do we detect? Variation in bird detectability by species, habitat and observer

Alison Johnston (British Trust for Ornithology)

There is a large variation in how easily birds are detected from their acoustic signals, by both humans and machines. Loud birds with a complex song and can be detected from a large distance, whereas quieter species will often only be detected when an observer or sensor is close to the bird. However, for all species, a higher proportion of individuals are detected closer to the observer. The function describing the reduction in the number of bird detections with increasing distance from an observer can be used to estimate detectability – the proportion of individuals present within a given area that are detected. In addition to the average volume of different species, many other factors can affect the detectability of birds and this can also vary within species. Habitat and

weather affect sound attenuation which alters detectability. The skill and experience of a human observer can also affect detectability. The combination of many of these factors can lead to considerable heterogeneity in acoustic detectability, both within and between species. Therefore, to accurately estimate the underlying patterns of bird distributions and trends, it is important to account for this heterogeneity in analyses. This is particularly relevant in large citizen science datasets, which often have large variation in the skill of participants and in many other survey characteristics. The growing use of broad-scale citizen science monitoring has increased the relevance of modeling and accounting for heterogeneity in detectability.

Probabilistic-based synthesis of animal vocalizations

Jordi Bonada (Universitat Pompeu Fabra, Barcelona)

While current efforts in realistic sound synthesis focus on imitating, by means of computational models, sounds produced by objects, musical instruments or the human voice, the synthesis of realistic non-human animal acoustic vocalizations lags significantly behind, unable to meet the demands in as varied areas as virtual reality, animation, robotics, animal assisted therapy, applied biology or psychology.

In this talk we will propose a general methodology to statistically model animal vocalizations from a signal model perspective. Current approaches mostly aim at obtaining a deeper knowledge of vocal tract morphology and phonation mechanisms for generating synthetic vocalizations. Although a deeper knowledge would certainly benefit a physically inspired model, we contend that what is really necessary for a general-purpose synthesizer is a good probabilistic model of the relevant time-varying sound characteristics.

Our proposal is to model and generate animal vocalizations with context-dependent Hidden Markov Models (HMMs). For this purpose, it is not sufficient to adapt standard methods used in HMM based speech synthesis. What has made this approach to work really well for speech is not just the statistical approach itself, but also to a great degree the acoustic parameterization and the specific signal models resulting from decades of focused research. Starting from this experience, we now have the opportunity and the challenge to explore the huge acoustic diversity existing in animal vocalizations.

Some species might require radically novel approaches. Birdsongs exhibit strong frequency modulations that require non-stationary analysis methods to accurately compute its acoustic features. Contextual description used for speech is not applicable to animal vocalizations, since the syntactic organization is very different. Nevertheless, advances in sound analysis and modeling of species-specific vocalizations can be theoretically straightforwardly translated to similar vocalizations of other species. This is a key feature that favors developing a general methodology.

– ~ –

Analysing the evolution of complex vocal traits: song learning precision and syntax in chaffinches

Rob Lachlan (Queen Mary University of London)

Like many animal signals, bird song varies considerably between species. But a surprising aspect of several bird groups is the high degree to which song varies within species too. This latter type of variation is ultimately a consequence of vocal learning, a ubiquitous feature of vocal development in oscine passerines that is also found in parrots and hummingbirds. Characterizing this variation is thus central to our understanding of vocal learning – its precision, its constraints, and its evolution. I will discuss these three aspects of learning in the chaffinch (*Fringilla coelebs*), a model species for vocal development, and show how, as with inter-specific variation, computational methods are becoming increasingly important in research in this field. (1) As a consequence of cultural evolution, microgeographic variation in songs provides a signature for the processes underlying learning. By fitting cultural evolutionary models to the output of comparisons of song similarity, I will show that it is possible to infer processes of learning from field data. (2) Song learning is constrained, however, by perceptual genetic predispositions that, in the case of chaffinches, impose syntactical regularities throughout a population. I will discuss a new method to quantify such syntactical structure derived from signal redundancy. (3) Chaffinches have colonized the Atlantic Island archipelagoes of the Azores, Madeira and the Canary Islands, forming genetically isolated populations, in which previous work had indicated that song had diverged. Using the methods described above, I will demonstrate how song learning precision has deteriorated, and syntactical constraints on learning have been relaxed in these island populations.

Mammal vocalisations: from quality to emotions

Alan McElligott (Queen Mary University of London)

Using the source-filter theory of call production as a framework for our research on mammal vocal communication, I will review our key findings in two quite separate fields of study. Firstly, in a sexual selection and evolution context (using fallow deer as the model species), I will detail the parameters of male calls that are linked to male competitive abilities and therefore mating success during the breeding season or rut. Secondly, in an animal welfare context (using goats as the model species), I will explain how vocalisations can potentially be used to identify when animals are in positive or negative emotional states. This latter research is now being extended to work on poultry vocalisations with a view to automating the process of monitoring animal emotional states.

– ~ –

Matrix factorization methods for environmental sound analysis

Emmanouil Benetos (Queen Mary University of London)

Matrix factorization (or spectrogram factorization) methods form a major part of current research in audio signal analysis, leading to the creation of systems that are robust, computationally efficient, and interpretable. In this talk I will present approaches for analysing environmental sounds based on spectrogram factorization methods; the first part of the talk will focus on automatic characterisation of acoustic environments using a temporally-constrained probabilistic model, applied to a corpus of train station recordings. I will then present a model for detecting overlapping events from audio, as well as an application of this methodology to the problem of acoustic identification of bird species. The final part of this talk will discuss ongoing work towards the creation of a morphological model of acoustic scenes suitable for generating soundscapes of varying complexity, and its applicability towards the evaluation of acoustic event detection systems.

The Bird Fancyer's Delight

Sarah Angliss (composer, roboticist and sound historian, London)

In the 18th century, musical manuals circulated showing songbird keepers how to teach their birds to sing human tunes. These treatises were known as the Bird Fancyer's Delight, sheets of music specially written to play to a pet bullfinch, linnet or canary in order that it would learn the tune and sing it back. The idea was to engineer primordial feathered recorders in the home, 100 years before the arrival of the phonograph and the advent of recorded sound. Musician and inventor Sarah Angliss explores to what extent this interplay was successful and looks for its modern day equivalent.

Listening in the Wild: poster abstracts

Vocal response of male serin, Serinus serinus, to interactive playback

Ana T. Mamede

Song overlapping and alternating in birds has been studied over the past few decades and, more recently, spawned some controversy over its communicative value. Proposed hypotheses to explain the function of this vocal behavior leads to quality and/or motivation signaling. Results of previous experiments showed that male serin react to songs with shorter inter-syllable intervals but not to frequency variations. We analyze vocal response to interactive playback attempting to rate alternate and overlap song stimuli influence. Analyzing overall differences in responses between experimental treatments, namely song length and interval between songs and syllables, we found a decrease in male song length with playback overlapping and alternating. On the other hand, the decrease in inter-syllabic range during alternating playback may indicate higher aggressiveness. The results suggest that singing in overlap and alternate can be considered a threat but male reaction is different, responding more aggressively to overlap stimuli.

The Role of Form in Modelling Auditory Scene Analysis

Susan L. Denham, Martin Coath, Sarah A. Collins

Separating out and correctly grouping the sounds of a communicating animal from the natural acoustic environment poses significant challenges to models of auditory scene analysis, yet clearly animals perform this task very effectively. Inspired by the important part that form plays in the segregation and recognition of visual objects, we consider the role of form in auditory scene analysis. By form in audition we mean the dynamic spectrotemporal patterns characterising individual sound events as well as their timing with respect to each other. We present a model capable of segregating and recognising complex natural communication calls within natural acoustic environments. The model consists of four key stages: 1) incoming sounds are processed using a model of the auditory periphery, followed by 2) a model of contrast gain control; 3) two neural networks, characterised by different timescales learn to respond preferentially to specific spectrotemporal patterns in the modified peripheral response, and 4) a second order network learns the relationships between the responses at the two different temporal scales; a moving average of the network output indicates detection of the target communication call. Songs of the cirl bunting (Emberiza cirlus) are used to investigate the capabilities of the model in detecting and classifying different song types in natural environments under various noise conditions, and its sensitivity to a range of temporal and spectral manipulations. We conclude that this approach shows great promise for developing specific sensors capable of learning to detect and classify target songs and calls in natural conditions in real-time.

Quantifying difference in vocalizations of bird populations

Colm O'Reilly

Evolution has caused populations of birds to change over time. As a population of bird migrates from other populations of the same species, this population can evolve dramatically. Bird vocalisations tend be the first characteristic to experience change. These vocalisations are a strong indicator of diversification when plumage patterns and other morphological information has not altered significantly. When vocal changes evolve dramatically, reclassification of a population as a new distinct species comes into question. Thus ornithologists are interested in a systematic and repeatable method to measure call and song similarity. The work in this poster is inspired by human speech dialect separation using pitch contours. Results from experiments using data from pairs of populations considered very similar, very different and between the two extremes are shown. Initial results are promisingly close to the accepted relative level of similarity, based on ornithologist's studies. Further work will investigate how much data is required to get an accurate measure of difference between two populations, and to use vector quantization to select pitch contour thresholds.

CHiME-Home: A Dataset for Sound Source Recognition in a Domestic Environment

Peter Foster, Sid Sigtia, Sacha Krstulovic, Jon Barker, Mark Plumbley

For the task of sound source recognition, we introduce `CHiME-Home', a novel data set based on 6.8 hours of domestic environment audio recordings. We describe our approach of obtaining annotations for the recordings. Further, we quantify agreement between obtained annotations. Our annotation approach associates each 4-second excerpt from the audio recordings with multiple labels, based on a set of 7 labels associated with sound sources in the acoustic environment. With the aid of 3 human annotators, we obtain 3 sets of multi-label annotations, for 4378 4-second audio excerpts. We quantify agreement between annotators by computing Jaccard indices between sets of label assignments. Observing varying levels of agreement across labels, with a view to obtaining a representation of `ground truth' in annotations, we refine our dataset to obtain a set of multi-label annotations for 1946 audio excerpts.

Large-scale decline of bats and bush-crickets revealed thanks to automatic acoustic monitoring scheme

Yves Bas, Christian Kerbiriou, Alienor Jeliazkov, Isabelle Le Viol, Jean-François Julien

A nation-wide acoustic monitoring program based on both car transects and point counts carried out by volunteers was launched in France in 2006. Data gathered on 3560-km car transects and 1270 point counts, surveyed twice a year, revealed a

negative trend for three common bat species whose decline was previously unsuspected: Common Pipistrelle (Pipistrellus pipistrellus), Leisler's Bat (Nyctalus leisleri) and Serotine Bat (Eptesicus serotinus). Useful data were also serendipitously collected on several species of bush-crickets (Orthoptera Tettigonioidea) thanks to their nocturnal activity producing ultrasonic songs. Using a new automatic identification process on the recordings, these data also revealed unexpected decline for two common species of bush-crickets: Great Green Bush-Cricket (Tettigonia viridissima) and Large Conehead (Ruspolia nitidula). During this same period, new technologies allowed to achieve full-night sampling, and thus to increase sampling efficiency, especially for elusive species of both groups. This led us to propose in 2014 a third protocol taking advantage of these new detectors. Using both current data and simulations, we compared the three different types of acoustic data collection (car transect, short point counts, and full-night point recordings), and their statistical power to detect alarming species trends (-30 % over 10 years). Results showed that car transects were optimal for monitoring most bush-cricket species, and some of the most mobile and large bat species, while full-night recordings would provide a better monitoring for most bat species, especially those which have a high activity rate along the night. The protocols therefore showed a very good complementarity and keeping up them should help avoiding any representativeness bias.

Assessing snore sounds recorded in the home via smartphone

Amy V. Beeston, Guy J. Brown , Xibo Wang

Obstructive sleep apnoea, associated with symptomatic snoring in the majority of patients, is currently assessed via overnight polysomnography in hospital. Recent work suggests an audio channel alone may be sufficient for diagnosis (e.g. Abeyratne et al, 2013). However, sleep in clinic appears unrepresentative of sleep at home, and single-night studies have proved unreliable (Malhotra et al, 2015). A home-based assessment is therefore in high demand, and poses an interesting challenge as recordings vary enormously in microphone quality, room-acoustics, and noise sources.

This paper describes a system for capture of snoring events in domestic environments via smartphone. At registration, users identify snoring risk factors (e.g., obesity) and potential noise sources (e.g., bed-partners, who might also snore). Overnight, users' audio recordings are segmented, compressed and streamed to a server for analysis. The paper illustrates a scene-analysis approach to snore segregation, and outlines our plan for determining objective and subjective measures of snoring severity. We also show how the audio data can be supplemented with heart-rate and motion recordings from a personal fitness tracker.

Automatic Acoustic Monitoring of Natural Systems: Towards the Detection and Classification of Bird Flight Calls

Justin Salamon, Juan Pablo Bello, Andrew Farnsworth, Steve Kelling

To understand the dynamic patterns of species occurrences across the breadth of their ranges, data must be collected both at fine resolutions and over broad spatial and temporal extents. Currently, human observers collect almost all species occurrence data. However, accumulating the knowledge necessary for identification is not a trivial task, sometimes requiring thousands of hours of effort. Additionally, not all individuals can acquire the necessary skills for detection; and the number of potential observers with sufficient expertise is limited. Finally, basic physiology and logistics prevent typical human observers from collecting data in all desired locations and times. Importantly, a very small number of observers are active at night, the time at which nearly all bird migration occurs. Thus human observers cannot adequately collect the needed data to describe the composition and magnitude of these massive movements of birds. The work presented here is a collaboration between NYU's Music and Audio Research Lab (MARL) and the Cornell Lab of Ornithology (CLO). The goal is to enable the large scale study of bird migration patterns through the automatic detection and classification of bird flight calls. A pilot study is currently underway using outdoor recording units that employ a simple energy-based activation system to capture potential flight calls and transmit them to a centralized server for further analysis. In this poster we present the results of our preliminary experiments on automatic flight call detection and species classification using data collected by these units. We describe our classification model, based on unsupervised feature learning, the data used in the experiments, and present some preliminary analysis results both for flight call detection in continuous recordings and for species classification from short audio clips. Finally, we also discuss some of the main challenges highlighted by our experiments such as model generalization in the presence of varying background noise and the need for high-precision (as opposed to high classification accuracy) models.

Using Identity Vectors for the Bird Individual Identification on the Close Set

Ladislav Ptacek, Zbynek Zajic, Jan Vanek, Ludek Muller

An Identity Vector (ivector) is the State-of-the-Art in the Speaker Recognition. A speaker is represented by the supervector of accumulated statistics of speaker's data with respect to Universal Background Model (UBM). The Factor Analysis (FA) decomposition is used to reduce a huge dimensionality of the supervector to a low dimensionality space vector – ivector. An ivector could be a final representation of the speaker or it is further processed by Probabilistic Linear Discriminant Analysis (PLDA) model to maximize the ratio of between- to within-class covariance in order to increase separability of given classes. The training of both the FA decomposition and PLDA model requires a huge amount of data. But ornithologist usually records about hundreds or thousands of songs, but FA and PLDA training requires millions and/or

much more recordings. Moreover the records have to be precisely annotated (species, individuals). To avoid described obstacle in our experiment the system (FA and PLDA) was trained on the human-speech data. The main scope of the research was a Bird Individual Identification on the Closed Set. We used 5,176 bird song records origin of thirteen chiffchaff individuals. The identification accuracy varies between 61.9% and 85.8%.

Dynamic Time Warping and Affinity Propagation Clustering for the categorisation of bird species: A case study

Simone Clemente, Marco Gamba, Daniela Pessani, Livio Favaro

We present an implementation of Dynamic Time Warping to calculate the pairwise acoustic dissimilarity of bird sounds. Diurnal calls of 26 bird species were recorded from February to July 2014 close to a large industrial plant of Fiat Chrysler Automobiles located within the Riserva Naturale Orientata delle Baragge (45°29'14"N; 8°07'45"E) in Northwest Italy. Recordings were collected with a RØDE NTG2 condenser transducer microphone (frequency response 20 Hz to 20 kHz, max SPL 131dB). The microphone was mounted on a RØDE PG2 Pistol Grip and connected to a TASCAM DR-680 digital recorder (44.1 kHz sampling rate). All species were recorded at a distance of between 5 and 10 m from the caller. Segments containing vocalisations (WAV format, 16-bit amplitude resolution) were selected from the original files and coupled with the information collected in the field on the vocalising species. Each segment could include the calls of one to four bird species. Overall, a spectrographic inspection provided us with a total of 2058 audio files that we used to calculate a pairwise dissimilarity matrix with a custom-built script in Python (Python Software Foundation). The matrix was then submitted to a clustering process in R (R Core Team) using the Affinity Propagation algorithm. The most robust clusterisation was selected using the Adjusted Rand Index (ARI), evaluated between successive clustering processes. The ARI showed higher values for the clustering solution with 71 leaves. Moreover, this clustering solution showed 90% agreement with the bird species or assemblage we observed in the field. Our results demonstrate that the combination of dynamic time warping and affinity propagation clustering is a powerful tool for categorisation of wild bird calls. This approach could be used to develop effective passive acoustic monitoring systems.

Listening in the Wild 2015

One-day research workshop

August 28th 2015

Queen Mary University of London

Organised by Dan Stowell with Bob Sturm and Emmanouil Benetos

dan.stowell@qmul.ac.uk

b.sturm@qmul.ac.uk

emmanouil.benetos@qmul.ac.uk

Supported by EPSRC Platform Grant EP/K009559/1