Workshop on

Auditory Neuroscience, Cognition and Modelling

Queen Mary University of London

February 17th, 2016

Main conference location: G.02 Lecture Theatre, Joseph Rotblat Building **(nr. 5 on the map)**

Coffee, lunch and poster session: The Shield Café **(nr. 4 on the map)**

Welcome to the Workshop on Auditory Neuroscience, Cognition and Modelling 2016

This workshop brings together cognitive scientists, neuroscientists and computer scientists working on sound, music and speech processing. We are pleased to welcome such a large number of participants, with widely varying backgrounds. This diversity is reflected in today's oral and poster presentations: not only in the topics of research, but also in the methods used. Behavioural studies, neuroimaging and computational modelling will all be addressed, and sometimes combined.

Also our keynote speakers each have a very different approach to their research in sound processing: Prof. Elvira Brattico, Dr. Jean-Julien Aucouturier and Dr. Richard E. Turner. We anticipate the keynote talks, short oral presentations and posters to provide an interesting overview of the current findings as well as the available methods in the field. Most of all, this workshop is an opportunity for everyone to meet likeminded researchers, and possibly set up new multidisciplinary collaborations.

We hope you will have an informative, inspirational and fun day!

The organising committee:
Marcus Pearce
Emmanouil Benetos
Yvonne Blokland

Thanks to everyone helping out during the day:
Session chairs: Katrin Krumbholz and Slim Essid
Behind the scenes: Sam Duffy, Sarah Sauvé, Léna Delval, Miriam Tenderini and Kat Agres
The Shield Café

This workshop is supported by:

EPSRC
Engineering and Physical Sciences Research Council

centre for digital music

Programme

09:30 Registration, coffee/tea

10:00 Welcome and introduction

10:10 Keynote 1: Elvira Brattico

11:10 Oral session 1

12:10 Poster session

12:30 Lunch, poster session continuation

13:30 Keynote 2: Jean-Julien Aucouturier

14:30 Oral session 2

15:30 Coffee/tea break, poster session continuation

16:00 Keynote 3: Richard E. Turner

17:00 Closing

10:10 - 11:10
Keynote 1

Elvira Brattico
Aarhus University

Automatic and conscious processing of musical sound features in the brain

Several features of the auditory environment are analysed and predicted even before the intervention of attention in an automatic and irrepressible way in order to facilitate response to salient and potentially dangerous events. Music capitalises on variations of low-level spectrotemporal features common to other auditory signals, and is also characterised by high-level sound schemata based on conventional agreement between members of a certain musical culture, which need to be learned via acculturation. In this talk I will review my recent neurophysiological and neuroimaging studies on the attentional resources required for encoding and predicting low- vs. high-level sound features in isolation or in a realistic music context. I will also discuss how music acculturation can strikingly modify the neural processes and structures involved in musical feature processing and prediction.

Bio:
Elvira Brattico (PhD in Psychology, University of Helsinki, 2007) is Professor of Neuroscience, Music and Aesthetics at the Center for Music in the Brain (MIB), Department of Clinical Medicine, Aarhus University and Royal Academy of Music, Aarhus/Aalborg, Denmark. She also holds adjunct professorships at the University of Helsinki and the University of Jyväskylä, Finland. Her background is multidisciplinary: she studied piano performance and philosophy in Italy, and cognitive neuroscience and brain research methods in Finland and Canada. She is a pioneer in applying computational music information retrieval methods to neurophysiological and neuroimaging methods to solve questions concerning music processing, such as how the brain represents musical features, why we enjoy music, how music shapes neural structures and functions, and how each of these processes are dependent on the characteristics of the individual. Prof.Brattico has published more than 100 papers, of which 68 appear in peer-reviewed international journals or conference proceedings, and 10 invited book chapters.

13:30 - 14:30
Keynote 2

Jean-Julien Aucouturier
CNRS/IRCAM

Real-time transformations of emotional speech alter speaker's emotions in a congruent direction

Recent research about emotion regulation and forward models have suggested that emotional signals are produced in a goal directed way, and monitored for errors like other intentional actions. We created a digital audio platform to covertly modify the emotional tone of participants voices while they talked, in the direction of happiness, sadness or fear. We found that, while external listeners perceived the audio transformations as natural examples of the intended emotions, the great majority of the participants remained unaware that their own voices were being manipulated. We take this to indicate that people are not continuously monitoring their own voice to make sure it meets a predetermined emotional target. Instead, as a consequence of listening to their altered voices, the emotional state of the participants changed in congruence with the emotion portrayed, as measured both by self-report and skin conductance responses (SCR). This we believe is the first evidence of peripheral feedback effects on emotional experience in the auditory domain. As such, this result reinforces the wider framework of self-perception theory; that we often use the same inferential strategies to understand ourselves as those we use to understand others.

Bio:
JJ Aucouturier is a CNRS researcher in IRCAM in Paris. He was trained in Computer Science, and held several postdoctoral positions in Cognitive Neuroscience in RIKEN Brain Science Institute in Tokyo, Japan and Université of Dijon, France. He is now building a music neuroscience lab in IRCAM, and interested in using audio signal processing technologies to understand how sound and music create emotions. Lab website: http://cream.ircam.fr

16:00 - 17:00
Keynote 3

Richard Turner
University of Cambridge

Probabilistic models for natural audio signals

In this talk I will present a family of probabilistic models specifically designed for audio analysis that are able to automatically adapt to match the statistics of the input. At the heart of the approach are modern probabilistic machine learning methods, which provide techniques both for tuning representations and also for handling noise and missing data through an explicit representation of uncertainty. I will show that the new methods provide superior representations of audio as evidenced by results on denoising, missing data imputation and audio synthesis problems. I will also show that there is a close connection between the adapted representations and those employed by the brain for auditory scene analysis. This suggests that probabilistic modelling might shed light on the computations being performed in the auditory brain.

Bio:
Richard Turner holds a Lectureship (equivalent to US Assistant Professor) in Computer Vision and Machine Learning in the Computational and Biological Learning Lab, Department of Engineering, University of Cambridge, UK. Before taking up this position, he held an EPSRC Postdoctoral research fellowship which he spent at both the University of Cambridge and the Laboratory for Computational Vision, NYU, USA. He has a PhD degree in Computational Neuroscience and Machine Learning from the Gatsby Computational Neuroscience Unit, UCL, UK and a M.Sci. degree in Natural Sciences (specialism Physics) from the University of Cambridge, UK. His research interests include machine learning for signal processing and developing probabilistic models of perception.

Oral presentations: session 1
11:10 - 12:10

High-level influences on auditory streaming

Alexander J. Billig, Matthew H. Davis, Robert P. Carlyon
MRC Cognition & Brain Sciences Unit, Cambridge, UK

The perceptual organisation of an auditory scene establishing which sound elements arise from which sources takes place to a large extent in accordance with Gestalt principles. For example, a repeated ABA- pattern of pure tones can be heard as a single integrated form with a galloping rhythm, or as segregated into two streams, each consisting of tones of a single frequency. The lower the rate of change of frequency between A and B tones, the more likely that integration is perceived. There is also evidence that higher-level factors, such as context and attention, can influence streaming. However whether such effects have a low-level perceptual locus, or occur instead at a response stage, is not clear. Here, we investigate the effects of intention and linguistic knowledge on streaming by collecting objective behavioural and neurophysiological measures alongside subjective reports.

In the first set of experiments, participants heard sequences of repeated words (e.g. stem stem ...) or of acoustically similar non-words (e.g. stensten ...). After several presentations they reported that the initial broadband /s/ sound in each syllable formed a separate stream; the percept then fluctuated bistably between integration and segregation. In addition to measuring these verbal transformations, we required listeners to detect occasional targets – syllables containing a gap after the initial /s/. This task provides an objective measure of streaming because it is harder to compare the timing of sounds falling into separate streams than those occurring in a single stream. Performance was better for sequences in which streaming caused the syllables preceding the target to transform from words into non-words, rather than from non-words into words. This indicates that linguistic information can affect low-level perceptual organisation.

In another experiment, participants listened to ABA- patterns, either neutrally or while trying to hear them with a particular perceptual organisation. In line with subjective reports in previous studies, participants reported hearing segregation most frequently when they tried to segregate the sounds, and least frequently when trying to integrate the A and B tones into one stream. Recordings of neural activity using electro- and magneto-encephalography indicated that this was not the result of a response bias: intentions were reflected

in the evoked response, with signatures of streaming in the neutral listening condition persisting when participants tried to promote a particular percept.

This work reveals that experience and intention, as well as the acoustic properties of sounds, can determine how auditory scenes are perceptually organised.

Multiple hypothesis testing on partial coherences: Graphical modelling of Neurological data in EEG/MEG

Deborah Schneider-Luftman
Imperial College London, UK

The understanding of connectivity in large-dimensional time series has been a topic of central importance in Neurology, and more precisely in Neurological imaging. The interest in these techniques is widespread across imaging techniques such as EEG and MEG but also across applications, notably in the study of sensory and auditory functions. The most important feature of neurological data analysis is connectivity between parts of the brain: how do various regions of interest interact? For this purpose, graphical modelling of time series presents themselves as ideal tools.

In the graphical modelling of brain data issued from EEG/MEG, we are interested in estimating the partial coherences between all channels and evaluate their statistical significance in order to establish the existence of edges between channels. This process involves aggregating results across entire frequency ranges, in order to obtain an overall result that can be fed into the construction of the graph. In this paper, we compare an approach that has been tried and tested in the literature - the Holm's stepdown procedure - to a methodologies that has otherwise never been used for this application.

EEG-based Emotion Detection in Music Listening

Rafael Ramirez, Zacharias Vamvakousis, Sergio Giraldo
Music and Machine Learning Lab, Universitat Pompeu Fabra, Barcelona, Spain

The study of emotions induced by multimedia stimuli has increased in recent years. This is due to the growing need for computer applications capable of detecting the emotional state of users and adapt to them accordingly. On the other hand, current affordable technology allows the recording of brain activity in real-time and the discovery of patterns related to emotional states. This paper describes an approach to detecting music-induced emotion from electroencephalogram (EEG) signals measured with a low-cost EEG headset. Thirty-seconds music fragments with different emotional content are presented

to subjects and while their response EEG activity is recorded. The resulting EEG data is processed in order to extract emotion-related features and machine learning techniques (linear discriminant analysis and support vector machines) are applied to classify emotional states into high/low arousal and positive/negative valence. The obtained results indicate that EEG data obtained with the low-cost EEG device contains sufficient information to distinguish among the emotional states induced by the music stimuli, and that machine learning techniques are capable of learning the patterns that distinguish these states.

Oral presentations: session 2
14:30 - 15:30

Investigating the role of auditory and cognitive factors for various speech-perception-in-noise situations in older listeners

Sarah Knight and Antje Heinrich
MRC Institute of Hearing Research, Nottingham, UK

Understanding the causes for speech-in-noise (SiN) perception difficulties is complex, and is made even more difficult by the fact that listening situations can vary widely in target and background sounds. While there is general agreement that both auditory and cognitive factors are important, their exact relationship to SiN perception across various listening situations remains unclear.

This study manipulated the characteristics of the listening situation in two ways: first, target stimuli were either isolated words, or words heard in the context of low- (LP) and high-predictability (HP) sentences; second, the background sound, speech-modulated noise, was presented at two signal-to-noise ratios. Speech intelligibility was measured for 50 older listeners (ages = 61-86; mean = 70) with age-normal hearing and related to individual differences in cognition (working memory, inhibition and linguistic skills) and hearing (PTA0.25-8kHz and temporal processing).The results showed that while the effect of hearing thresholds on intelligibility was rather uniform, the influence of cognitive abilities was more specific to particular listening situations. Furthermore, the effect of cognition on speech perception was modulated by educational attainment.

By revealing a complex picture of relationships between intelligibility and cognition, these results may help us understand some of the inconsistencies in the literature as regards cognitive contributions to speech perception. They also

suggest ways to maximise and minimise the influence of cognition on speech perception by manipulating both listener and stimulus characteristics.

Contextual effects on the neural encoding of speech sounds

S. Rutten[1], R. Santoro[1], A. Hervais-Adelman[1], E. Formisano[2,3], N. Golestani[1]
[1] *Brain and Language Lab, Faculty of Medicine, University of Geneva, Switzerland*
[2] *Faculty of Psychology and Neuroscience, Department of Cognitive Neuroscience, Maastricht University, the Netherlands*
[3] *Maastricht Brain Imaging Center (MBIC), Maastricht University, the Netherlands*

Our rich and constantly changing auditory environment requires a flexible and dynamic auditory processing system. At the single cell level it has been shown that animal auditory receptive fields are responsive to contextual demands. It is unknown however, how these findings translate to humans and importantly how they translate to naturalistic contexts. We conducted a 7 Tesla fMRI study in which participants listened to speech sounds while performing two different tasks, one linguistic and one paralinguistic, on the very same stimuli. With the use of computational modeling, we mimicked the filtering of sounds by the cochlea as well as acoustic sound decomposition within the auditory cortex. This allowed us to model task-specific voxels response profiles along different acoustic dimensions, i.e. frequency, spectral modulations and temporal modulations. We found that performance of the two tasks evoked differential neural encoding of the very same sounds in different auditory areas. In the posterior auditory areas we found tuning of the response profiles for center frequencies of about 500 Hz during the linguistic task whereas for the paralinguistic task this was found for center frequencies of about 800 Hz. During the linguistic task we additionally found neural tuning for temporal modulation rates at 8 Hz and 40 Hz. Due to the fact that successful task performance requires participants to focus on distinctive acoustical features of the sounds, we further aim to better explain the voxels response profiles by employing acoustical models that differentially emphasize task-relevant acoustical features.

Phonological Model for Automatic Recognition of Continuous Speech

Vipul Arora, Aditi Lahiri and Henning Reetz
University of Oxford, UK and Goethe University Frankfurt, Germany

We aim at improving the modern automatic speech recognition (ASR) systems by incorporating principles derived from phonological knowledge and neuro-

linguistic experiments on how humans perceive and process speech. While most modern systems use phonetic units as output of their acoustic model, followed by decoding them into words and sentences, we propose the use of phonological features in place of phonetic units. The incoming speech signal is permeated with enormous variations which humans tackle effortlessly. Our phonological model, known as Featurally Underspecified Lexicon (FUL) model, explains how these variations can be taken care of very efficiently with the help of phonological features as the underlying representation in the human brain.

Our evidence for this comes from two major sources: neuro-linguistic experiments and study of different languages across the world in terms of their phonological structure as well as evolution. The EEG data from various neuro-linguistic studies [1] support that certain phonological features are underspecified in our brain and can be recognised easily even when they change to other phonological features in speech; while certain other phonological features are specified and hence, are not recognised if they are changed. An example to illustrate this point is that 'greem-bag' in speech is recognised as 'green-bag', while 'lane-duck' does not activate 'lame-duck'. This asymmetry between stimulus and activation is the key evidence for underspecified FUL model. Moreover, this phenomenon has been observed universally across different languages.

The incorporation of this phonological knowledge into ASR systems will make them handle the problem of co-articulation (e.g., 'green-bag' uttered as 'greem-bag') and reduction (e.g 'did you' pronounced as 'didja'). Our method aims at handling these problems in a more principled way, rather than over-depending on statistical training as done in modern ASR systems. Our approach will optimise the system so as to minimise the amount of training required, and also, will make the system adaptable to different languages with minimal extra training.

We have implemented the use of phonological features for digit recognition task over TIdigit database. Now, we extend it to sentence recognition by blending our phonological model into state-of-the-art Kaldi ASR toolkit, so as to test the efficacy of our approach to handle the aforementioned problems in continuous speech.

[1] Lahiri, Aditi &Reetz, H. Distinctive Features: Phonological underspecification in representation and processing. Journal of Phonetics 38 (2010) 44-59.

Poster presentations
12:10 - 13:30

---

1. Shared acoustic codes underlie emotional communication in Music and Speech – Evidence from Machine Learning

Eduardo Coutinho
Department of Music, University of Liverpool
Department of Computer Science, Imperial College London

Music and speech exhibit salient similarities in the process of emotional communication in the acoustic domain. A central aspect of this overlap is the existence of shared acoustic patterns that, at least to a certain extent, communicate perceptually similar emotions in both domains. From an Affective Sciences points of view, determining the degree of overlap between both domains is fundamental to understand the shared mechanisms underlying such phenomenon. From a Machine learning perspective, the overlap between acoustic codes for emotional expression in music and speech opens new possibilities to enlarge the amount of data available to develop music and speech emotion recognition systems. In this work, I make use of Machine Learning techniques to evaluate the similarities between the acoustic-emotional patterns in both domains. I will describe a comparative framework that includes intra- (i.e., models trained and tested on the same modality, either music or speech) and cross-domain experiments (i.e., models trained in one modality and tested on the other). In the cross-domain context, we evaluated two strategies the direct transfer between domains, and the contribution of Transfer Learning techniques (feature-representation-transfer based on Denoising Auto Encoders) for reducing the gap in the feature space distributions. The results show an excellent cross-domain generalisation performance with and without feature representation transfer in both directions. In the case of music, cross-domain approaches outperformed intra-domain models for Valence estimation, whereas for Speech intra-domain models achieve the best performance. This is the first demonstration of shared acoustic codes for emotional expression in music and speech in the time-continuous domain.

---

2. Analysis of spectral correlates of violin timbre quality in relation to experts' subjective ratings

Ewa Lukasik
Institute of Computing Science, Poznan University of Technology, Poland

Timbre is a multidimensional tonal quality of the sound that has its stationary (spectrum) and dynamical (changes of spectrum, of intensity, of pitch as well

as sound attack and decay). One of the most mysterious acoustic phenomenon is timbre of the violin sound. It has been anaysed from the times of Helmholtz and is still of interest of many contemporary researchers. The most developed technical means of the epoch have been used to analyse it and the results have been confronted with human perception and cognition. We have now well developed computational and psychoacoustical tools and methods to analyse and explain various phenomena related to the violin sound perception. However nowadays we expect much more from neuro- and cognitive sciences in an explanation of human mental processes associated with triggering of delight over the sound of the violin.

In the proposed presentation we are going to discuss sound spectral features considered as representative for the quality violin timbre and their relationship to violinist experts' subjective ratings. Violin sound quality evaluation has been performed in specific conditions of violinmaking competition. During the competition the timbre is rated as one of several sound quality criteria (intensity, responsiveness, quality of montage and overall quality) and may be expressed only on a numerical scale from 1-5 (for each string). The feature set consists of spectral audio descriptors from the Timbre Toolbox[1] related to specific frequency bands (proposed by e.g. Duennwald[2] and Fritz[3]), were resonances and anti-resonances important for quality violin sound are located. In this way some experts' preferences may be discovered. However it would be advantageous to know what internal criteria the jurors use while judging the violin timbre. Some of planned investigations in this direction will be presented.

[1]G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams. The Timbre Toolbox: Extracting audio descriptors from musical signals. J. Acoust. Soc. Am., 131, 2011, 2902 2916.

[2]H. Duennwald. Deduction of objective quality parameters on old and new violins. Catgut Acoust. Soc. J. Vol. 1, No. 7 (Series II), May 1991, 1-5.

[3]C. Fritz, A. F. Blackwell, I. Cross, J. Woodhouse & B.C.J. Moore, Exploring violin sound quality: Investigating English timbre descriptors and correlating resynthesized acoustical modifications with perceptual properties. J. Acoust. Soc. Am. 131, 2012, 783794.

---

3. Analysis of envelope following responses to natural vowels using a Fourier analyzer

Frederique J Vanheusden, Steven L Bell and David M Simpson
University of Southampton, UK

Introduction. Objective methods to detect responses to speech stimuli based on the electroencephalogram (EEG) have been proposed in literature. These could be used to evaluate whether hearing aids are giving access to speech

information and may ultimately improve hearing aid fitting, and could be an improvement to standard current protocols assisting hearing aid fitting. One suggestion has been measuring envelope following responses (EFRs) to complex speech stimuli[1]. As these stimuli have undergone significant modification, it is desirable to compare the robustness of these algorithms using natural speech stimuli. This study aimed to compare the robustness of detecting EFRs for natural speech to current standard objective measurements.

Experiments. So far, the study comprised 5 volunteers with normal hearing thresholds. Neurophysiological responses were measured using a 32-electrode EEG system (BioSemiActiveTwo, sampling rate 16 kHz). Subjects were presented by click stimuli (stimulus rate: 33.3 Hz) for 4 minutes. After this, four word stimuli of 1 second length were presented in two polarities (male speaker, 100 epochs per polarity). Stimuli were structured as /h/-vowel-/d/. All stimuli were presented through a UMC Fireface soundcard via ER-2 earphones to both ears.

Methods. For click-ABR analysis, EEG signals were bandpass filtered (30-3000 Hz). After artefact rejection (>20 V), individual epochs were stacked and the coherent average calculated. Envelope-following responses to vowels were determined using a Fourier Analyzer (FA)[1]. EEG data were bandpass filtered (1-500 Hz). f0 reference sinusoids were derived from the phase angle of the speech envelope. The amplitude of the FA spectrum was derived by multiplying these references to coherent averages of the EEG response. Results were compared against spectral amplitudes to neighbouring frequencies (f0 $\pm$ 50Hz).

Results and Conclusion. All participants showed a clear click ABR response at Cz (wave V Fsp: 13.837$\pm$5.95, mean$\pm$SD). Initial results for EFR show significantly greater responses to f0 and/or its harmonics compared to neighbouring frequencies around the vertex (F-test: 3 and 50 degrees of freedom). This is in agreement with previous studies measuring EFRs using a standard 3-electrode ABR system. More lateral electrodes did not show significantly higher spectral amplitude for f0, possibly due to a larger distance from the source producing the brainstem response. Current work thus indicates the potential of recording EFRs in natural speech.

[1]Aiken SJ, Picton TW (2006). AudiolNeurotol 11;213-232

---

4. Feature Extraction Based on Auditory Image Model for Noise-Robust Automatic Speech Recognition

X. Yang[1], M. Karbasi[2], S. Bleeck[1], and D. Kolossa[2]
[1]*Institute of Sound and Vibration Research, University of Southampton, UK*
[2]*Institute of Communication Acoustics, Ruhr-University Bochum, Germany*

Automatic speech recognition (ASR) systems are still much more susceptible to the quality of speech signal than human speech recognition (HSR). As soon as the signal-to-noise ratio (SNR) decreases from 40 dB to 0 dB, the recognition rate usually drops rapidly from above 90% to below 20%, due to the mismatch between the clean training data and the noisy testing data. On the contrary, human listeners that have normal hearing can understand speech perfectly well even at negative SNRs. One reason for the difference is that the human auditory system has the ability to capture more features of the speech in the noisy signal. This inspired us to develop the feature extraction method for an auditory model, the Auditory Image Model (AIM; Bleecket al., 2004), to improve the robustness of a speech recognizer, JASPER (Kolossaet al., 2010), which is based on Hidden Markov models (HMMs). The proposed features for ASR were extracted from the tempo-spectral representations generated by AIM, which enhances the periodicity information of voiced speech. As an evaluation, the ASR system was trained and tested on the GRID (Cooke et al., 2006) corpus, for three noise types (white noise, speech-shaped noise, and babble noise) at different SNRs (-6 - 40 dB). Compared with the widely used Mel-Frequency Cepstrum Coefficients (MFCCs), AIM features led to higher recognition rates in almost all conditions, and the performance decreased more slowly with the decay of SNR.

Bleeck, S., Ives, T., & Patterson, R. D. (2004). AIM-MAT: the auditory image model in MATLAB. ActaAcustica United with Acustica, 90(4), 781-787.

Kolossa, D., Chong, J., Zeiler, S., & Keutzer, K. (2010) Efficient manycore CHMM speech recognition for audiovisual and multistream data In, INTERSPEECH (pp. 2698-2701).

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America, 120(5), 24212424.

---

5. Adaptive Frequency Neural Networks for Dynamic Pulse and Metre Perception

Andrew Lambert[1,2], Tillman Weyde[1], and Newton Armstrong[2]
[1] *Department of Computer Science, City University London, UK*
[2] *Department of Music, City University London, UK*

Beat induction, the means by which we listen to music and perceive a steady pulse, is achieved via a perceptual and cognitive process. Often there is expressive shaping of the music such as tempo change and rubato that affects our perception of the pulse and metre. Nevertheless most people are able to tap along to a wide variety of rhythms and tempos accurately. Modelling the phenomenon of beat induction computationally, especially when processing expressively timed rhythms, is an open problem that could lead to more

accurate beat tracking methods.

Large et al. (2010) have proposed an oscillating neural network model for metre perception based on the neuro-cognitive model of nonlinear resonance. Nonlinear resonance models the way our entire nervous system resonates to rhythms we hear by representing a population of neurons as a canonical nonlinear oscillator. A Gradient Frequency Neural Network (GFNN) consists of a number of these oscillators distributed across a frequency range. The resonant response of the network adds rhythm-harmonic frequency information to the signal, which can be interpreted as a perception of beat and metre. GFNNs have since been applied successfully to a range of 'difficult' music perception problems including those with syncopated and polyrhythmic stimuli (see Angelis et. al., 2013; Velasco and Large, 2011).

In this paper we consider the reaction of perceptual models to tempo changes. We show that GFNNs perform poorly when dealing with tempo changes in the stimulus. GFNNs generally require a high concentration of oscillators per octave to capture relevant frequency information. This can lead to an increase in noise in the network, which worsens as the pulse frequency fluctuates.

We present a novel Adaptive Frequency Neural Network (AFNN) that applies Righetti et al.'s (2006) Hebbian learning rule to the oscillator frequencies in the network. The frequencies adapt to the stimulus through an attraction to local areas of resonance. A secondary elasticity rule is also added, which attracts the frequencies back to their original value. These two new behaviours increase each oscillator's entrainment basin, and allow for a great reduction in the dimensionality of the network. As a result, there is a better response to stimuli with both steady and varying pulses.

Angelis, Vassilis, Simon Holland, Paul J. Upton, and Martin Clayton. Testing a Computational Model of Rhythm Perception Using Polyrhythmic Stimuli. Journal of New Music Research 42, no. 1 (2013): 4760.

Large, Edward W., Felix V. Almonte, and Marc J. Velasco. A Canonical Model for Gradient Frequency Neural Networks. Physica D: Nonlinear Phenomena 239, no. 12 (June 15, 2010): 90511.

Righetti, Ludovic, Jonas Buchli, and Auke Jan Ijspeert. Dynamic Hebbian Learning in Adaptive Frequency Oscillators. Physica D: Nonlinear Phenomena 216, no. 2 (2006): 26981.

Velasco, Marc J., and Edward W. Large. Pulse Detection in Syncopated Rhythms Using Neural Oscillators. In 12th International Society for Music Information Retrieval Conference, 18590. Miami, FL, 2011.

6. Compensation for spectral and temporal envelope distortion caused by transmission channel acoustics

Cleo Pike[1], Amy V Beeston[2], Tim Brookes[1], Guy J Brown[2], and Russell Mason[1]
[1] *Institute of Sound Recording, University of Surrey, UK*
[2] *Department of Computer Science, University of Sheffield, UK*

Recognition of a sound's identity depends on accurate perception of its spectral and temporal envelopes, both of which are distorted in everyday listening spaces by a multitude of spectrally-altered and temporally-delayed reflections of the original sound. For example, room reflections might distort the frequency regions in which resonance cues signal the formant /e/, such that an /i/ is recognised instead (Watkins 1991,Pike et al. 2014). Similarly, room reflections might obscure dips in the temporal envelope which would otherwise cue the presence of an unvoiced plosive such as /t/ (Watkins 2005, Beeston et al. 2014). However, despite these distortions to the signal, human recognition of a sounds identity remains remarkably robust in most rooms.

A number of auditory mechanisms are thought to reduce the perceptual effects or confusions arising from acoustic distortions caused by room reflections. For instance, the 'precedence effect' is a perceptual mechanism that immediately suppresses our perception of room reflections so that we can accurately identify the direction from which a sound source originates. This prevents confusion that might arise through multiple reflected copies of the source arriving from different directions. Further mechanisms appear to allow us to compensate for the distorting influence of reflections on the spectral and temporal envelope. These mechanisms are less well-known than the precedence effect and appear to act more slowly, building up with a listeners experience of the space. Recent work investigating compensation for spectral envelope distortion caused by reflections (Pike et al. 2014, Watkins 1991) and compensation for temporal envelope distortion caused by reflections (Beeston et al. 2014, Watkins 2005) suggests that common mechanisms may underlie our robustness to both types of distortion. These mechanisms may also share a basis with the precedence effect.

Our poster sets out a skeleton conceptual model of the compensation process for both spectral envelope and temporal envelope distortion, illustrating similarities between the two compensatory processes. The model collates findings from previous listening experiments both by the authors and by other researchers, and identifies blind spots where insufficient data exists at present. This leads us to propose and discuss further behavioural, neuroscientific and computational experiments designed to test for common neural and psychological mechanisms behind compensation for spectral and temporal envelope

distortion.

Beeston, A. V., Brown, G. J. and Watkins, A. J. (2014). Perceptual compensation for the effects of reverberation on consonant identification: Evidence from studies with monaural stimuli. J AcoustSoc Am, 136(6), 30723084.
Pike, C., Brookes, T., and Mason, R. (2014). Auditory compensation for spectral colouration, 137th Convention of the Audio Engineering Society, Los Angeles, USA. 9-12 Oct, preprint 9138
Watkins, A. J. (1991). Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. J AcoustSoc Am, 90(6), 2942-2955.
Watkins, A. J. (2005). Perceptual compensation for effects of reverberation in speech identification.J AcoustSoc Am, 118(1), 249262.

---

7. Using auditory brainstem responses (ABRs) to measure hearing loss-induced increases in neural gain and its implications with tinnitus

A.J Hardy, J. de Boer, and Katrin Krumbholz
MRC Institute of Hearing Research, Nottingham University, UK

The homeostatic plasticity model posits that tinnitus is triggered by an increase in neural gain due to a reduced input from a damaged periphery. Evidence for this in humans has come from click-evoked auditory brainstem responses (ABR) (Roland Schaette & Kempter, 2006). It is thought that the wave I to V amplitude ratio reflects neural gain, with wave I reflecting activity at the auditory nerve and the wave V at the upper brainstem. Schaette (2011) found that tinnitus patients had a reduced wave I but the wave V was the same as normal hearing participants, indicating that there was an increase in neural gain. However, wave I is mainly dependent on high frequencies, whereas wave V integrates over low and high frequencies. As a result high frequency hearing loss,which is commonly associated with tinnitus can affect the wave amplitudes differently, which could cause the increased wave I/V ratio. We aim to address this confound by measuring frequency-specific ABRs, which are obtained by restricting the response from the cochlea using intense high pass masking noise with a variable cut off frequency. Responses recorded with different cut off frequencies are subtracted to produce a response which contains contributions only from a band-limited frequency region. However, we found that wave I was barely detectable using this approach, in particular low frequency responses. We also found that a very high stimuli sound level was required and therefore a very intense noise would be needed to effectively mask it. Therefore our aim is to develop a method of recording frequency specific ABRs with improved signal to noise ratios but without exposure to very high intensity sound that is used in traditional ABR recordings.

Our first step is to replace the eliciting click with a rising chirp stimulus which starts at a low frequency and rises to a high frequency over a short period of time. The chirp causes the high and low frequency auditory nerves to fire simultaneously, by delaying the high frequencies relative to the low frequencies. A second step is to use low-pass filtering of the stimulus, which will reduce the high pass noise level required to eliminate contributions outside the targeted frequency band. If there is a neural gain increase we will expect to see an increase in wave I to V ratio at hearing loss regions in the tinnitus group compared to normal/hearing matched groups but not in regions where hearing is intact.

Schaette, R., & Kempter, R. (2006). Development of tinnitus-related neuronal hyperactivity through homeostatic plasticity after hearing loss: A computational model. European Journal of Neuroscience, 23(December 2005), 31243138. doi:10.1111/j.1460-9568.2006.04774.x
Schaette, R., & McAlpine, D. (2011). Tinnitus with a Normal Audiogram: Physiological Evidence for Hidden Hearing Loss and Computational Model, 31(38), 1345213457. doi:10.1523/JNEUROSCI.2156-11.2011

---

8. A mobile-based platform for evaluating localisation of virtual sound sources

Mark Steadman and Lorenzo Picinali
Dyson School of Design Engineering, Imperial College London, UK

Over the last few decades, hearing aid technology has advanced dramatically. However, many of the functions of modern devices are often underused or overlooked entirely. The aim of the Horizon 2020 3D Tune-In project is to use game design techniques to create a series of applications aimed at enabling hearing aid users to optimise the use of their devices through education, training and calibration.

As part of this project, a mobile-based virtual reality platform was developed at Imperial College London in order to investigate factors influencing the effectiveness of binaural audio rendered using non-individualised head-related transfer functions (HRTFs). The platform will be used to evaluate the ability of human listeners to localise virtual sound sources and to investigate the effectiveness of gamification in a sound localisation training paradigm.

This research is being coordinated by Dr Lorenzo Picinali in the newly-formed Dyson School of Design Engineering at Imperial College London.

9. A model-based EEG approach for investigating the hierarchical nature of continuous speech processing

Giovanni M. Di Liberto, Michael J. Crosse, and Edmund C. Lalor
University of Dublin, Trinity College, Ireland

That cortical sensory systems are organized in a hierarchical structure is reasonably well established. In the context of human speech it has been suggested that such an organization could explain how acoustically variable inputs can be perceived as categorical speech units. A number of studies have been conducted to reveal the precise mechanisms that underlie this hierarchical system; however the analysis methodologies have limited the stimuli to unnaturalistic discrete units of speech, such as isolated syllables or words.

An approach for indexing the neurophysiology of this hierarchical processing in the context of natural, continuous speech has been recently introduced (Di Liberto et al., Current Biology, 2015). Specifically, the relationship between continuous speech and low-frequency EEG responses was estimated using a multivariate regression model based on different speech representations. This mapping was shown to be best described when speech was represented using both its low-level spectrotemporal information and a categorical labeling of its phonetic features. While this approach results in a quantitative measure of scalp neural activity related to phonetic features, it remains unclear to what extent this measure reflects speech-specific processing.

Here, we outline an experiment aimed at investigating the speech-specific nature of our model-based neural measure. The intelligibility of 10-s speech stimuli was degraded using noise vocoding. Each vocoded stimulus was presented twice with an intervening presentation of the original clean speech version of the same stimulus. As such, the second presentation of the vocoded stimulus was primed by the clean speech and was found to be significantly more intelligible on a match-to-sample task. Our model-based neural measure was found to be significantly correlated with a behavioral measure of intelligibility, suggesting that we have isolated a dependent measure of speech-specific processing at the phonetic level.

10. Towards a Library of Musical Core-Signals

Clara Hollomey
Glasgow Caledonian University, UK

Musical timbre is not yet clearly understood and this lack of knowledge severely impedes research on music perception and cognition as well as advances in music information retrieval. Research is clearly complicated by the multidimensionality timbre is assumed to possess in contrast to most other psycho-

physical signal descriptors such as pitch, loudness and sound duration. While in perceptual terms, spectral information is thought to clearly outweigh temporal features, recent research has once more suggested that it is the combination of spectral and temporal cues that is being processed in the brain to yield the sensation of musical timbre.

Research in that area is likely to be data-heavy and suffers from the lack of a clear definition for timbre. Finding a way to simplify musical signals in a reproducible way would clearly facilitate further studies. Musical sounds can be generated in a large variety of ways. With many questions unsolved from their origin to their reception by the hair cells of the inner ear, the application of a source-filter model similar to those heavily in use in the field of speech communication, seems not feasible for the moment.

Approaching the problem the other way round, namely from the perspective of the actual spectrum arriving at the human ear, yields the generation of signals with known theoretic properties causing a similar sensation to actual recordings of musical instruments. Arguably, since the advent of samplers in the music industry, a vast amount of such simplified music signals exist, but they are usually perceptually inconsistent and their exact signal-theoretic properties are usually unknown. Moreover, their representativeness with regards to the full range of a musical instrument as well as to the type of listener is not asserted.

In order to establish such a library, one would ideally know about the number of partials necessary for a perceptually close reconstruction, their frequencies, the amount of frequency deviation occurring due to varying sound pressure, pitch, delay and dispersion, the possible duration and the average temporal shape, as well as the relative importance of phase.

A series of listening tests has been conducted to answer these questions and signals have been generated based on their outcomes. This library of perceptually justified core signals can be useful for many technical applications such as source separation, crosstalk cancellation and audio bandwidth extension and will be used for further research in the field of music perception.

---

11. Functional neural modelling of just noticeable difference in interaural time detection for normal hearing and bilateral cochlear implant users

Andreas N. Prokopiou, Jan Wouters, and Tom Francart
ExpORL: Research Group Experimental Oto-rhino-laryngology, KULeuven, Belgium

Sound localisation is mediated by interaural time differences (ITD) and interaural level differences (ILD). Cochlear implant (CI) users of current clinical systems do not have access to ITD cues because of the loss of phase infor-

mation after sound processing. This loss of phase is partly caused by the fixed period of the carrier pulse train. However, recent studies have shown that it is possible for bimodal and bilateral CI users to perceive ITDs from interaural phase shifts of the envelope of modulated pulse train stimuli, such as the ones generated by clinical processors. Furthermore the shape of the temporal envelope determines ITD sensitivity. This creates a wide parameter space which has not been fully explored. To aid in the study of this phenomenon a computational model was developed. It amalgamates existing work on acoustic stimulation and electrical stimulation of the auditory nerve with a novel neurometric-psychometric comparison which aims at estimating the just noticeable difference (JND) in ITD for both normal hearing subjects and CI users. The acoustical and the electrical stimulation of the auditory nerve constitute the peripheral processing part of the model. In both cases the output is a train of action potentials generated by the auditory nerve. The model used for electrical stimulation is an already existing model, developed by Bruce et al. [IEEE Tr Biomed. Eng. 1999] and extended by Goldwyn et al. [J Neurophysiol 2012]. Temporal aspects of the generated spike train, such as neuron refractory period, adaptation and response latency to the electric pulse are well predicted by this model. The model used for acoustic stimulation is also an already existing model developed by Zilany et al. [J. Acoust. Soc. Am. 2014]. It is a cascade of phenomenological description of the major functional components of the auditory-periphery, from the middle-ear to the auditory nerve. The neurometric-psychometric comparison developed takes as inputs the action potentials generated by the left and right auditory nerve stimulation models, either electrically or acoustically, thus permitting the modelling of normal hearing, bilateral CI and bimodal listeners' psychometric estimates, specifically the JND in ITD. This is especially useful, as estimated JND values can be validated against experimental investigations of the human binaural performance, and assess the model's performance. Furthermore the JND estimation can be directly used as a performance metric of novel temporal enhancement stimulation strategies, and as such the model can readily be applied as a test-bench for developing strategies for bimodal and bilateral CIs.

---

12. Sensitivity to the statistics of rapid, stochastic tone sequences

Sijia Zhao[1], Marcus Pearce[2], Fred Dick[3], and Maria Chait[1]
[1] *Ear Institute, University College London, UK*
[2] *Queen Mary University of London, UK*
[3] *Birkbeck-UCL Centre for NeuroImaging, University of London, UK*

Background. Accumulating work suggests that the human brain is remarkably sensitive to patterns in sound. However, much of the work to date has

focused on regular patterns. Here we used psychophysics and EEG to investigate what statistical information listeners are tracking in the random pattern, the time-scales associated with these processes, and their underlying brain mechanisms.

The experimental paradigm measured behavioural and neural responses to temporally jittered spectral transitions in the random pitch pattern of one-pip sequences. To detect these transitions, listeners had to integrate information over different timescales. By measuring reaction time (RT) and d' to transitions, we infer what - and how much - information about the signal listeners extract.

Methods. In the behavioural experiments listeners responded to transitions that were embedded within stochastic sequences. Sequences were created by 80 50-ms tone-pip whose frequency was drawn randomly with replacement from a fixed pool of 20 log-spaced values between 200-2k Hz. 50% of trials contained a change in statistics partway through the stimuli, namely a reduction in frequency pool size (from 20 to 10). Conditions differed in the spectral distribution of the reduced tone pool: 10 tones sampled equally from the entire frequency range (All Bands), 10 highest frequency tones (High), 10 lowest tones (Low), the 10 middle tones (Medium), or the 10 'edge' frequencies (5 highest & 5 lowest; Edge). The stimulus set also included STEP stimuli, with the transition being a simple step change in frequency; this condition was used to estimate basic audiomotor RT. In the EEG experiment (N = 11), passive nave listeners listened to High, Medium, Edge and no-change control stimuli while watching silent videos.

Results. RT and d measures showed detection was fastest and most accurate to High and Low spectral transitions (average detection ∼11 tones post-transition), followed by Medium (∼15 tones) and then Edge (∼18 tones). We compared behavioural data to both statistical simulations and a variable-length Markov chain model. EEG analysis suggested that sustained global field power diverged near the point where statistical information allows for reliable transition detection, consistent with other work using similar paradigms.

Conclusion. Performance revealed that human listeners were tuned to the statistical structure of rapid, random, tone-pip sequences. The tracking of statistical structure and detection of change in statistics occurs automatically, irrespective of attentional focus.

---

13. A meta-analysis and systematic review of perceptual studies of high resolution audio discrimination

Joshua D. Reiss
Centre for Digital Music, Queen Mary University of London, UK

There is considerable debate over the benefits of recording and rendering high

resolution audio, i.e., systems and formats that are capable of rendering beyond CD quality audio, i.e., more than 16 bits and/or more than 44.1 kHz sample rate. We undertook a systematic review and meta-analysis to assess the ability of test subjects to perceive a difference between high resolution and standard, 16 bit or 44.1 kHz audio. Fifty publications describing high resolution audio perception experiments were reviewed, and all published experiments for which sufficient data could be obtained were subjected to a meta-analysis. Overall, the results showed a small but statistically significant ability of test subjects to discriminate high resolution content, especially when test subjects received extensive training. Potential biases in studies, effect of test methodology, experimental design and choice of stimuli were also investigated. Based on the analysis, conclusions are provided concerning the nature of our ability to perceive high resolution audio content, and recommendations are given concerning how best to evaluate and investigate this perception.

---

14. Does adaptation sharpen frequency representation in auditory cortex?

Oscar Woolnough, Jessica de Boer, Katrin Krumbholz, Rob Mill, and Chris Sumner
MRC Institute of Hearing Research, Nottingham, UK

Background. Adaptation, the reduction in neural responses to repeated stimuli, is a ubiquitous phenomenon throughout sensory systems and is an important aspect of neural coding. Here we test the hypothesis that prolonged adaption increases the frequency selectivity of cortical responses in human EEG recordings, and present a neural model to explain the responses.

Method. Pure tone adapter-probe sequences were presented in which a single probe tone of one frequency was preceded by adapters of a second frequency (frequency difference: 0-, 600-, 1800- cents). Adapters consisted of either a train of 100ms tones (25ms gap), varying in number (1, 2, 3, 6, 9, 15), or a single adapter which varied in duration (100, 225, 350, 725, 1100, 1850-ms). Adaptation was characterized as a reduction of the probe response relative to the response to the probe alone.

Results. In all cases the response to the probe was most strongly reduced when the adapter frequency was closest to the probe frequency. However, the degree of adaptation depended on the temporal arrangement of adapters. At all frequency differences, adaptation of the probe initially grew with increasing numbers of adapters, or adapter duration. At long adapter durations (>400ms), regardless of frequency difference, adaptation was reduced relative to peak values. Adaptation also reduced with >3 repeated adapters, but only when there was a difference in frequency between the adapter and probe. Effectively, the tuning of adaptation became sharper with increasing numbers of

adapters, but not with adapter duration.

EEG adaptation tuning was explained by an extension to a model proposed to explain frequency specific adaptation in single neurons (Mill et al. 2011). The model was a two-layer network with convergent inputs, independently adapting synapses and separate non-adapting inputs for onset responses. This model is able to quantitatively reproduce the observed non-monotonic adaptation and sharpening of tuning observed in our EEG responses, and the effects of repeated and prolonged adapters.

Conclusion. The results suggest that adaptation from repeated, but not prolonged stimulation leads to a sharpening of the frequency tuning in auditory cortex, which may serve to emphasize the representation of stimuli that are different in frequency to those that precede it. This may be a natural consequence of a hierarchical network of neurons with independently adapting inputs.

Mill R, Coath M, Wennekers T, Denham SL (2011) Aneurocomputational model of stimulus-specific adaptation to oddball and Markov sequences. PLoS ComputBiol 7.

---

15. Perceiving auditory streams for instrumental and vocal music: the effects of prior knowledge and frequency separation

Sandra Quinn and Eliza Mclaughlin
Division of Psychology, University of Abertay Dundee, UK

When listeners are unaware of the presence of a target melody (a nursery rhyme), recognition of the target from a non-target melody (a different nursery rhyme) improves as the frequency separation between the melodies increase and the two patterns are segregated. However, when listeners are told the name of the target in advance, recognition occurs even under circumstances where the melodies are presented in a similar frequency range. It is possible identifying the target relied on an enhanced recognition process that was stimulated by activation of the schema rather than stream segregation. Additionally, activating the schema for the target (e.g. a nursery rhyme that was presented instrumentally), may have provided the listeners with access to the words stored in long-term memory. In doing so, listeners may have accessed information that enhanced their performance when the two melodies were physically similar. However, it is unclear whether stream segregation and/or the vocal content were responsible for recognition in the latter study.

To address this, we used a task where listeners were told the name of a familiar target melody in advance and were asked to detect the target from a non-target melody. In condition 1, listeners were presented with instrumental versions of the melodies where the notes in the target were presented in a similar pitch range as the non-target or transposed above or below the non-

target. Condition 2 was the same as condition 1, except that the melodies were sung by a male vocalist. In both conditions, listeners were asked to determine whether the target was higher or lower in pitch than the non-target.

As the frequency separation between the target and non-target melodies increased, detecting the target improved in both conditions. However, listeners showed greater variability in detecting the target from the non-target when the melodies were presented instrumentally than vocally, suggesting that listeners were more certain of their responses in condition 2.

Therefore, melody recognition may not require an additional segregation process when the schema for a melody is activated. Instead the former studies may rely on two independent processes: one involving stream segregation and recognition and another involving recognition alone. Additionally, having access to the vocal content of a melody is unlikely to explain the results described at the outset. Instead, our results demonstrate that listeners become increasingly confident in recognising the vocal target, suggesting that vocalisations may provide some additional benefit to melody recognition.

---

16. Stimulus predictability dynamically modulates neural gain in the auditory processing stream

Ryszard Auksztulewicz[1,2], Nicolas Barascud[3,4], Gerald Cooray[2,5], Maria Chait[3], and Karl Friston[2]

[1] *Oxford Centre for Human Brain Activity, University of Oxford, UK*
[2] *Wellcome Trust Centre for Neuroimaging, University College London, UK*
[3] *Ear Institute, University College London, UK*
[4] *Ecole Normale Suprieure, Paris, France*
[5] *Clinical Neurophysiology, Karolinska University Hospital, Stockholm, Sweden*

Recent MEG work demonstrates that regularly repeating tone sequences evoke neural activity characterised by higher amplitude of sustained fields (SF) than random sequences [1]. Furthermore, SF amplitude depends on alphabet size of the sequence. Here, we aimed to provide a mechanistic explanation of the observed SF amplitude shifts. Specifically, we tested whether (i) pattern regularity and alphabet size affect the amplitude of induced high-frequency (broadband) gamma oscillations; (ii) these effects can be modelled as slowly fluctuating changes of neural gain in a biophysically realistic network; (iii) modulations of neural gain can explain the observed SF dynamics.

We source-localised MEG activity evoked by acoustic pattern onset using multiple sparse priors. Peak voxels in bilateral superior temporal gyri (STG) and inferior parietal cortices (IPC) were used to extract individual subjects source-level time-series, separately for each condition (regular/random sequences; alphabet size 5/10/15). Based on these virtual electrode data, time-frequency maps were estimated using a multitaper method (frequency

range: 8-128Hz) and entered into a flexible-factorial GLM. We identified significant main effects of sequence regularity and alphabet size on the amplitude of induced beta and gamma oscillations in the IPC, and an interaction effect on induced gamma power in the STG.

To model these effects in terms of the underlying neurophysiology, we used dynamic causal modelling (DCM) for cross-spectral density. DCM allows to explain the observed neural responses using biologically realistic mean-field models of coupled dynamical systems. Here the observed grand-averaged spectra were modelled in a network of two interacting cortical sources in the STG and IPC. Each cortical source was modelled as a canonical cortical microcircuit. We estimated slowly fluctuating changes in intrinsic connectivity (modelling neural gain) by fitting the model to a series of consecutive, partially overlapping time windows during acoustic stimulation.

To establish whether modulations of neural gain can explain the observed SF dynamics, the time-resolved DCM parameter estimates were treated as predictors in a stepwise multilinear regression of the sensor-level SF interaction effect between stimulus regularity and sequence size. Finally, we directly simulated the SF effects observed in sensor-level data by passing sustained sensory input through a network model with parameters established in the DCM analysis. The output of this simulation had a similar dynamic and amplitude as the SF effects observed independently. This provides evidence that the slow-fluctuating changes in synaptic efficacy combined with sustained input (e.g. tone sequences) can result in large sustained effects on neural activity.

[1] Barascud, N., Pearce, M. T., Griffiths T, Friston, K & Chait M. (in press). Brain responses in humans reveal ideal-observer like sensitivity to complex acoustic patterns. Proceedings of the National Academyof Sciences.

---

17. Automatic identification of musical schemata via symbolic fingerprinting and temporal filters

Andreas Katsiavalos, Tom Collins, and Bret Battey
De Montfort University, Leicester, UK

In the books A Classic Turn of Phrase, and Music in the Galant Style, Gjerdingen defines a schema as a combination of scale degree sequences in melody and bass, underlain with specific harmonies and occurring with certain metric strengths. During the galant and Classical periods, schemata were perceptually relevant, if abstract, units of musical expression for composers and audiences, and remain so for enculturated listeners today. When identifying schemata automatically (either according to Gjerdingen's definition or in music cognition more generally), challenging aspects include that schemata do not occur on the musical surface but require the extraction of abstract rep-

resentations, and that the relevant features may be interspersed by multiple synchronous and asynchronous musical events.

Initially we encoded a set of musical schemata based on Gjerdingen's books, to establish a practical ground truth and to explore possible syntaxes for representing schemata digitally. Next we developed several algorithms with the aim of identifying all instances of a particular schema (called the "Meyer") in a given piece of music. The algorithms comprise the estimation of global key and hence scale degree for each note, followed by symbolic fingerprinting to establish matches to the query sequences in melody and bass. The algorithms vary in the number and nature of filters that are applied to these potential matches' temporal characteristics, in order to try to differentiate true positives from false identifications.

Preliminary results indicate that the algorithms have high recall (they tend to find Gjerdingen's annotated schemata) but less-than-perfect precision (they find approximately two to three times as many supposed occurrences). We are yet to implement a functional-harmonic filter and conduct rigorous evaluation of the temporal filters, both of which have the potential to improve precision. Testing these filters and their effect on the identification algorithms' performance is already shedding light on the cognitive underpinnings of music, enabling us to infer properties that give rise to the perception of musical schemata.

---

18. Can the non-human primate core-belt model be applied to the human auditory cortex? Evidence from functional and structural MRI at 7 Tesla

Julien Besle[1], Olivier Mougin[2], Rosa Sanchez-Panchuelo[2], Penny Gowland[2], Richard Bowtell[2], Sue Francis[2], and Katrin Krumbholz[1]
[1] *MRC Institute of Hearing Research, Nottingham, UK*
[2] *Sir Peter Mansfield Magnetic Resonance Centre, University of Nottingham, UK*

Anatomical and electrophysiological studies in non-human primates (NHP) have subdivided the auditory cortex (AC) into core and belt areas showing distinct structural and functional properties. It remains unclear however whether analogous subdivisions exist in the human AC and how they are laid out on the supratemporal plane. Post-mortem and in-vivo anatomical evidence shows that a core area with similar properties as in the NHP is located along Heschls-gyrus (HG), whereas tonotopic mapping using fMRI suggests that the main axis of the core and belt areas should cross HG. Recent models of the human AC aiming at reconciling these data with the NHP model have proposed that core areas might in fact be at an angle with HG and that tonotopic gradients in difference sub-areas of the core and belt might not be aligned. Here we

measured tonotopy, frequency selectivity and myelination in 12 subjects at ultra-high magnetic field (7T) and specifically tested the recently proposed models.

For the functional mapping (sparse 2D GRE EPI, 1.5 mm resolution, phase-corrected for B0-related distortions), we estimated best frequency and frequency selectivity using trains of narrowband noises at 7 single centre frequencies. For structural mapping of myelination, we estimated the R1 longitudinal relaxation rate (3D MP2RAGE, 0.6 mm resolution) and the magnetization transfer ratio (3D MTRAGE, 0.7 mm resolution). All structural and functional measures were projected onto a flattened model of the supratemporal cortex, segmented from the high-resolution processed MP2RAGE volume. Structural measures were corrected for cortical thickness and curvature. Group-averaged maps were created using spherical registration.

In all subjects/hemispheres, we identified 2 central tonotopic gradients (high-to-low-to-high) oriented, each oriented at 70° relative to HG (and 140° with each other), confirming that the main tonotopic gradients are oriented across HG, although they are not exactly aligned. In most hemispheres, a core area of high myelination and high selectivity was found aligned with HG (only in its medial part for myelination), with no evidence for an angle between core and HG. On the group-averaged maps, the area of higher myelination/selectivity corresponded mostly to the most anterior central gradient that was previously identified with R in the NHP model. In individual maps however there was considerable inter-subject variability, with higher myelination and selectivity corresponding to either or both of the central gradients. Overall, these results suggest that even a tweaked version of the NHP model cannot be applied to the human auditory cortex.

---

19. Validation of a new open-source platform for real-time emotional speech transformation

Laura Rachman
STMS CNRS UMR9912/IRCAM/UPMC, Paris, France

Our voice gives us a powerful tool to convey the emotional state we are in. Moreover, people are typically very good at recognizing emotions in other voices. And while emotion research in the audio domain is catching up on the visual domain, tools to control the emotional content auditory stimulus material are still scarce. We present here a new open-source platform that can manipulate neutral speech to make it sound more emotional. The platform provides four different audio effects: pitch shift, vibrato, inflection and filtering. By combining these effects in different configurations, we created three emotional transformations: happy, sad and afraid. Not only can these audio effects be applied to prerecorded speech signals, but with a latency of less than

20 milliseconds they can also be used for real-time speech manipulation using live input from a microphone.

In a series of validation experiments we presented recordings of semantically neutral utterances transformed by our software tool. Results showed that synthesized emotional expressions were recognized at levels above chance rate in the French, English, Swedish, and Japanese languages. Moreover, the naturalness of the transformed voices was comparable to natural speech.

For future research applications in psychology and neuroscience, this new tool can provide a high level of systematic control over the acoustical and emotional content of stimuli, or a means to study social communication in real-world conversation contexts, opening new doors to a wide variety of novel experimental paradigms.

---

20. The reduced GABA concentration with absolute pitch possessors revealed by Magnetic resonance spectroscopy

Tomoya Nakai[1,2], Hiroaki Maeshima[1], Chihiro Hosoda[1,4], Kazuo Okanoya[1,3,5*]
[1] *Graduate School of Arts and Sciences, The University of Tokyo, Japan*
[2] *Japan Society for the Promotion of Science*
[3] *Center for Evolutionary Cognitive Science, The University of Tokyo, Japan*
[4] *PRESTO, Japan Science and Technology Agency (JST), Tokyo, Japan*
[5] *Riken Brain Science Institute*
*Corresponding author

Absolute pitch (AP) is the capacity to identify the pitch of any tone in the absence of an external reference. A previous study has shown the leftward asymmetry in the auditory cortex for AP possessors (Schlaug et al., 1995). AP possessors also showed distinct activation patterns in the auditory cortex (Schulze et al., 2009). It is possible that such anatomical/functional difference is accompanied with physiological difference. GABA (Gamma-aminobutyric acid) is known as inhibitory neurotransmitter in the adult human brain, and recent neuroimaging techniques revealed strong correlation between GABA concentration and the peak gamma frequency in the visual cortex (Muthukumaraswamy et al., 2009, Edden et al., 2009). In the present study, we measured the concentration of GABA in the bilateral auditory cortex by using Magnetic resonance spectroscopy with frequency-selective editing pulses, called as MEGA-PRESS (MEshcherGArwood Point RESolved Spectroscopy) sequence (Mescher et al., 1998). We recruited 23 participants who had a normal hearing (aged from 18 to 21 years, 9 females, 2 left-handers). We performed an absolute pitch (AP) test (Keenan et al., 2001), where the participants listened 13 pure tones from F#4 (370 Hz) to F#5 (740 Hz), and each tone was presented four times. Based on the result of the AP test, 10 participants were classified as partial or perfect AP possessors (average music training, 9.1 years), while

13 participants were classified as Non-AP possessors (average music training, 1.3 years). We found a moderate negative correlation between GABA concentration and AP test scores in the auditory cortex in both hemispheres (correlation coefficient in the left auditory cortex: r = -036, right: r = -0.33). By using two-way repeated measures analysis of variance, we found significant main effect of AP (AP/Non-AP). The main effect of hemispheres (Left/Right) and interaction were not significant. These results suggest that AP possessors have less GABA concentration in their auditory cortex compared to non-AP possessors, indicating the importance of GABAergic inhibition in the auditory cortex for pitch recognition.

---

21. Rising to the challenge: modelling transfer learning of polyphonic musical structure

Reinier de Valk and Tillman Weyde
Music Informatics Research Group, Department of Computer Science, City University London, UK

We describe an experiment in which a classifier neural network model for voice separation is applied to datasets of different polyphonic complexity. The model operates on symbolic data and learns to assign notes to voice classes. We use a set of ten four-voice sixteenth-century pieces in MIDI format, approximately 6000 notes in size, and divide it into three subsets of decreasing polyphonic complexity. The *imitative* subset contains pieces whose polyphonic structure is governed by motivic imitation; the *semi-imitative* subset contains pieces that contain imitation but whose structure is not governed by it, and the *non-imitative* subset contains pieces that contain no imitation.

In an initial experiment we then train and evaluate the model on each subset using piece-wise cross-validation. The model is evaluated in two modes. In *test mode*, like in training, the feature vectors are calculated using the correct voice information. In *application mode*, which corresponds to the real-world situation where no voice information is known, the feature vectors are calculated using the voice information generated by the model.

As expected, the model shows the highest accuracy on the least complex subset, and the lowest on the most complex. We then train the model on each subset t and evaluate it on each of the two remaining subsets v. We find that in test mode, the performance ranking of the model only depends on the complexity of the training data, and, unexpectedly, not on the complexity of the evaluation data. When t is more complex than v, the model trained on t and evaluated on v always performs better on v than the model trained on v itself (i.e., the model from the initial experiment). This effect is also witnessed in application mode, where it is less pronounced.

The model can thus in some cases transfer more relevant information about

v from t than it can extract from v itself. From a machine learning point of view, this is a somewhat surprising result, as we would expect best performance by training and testing on the same type of data. From a cognitive point of view, however, it makes sense: it indicates that learning from challenging material is more effective. We plan further studies including larger datasets to investigate how widespread this effect is, how it compares to human cognition, and how it can be understood in more detail.

---

22. Comparison of reaction time measurement and Yes/no question paradigm regarding the perception of spatial coherence

Hanne Stenzel and Philip J. B. Jackson
Centre for Vision, Speech and Signal Processing, University of Surrey, UK

Studies from multimedia research on perceived audio-visual spatial coherence and the degree of accepted spatial mismatch have generally employed either impairment scales as suggested by ITU or yes/no question paradigms to generate a psychometric curve. These methods inherently draw the attention of participants towards the percept of coherence; this form of priming is reported to bias the measurement of coherence. In contrast, studies in neuroscience and psychoacoustics often use measurement of reaction time to define cognitive load. Yet, the typical flashing-light-and-noise-burst stimuli lack ecological validity. To study the perception of coherence for applications in TV and cinema in an unbiased manner, experiments with ecologically-valid, speech stimuli were conducted to measure the psychometric curve on perceived coherence together with measurements of reaction times. Preliminary results will be presented at the workshop.

---

23. EEG-powered Soundtrack for Interactive Theatre

Grigore Burloiu[1], Alexandru Berceanu[2], and Cătălin Creţu[3]
[1] *University Politehnica Bucharest, Romania*
[2] *UNATC Bucharest, Romania*
[3] *National University of Music Bucharest, Romania*

The advent of accessible EEG headsets such as the Emotiv EPOC has made possible the real-time control of live performance environments through BCIs. Such a project is INTER@FAŢA, an interactive theatre piece where generative musical algorithms are directed by the interplay between a pair of EPOC signal bundles, sourced from a performer and an audience member.

We developed a bespoke C++ application that captures the data streams from the Emotiv API and also performs spectral analysis on the AF3 and AF4 pre-fontal cortex sensor output to compute alpha and beta wave activity. The application then relays the combined data frames through OSC to

the Max/MSP programming environment, where specific parts of the frames control certain musical algorithms at different places in the script.

The input frames thus consist of the following streams: 14 raw EEG amplitude signals from the headset sensors, 5 affective indices based on proprietary Emotiv algorithms, and the difference of valence between AF3 and AF4, which has been found to indicate positive or negative emotion. Algorithm control can either be direct (e.g. a proportional relationship between the raw EEG signals from the audience member and a bank of sinusoidal generator amplitudes) or the result of a relationship between different stream trends (e.g. the triggering of a sampling process at the moments where the AF3-AF4 valences of the performer and spectator start increasing in tandem). The musical algorithms are based on continuous control of low (frequency, amplitude, timbre) or high-level (pitch, tempo, duration) musical parameters, as well as sampling and granular-based generative techniques. For the sake of clarity and according to theatrical design, the number of algorithms running in parallel is limited to two or three for the majority of the script. These are complemented in the final section of the piece by real-time lighting colour effects controlled by the incoming AF3-AF4 valence values.

Throughout two series of local and national performances, there emerged a number of recurring musical/EEG highlights between several instances of the piece. Our analysis of recorded AF3-AF4 valence data revealed peaks coinciding with intense performative moments or marked alternations of performance style. While our measurements are far less accurate than those produced in controlled conditions, they have shown to be qualitatively dependable, and reflective of the deep structure of the piece, as reproduced in a number of performances.