SPEECH IN NOISE WORKSHOP

Abstracts

Toulouse, FR | 9-10 January 2020

The 12th Speech in Noise Workshop is organised by selfless volunteers::

- *Chris James*
- *Julien Pinquier*
- *Lionel Fontan*
- *Christian Füllgrabe*
- *Bernard Fraysse*

Coordinator: Etienne Gaudrain, Lyon Neuroscience Research Center, CNRS, Lyon, France.

Contact information: info@speech-in-noise.eu

The Speech in Noise Workshop is generously supported by:

Talks

Thursday 9 January 2020, 09:15—09:45

Open challenges for driving hearing device processing: lessons learnt from automatic speech recognition

Jon P. Barker[1], Michael A Akeroyd[2], Trevor Cox[3], John Culling[4], Simone Graetzer[3], Graham Naylor[2], Eszter Porter[2]

1. University of Sheffield, United Kingdom | 2. University of Nottingham, United Kingdom | 3. University of Salford, United Kingdom | 4. University of Cardiff, United Kingdom

Recent advances in machine learning raise the prospect of a new generation of hearing devices able to address the speech-in-noise problem. However, the exact path to this goal remains unclear. In contrast, in the field of automatic speech recognition, new machine learning techniques are transforming speech-in-noise performance. The speed of this progress has been enabled, in part, by a research tradition of `open challenges'. This talk explains how such challenges operate to drive speech recognition research and how a similar methodology could benefit hearing device development.

To motivate the challenge methodology, the talk will first present the recent CHiME-5 [1] (and ongoing CHiME-6) speech recognition challenges. These have focused on conversational speech in a dinner party scenario using audio captured using multiple microphone array devices and in-ear binaural microphones. The talk will look at how these challenges have fostered collaboration between research groups specialising in different aspects of the problem, and how they have encouraged system components to be shared leading to further advances. Some components of these ASR solutions, which include de-reverberation, multi-channel signal processing and source separation components, may be directly relevant for hearing device processing, but they remain un-evaluated in a hearing context.

The talk will then present our new project, Clarity [2], which will deliver open challenges specifically designed for hearing aid signal processing and hearing aid speech quality/intelligibility prediction. These tasks are very different from speech recognition, but share common features that motivate a challenge-driven approach. The talk will outline our initial plans, which have also been inspired by other listener-directed challenges such as Blizzard [3], Hurricane [4], Reverb [5] and SISEC [6]. Our plans are at a very early stage and we are actively seeking input and feedback from the speech-in-noise community.

References:
[1] J. Barker, S. Watanabe, E. Vincent, J. Trmal., "The fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines", Interspeech, 2018
[2] M. Akeroyd, J.Barker, T. Cox, J.Culling, S.Graetzer, P.Naylor, E.Porter, www.claritychallenge.org
[3] S. King "Measuring a decade of progress in Text-to-Speech" Loquens, Vol 1, No 1, 2014
[4] M. Cooke, C. Mayo, C. Valentini-Botinhao. "Intelligibility-enhancing speech modifications: the Hurricane Challenge". Interspeech. 2013
[5] K. Kinoshita et al; "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research" EURASIP Journal on Advances in Signal Processing, doi:10.1186/s13634-016-0306-6, 2016
[6] Fabian Robert-Stöter, Antoine Liutkus, Nobutaka Ito. The 2018 Signal Separation Evaluation Campaign. LVA/ICA, Surrey, UK.

Predicting real-life listening abilities in presbyacusic adults: The use and role of auditory supraliminary and cognitive measures in audiological practice

Christian Füllgrabe
Loughborough University, UK

It is well known that many hearing-impaired (HI) people, who seek help from a hearing-aid (HA) audiologist/dispenser, are not satisfied with the speech perception performance achieved with their HAs in everyday life. Assuming that HA fitting is done appropriately, the discrepancy between listening behavior in the laboratory/clinic and that in real-life communication situations must therefore stem from the fact that the tests conducted by the HA audiologist/dispenser are insufficient and/or of poor ecological validity. Three studies were conducted to shed some light on these hypotheses.

First, an online survey of the current practice of speech audiometry in France, where the evaluation of speech intelligibility as part of the audiological assessment is required by law, was designed. Two hundred and ninety-seven HA dispensers completed the questionnaire, providing information about the type and frequency of use of diagnostic and prognostic tests performed during the audiological assessment of adults, including the most common forms of speech audiometry in terms of speech material and background sounds (Rembaud, Fontan, & Füllgrabe, 2017).

Then, two behavioral studies were conducted on 100 older (aged 60-85 years) presbyacousic adults. All had normal cognitive functioning based on MMSE scores and had worn bilateral HAs for at least six months. In addition to the most frequently used tests of speech audiometry (as identified by the online survey), a measure of sensitivity to binaural temporal-fine structure information (i.e., the TFS-AF test; Füllgrabe, Harland, Sęk, & Moore, 2017) and two cognitive tasks, probing respectively working memory

and inhibition, were administered. Real-life listening abilities were assessed using the 15iSSQ, a shorter version of the SSQ, composed of 15 items, covering speech perception, spatial hearing, and qualitative and cognitive aspects of hearing (Moulin, Vergne, Gallego, & Micheyl, 2019).

Performances on the different tasks as well as their relationship with the self-report of everyday listening will be discussed in the presentation.

References:

Füllgrabe, C., Harland, A. J., Sęk, A. P., & Moore, B. C. J. (2017). Development of a method for determining binaural sensitivity to temporal fine structure. International Journal of Audiology, 56(12), 926-935.

Moulin, A., Vergne, J., Gallego, S., & Micheyl, C. (2019). A new Speech, Spatial, and Qualities of hearing scale short-form: Factor, cluster, and comparative analyses. Ear and Hearing, 40(4), 938-950.

Rembaud, F., Fontan, L., & Füllgrabe, C. (2017). L'audiométrie vocale en France: état des lieux [Speech audiometry in France: status quo]. Les Cahiers de l'Audition, 6, 22-25.

Thursday 9 January 2020, 10:45—11:15

Speech perception and listening effort in age-related hearing loss

Stephanie Rosemann, Julia Pauquet, Christiane M. Thiel
Biological Psychology, Carl-von-Ossietzky Universität Oldenburg, Germany

Age-related hearing loss involves the decrease in hearing abilities for the high frequencies and therefore leads to impairments in understanding and processing speech, particularly in difficult listening situations. Growing evidence suggests that the decrease in hearing abilities is associated with an increased listening effort, decreased cognitive functioning as well as alterations in neural activity.

This talk therefore concentrates on functional magnetic resonance imaging studies in elderly hearing-impaired participants considering speech perception under difficult listening conditions. Differences in speech processing between hearing-impaired and normal-hearing participants under different stimulation conditions (auditory-only, visual-only, audio-visual congruent and audio-visual incongruent) as well as varying difficulty levels (high and low cognitive load as well as high and low listening effort) will be demonstrated. Further, the influence of listening effort experienced in daily life and cognitive abilities, for instance cognitive flexibility, will be discussed. Findings support the hypothesis that age-related hearing loss leads to widespread changes in neural activity that are related to the decreased auditory input but also to the increased experienced listening effort.

A brief introduction to multichannel noise reduction with deep neural networks

Romain Serizel
Université de Lorraine, Loria, France

Over the past decade deep learning has become the state-of-the-art in many applications including several tasks of speech and audio processing. It has recently been applied to multichannel speech enhancement, outperforming most of the classical approaches. In this presentation, I will present a short overview of some deep learning architectures that are currently used. I will then describe the problem of multichannel speech enhancement and a solution to this problem: the multichannel Wiener filters. Finally, I will present recent works that capitalize on both multichannel filters resulting from decades of work in signal processing and on the modeling power of deep learning to design deep learning based multichannel speech enhancement algorithms that are now the state-of-the-art in the domain.

Voice pitch perception in cochlear implant users with a spectro-temporally enhanced dual filter-bank sound coding strategy

Damir Kovačić
University of Split, Croatia

Chris James
Cochlear France SAS, Toulouse, France

Despite their great clinical success in partial restoration of hearing and speech understanding in the deaf and hard-of-hearing, current cochlear implants (CI) still do not provide adequate spectral and temporal cues for voice perception. These limitations could particularly affect speech recognition with competing talkers, and may represent one of the key factors for substantial variability in the efficacy of CI.

We present STEP, a novel experimental multi-channel sound coder with enhanced spectro-temporal processing. STEP employs an FFT-based approach with dual filter banks, the first with narrow, good quality filters for spectral processing and the second with a parallel bank of wide filters to reinforce temporal cues. STEP coding was assessed in two experiments investigating voice pitch perception by 16 Nucleus® CI users, where we controlled the modulation characteristics and varied the carrier rate. We used a fixed set of threshold and comfortable stimulation levels for each subject, obtained from clinical MAPs. In the first experiment, we determined equivalence for voice pitch ranking and gender identification between the clinical ACE strategy and STEP for fundamental

frequencies (F0) between 120 Hz and 250 Hz. In the second experiment, loudness as a function of the input amplitude of speech samples, was determined for carrier rates of 1000, 500, and 250 pps per channel. Then, using equally loud sound coder programs, we evaluated the effect of carrier rate on voice pitch perception.

Voice pitch perception was heterogeneous across subjects from no ranking ability to good ranking. Conversely, nearly all subjects could identify voice gender at a level significantly above chance. Overall, carrier rate did not have a substantial effect on voice pitch ranking or gender identification as long as the carrier rate was at least twice the fundamental frequency, or if stimulation pulses for the lowest, 250 pps carrier were aligned to F0 peaks. Also, in the overlap region of male/female F0s, gender categorization was also dependent on speaker gender.

These results indicate that carrier rates as low as 250 pps per channel are enough to support functional voice pitch perception based on temporal cues at least when temporal modulation and pulse timings in the coder output are well controlled. The results also suggest that further investigation of the use of spectral cues for gender identification by CI subjects is warranted. Finally, we will present some data on speech-in-noise performance with STEP.

Funding: Research was supported by IIR Grants #2098 and #853 from Cochlear© to DK.

Thursday 9 January 2020, 13:30—14:30

Keynote lecture

Music in the ear and in the brain

Barbara Tillmann
CNRS, Lyon Neuroscience Research Center, France

In my presentation, I will review two strands of research in the domain of music cognition that can provide further insights for the investigation of auditory perception in hearing-impaired listeners. In the first part, I will review music cognition research using an implicit investigation method that provided evidence for sophisticated musical knowledge in (normal-hearing) nonmusician listeners. I will then present our recent study using this investigation method to further analyze music perception capacities in postlingually deafened cochlear implant users. In the second part, I will focus on temporal processing and predictions within a theoretical framework of temporal attention and dynamic attending. We have shown that musical primes with regular rhythmic structures can improve speech processing in sentences presented after the primes. This benefit has been shown in normal-hearing adults and children as well as in listeners with developmental language disorders (i.e., dyslexia, SLI). In one study, we use reg-

ular rhythmic primes to boost the efficacy of syntax processing training programs in speech therapy sessions for children with cochlear implants. Together, music cognition research can provide perspectives for testing and training music and speech processing in hearing-impaired listeners via implicit methods and rhythmic stimulation programs.

Thursday 9 January 2020, 14:30—15:00

Deep neural networks for predicting human auditory perception

Bernd Meyer
University Oldenburg, Germany

Deep learning resulted in a major boost of speech technology and enabled devices with a relatively high robustness in automatic speech recognition (ASR). For some use cases, the underlying algorithms have become so robust that their degradation in presence of noise is similar to the perception of noisy speech of human listeners. In my talk I will provide examples of models for speech intelligibility, perceived speech quality, and the subjective listening effort derived from deep neural networks that are based on estimates of phoneme probabilities calculated from acoustic observations. In some cases, these algorithms outperform baseline models despite the fact that they operate on a mixture of noise and speech – in contrast to other approaches that often require separate noise and speech inputs. This implies a reduced amount of a priori knowledge for these algorithms, which could be interesting for applying them in the context of hearing research, e.g., for continuous optimization of parameters in future hearing devices. The underlying statistical models were trained with hundreds or thousands hours of speech and are harder to analyze in comparison to many established models; yet they are not black boxes since we have various methods to study their properties, which I will briefly outline in the talk.

Thursday 9 January 2020, 15:00—15:30

Perceptual learning of vocoded speech with and without contralateral hearing: Implications for the rehabilitation of cochlear implant subjects

Olivier Macherey
CNRS, Marseille, France

An increasing number of cochlear implant (CI) candidates show residual acoustic hearing, usually in their contralateral ear. A small proportion of them even has normal or quasi-normal hearing. This patient population may react very differently to CI implantation compared to bilaterally-deaf patients, mainly because they may not rely mostly on their CI to understand speech. From their first activation, they are receiving very dif-

ferent information in each ear, and there is currently no data on the type of training or rehabilitation strategy they should follow: do they need to spend some time listening to their CI alone, without contralateral hearing to maximize their speech intelligibility? Or, on the contrary, does their residual hearing help them train to recognize speech with their CI more quickly?

To get a first insight into this question, we tested 60 normal-hearing listeners in an auditory perceptual learning experiment. Each subject was randomly assigned to one of three groups of 20 referred to as NORMAL, LOWPASS and NOTHING. The experiment consisted of two test phases separated by a training phase. In the test phases, all subjects were tested on recognition of monosyllabic words passed through a six-channel "PSHC" vocoder presented to a single ear. In the training phase, all subjects were also presented with the same vocoded speech in one ear but the signal they received in their other ear differed across groups. The NORMAL group was presented with the unprocessed speech signal; the LOWPASS group with a low-pass filtered version of the speech signal (filter cut-off of 250 Hz) and the NOTHING group with no sound at all. These three groups aim to simulate groups of CI subjects having normal contralateral hearing, residual low-frequency hearing and no residual hearing, respectively. The training phase consisted of listening to a 30-minute audio book with subtitles displayed on a computer screen.

A mixed-effect ANOVA showed a significant effect of training. All subject groups performed better after than before the training phase. Furthermore, there was a significant interaction between training and group. Further analysis showed that the amount of improvement was significantly smaller for the NORMAL than for the LOWPASS and NOTHING groups.

This shows that having normal contralateral hearing reduces or slows down perceptual learning of vocoded speech but that having instead an unintelligible contralateral signal does not have any effect. Potential implications for the rehabilitation of CI patients with partial or full contralateral hearing will be discussed.

Binaural integration for speech in noise and sound localization: impact of brain plasticity following unilateral hearing loss

Pascal Barone, Nicolas Vannson
CNRS CerCo, Toulouse, France

Kuzma Strelnikov
CHU Purpan, Toulouse, France

Olivier Deguine, Mathieu Marx
Service Oto-Rhino-Laryngologie et Oto-Neurologie, Hôpital Purpan, Toulouse, France

In patients with unilateral hearing loss (UHLp), binaural processing is obviously disrupted and spatial localization of the sound source is impaired as well as the ability in understanding speech in noisy environments. At the brain level, a limited number of studies have explored the functional reorganisation that occurs in the adult after a unilateral deafness. We conducted an original study aimed at investigating in UHLp the relationships between the severity of unilateral hearing loss, the resulting deficit in binaural processing and the extent of cortical reorganisation across the auditory areas.

We have recruited 14 UHL patients (hearing loss 37-120 dB HL) and aged-matched hearing controls. All subjects were evaluated for free-field sound localization abilities and speech in noise comprehension (French Matrix test). All subjects went through a fMRI protocol to evaluate the activation pattern across auditory areas during a natural sounds discrimination task. First, a the neuronal level in the auditory cortex we observed a lack of suppression/occlusion mechanisms that characterize binaural integration. Second, the brain imaging analysis clearly demonstrated that in non-primary areas (NPAC), UHL induces a shift toward an ipsilateral aural dominance. Such reorganization, absent in the PAC, is correlated to the hearing loss severity and to lower spatial localization ability performances. Second, a regression analysis between brain activity and patient's performances, clearly demonstrated a link between the spatial ability deficit and a functional alteration that impacts specifically the posterior auditory areas known to process spatial hearing. On the contrary, the core of the auditory cortex appeared relatively preserved and maintains its normal implication in processing non-spatial acoustical information.

Altogether our study adds further evidences of a functional dissociation in the auditory system and shows that binaural deficits induced by UHL affect predominantly the dorsal auditory stream.

Towards microscopic intelligibility modelling

Ricard Marxer
Université de Toulon, Aix Marseille Univ, CNRS, LIS, Marseille, France

Existing intelligibility models successfully estimate word recognition rates in broadly stated noise conditions. These predictions may be characterized as macroscopic since they represent aggregates, averages over many listeners and many stimuli. Many macroscopic intelligibility models have been proposed in the past, mainly in the context of communication channel assessment. Articulation index (AI), speech-transcription index (STI) and short-time objective intelligibility (STOI) are all predictors of the intelligibility of distorted speech signals. We hypothesize that by employing data-driven techniques we can predict individual listener behaviors at a sub-lexical level. These microscopic models should be capable of making precise predictions of what a specific listener might hear in response to a specific speech signal. Furthermore we expect the development of such models to provide insights or validate our understanding of speech-in-noise perception.

Microscopic approaches to intelligibility prediction have started receiving some attention in recent years due to their potential in facilitating hearing research. We present work preparing the terrain for the development of such microscopic intelligibility models. This overview covers studies ranging from data acquisition and definition of evaluation metrics, to the analysis of collected experimental responses and intelligibility-related ASR performance.

At first, we limit our field of study to single word recognition. We review the proposal of tasks and methods to evaluate microscopic intelligibility models in such a setup. We present a corpus of noise-induced British English speech misperceptions, mimicking the existing consistent confusion corpus in Spanish. We then discuss the language factors that influence the elicitation of such confusions, and focus on the linguistic effects such as word usage frequency, or acoustic factors such as the type of masking noise. On the modeling side we start by showing how intelligibility measures such as STOI can be a better predictor of word-error rates of ASR systems than SNR. Finally we review recent existing work on using ASR-style modelling to predict fine-grained speech perception responses.

Colin Cherry Award 2019

Efficacy of audiological rehabilitation: a randomized controlled trial

Sara Magits[1], Linus Demeyere[1], Ellen Boon[2], Ann Dierckx[2], Nicolas Verhaert[2], Tom Francart[1], Jan Wouters[1], Astrid van Wieringen[1]

1. Department of Neurosciences, Research Group Experimental ORL, KU Leuven, Leuven, Belgium | 2. Department of Otorhinolaryngology, Head and Neck Surgery, University Hospitals Leuven, Leuven, Belgium

Hearing impairment (HI) presents a burden on the daily life of people as it is shown to be associated with considerably higher risk of social isolation, depression and declining cognitive functions. In addition, the prevalence of HI increases with advancing age, affecting nearly one-third of persons above 65 years. If HI remains untreated, the negative consequences can severely affect long-term health and quality of life. Auditory rehabilitation (AR) to improve auditory skills and prevent or decrease participation restrictions could alleviate these negative consequences. Some persons with HI are entitled to AR in the clinic, but many people do not benefit from AR. Moreover, the efficacy of AR remains debated, at least partly driven by an important need for high-quality evidence.

In this study, our aim is to assess the efficacy of AR for middle-aged (45-64y) and older adults (65-75y) with HI in a randomized controlled trial. Therefore, we have developed the LUISTER AR scheme and implemented it as an application installed on a tablet. It consists of multiple assessment tests and auditory-cognitive training tasks. The primary goal of the LUISTER AR scheme is to improve speech perception abilities in noise.

Currently, the LUISTER AR application is being evaluated in comparison to a placebo AR application (active control group) in a randomized controlled trial with middle-aged (n=20) and elderly (n=14) cochlear implant users. Data logging of the different parameters of the application allows for in-depth analysis of the efficacy of the LUISTER AR scheme.

We will present the development and implementation of the LUISTER AR scheme. Additionally, baseline results of the randomized controlled trial, primarily speech perception in noise and cognitive functioning, will be discussed for the persons with HI and their normal hearing peers. In conclusion, preliminary results from the LUISTER AR scheme evaluation, such as on-task improvement and data logging of the intensity and frequency of training, will be presented.

Relating speech perception in noise to temporal-processing auditory capacities during childhood

Laurianne Cabrera
CNRS, Paris, France

Temporal cues (e.g., amplitude modulation, AM) play a crucial role in speech intelligibility for adults. How the ability to track these temporal cues develops and interacts with the development of speech perception is however unclear. Although aspects of AM processing appear to be mature as early as 3 months of age, children' ability to detect AM continues to improve until 10 years of age. The present study explored whether the development of AM processing is related to sensory development or to changes in processing efficiency and how this ability relates to speech intelligibility in noise during childhood.

Eighty-three children with normal-hearing from 5 to 11 years and 22 young adults completed three 3IAFC adaptive tasks. The first psychophysical task assessed AM sensitivity using pure tone carriers and three modulation rates (4, 8, 32 Hz). A second task assessed susceptibility to AM masking by comparing AM detection thresholds at the same modulation rates using three carriers varying in their inherent AM fluctuations: tones, narrowband noises with small inherent AM fluctuations and noises with larger fluctuations. Finally, a third XAB task was designed to measure consonant identification thresholds in speech-shaped noise using fricative consonants contrasting on either place of articulation or voicing.

Results showed that between 5 and 11 years, AM detection thresholds improved and that susceptibility to AM masking slightly increased. However, the effects of AM rate and carrier were not associated with age, suggesting that sensory factors (tuning of AM filters, susceptibility to AM masking) are mature by 5 years. Increased AM masking with age result from worse thresholds with tone carriers at 5 years as if tone carriers were "noisier".

These changes in AM sensitivity and masking during childhood may reflect a more efficient use of AM cues with age, due to a reduction in internal noise and/or optimization of decision strategies. Subsequent computational modelling indicated that a reduction in internal noise by a factor 10 better accounted for these developmental trends.

The development of temporal processing during childhood may be in terms of changes in processing efficiency. Finally, children's consonant identification thresholds in noise decreased with age and were somewhat related to AM sensitivity. Thus, increased efficiency in AM detection may support better use of temporal information in speech during childhood.

Using automatic speech recognition to improve hearing-aid fitting

Lionel Fontan
Archean Labs, Montauban, France

Because people with age-related hearing loss (ARHL) generally experience difficulties in understanding speech, tests of speech identification are often used by audiologists and hearing-aid (HA) dispensers to evaluate the benefits of rehabilitation through hearing-aids.

However, these tests are fairly time-consuming, which can lead to an increase in fatigue (and thus potentially to a decrease in performance) for the older listeners. Moreover, the listeners' speech-identification performances are likely to be influenced by their familiarity with the speech materials, which, ideally, should be refreshed for every test condition. These issues make it impossible to test all the HA settings that might yield optimal speech intelligibility for the listener.

Automatic speech recognition (ASR) systems could overcome these shortcomings, by providing fast and objective measures of speech intelligibility, provided that the perceptual consequences of ARHL can be accurately simulated by signal-processing algorithms.

In this presentation, we report on a series of proof-of-concept experiments that compared human speech-identification performances with the performance of an ASR system fed with speech signals simulating three of the perceptual consequences of ARHL (Nejime & Moore, 1997): elevation of hearing thresholds, loss of frequency selectivity, and loudness recruitment. For both young, normal-hearing listeners (Fontan et al., 2017), and older, hearing-impaired listeners (Fontan, Cretin-Maitenaz, & Füllgrabe, in revision), strong correlations between human and ASR performance were observed, indicating that trends in speech intelligibility can be predicted.

The system was later used in combination with a HA simulator (Moore, Füllgrabe & Stone, 2010) to find optimal HA amplification gains (in terms of predicted intelligibility) for 24 older listeners with ARHL. Participants' aided speech-intelligibility scores and subjective judgements of speech pleasantness were found to be significantly higher when applying the ASR-based amplification gains than when applying a baseline fitting rule (Moore, Glasberg, & Stone, 2010).

References
Fontan, Cretin-Maitenaz, & Füllgrabe. (In revision). Predicting speech perception in older listeners with sensorineural hearing loss using automatic speech recognition. Trends in Hearing.
Fontan, Ferrané, Farinas, Pinquier, Magnen, Tardieu, Gaillard, Aumont, & Füllgrabe. (2017). Automatic speech recognition predicts speech intelligibility and comprehension for listeners with simulated age-related hearing loss. JSLHR, 60(9), 2394–2405.
Moore, Füllgrabe, & Stone. (2010) Effect of spatial separation, extended bandwidth, and compression speed on intelligibility in a competing-speech task. JASA, 128(1), 360–371.
Moore, Glasberg, & Stone. (2010). Development of a new method for deriving initial fittings for hearing aids with multi-channel compression: CAMEQ2-HF. IJA, 49(3), 216–227.
Nejime, & Moore. (1997). Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise. JASA, 102(1), 603–615.

Exploring listeners' speech modification preferences

Olympia Simantiraki
University of the Basque Country, Spain

Martin Cooke
Ikerbasque (Basque Science Foundation), Spain

Listening to synthetic or artificially-produced speech under adverse conditions is an everyday phenomenon. Many algorithms have been proposed for augmenting the speech signal before reaching the listener. Near-end listening enhancement algorithms can achieve significant improvements in speech understanding compared to unprocessed speech in adverse conditions (Taal & Jensen, 2013; Schepker et al, 2015). Other factors such as listening effort and naturalness are also important when intelligibility is close to ceiling. One means to explore these supra-intelligibility factors is through listener preferences. Earlier studies have measured listener preferences via subjective scales (Moore et al, 2007; Adams & Moore, 2009) or by allowing listeners to modify speech properties in real-time (Wingfield & Ducharme, 1999; Zhang & Shen, 2019; Simantiraki & Cooke,2019).

Using the virtual adjustment tool proposed in Simantiraki & Cooke, 2019, we conducted several experiments to explore the effects of speech properties on listening preferences and intelligibility. Participants were permitted to change a speech feature during an open-ended adjustment phase, followed by a test phase in which they identified speech presented with the feature value selected at the end of the adjustment phase. This technique generates information about the feature value that listeners subjectively feel allows comfortable speech recognition performance as well as the actual intelligibility, and the time required to make the adjustment.

Experiments with native normal-hearing listeners measured the consequences of allowing listeners to change spectral slope, the location of a spectral band of speech, speech rate and mean F0. Speech stimuli were presented in both quiet and masked conditions. As the noise level increased, compared to the original values, listeners (i)

chose increasingly flatter spectral tilts; (ii) moved spectral bands to higher frequencies and (iii) preferred slower speech rates. However, the mean of F0 was unaffected by noise and was always lower than the original value. These outcomes are largely consistent with earlier findings of the effect of corresponding modifications on intelligibility, but provide additional information in cases where intelligibility is at ceiling levels.

References:
Taal and Jensen (2013) Interspeech 3582–3586
Schepker et al (2015) J. Acoust. Soc. Am. 138, 2692–2706
Moore et al (2007) Int. J. Audiol. 46, 154–160
Adams and Moore (2009) J. Am. Acad. Audiol. 20, 28–39
Wingfield and Ducharme (1999) J. Gerontol. B Psychol. Sci. Soc. Sci. 54B, P199–P202
Zhang and Shen (2019) Interspeech 1383–1387
Simantiraki and Cooke (2019) ICA 5736–5738

Posters

01 Sound-in-noise recognition: An international study on a language-independent school-entry hearing screening test

Elien Van den Borre, Sam Denys
ExpORL, KU Leuven, Belgium

Lea Zupan
General Hospital Celje, Slovenia

Wouter Dreschler
UMC, Amsterdam, Netherlands

Jan de laat
UMC, Leiden, Netherlands

Astrid Van Wieringen, Jan Wouters
ExpORL, KU Leuven, Belgium

Hearing loss is one of the most common congenital impairments, occurring in 1 to 3 per 1000 newborns. The incidence of acquired hearing loss at later age is not exactly known. Therefore, in 2012, the European Federation of Audiology Societies formally recommended the implementation of preschool and school-age hearing screening, in addition to the systematic hearing screening of newborns to detect those children and prevent secondary impairments, such as language disorders. With systematic screening and data registration, results of school-age and newborn hearing screening databases could be linked, allowing accurate quantification of the incidence of acquired childhood hearing loss. When using the same reference test across European countries, between-country comparisons will be possible as well.

At this moment, research is being conducted to develop a language independent sound-in-noise test, the Sound Ear Check (SEC). This is an automated adaptive self-test on tablet based on recognition of masked ecological sounds. The SEC has already been evaluated in adults, and shows promising results. A reference curve with a steep slope of 18%/dB was obtained, resulting in a test with a high measurement precision of 1 dB. Significant correlations with both pure tone thresholds ($r = 0.70$) and the Digit Triplet Test ($r = 0.79$) speech-in-noise test were found in adults. Sensitivity and specificity values of about 80% were obtained.

The current follow-up study aims to investigate the feasibility of the test in young school-age children, at the age of school-entry (5-6 years) and to investigate the test's reliability as well as its sensitivity and specificity for both conductive and sensorineural hearing loss. In collaboration with study partners from different European countries, the language- and culture-independency was estimated as well.

Unfortunately this author could not make it to present their poster. We have left the abstract for your information.

02 Using a spatial speech-in-noise test to assess advanced hearing-aid features

Bhavisha Parmar, Jennifer K. Bizley
University College London, UK

Debi Vickers
University of Cambridge, UK

The Spatial Speech Test was developed to simultaneously assess relative localisation and word identification in the presence of multi talker babble.

During the task, listeners hear two sequentially presented words from adjacent speakers with a 30° separation. Participants were instructed to select the two words that they heard in the correct order and to report the direction of the location shift between the first word and the second word. The task is performed in the presence of multitalker babble and at an individually determined signal-to-noise ratio.

This assessment method has been piloted on normal hearing listeners, cochlear implantees and hearing aid users. Current findings have shown that the spatial location of the words have a significant effect on relative localisation performance for both normal-hearing and hearing-aid users. Hearing-aid user's relative localisation performance was significantly worse than normal-hearing listeners. Hearing aid user's had significantly poorer spatial unmasking compared to NH listeners. Most recently, the adapted Spatial Speech Test was used to assess the effects of Oticon's OpenSound Navigator feature on relative localisation and word identification performance in the presence of multi-talker babble.

03 Effects of asymmetric envelope compression on speech intelligibility and binaural unmasking

Emily Burg[1], Tanvi Thakkar[1], Sean Anderson[1], Matthew Winn[2], Ruth Litovsky[1]
1. University of Wisconsin-Madison, USA | 2. University of Minnesota, USA

The temporal envelope of a speech signal plays an instrumental role in speech understanding and perceptual segregation of speech and noise. Previous work has shown that compressing the envelope of vocoded speech (effectively reducing the dynamic range) can decrease speech understanding, but it is unknown whether dynamic range affects binaural integration and auditory source segregation. In order to investigate this, we examined the effect of reduced dynamic range on binaural unmasking in normal-hearing (NH) individuals. Specifically, we were interested in the degree of similarity in dynamic range across ears, because asymmetries in dynamic range are common in individuals with cochlear implants (CIs) and may contribute to their limited access to binaural benefits. We chose a binaural unmasking paradigm to assess binaural processing while avoiding effects that stem from simply attending to whichever ear has the better signal-to-noise ratio. The temporal envelope of the signal was compressed independently in one or both ears to simulate variations in dynamic range. Target stimuli consisted of female-talker IEEE sentences and masker stimuli consisted of female-talker AzBio sentences, all in English. Stimuli were processed using a 16-channel vocoder and presented over headphones at 65 dBA. In one condition, target speech was presented in quiet. In a second condition, target and masker speech were presented to one ear. In a third condition, a copy of the masker was also presented to the contralateral ear. When compared to performance in condition two, this third condition has been shown to elicit binaural unmasking in NH individuals, with improvements in speech reception thresholds ranging from 3-5 dB. We predicted that speech intelligibility in quiet and in noise would decrease monotonically with decreasing dynamic range. For bilateral conditions, we predicted that reducing dynamic range symmetrically in both ears would result in more unmasking than reducing it asymmetrically in one ear alone, because binaural integration requires similar representation and fidelity of signals across ears. Preliminary data show that speech intelligibility decreased with reduced dynamic range. Additionally, symmetrical reductions in dynamic range elicited more unmasking than asymmetrical reductions. This indicates that unmasking is not simply limited by the ear with the smallest dynamic range, and that asymmetries in dynamic range may be one factor contributing to the poor binaural processing demonstrated by individuals with bilateral CIs.

04 The effects of background noise, noise reduction and task difficulty on recall

Andreea Micula[1,2], Elaine H. N. Ng[2], Fares El-Azm[2], Jerker Rönnberg[1]
1. Linköping University, Linköping, Sweden | 2. Oticon A/S, Copenhagen, Denmark

The Sentence-final Word Identification and Recall (SWIR) test was designed to investigate the effect of hearing-aid signal processing on memory for highly intelligible speech in noise. Previous findings suggest that people with high working memory capacity benefited more from advanced hearing-aid signal processing than people with low working memory capacity. However, people with low working memory capacity showed benefit when a less demanding version of the test was used. Thus, excessive task difficulty could have prevented capturing potential benefit from signal processing.

The aim of this study was to continue developing the SWIR test by manipulating task difficulty, as well as to investigate the effect of task difficulty predictability on recall. Moreover, the effects of noise and noise reduction on recall were investigated. Lastly, the correlation between working memory capacity and recall performance was analyzed for each task difficulty level.

Thirty-two experienced hearing-aid users with symmetrical moderate sensorineural hearing loss participated in this study. The SWIR test task consists of listening to lists of sentences and recalling the last word of each sentence after the list is finished. The SWIR test was administered with noise reduction on and off in competing speech and speech-shaped noise. The task difficulty was manipulated by varying the list length (three, five, seven and nine sentences per list). Half of the participants were informed about list length in advance (predictable task difficulty), while the other half were not (unpredictable task difficulty). Working memory capacity was measured using the Reading Span test.

The results revealed that recall performance was improved when noise reduction was on. However, this improvement was only significant in competing speech when task difficulty was unpredictable. Analysis of the probability of first recall suggested that there was a higher tendency to begin recall with the first list item when noise reduction was off when task difficulty was unpredictable in competing speech. When noise reduction was on, recall tended to start with the last list item. Thus, this finding may be attributed to potential effects of noise reduction on recall strategies. The results also showed a significant positive correlation between working memory capacity and recall performance on lists of five, seven and nine sentences. This finding suggests that the procedure of the test could be modified to be adaptive to individual cognitive capacity.

05 A new set of superimposed speech features to predict a priori the performance of automatic speech recognition systems

Sébastien Ferreira[1,2], Jérôme Farinas[1], Julien Pinquier[1], Julie Mauclair[1], Stéphane Rabant[2]
1. IRIT, University of Toulouse, France | 2. Authôt, Ivry-sur-Seine, France

The aim of this exploratory work is to predict, a priori, the quality of the automatic transcription in the case of speech mixed with music. In order to make this prediction, we need to quantify the impact of music (considered as noise in our study) on speech, before decoding by an Automatic Speech Recognition (ASR) system. Generally, the estimate of noise level in a speech signal exploits the bimodality of the noisy speech distribution. When the studied noise is music, the distribution has more than two modes, which makes noise level estimation unviable.

We propose a new set of features.

Entropy modulation (Pinquier et al., 2002, ICSLP) detects how much a signal is considered speech by measuring the lack of order of the signal: the music signal has a more orderly structure than speech. A voiced/unvoiced signal duration ratio is computed to measure f0 detection anomalies due to background music. The dip test (Hartigan et al., 1985, The Annals of Statistics, vol. 13, p. 70-84) measures the unimodality of f0 distribution. When music is mixed with speech f0 distribution becomes less unimodal. The excitation behavior of linear prediction residuals (Ferreira et al., 2019, SPIN), which is originally a reverberation measure quantifies the “strength of voicing” of a voiced signal: for the case of reverb, the superposition between phonemes modifies this value.

The experiment was conducted on the Wall Street Journal (WSJ) corpus. Six pieces of music of different styles from RFM directory of the MUSAN corpus (Snyder et al., 2015, arXiv:1510.08484) were mixed with WSJ at three levels of the signal-to-noise ratio (5, 10 and 15 dB). The mean absolute error (MAE) of Word Error Rate (WER) prediction obtains 9.17 at the utterance level. The main goal is to inform users as soon as possible about the quality of the automatic transcription of their audio documents. In most cases, records submitted by transcription services users exceed 3 min. When 20 utterances are used (around 140s), MAE of WER prediction achieves 3.82. This experiment indicated that our set of features was probably well correlated with the recognition error of the ASR system, in this case of speech mixed with music.

06 Listening effort in young children with cochlear implants

Amanda Saksida, Sara Ghiselli, Enrico Muzzi, Eva Orzan
Institute for Maternal and Child Health, IRCCS "Burlo Garofolo", Trieste, Italy

In humans, effortful listening evokes stress and fatigue. In constantly effortful environments (such as classrooms), well-being, learning abilities and academic success are compromised. Such listening environments are much more challenging for severely/profoundly deaf children with cochlear implants. With technological advancement and early implantation, listening skills are thought to have significantly improved. However, it is sometimes difficult to measure listening skills and effort, especially in very young children. Behavioral measures of effort and fatigue cannot be obtained in young/uncooperative children, whereas objective measures are either difficult to assess (cf. EEG) or not reliably correlating with objective behavioral measures (cf. cortisol levels). Recent studies have thus explored pupillometry as a possible objective measure. Available results indicate that pupil dilates in response to effort, which corresponds to pupillary reaction to attention or cognitive load. No study has been, however, done on infants' or preschool children's response to various signals and noise levels. The aim of this study is to explore pupillary response to signal in noise in young bilaterally implanted children with congenital hearing loss.

Present study explores pupillary behavior of 14 children (1.5-4-y-old) listening to speech (rhymed verses in their mother tongue, Italian) and music (excerpts from the "Happy song") at a constant intensity level (60 dB spl) and at various noise levels (babble noise), in an ecological environment (Ambisonics semi-sphere), with one or both cochlear implants switched on. During the session, children watched a repeating excerpt from an animated film.

Preliminary results show that in response to noise, pupil dilates more in noisy conditions (SNR 0 dB) compared to low background noise (SNR 10 dB), but not compared to silence. Possibly, this result reflects the fact that listening in low background noise is the most common everyday experience for these children. Pupil also dilates more in presence of music and speech compared to silence (with low or without background noise), opening the possibility to use pupillometry as a potential measure of auditory signal detection. The overall pupil dilation changes in different cochlear implant configurations (bilateral, left only, right only), increasing significantly when only left implant is turned on. This result may be affected by the sequence of implantation (left implant followed the right one in 13 out of 14 children), or, alternatively, by the general right ear superiority observed in humans. Further research is, however, needed to better understand these results.

07 Fully convolutional Wasserstein autoencoder for speech enhancement

Clement Laroche, Rasmus Olsson
Jabra (GN audio), Ballerup, Denmark

Speech enhancement methods traditionally operate in the time-frequency domain and/or exploit some higher-level features such as Mel-Frequency Cepstral Coefficients. By default, these approaches are discarding the phase and are only partially using the input data. The signal reconstruction necessitates to estimate the phase which is a difficult problem on its own. To overcome this limitation, new speech enhancement techniques operate in the time domain. These data-based approaches based on deep learning directly map the raw input signal waveform to the enhanced speech signal. The success of the Wavenet in speech synthesis promptly motivated the use of generative models for speech enhancement. Similarly, Generative Adversarial Networks (GAN) have shown good performance in denoising by effectively suppressing additive noise in raw waveform speech signals.

In order to benefit from the powerful acoustic modeling capabilities of the recent generative approaches, we propose to use a Wasserstein Auto-Encoder (WAE) for building a generative model of the clean speech. More specifically, this model uses a fully convolutional encoder and decoder architecture with skip connections. The generative model we propose do not suffer from instability during the training phase.

We apply our network to the problem of denoising to remove the additive noise from the target signals. Our model is trained on pairs of noisy and clean audio examples and at test-time, it predicts clean sample from a noisy signal. Our approaches outperformed other speech enhancement approaches and it demonstrates the effectiveness of convolutional autoencoder architectures on an audio generation task.

08 Development and testing of a simulated gaze-directed beamformer

John Culling
Cardiff University

Patrick Naylor, Emilie D'Olne
Imperial College London

Head-mounted multi-microphone beamforming systems offer opportunities to improve signal-to-noise ratio in complex listening environments for hearing-impaired listeners. However, conventional microphone arrays are bulky and must be directed by unnaturally large head movements. The advent of MEMS (micro-electrical mechanical systems) microphones and small discreet eye trackers has led to renewed interest in

beamforming systems. MEMS microphones are small enough to be discreetly mounted in spectacle frames, and similarly miniaturised eye-tracking systems could be used to steer the beam, reducing the required head movement. We created a prototype 8-microphone beamforming array on the frames of a pair of glasses and mounted it on an acoustic manikin. Head-related impulse responses were measured for each microphone for 48 source directions on the horizontal plane. These impulse responses were used to calculate the MVDR (minimum-variance distortionless response) beams that could be created for different source directions. The beam specifications can be used to predict the benefit of beamforming and to simulate beamforming in experiments. Speech-importance-weighted beams were used to predict the effective beam depth and width when listening to speech, showing that off-beam sound sources will be effectively attenuated by about 6 dB at 30 degrees and 8-10 dB beyond 60 degrees. Digital filters were designed that represented the frequency response of the beamforming system for each direction. The filters will be used in two ways in order to evaluate the potential benefits of such a system. First, the filters will be used to simulate specific fixed listening situations to corroborate the predicted improvements in speech reception and compare them with binaural listening. Second, a Simulink model has been developed that can dynamically select filters from a look-up table and filter multiple sources in real-time. This model can be driven by input from an eye tracker in order to simulate a complete gaze-directed system in use and evaluate its usability in realistic listening scenarios.

09 How consistently do speakers apply the Lombard speech clarification effect over time?

Chen Shen, Esther Janse
Radboud University, Nijmegen, Netherlands

People with hearing impairment may experience that their interlocutors are willing to clarify their speaking style upon request, but then quickly return to their habitual speech behaviour as the conversation continues. Whereas the acoustic-phonetic changes that come with Lombard speech (i.e., speech produced in noise) and clear speech have been described extensively, it is unclear how consistent speakers' acoustic-phonetic modifications are over a period of time. We investigated the acoustic-phonetic differences between speakers' habitual speaking style and their speaking style in a condition where they were presented with loud noise and were also instructed to speak clearly. Our research question was whether acoustic differences in articulation rate, pitch median, pitch range, and spectral tilt between habitual and Lombard/clear speaking style would change over the course of a sentence list. Conceivably, speakers could get tired of raising their voice and of their careful speaking style, such that the differences between speaking styles would decrease over trials. Alternatively, speakers may need some practice to show the full Lombard/clear speech effect.

Seventy-eight participants read out 48 sentences in both their habitual speaking style, and in a condition where they were instructed to speak clearly while hearing loud speech-shaped noise over headphones (Lombard/clear style). The list of 48 sentences was randomised four times to create four lists assigned to different participants, such that trial effects could be isolated from sentence (i.e., item) effects.

Results from linear mixed-effects models indicate that trial main effects were present in three of the four acoustic measures (i.e., for articulation rate, pitch median, and spectral tilt). Across all four acoustic measures, sentence trial interacted with speaking style. More specifically, acoustic differences between habitual and Lombard speech increased over trials, which was sometimes due to speakers becoming 'sloppy' in their habitual style over trials, e.g., faster articulation rate and smaller pitch range towards the end of the list. However, speakers also enhanced some of their Lombard style modifications over Lombard trials, e.g., higher pitch median and flatter spectral tilt at later trials. Thus, despite the higher vocal effort required to produce Lombard speech, speakers in our study were able to not only maintain but even enhance their Lombard speech modifications over trials. Research with actual conversation is needed to investigate to what extent our observations generalise to more demanding speaking tasks, and to see how clear/Lombard speaking style may change over the course of a conversation.

10 Using sinewave speech to investigate the locus of informational interference during speech-in-noise perception

Sarah Knight, Sven Mattys
University of York, UK

Speech-in-noise research typically distinguishes between energetic masking (EM: interference between target and masker at the periphery) and informational masking (IM: interference higher in the auditory pathway). IM can itself be broken down into low-level and high-level IM. We use the term "informational interference" (inf-int) to refer to high-level IM involving linguistic and cognitive factors, and which is influenced by long-term knowledge (e.g. familiarity with the language spoken by competing talkers).

Unlike EM, inf-int is poorly understood, in part because it is extremely difficult to manipulate inf-int without altering EM or lower-level IM. The current study aims to isolate one way in which inf-int may arise: awareness of a masker being speech as opposed to non-speech. This is achieved by taking advantage of the perceptual properties of sinewave speech (SWS). SWS is produced by extracting the first few speech formants of a natural utterance and replacing them with time-varying sinusoids reproducing their frequency and amplitude variations. SWS is not usually initially perceived as speech, but through training can be made at least partially intelligible.

Young normal-hearing listeners (N = 54) completed 3 tasks (pre-exposure, exposure, post-exposure). The pre- and post-exposure tasks were speech-in-noise tasks, with target sentences presented in a masker of SWS and amplitude-modulated white noise. During the intervening exposure phase, one listener group was trained to understand similar SWS/white noise stimuli. A second group listened to the same SWS stimuli, but were told that they were hearing randomly-generated machine noise, and performed a simple same/different task. As a result, all listeners heard the same stimuli (thus controlling for EM and lower-level IM), but only listeners in the trained group were aware of linguistic content in the masker, and only post-exposure. By comparing pre- and post-exposure performance across the two groups, it is therefore possible to isolate any effects on speech-in-noise perception arising specifically from an awareness of the masker being speech. Results will be presented and interpreted in light of the literature on informational masking.

11 Identifying overlapping vowel-consonants following hearing loss: Machine learning of neural representations

Samuel S. Smith[1], Mark N. Wallace[1], Joel I. Berger[2], Michael A. Akeroyd[1], Christian J. Sumner[3]
1. University of Nottingham, UK | 2. University of Iowa, USA | 3. Nottingham Trent University, UK

A moderate hearing loss can pose major challenges for speech identification, principally in noisy environments. This is despite most features of speech remaining audible. It is not yet clear how the neural representation of speech changes following hearing loss. Here, in order to quantify this, we present a framework that integrates neural data gathered from animals with either normal hearing or a noise induced hearing loss with a probabilistic classifier.

Neural responses to the presentation of a target vowel-consonant (VC) overlapped by a distractor VC, with distractor lags between -262.5 ms and +262.5 ms, were recorded from the inferior colliculus of anaesthetised guinea pigs. Animals either had normal hearing (NH) or were exposed to high intensity sound (8-10 kHz at 115 dB SPL for 1 hour) imposing a moderate hearing loss (HL) at frequencies above 4 kHz. A machine learning classifier (naïve Bayes) was implemented to predict auditory perception. The classifier was both trained and tested on neural data recorded from either NH or HL animals.

The classifier, trained on NH responses to VCs in quiet, was set to identify VCs in quiet with an accuracy of 95%. In line with human behaviour, the classifier's identification of VCs in quiet did not much reduce for the HL data (93%). Crucially, and again in line with human behaviour, the classifier's identification of overlapping VCs was significantly worse (up to 20%) for neural representations from animals with HL in comparison to NH. It was determined that these findings were not solely attributable to a simple loss of auditory information in the frequencies most affected by hearing loss.

However, it is likely that people listening to speech in a noisy background are able to utilise prior knowledge of interfering sounds. When the classifier employed knowledge of distracting VCs (i.e. the classifier was trained on representations of overlapping VCs), not only was performance improved in all cases, but the difference between NH and HL was greatly diminished. One possible interpretation of this is that prior information about interfering sounds can reduce the impact of poorer neural coding following HL.

Overall, this work offers evidence for a degraded representation of speech in complex acoustic backgrounds, at the midbrain level, following hearing loss. Applying a machine learning classifier to the neural representation of speech sounds appears to be a promising method for understanding real-world problems associated with hearing loss.

12 Masked speech perception: the effect of age and language background

Linda Taschenberger, Outi Tuomainen, Ryan J. Oakeson, Valerie Hazan
University College London, United Kingdom

Most communication in everyday life takes place in less than ideal acoustic conditions. The presence of background noise or competing voices affects some populations more adversely than others. For example, Goossens et al. (2017, Hear. Res. 344:109) found that, although normally hearing older adults' self-reported hearing abilities do not differ from younger and middle aged adults', their speech reception thresholds (SRT) are poorer. The researchers found age (rather than hearing impairment) to explain a significant part of this decline in performance. Here, we expand their study to younger and non-native listeners.

To measure self-assessed speech intelligibility in everyday situations, the Speech, Spatial and Qualities of Hearing (SSQ; Noble et al., (2013), Int. J. Audiol. 52(6):409) questionnaire was used. To measure SRTs, we ran an adaptive coordinate response measure task [CRM; adapted from Bolia et al. (2000), JASA 107(2):1065] for simple sentences (e.g. "show the dog where the red six is") presented in (a) a speech-shaped noise masker (SSN), and (b) a 3-talker-babble masker (BABB). Participants were (i) native English speakers (8-85 years, N=114, 60 female) divided into six age groups [young children (CH-Y), older children (CH-O), young adults (YA), middle aged adults (MA), younger older adults (OA-Y) and older adults (OA-O)], and (ii) non-native speakers of English (N=19, 18 female, mean age 24.89). All had normal audiometric thresholds up to at least 4 kHz or self-reported normal hearing (for L2 speakers) and anyone over 65 was screened for cognitive impairment.

The results showed similar self-assessed speech intelligibility scores in the SSQ for all L1 groups. On the contrary, L2 listeners significantly differed from the L1 groups, reporting lower abilities in everyday speech understanding ($p<.05$). In the CRM, the effect of masker type was significant ($p<.001$) with better SRTs for SSN (M=-5.32 dB) than for BABB (M=-

1.68 dB). The interaction between noise type and group was also significant ($p<0.042$): for BABB, but not for SSN, better thresholds were obtained for YAs compared to all other age groups apart from MAs ($p<.05$). Conversely, L2 listeners differed from all L1 groups in SSN (all $p<.001$), and from YAs and MAs in BABB ($p<.010$).

In line with Goossens et al. (2017), we found that when background noise is more cognitively demanding, there is a larger decline in speech perception in OAs. Additionally, we also found children (8-16) to show reduced speech perception abilities in this masker type. Overall, non-native speakers' performance was most affected by both types of masker.

13 The influence of a physiologically inspired complex compression scheme on perceived listening effort for speech in noise

Saskia M. Waechter, Vinzenz H. Schönfelder, Sarah Voice, Nicholas R. Clark
Mimi Hearing Technologies GmbH, Research Department, Berlin, Germany

Objective: The primary goal of this study was to assess whether the required listening effort for speech recognition can be decreased by processing a clean speech signal with a complex compression scheme consisting of an instantaneous feed-forward and delayed feedback component mimicking the early stages of the healthy human auditory system ("Mimi-processing"). Listening effort measures were compared between processed and unprocessed sentences. A global equal-RMS constraint was imposed to avoid the influence of level-boost.

Methods: Perceived listening effort was assessed for 30 participants between the ages of 21 to 60 years old (mean = 32.5 ± 10.7 years SD) with the ACALES procedure (Krueger et al., 2017). Participants had average PTA4s of 9.4 dBHL (SD= 5.5 dBHL) in their better ear. The ACALES method employs a rating scale which is applied in an adaptive procedure to measure perceived listening effort for a wide range of (individualised) SNRs without resulting in ceiling effects. Participants rated their effort from 1='(almost) no effort' to 13='Extremely effortful' or 14='Only Noise'. Sounds were presented binaurally via Etymotic ER-1 insert earphones. The cohort was divided into three groups for which three different noise types were assessed, namely speech-shaped noise (SSN), multi-talker babble (MTB) and Cafeteria noise.

For each condition and participant, a two-slope function was fitted to the data points and the SNR-distance between the fitted functions of two different listening conditions at equal ratings is the measure of interest. The mean SNR-distance across ratings was calculated per participant and provides a value for how much the SNR can differ between two conditions and yet provide equal average effort ratings.

Results: SNR-differences [dB] between processed and unprocessed stimuli were significantly different from zero ($p<0.001$) with mean Mimi-processing benefits of 2.55 dB (SSN), 2.31 dB (MTB noise) and 2.22 dB (Cafeteria noise). This means that after Mimi-processing, speech stimuli with SNRs reduced by 2.22dB - 2.55 dB (noise dependent) are rated at equal listening effort by the average listener compared to unprocessed stimuli.

Conclusions: These results indicate that the Mimi-processing algorithm can decrease the perceived listening effort for speech presented in noise. This work provides a promising foundation upon which further improvements of the processing parameters may be implemented to increase speech intelligibility in noise.

14 Cognitive factors contributing to speech-in-noise comprehension: insights from a thousand young, normally-hearing listeners

Alexis Hervais-Adelman, Robert Becker
University of Zurich, Switzerland

The ability to comprehend speech under acoustically challenging conditions varies widely across individuals. This variability is typically attributed to cognitive factors that may have a role in supporting the listening effort required to comprehend speech experienced in adverse listening conditions. This notion has been formalised in a number of models that focus on the cognitive factors supporting speech comprehension in hearing impairment.

Ease of Language Understanding model (ELU, Rönnberg et al., 2013, Front. Syst. Neurosci. 7:31) posits that comprehending noisy speech specifically places demands upon processing in working memory. The Framework for Understanding Effortful Listening (FUEL, Pichora-Fuller et al., 2016, Ear Hear. 37:5S) places a decision making mechanism at the heart of the cognitive processes implicated in speech-in-noise comprehension, which weighs the demands of the task against the potential for success, executing this decision before task engagement and resource allocation. These models provide compelling bases upon which to consider the role of working memory and cognitive flexibility in adverse listening situations. However, evidence supporting these models in younger, normally-hearing, listeners remains scant (reviewed by Füllgrabe & Rosen, 2016, Front. Psychol. 7:1268). Here we use a large dataset, to probe the relationship between speech-in-noise perception ability and a battery of cognitive factors.

The Human Connectome Project (HCP, Van Essen et al., 2012, NeuroImage 62:2222), is a neuroimaging and behavioural dataset of over a thousand young (age range 22 – 37 years, Mean = 28.80, SD = 3.69), healthy participants with no self-reported history of hearing impairment. The HCP provides data on word in noise recognition (derived using the NIH words-in-noise test, hereafter “WIN”) and a battery of several dozen cog-

nitive indicators. We used these data to examine the relevance of cognitive factors to WIN using simple linear correlations. After correcting for age, we identify a number of significant ($p<.05$, Bonferroni corrected for multiple comparisons) cognitive predictors of WIN, including working memory ($r=.186$), crystallised intelligence ($r=.220$) and fluid intelligence ($r=.178$).

While the implication of crystallised intelligence, which contains measures of vocabulary, may not be altogether illuminating, the fact that both working memory and fluid intelligence are significantly associated with speech-in-noise comprehension performance lends support to the central tenets of both the ELU and FUEL. Importantly, although the proportion of variance explained by these particular cognitive factors may be low (working memory: 3.5%, fluid intelligence: 3.2%), this establishes that they are relevant even in younger listeners, although other factors are evidently also in play.

15 Speech in noise perception in childhood: Role of modulation filtering and processing efficiency

Irene Lorenzini[1], Christian Lorenzi[2], Laurianne Cabrera[1]
1. Integrative Neuroscience and Cognition Center, Université Paris Descartes, CNRS, France | 2. École Normale Supérieure, Université Paris Sciences et Lettres, Paris, France

The present study explored the relationship between the capacity to detect amplitude modulation (AM) and speech-in-noise (SIN) identification during childhood. Auditory models suggest that AM detection is not only constrained by the filtering properties of sensory mechanisms in the modulation domain, but also by "processing efficiency", the ability to make optimal use of the available sensory information. Behavioral tasks were designed to assess the development of modulation filtering and processing efficiency of AM cues and its relationship with SIN between 6 and 8 years of age.

Eighty-two children first completed a 2-alternative-forced choice task (AFC) using an adaptive procedure estimating AM detection thresholds for an 8-Hz sinusoidal AM. In this task, the AM carrier was varied in 2 conditions to assess: i) AM sensitivity using a 500-Hz sine tone (No Masking), and ii) AM masking using a 4-Hz wide narrowband noise centered at 500 Hz with small envelope fluctuations (Masking). Second, a "double-pass technique" evaluated the consistency of children's responses for AM detection using a constant-stimuli procedure. Then, AM detection performance in the Masking condition was measured at threshold for 200 trials repeated twice (2 passes) using a 2-AFC task. Percentage of Correct AM detection in each pass (PC) and Percentage of Agreement between the 2 passes (PA) were used to estimate within-listener consistency, a proxy of AM processing efficiency related to internal noise. Finally, children completed an XAB adaptive task measuring consonant identification thresholds in noise using fricatives and stops contrasting over three phonetic features (voicing, place, and manner). Additionally, children completed two standardized tests assessing receptive vocabulary and non-verbal reasoning.

Results showed that AM detection thresholds obtained with both carriers did not significantly improve from 6 to 8 years (ps > .37) and all children were similarly affected by AM masking ($p < .001$). When children were tested at threshold, both PC and PA increased with age (ps = .03). Thus, AM filtering is not affected by age, but aspects of processing efficiency are. Regarding SIN, thresholds significantly improved with age ($p = .02$) and were affected by phonetic feature ($p < .001$). Backward regression analyses showed that AM masking associated with vocabulary scores significantly predicted data for Manner (8.9%), PC and PA predicted to a small extent SIN data for Voicing (adjusted $R2 = 5.8\%$) and vocabulary predicted data for Place (5.1%). Overall, processing efficiency, modulation filtering and linguistic level determine SIN identification in childhood.

16 Effortful listening under the microscrope: Examining relationships between the task-evoked pupil response and the experience of effort and tiredness from listening

Ronan McGarrigle, Lyndon Rakusen, Sven Mattys
Department of Psychology, University of York, UK

Effort and tiredness from listening are common complaints from individuals suffering from declines in hearing and/or cognition. In recent years, pupillometry has emerged as a possible objective tool for measuring the mental effort associated with listening in adverse conditions. However, the precise relationship between changes in the task-evoked pupil response (TEPR) and the subjective experience of listening-related effort and tiredness remains unclear. Data from two experiments are presented that seek to examine the relationship between TEPR and perceived effort and tiredness by measuring covariance in these measures over the course of a sustained effortful listening task. For Experiment 1, we sought to replicate the effect of SNR on TEPR and self-report indices of effort and tiredness from listening during a competing talker task. Results suggest that a more adverse signal-to-noise ratio results in both larger TEPRs and increased self-report effort and tiredness ratings. However, the data also suggest that the effect of SNR on tiredness from listening ratings may be masked when using a single 'pre-post' measure of perceived tiredness. Experiment 2 sought to simulate a more sustained (and thus, fatiguing) challenging listening condition to examine these relationships in more detail. Results suggest that tiredness from listening ratings are related to both perceived effort and TEPR at the level of the individual; higher tiredness ratings are associated with higher effort ratings and smaller TEPRs. No relationship was found between TEPR and perceived effort ratings. Perceived performance was negatively related to both perceived effort and tiredness, suggesting a possible link between the feeling of tiredness from listening and self-efficacy. Theoretical implications of the data are discussed within the Framework for Understanding Effortful Listening (FUEL).

17 The presence of a social other motivates to invest effort while listening to speech-in-noise

Hidde Pielage[1], Adriana A. Zekveld[1], Gabrielle H. Saunders[2], Niek J. Versfeld[1], Thomas Lunner[2], Sophia E. Kramer[1]

1. Amsterdam UMC, Vrije Universiteit Amsterdam, Otolaryngology Head and Neck Surgery, Ear & Hearing, Amsterdam, NL | 2. Eriksholm Research Centre, Snekkersten, DK

Background: Mental effort has been gaining attention as an important facet of listening. A relevant factor influencing mental effort is motivation, which in turn can be influenced by reward. Reward has been found to enhance the mental effort that is spent while listening, as shown by an increased peak pupil dilation. Furthermore, social interactions have been suggested to be rewarding and may also increase the motivation to spend effort while listening. However, how social aspects influence listening effort has not been examined until now. In this study, we examined the influence of a social presence on listening effort.

Objectives: The aim of this study was to modify existing speech-in-noise paradigms to assess whether a social presence influences the amount of effort spent while listening. More specifically, we aimed to assess if doing a speech-in-noise task together with another individual, rather than alone, affected the task-evoked pupil dilation response. Furthermore, we examined if any potential effects were influenced by the difficulty of the task and the requirement to repeat the sentence.

Methods: 34 Young, normal-hearing participants (10 males, 24 females) listened to Dutch sentences that were masked with a stationary noise masker and presented through a loudspeaker. The participants' task alternated between repeating sentences (active condition) and not repeating sentences (passive condition). The participant did this either alone or together with another individual in the booth. When together, they repeated sentences in turn. The participant and the other individual did not know each other before the study. Participants performed the task at three intelligibility levels (20%, 50% and 80% sentences correct) in a blockwise fashion. During testing, pupil size was recorded as an objective outcome measure of listening effort.

Results: Both task difficulty and doing the task in the presence of another individual significantly increased peak pupil dilation (PPD). There was no interaction between task difficulty and the presence/absence of another individual on PPD. Furthermore, PPD was significantly lower in the passive conditions. This effect interacted with intelligibility. Lastly, performance on the listening task was affected by task difficulty, but not the physical presence/absence of another individual.

Conclusion: Increased PPD values suggest an increase in mental effort during listening when another participant is present, but only in the active condition (i.e. when the participants had to repeat the sentence). The effect of a social presence on pupil dilation seems to be independent of task difficulty.

18 Glimpses of what? Effects of varying the substrate while keeping the spectro-temporal mask constant

Martin Cooke
Ikerbasque, Vitoria, Spain

Maria Luisa Garcia Lecumberri
University of the Basque Country, Vitoria, Spain

Speech can be generated by sampling a base signal (the substrate) at the locations defined by a subset of spectro-temporal regions (the mask). Glimpse resynthesis performed in this way can lead to highly-intelligible speech. To explore what properties of the underlying substrate are important for successful speech perception, the current study examined the effect of changing the substrate while keeping the mask constant.

For each of 240 sentences in the Spanish Harvard Corpus, glimpse resynthesis was applied to a mask formed from spectro-temporal regions where the speech was more energetic than a speech-shaped noise masker when mixed at 0 dB SNR. In each case, the resulting mask was applied to 7 distinct substrates which varied in the amount of information they contained from the original speech. Two substrates contained the complete speech signal, viz. the speech signal itself, and the speech-plus-noise mixture. Two substrates contained partial speech information: either the temporal fine structure came from the mixture and the envelope from the masker or vice versa. Three further substrates -- speech-shaped noise, wideband polyphonic music, and speech from a different language (English) -- contained no information at all from the original speech. Some 26 Spanish normal-hearing listeners identified 5 keywords per sentence in the 7 constant-mask conditions, and in an additional condition involving the original speech-plus-noise mixture employed without a mask.

The keywords correct score for the non-masked mixture was 87%. Expressed as proportions of this score, listeners performed equivalently for masks applied to the speech signal (1.02) and the mixture (0.97). Removal of amplitude information in the mix led to a modest loss in intelligibility (0.91) while removal of mixture fine structure produced a larger fall (0.83). However, replacing the substrate by a noise masker led to no further drop in intelligibility (0.81). Listeners were also able to identify substantial numbers of keywords in the music substrate (0.61), but performance fell to near chance when the substrate was a different-language masker (0.04). These findings reveal that the mask alone contains a remarkable amount of information to support speech perception, but that properties of the substrate can interfere with listeners' ability to treat the mask as conveying speech cues. Intriguingly, listeners reported that they were totally unaware of the nature of the musical substrate, and were generally unable to recognise any words in the different-language masker nor identify its language.

19 Relationship between objective and subjective evaluation of heavily-distorted speech signals

Mitsunori Mizumachi
Kyushu Institute of Technology, Fukuoka, Japan

Speech applications on smartphones and smart speakers are widely spread in our daily lives. In most cases, distortion-less speech signals are assumed as their input signals, so that they do not work well in the real world, where the speech signals are distorted by acoustic interferences such as background noises and room reverberation. Noise reduction and dereverberation are indispensable for practical speech applications. Noise reduction and dereverberation aim to decrease the speech distortion caused by additive and convolutive acoustic interferences, respectively. Such objectives can be achieved by a wide variety of linear and nonlinear signal processing techniques. In general, the nonlinear methods achieve speech enhancement effectively and efficiently but cause annoying nonlinear distortion on the output target signals. Perceptually-oriented, less-distorted nonlinear speech enhancement is one of the recent trends in acoustic signal processing. Evaluation of speech distortion has also been an important issue for several decades. The relationship between the objective and subjective evaluation of speech distortion has been investigated under controlled experimental conditions. The effect of the speech distortion caused by highly nonlinear signal processing is, however, not well known under realistic, complex, heavily-distorted acoustic conditions. In this poster, the relationship between objective and subjective evaluation is discussed both for linear and nonlinear speech distortion in the severe acoustic conditions. Noisy speech samples were prepared with stationary and non-stationary noises in very noisy conditions where the signal-to-noise ratios (SNRs) were below 0 dB. The noisy speech samples were enhanced by linear and nonlinear signal processing, respectively. A listening test was carried out to quantify the subjective impression on speech distortion using the five-scale Mean Opinion Score (MOS). The relationship between the objective scores calculated by using some speech distortion measures and the subjective MOS is summarized in the viewpoints of types of signal processing, temporal characteristics of noise signals, and SNR.

20 The effect of intensity on an EEG-based objective measure of speech intelligibility

Eline Verschueren, Jonas Vanthornhout, Tom Francart
ExpORL, Dept. Neurosciences, KU Leuven, Belgium

Recently an objective measure of speech intelligibility, based on neural responses, has been developed in normal hearing listeners. However, the population for whom this method is being developed are patients with a hearing loss, requiring overall higher speech intensities to reach a good speech understanding level. As these higher intensities could influence the outcome, we investigated the influence of stimulus intensity on this objective measure. Similar to literature investigating the effect of intensity on cortical responses to non-speech stimuli, we hypothesized that increasing stimulus intensity would (1) increase the amplitudes and (2) decrease the latencies of the peaks in the neural response.

We recorded the electroencephalogram (EEG) in 20 normal-hearing participants while they listened to a narrated story. The story was presented at intensities varying from 10 to 80 dB A. To investigate the brain responses, we analyzed neural tracking of the speech envelope because the speech envelope is known to be essential for speech understanding. Envelope tracking can be measured by reconstructing the envelope from EEG using a linear decoder and by correlating the reconstructed envelope with the actual envelope. We investigated the delta (0.5-4 Hz) and the theta (4-8 Hz) band at each intensity. We also investigated the latencies and spatial components of the responses in more detail using temporal response functions.

Preliminary results show that when presenting the stimuli at higher intensities, response latencies of the peaks between 0 and 250ms shorten, similar to literature using non-speech stimuli. The amplitudes of these peaks, on the other hand, behave opposite to literature, decreasing with increasing stimulus intensity. Despite these latency and amplitude changes as a function of intensity, we can still objectively measure speech intelligibility. These results indicate that the objective measure of speech intelligibility can reliably be used in patients requiring higher stimulus intensities to enhance their speech understanding.

Acknowledgements: Research of Eline Verschueren (1S86118N) and Jonas Vanthornhout (1S10416N) is funded by a PhD grant of the Research Foundation Flanders (FWO). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 637424 to Tom Francart). Further support came from KU Leuven Special Research Fund under grant OT/14/119.

21 Perceptual adaptation to speech variation when listening to vocoded speech: Preliminary results

Olivier Crouzet[1], Etienne Gaudrain[2,3], Deniz Başkent[3]
1. LLING UMR6310 - Université de Nantes / CNRS, France | 2. CNRS, Lyon Neuroscience Research Center, France | 3. University of Groningen, University Medical Center Groningen, Netherlands

Background: When processing speech signals, listeners adapt to several sources of variation. Some are associated with 'voice' information (vocal characteristics, speaker identity, gender, emotional state...) while others relate to phonological categories (phonemic classes / distinctive features). It has been shown that variations in acoustic context or individual voices can significantly influence speech identification performance in normal-hearing listeners. Alterations in acoustic context can shift phonological boundaries (Ladefoged & Broadbent, 1957, JASA 29:98; Sjerps, McQueen & Mitterer, 2013, AP&P 75:576) while increasing variation in multi-speaker setups can hinder speech identification performance (Goldinger, 1996, JEP:LMC 22:1166). Chang & Fu (2006, JSLHR 49:1331), comparing normal-hearing (NH) listeners processing vocoded speech with cochlear-implanted (CI) listeners, showed that CI participants experienced consistent difficulties in multi-speaker conditions, whereas the effect was more variable in NH listeners, depending on vocoder parameters. Therefore, some of the speech recognition difficulties that CI listeners experience on a daily basis may have their origins in limitations for the appropriate 'separation' of 'voice' properties from 'speech' cues. As stated by Chang & Fu (2006), 'acoustic variations among talkers may be confused with acoustic differences between phonemes' as a consequence of the low spectral resolution of the CI. In order to get a better understanding of perceptual adaptation effects with vocoded-speech, a replication of the seminal work by Ladefoged & Broadbent (1957) was conceived.

Methods: 15 Dutch monosyllabic word-pairs were selected (e.g. [kip] 'tilted' vs. [kIp] 'chicken'). For each vowel contrast (word pairs), an acoustic continuum was generated on the basis of the actual formant frequencies. Each of these monosyllabic words was concatenated at the end of a short fixed carrier sentence (eng. tr. 'Please say what this word is:') which was submitted to various acoustic manipulations (either changes in the F1~F2~F3 formant space along the vowel continuum dimension or changes in VTL relating to formant frequencies as well). The aim of this study was to investigate the influence that mechanisms of adaptation to changes associated with two different sources of voice variation may exert on phonological classification (specifically vowel identification) when speech signals are processed through channel vocoding.

Results: Preliminary results have been collected from an online experiment in order to estimate the size of the effects, to test statistical modelling approaches and to perform power analyses through simulations for the final data collection. These preliminary results along with simulations for power analyses will be discussed at the conference.

22 Pupil dilation during speech production in noise is modulated by masker type

Maximillian Paulus, Valerie Hazan, Patti Adank
University College London, UK

Pupil dilation is affected by both sensory and motor events. In speech perception, pupil dilation can indicate increased listening effort, as induced for instance by decreasing the signal-to-noise ratio. Moreover, a larger pupil dilation has been observed for speech masked by a competing speaker, compared to speech masked by stationary noise. In speech production, pupil dilation has shown to be modulated by language processing, but the literature is sparse. For instance, it is unclear whether masker type similarly affects pupil dilation when speech is produced in noise.

In the current study, twenty-four normal-hearing participants were tasked to read out loud sentences that were presented on a screen. Simultaneously, background noise was played, which consisted of either a competing speaker, speech-shaped stationary or fluctuating noise at similar loudness levels. In a control condition, no background noise was played. Pupil size was recorded for the duration of each trial and acoustic features such as intensity and spectral tilt were extracted from the verbal response.

Our results confirmed previous findings showing that intensity is increased and spectral tilt decreased for speech produced in stationary masking when compared to fluctuating and competing-speaker masking. Conversely, we found larger pupil dilation for competing-speaker masking, similar to results from speech perception studies. Despite larger vocal effort found for stationary masking, pupil dilation was more sensitive to the cognitive effort imposed by the competing speaker. Since pupil dilation has been shown to be affected by movement intensity, this finding appears to indicate that cognitive effects outweigh articulatory effects. The results demonstrate the potential use of pupillometry in speech production studies.

23 Babble noise augmentation for phone recognition applied to children reading aloud in a classroom environment

Lucile Gelin[1,2], Morgane Daniel[1], Thomas Pellegrini[2], Julien Pinquier[2]
1. Lalilo, Paris, France | 2. IRIT, Paul Sabatier University, CNRS, Toulouse, France

Current performance of speech recognition for children is below that of the state-of-the-art for adult speech (Shivakumar et al. 2018, arXiv:1805.03322; Hadian et al. 2018, Proc. Interspeech, 12-16). Young child speech is particularly difficult to recognise, and substantial corpora are missing to train acoustic models. Furthermore, in the scope of our reading assistant for 5-7-year-old children learning to read, models need to cope with slow reading rate, disfluencies, and classroom-typical babble noise.

In this work, we aim at improving a phone recognition system's robustness to babble noise, to be able to give accurate feedback to children reading aloud despite the noisy environmental conditions. We use a data augmentation method that consists of mixing the speech recordings with babble noise recordings at target Signal-to-Noise (SNR) ratios of 2, 5, 10 and 15 dB. The speech recordings are part of our in-house speech dataset, gathered directly in schools or via the Lalilo platform. The noisy recordings come either from the DEMAND corpus (Thiemann et al. 2013, Zenodo.1227121), where babble noise is composed of adult voices and is constant, or from our in-house noise (IHN) corpus, containing real-life classroom environments, where babble noise comes mostly from children and is much more irregular. The evaluation set is comprised of recordings where children read isolated words, in a classroom environment, with SNRs varying between -10 and 50 dB: mean SNR of 23.8 dB and standard deviation of 10.9 dB.

To build a phone recognition system, we used a model trained on the Commonvoice French adult corpus, to do transfer learning (TL) with our small children corpus. The TL method (Shivakumar et al. 2018, arXiv:1805.03322) takes the source adult model and re-trains it with child data, with higher learning factors for output layers. We use separately clean, clean+DEMAND-augmented and clean+IHN-augmented child data as the target data for transfer learning. We show that adapting an adult model trained on clean speech with noise-augmented child data improves the system's global performance on our evaluation subset. When measuring performance as a function of the SNR, we observe that noise augmentation highly reduces the error rate for very noisy recordings (SNR < 10 dB) and does not degrade performance for clean recordings. Transfer learning with babble noise augmented child data thus enabled an improvement in the child speech recognition systems' robustness to classroom-typical babble noise, a necessary quality for vocal reading assistants.

24 Speech perception in noise and auditory working memory in vocalists, violinists and non musicians

Priyanka Vijaya Kumar, Rajalakshmi Krishna
All India Institute of Speech and Hearing, Mysore, Karnataka, India

Music is a highly complex sensory stimulus and is structured in several dimensions. This richness makes music an ideal tool to investigate the functioning of the human brain. Since there are many different training methods used to develop musical expertise (e.g. vocal or instrumental), these differences could lead to varying auditory processing abilities of acoustic signals. The current study aims to see if there are any differences in speech perception in noise and auditory working memory between vocalists, violinists and non-musicians. 30 participants from each of the group were subjected to speech perception in noise test (QuickSIN) and two auditory memory tests (forward and backward digit span tests). This study also aimed to study the effect of years of musical experience on the above mentioned auditory processing skills by regrouping the same 30 participants in each group (i.e., the vocalists and the violinists) into 3 sub-groups consisting of 10 participants with different music expertise (10 participants in the junior level, 10 participants in the senior level and 10 participants in the vidwath level — these levels are the levels of musical expertise/proficiency which are obtained after clearing the theory and practical exams prescribed by the governing bodies of music board in Karnataka, India).

Overall results revealed that in all the auditory processing tests (speech perception in noise and auditory working memory) musicians (both vocalists and violinists) outperformed the non-musicians. However, no significant difference was noticed between violinists and vocalists. The results of the study are in congruence with other literature report indicating musical experience as an important factor inducing enhancements in the overall auditory perceptual abilities. Further, the study results lead to the possible speculations that type of music (vocal vs. instrumental) does not influence music induced differences in the auditory processing skills. Similarly, there was no significant difference observed in the performance of the musicians with respect to the years of musical experience both in the violinists and vocalist groups.

25 Factors affecting the subjective impression of speech intelligibility

William M. Whitmer, David McShefferty
Hearing Sciences - Scottish Section, University of Nottingham, UK

From laboratory investigations to clinical interventions, we often assume that our objective measures of speech intelligibility have perceptual relevance. By the same assumption, an objective equivalence, such as the signal-to-noise ratio (SNR) where 50% of the keywords in a sentence can be correctly repeated (SNR50), should have perceptual equivalence, such as being judged to be equally clear. We previously found, however, that when presented with a sentence in same-spectrum noise (SSN) and two-talker babble at their respective individual SNR50s, participants were approx. 30% more likely to choose the sentence in two-talker babble being clearer, despite being at a lower, objectively equivalent SNR50. In the current study, we explore whether this clarity bias persists under different conditions, and if it is due to (1) the presence of speech in the masker, and/or (2) unmasked segments of the target signal that do not contribute to speech intelligibility. Thirty-six adults of varying hearing ability are first asked to repeat back sentences presented at various SNRs in three noise types: SSN, n-talker babble (n = 1, 2, 4, 8 & 16), and SSN modulated with the envelope (low-pass Hilbert transform) of n-talker babble. Individual SNR50s are estimated for each of the 11 noise conditions. Participants are then asked which of a pair of stimuli presented at their respective SNR50s are clearer. Pairs consist of either (a) n-talker babble and SSN, (b) n-talker modulated SSN and SSN, or (c) n-talker babble and n-talker modulated SSN. If clarity bias is due to speech in the masker, the bias should be reduced when the carrier signal is noise modulated by the babble. If clarity bias is due to unmasked segments, the bias should persist with a babble-modulated noise carrier, but decrease with increasing number of talkers in the babble. Results will be analysed for both objective and subjective disparities, and discussed in terms of the effect of noise type on subjective reports of speech intelligibility.

Funding: This work was supported by the Medical Research Council [grant number MR/S003576/1]; and the Chief Scientist Office of the Scottish Government.

26 The development of a multimodal communication behaviour capture system

A. Josefine Munch Sørensen[1,2], W. Owen Brimijoin[1]
1. Facebook Reality Labs, Redmond, WA, US | 2. Hearing Systems, Dept. of Health Technology, Technical University of Denmark, Kgs Lyngby, Denmark

Capturing precisely quantified conversational interactions in ecologically valid scenarios is of tremendous value for studying the nature and adaptation of communication behavior in challenging environments. Such data could, e.g., be used in evaluations of audio processing strategies, estimating the effects of hearing assistance devices on spoken interaction, and any number of behavioral listening models. To this end we assembled a high-precision behavioral capture system for recording voice, pupil dilations, gaze, and head and torso motion from several people simultaneously. The capture facility is also capable of reproducing realistic audio backgrounds using a 52-channel loudspeaker array. In order to relate events in one modality to others, data had to be highly synchronous, which posed a hardware challenge. To synchronize voice capture, noise playback and motion capture, two sound cards (one for voice capture, one for the loudspeaker array system), and a motion capture lock sync box were slaved by a single master clock via word clock for audio and genlock for video, and timecode was used as a common time reference. Due to CPU and software restrictions, gaze and pupil dilation capture was made on one computer for each eye tracker, and therefore had to be post-synchronized using recordings of an audio click train sent from one of the sound cards to each of the eye tracking computers. We will present how we designed and constructed the system, how we calibrated it, and what software was used and developed for the purpose. Moreover, we show some preliminary data captured from a three-person free conversation in spatially reconstructed cafeteria noise presented at various levels. We demonstrate how we use categorizations of conversational turn-taking to find points of interest in the data to investigate the transitions in all modalities leading up to, during, and immediately following a turn-taking.

27 Do state-of-the-art TTS synthesis systems demand high cognitive load under adverse conditions?

Avashna Govender, Cassia Valentini-Botinhao, Simon King
Centre for Speech Technology Research, University of Edinburgh, UK

Text-to-speech (TTS) synthesis is increasingly being deployed in many real-world applications. Yet, there is a lack of evaluation studies that measure how listeners cope with listening to synthetic speech produced by such technologies in noisy environments.

Significant improvements in TTS have been made since the adoption of Deep Neural Networks (DNNs). It is now possible to produce synthetic speech that is as intelligible as human speech and has high naturalness when listening in quiet. However, there is little knowledge on how these measures are affected when listening to synthetic speech under adverse conditions. Furthermore, there is little understanding of how synthetic speech interacts with the human cognitive processing system in noise.

In our previous work (Govender et al., 2019, Proc. 10th ISCA Sp. Synth), we investigated the cognitive load of synthetic speech in speech-shaped noise. Pupillometry and self-reported measures were used to measure cognitive load. Perception studies have shown that the pupil response indexes the amount of mental effort allocated by the human cognitive processing system when performing a task (Kahneman et al., 1966, Science, vol. 154, no. 3756, pp. 1583–1585 and Beatty et al., 1966, Psychonomic Science, vol. 5, no. 10, pp. 371–372). Therefore in our experiments, we used pupillometry to measure the pupil response of a listener whilst listening to speech through headphones and thus quantifies listening effort. The stimuli used were synthetic speech mixed with speech shaped noise at -3 dB and -5 dB SNR. Results were compared with the human speech recordings that were used to create the TTS voice. The results showed that in both conditions human speech was easier to listen to than TTS.

In addition, our work indicated that the contribution to an increased cognitive load could be due to the use of a conventional statistical parametric synthesis system (SPSS) which is derived from the source-filter model. In these systems, spectral and excitation features are predicted in separate streams which could potentially destroy correlations that exist between them.

This work aims to confirm whether this is true by using a TTS system trained using a sequence-to-sequence DNN model with an attention mechanism similar to Tacotron 2 (Shen et al., 2018, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing). In this way all speech features are predicted in a unified manner and thus any correlations that exist remain intact. We use a neural vocoder based on WaveRNN (Kalchbrenner et al., 2018, Proc. ICML, pp. 2415–2424) which is capable of reconstructing speech with high fidelity.

We expect to see an improvement in terms of reduced cognitive load in comparison to the previously used conventional statistical parametric synthesis system.

28 Listening effort in normal-hearing listeners with a cochlear implant vocoder simulation using subjective ratings and EEG measurements

Angelika Kothe[1,2], Amy J. Hall[2], Michael Schulte[3], Kamil Adiloglu[4], **Axel H. Winneke**[5]
1. Jade University of Applied Sciences, Oldenburg, Germany | 2. Fraunhofer IDMT-HSA, Oldenburg, Germany | 3. Hörzentrum Oldenburg GmbH, Oldenburg, Germany | 4. HörTech gGmbH, Oldenburg, Germany | 5. Fraunhofer IDMT, Oldenburg, Germany

Many hearing impaired individuals experience understanding speech in noisy environments (SpiN) to be exhausting. Being exposed to a challenging listening situation for a prolonged period of time can cause fatigue. This can lead to a chronic burden, which is a disadvantage in everyday life and in the workplace.

In the context of hearing devices and listening effort, less research has been conducted with cochlear implant (CI) patients. This motivated the current study to look more closely at how listening to SpiN with a CI affects listening effort. Since recruitment of sufficiently large number of CI users is more difficult to realize and because of the heterogeneity of the CI patients, in this project we equipped normal hearing participants with behind-the-ear (BTE) hearing aid dummies and a vocoder (Bräker et al., 2009, Z. Audiol. 48:158) to acoustically simulate a CI. Participants perform the adaptive categorical listening effort scaling task (ACALES, Krüger et al., 2017, JASA 141:4680) to obtain subjective ratings on experienced listening effort at various SNRs. A multi-channel electroencephalogram (EEG) is recorded simultaneously in order to quantify listening effort on a neurophysiological level. The EEG parameter of interest here are changes in the alpha-frequency band. The study investigates the effect of spectral and temporal degradation of a speech signal (i.e. CI simulation) in noise as well as the effect of noise suppression on listening effort. Noise suppression is performed using a minimum variance distortionless response (MVDR) beam former to reduce the impact of a noise source located behind the participant at 135° on the speech source at 0°. The whole signal processing chain including the noise suppression and vocoder is implemented in the real-time capable master hearing-aid (MHA, Grimm et al., 2006, Acta. Acust. united Ac. 92:618) platform and all measurements are conducted using real-time audio.

Despite the limitations of a simulation study, the project will shed light on listening effort and CIs, how listening effort is reflected in the EEG and how spatial noise suppression can aid CI users in terms of reducing listening effort when listening to SpiN.

29 Why do you listen to that program? – An analysis of hearing aid program usage in real life

Nadja Schinkel-Bielefeld[1], Jana Welling[1], Ann-Elisabeth Krug[2], Rosa-Linde Fischer[1]
1. Sivantos GmbH, Erlangen, Germany | 2. Akademie für Hörakustik, Lübeck, Germany

Hearing aids usually have an automatic program that detects the acoustic situation and changes hearing aid settings accordingly. For specific situations or special demands the hearing care professional can configure additional hearing programs, so that the patient can switch between them at will. However, many hearing aid wearers have only a single program. Possible reasons could be that a further program is not necessary or beneficial or that hearing aid wearers refuse to switch because choosing a suitable program might be too difficult.

In order to analyze when and for what reasons subjects switch between hearing programs we performed a study with 10 hearing impaired subjects (mean age 71 years, PTA4 = 42dBHL), who were bilaterally fitted with Signia Pure 13 7Nx hearing aids for a three weeks home trial. During this time, we asked them to fill out a questionnaire whenever they switched between hearing programs. Both was done via an Ecological Momentary Assessment app on a smart phone which we provided to the subjects. In addition, all program changes and objective information about the acoustic situation were recorded. To ensure that all programs were tested, the hearing aids switched automatically into a random program every morning. Subjects could switch back immediately, if desired.

We found that the universal program was listened to about twice as much as any other program and subjects never indicated that switching into the universal program deteriorated the listening experience. This shows that this program is, as intended, a safe option for all listening situations that never fails completely. However, we also see that the other programs were used and in the exit interviews each subject stated that they liked at least one other program except the universal program. While subjects indicated that the main reason for program switching is the change of a listening situation, this could often not be seen in the objective data. Thus, a subjective change is not necessarily represented in the objective data available from the hearing aid. Furthermore, the selection of programs for certain situations or listening goals differed greatly among subjects. This makes it difficult to create an automatic program that fulfils all individual needs and shows why special programs can be beneficial for hearing impaired.

30 *This poster is now numbered* **57**

31 Simultaneous EEG and pupillometry as indicators of listening effort for enhanced speech in adverse conditions

Amy J. Hall, Jan Rennies, Axel Winneke
Fraunhofer IDMT, Oldenburg, Germany

Understanding speech in noisy conditions requires additional cognitive processing or listening effort compared to quite environments. AdaptDRC is a near-end-listening enhancement algorithm developed to improve intelligibility (Schepker et al., 2013, Interspeech 3577) by altering the frequency content and dynamic range of a speech signal, dependent on the environmental noise. As more advanced speech enhancement technologies make their way into public use, quantifying their effects on listening effort has become an important area of research.

There are many methods for capturing changes in listening effort, each with their own theoretical bases and physiological underpinnings. Ongoing debate considers their respective methodological advantages and limitations. By measuring subjective ratings, pupil size, and EEG data concurrently, we may investigate how they relate to each other.

Native English speakers with normal hearing aged 18-30 listened to English OLSA sentences presented in cafeteria noise at -5, 0, +5 dB SNR, in which the speech is unprocessed or processed with AdaptDRC. Participants rated the subjective effort on an adapted version of the adaptive categorical listening effort scale (Krueger et al., 2017, JASA 141:4680). EEG data were collected using a 24-channel EEG system. Pupil size was recorded using the Eyelink 1000 plus. Further, participants each completed a short cognitive task battery including measures of working memory and inhibition.

In this poster we present preliminary data showing the relative sensitivity of subjective ratings, pupil size and EEG alpha power, and how they relate to each other for measuring changes in listening effort for speech in noise. All methods are anticipated to indicate changes in effort, as manipulated by SNR and speech enhancement, but the shapes of the responses may vary. We also examine the relationship between the identified markers of effort and individual differences in cognitive abilities.

32 A generalised recurrent sequence to sequence model for robust and efficient speech recognition

George Sterpu, Christian Saam, Naomi Harte
Trinity College Dublin, Ireland

Sequence to sequence neural network architectures have been applied successfully to sequence modelling tasks such as speech recognition. Two popular variants are based on recurrent transformations with cross-modal attention, or on self-attention layers. However, these models share the limitation of requiring full sentences as inputs, which prevents the online, or real-time, usage of the algorithm.

It is commonly observed that the alignment between the decoded graphemes and the acoustic features is loosely monotonic in such networks. This means that speech signals have mostly short term dependencies between linguistic and acoustic units, and that the full sentence models are computationally and mathematically inefficient. Several studies attempted to limit the temporal context to a fixed size window, but reported a degradation in recognition performance.

In this work we describe an end-to-end differentiable neural model that learns to cluster together contiguous acoustic representations into segments. Similar to the Adaptive Computation Time (ACT) algorithm, our model is equipped with a halting unit signalling the end of an acoustic segment. Conceptually, this can be viewed as an end-to-end alternative to the segmental features originally proposed for Hidden Markov Models.

We show that our model is a generalisation of the traditional sequence to sequence model. The low overhead halting unit can make the encoding process stop after every input timestep, leading to an encoder representation identical to the one of the traditional model. This is the default behaviour in the absence of any segmentation incentive, where only the cross-entropy between predictions and targets is optimised.

We explore a set of strategies that encourage the encoder to aggregate multiple timesteps, while maintaining the decoding accuracy. We also identify common patterns in segmental architectures that inhibit segment discovery by design, and show their connection with ACT and related approaches in speech and text processing. In some cases, the reduction in number of timesteps is of 50% or more.

Automatic segmentation has the potential of replacing pyramidal encoding strategies that downsample the input empirically by a constant factor with each layer in the stack. The advantage of segmental representations is their possibly higher correlation with the acoustic units, and the interpretability properties offered by the discovered segment boundaries. Our on-going work is focused on the analysis of the segments, and on the integration with a segmental decoder enabling online recognition. We conclude that, when designed appropriately, neural networks can learn to cluster acoustic representations in an unsupervised way.

33 Fast speech intelligibility estimation using a neural network trained via distillation

Trevor Cox[1], Yan Tang[2], Will Bailey[1]
1. University of Salford, UK | 2. University of Illinois, USA

Objective measures of speech intelligibility have many uses, including the evaluation of degradation during transmission and the development of processing algorithms. One intrusive approach is to use a method based on the audibility of speech glimpses. The binaural version of the glimpse method can provide more robust performance compared to other binaural state-of-the-art metrics. However, the glimpse method is relatively slow to evaluate and this limits its use to non-real time applications. We explored the use of machine learning to allow the glimpse-based metric to be estimated quickly.

Distillation is an established machine learning approach. A complex model is used to derive a simpler machine-learnt model capable of real-time operation. The simpler student model is trained on synthetic data generated from the complex teacher model, thereby distilling knowledge from teacher to student. In this case the teacher is the slow glimpse-based model, and the student an artificial neural network. Once the neural network is trained, the student rapidly estimates the glimpse-based speech intelligibility metric. It is fast enough to allow real-time operation as an intelligibility meter in a Digital Audio Workstation.

A shallow artificial neural network with a relatively simple structure is found sufficient. The inputs to the network are cross-correlations between Mel-frequency cepstral coefficients (MFCCs) for the clean and noisy speech. Only the largest value of the cross-correlation for the left and right ear signals are used as inputs, to simulate better-ear binaural listening. Even for this lightweight artificial neural network, a large amount of training data is necessary to make the distillation robust. 1,200 hours of audio samples were used containing speech from a wide range of sources (SALUC, SCRIBE and r-spin speech corpora and librivox audiobooks). Maskers included speech-shaped noise, competing speech, amplitude-modulated noise, music and sound effects. The signal-to-noise ratio ranged from -20 to 20 dB. Performance is evaluated using test data set not used in training. A comparison between the estimated speech intelligibility to the full glimpse-based model gave an r2 of 0.96 and a mean square error of 0.003.

34 The influence of a physiologically inspired complex compression scheme on speech intelligibility in noise

Saskia M. Waechter, Vinzenz H. Schönfelder, Sarah Voice, Nicholas R. Clark
Mimi Hearing Technologies GmbH, Research Department, Berlin, Germany

Objective: The primary goal of this study was to assess whether speech intelligibility in speech-shaped background noise can be improved by processing the clean speech signal with a complex compression scheme consisting of an instantaneous feed-forward and delayed feedback component mimicking the early stages of the healthy human auditory system ("Mimi-processing"). Speech intelligibility measures were compared between processed and unprocessed sentences. A global equal-RMS constraint was imposed to avoid the influence of level-boost.

Methods: Speech intelligibility was assessed for 35 native German speakers (24-68 years old) with an adaptive speech reception threshold (SRT) test, which provides an estimate of the required signal-to-noise ratio to achieve 50% correct word identification. Participants had average PTA4s of 9.9 dBHL (SD=7.1 dBHL). SRTs were measured with the German Oldenburg Matrix sentence test (OLSA). Sounds were presented monaurally via Etymotic ER-1 insert earphones. For a sub-cohort of 12 participants, an additional third condition was assessed in which speech was processed with an 'equivalent-equaliser' (equivalent-EQ).

For each condition, a psychometric function was fitted to individual datasets with the psignifit toolbox, which implements a maximum-likelihood method for estimating psychometric parameters. In this way, SRTs were estimated for all participants and conditions, and the results were analysed as the difference-SRT [dB], averaged across participants, between the unprocessed condition and the respective test condition of interest.

Results: *Full-cohort results:* Mimi-processing resulted in statistically significantly improved SRTs [$t(34)=19.78$, $p<0.0001$] with a mean SRT-improvement of 2.77 dB compared to unprocessed speech as assessed with a two-sided one-sample t-test on paired observations. *Sub-cohort results:* The sub-cohort data (n=12) indicated statistically significant SRT-differences between the conditions unprocessed, Mimi-processed and equivalent-EQ [ANOVA: $F(2,22)=219.73$, $p<0.0001$]. Post-hoc analysis with multiple t-tests and Bonferroni correction for multiple comparisons revealed that SRTs of the equivalent-EQ condition were significantly worse than unprocessed SRTs (SRT=-1.31 dB) and SRTS of the Mimification condition was significantly better than unprocessed SRTs.

Conclusions: These results indicate that the Mimi-processing algorithm can improve speech intelligibility for speech presented in noise. This benefit seems to be a result of its unique compression scheme and does not solely emerge from frequency dependent energy shifting as represented by the equivalent-EQ condition. This work provides a promising foundation upon which further improvements of the processing parameters may be implemented to increase speech intelligibility in noise.

35 Can visual capture of sound separate auditory targets from noise?

Chiara Valzolgher[1], Roberta Sorio[2], Giuseppe Rabini[3], Alessandro Farnè[1], Francesco Pavani[3]
1. Integrative, Multisensory, Perception, Action and Cognition Team (IMPACT), Lyon Neuroscience Research Center, France | 2. Department of Psychology and Cognitive Sciences (DiPSCo), University of Trento, Italy | 3. Center for Mind/Brain Sciences (CIMeC), University of Trento, Italy

Speech recognition in noise improves when two competing sound sources are spatially separated. This phenomenon is termed Spatial Release from Masking (SRM) and it can arise for a combination of masking and attentional selection. As the physical distance between the two sound sources increases, the head-shadow effect progressively changes the ratio of signal and noise at each ear, affecting energetic masking of the target sound. In addition, as distance in external space increases, the two concurrent auditory events can be segregated more easily for attentional selection. This cognitive mechanism prioritises information coming from the relevant source, while inhibiting the competing signal. Studies across sensory systems (e.g., vision or touch) have repeatedly shown that separating concurrent streams of information in space helps selective attention mechanisms. Here, we investigated to what extent illusory changes in the perceived position of the target sound in external space could influence the SRM phenomenon. Specifically, we used a multisensory paradigm to illusory increase or decrease the perceived separation between speech and noise, exploiting a visual capture of sound phenomenon known as 'ventriloquist effect'. In each experimental trail, normal-hearing participants (N=20) performed two tasks: hearing-in-noise and sound localization. The hearing-in-noise task entailed repeating aloud a sequence of 5 spoken digits, delivered from unseen speakers in front space, while ignoring concurrent noise delivered from a fixed visible speaker to the left of the apparatus. The sound localisation task entailed pointing to the perceived position of the sound, after digits identification. Crucially, all target sounds were delivered together with a visual stimulus that changed in brightness as a function of the target-sound's envelope. With respect to the auditory target, the visual stimulus either originated from the same location (audio-visual congruent, AVcon) or was located 15 degrees to the left or right (audio-visual incongruent, AVinc). Results showed that AVinc conditions induced visual capture of sounds: target sounds were perceived closer to noise when the visual stimulus was presented leftwards, and farther away from noise when the visual stimulus was presented rightwards. We also obtained SRM for auditory targets delivered to the right of the participant's body midline (i.e., opposite hemispace with respect to noise), compared to auditory targets delivered the left. However, SRM was unaffected by AV conditions, revealing no measurable effects of the multisensory illusion on SRM. This indicates a greater role for energetic masking compared to attentional selection, in contributing to SRM in our experimental paradigm.

36 Spatial release from masking in children with auditory processing disorder in virtual and real environments

Katharina Zenke, Stuart Rosen
University College London, UK

Auditory Processing Disorder (APD) is a developmental disorder characterized by difficulties in listening to speech in noise despite normal audiometric thresholds. It is still poorly understood and much disputed and there is need for better diagnostic and intervention tools for young children, not least because of suspicions that APD leads to learning difficulties in language and literacy, and hence to poor school performance.

One promising avenue of research is the claim that around 20% of children referred for APD assessment are found to have a reduced spatial release from masking (SRM). Current clinical tests measure the SRM in virtual auditory environments generated with head-related transfer functions (HRTFs) from a standardized adult head. Adults and children, however, can be very different in head dimensions and mismatched HRTFs are known to affect localization accuracy. HRTFs in children have not been systematically measured so far and it is unclear whether HRTF mismatch also impacts speech perception, especially for children with APD due to their problems with processing auditory information.

In our current study, we measured individual HRTFs in children with diagnosed APD and typically-developing children aged 7 to 12 years. The SRM was measured for target sentences and two symmetric speech maskers in virtual auditory environments generated from these individualized HRTFs or HRTFs of an artificial adult head as well as in a real anechoic environment. In order to assess the influence of spectral pinna cues, we also measured speech reception thresholds for HRTFs gained from a spherical head model that only contains interaural time and level differences. Preliminary findings suggest differences in speech reception thresholds between listening conditions and slightly better overall performance of typically-developing children but similar amounts of SRM for all conditions. Both groups of children obtained significantly worse speech reception thresholds and smaller SRM as normal-hearing adults in our previous study.

We hope our results will help to determine the relevance of individualized spatial cues for SRM and further clarify the nature of spatial processing difficulties in children with APD.

37 Comparison of simultaneous measures of pupil dilation, verbal response time and subjective evaluation of listening effort

Chiara Visentin[1], Chiara Valzolgher[2], Paola Potente[2], Francesco Pavani[2], Nicola Prodi[1]
1. Department of Engineering, University of Ferrara, Ferrara, Italy | 2. Center for Mind/Brain Sciences (CIMeC), University of Trento, Rovereto, Italy

Listening effort (LE) describes the amount of cognitive and attentional resources required to perform a listening task. Over the last decade, estimates of LE have complemented more traditional measures of task performance accuracy (i.e. speech intelligibility), on account of the need to consider both auditory and cognitive factors when investigating listening processes. Although a direct and comprehensive measure of LE is currently unavailable, several proxy indices have been proposed. Correlations between these quantities are uncommon, suggesting that they may probe different underlying cognitive dimensions. To date it is still unclear how to best quantify LE, and if and how different proxy measures should be combined to provide the most comprehensive description of this complex construct.

This study aimed to identify changes in LE during a speech reception task, using pupil dilation (physiological measure), verbal response time in a single task paradigm (behavioral measure) and self-ratings (subjective measure). Normal hearing adults (N=24) were presented with the matrix-sentence test in the Italian language, using three signal-to-noise ratios (SNR: -3, -6, -9 dB). Direction of selective attention was manipulated across blocks, by making the origin of the speech signal either blocked or random across the two sound sources. In addition, the role of visual cues in detecting the origin of the auditory signal was assessed by visually marking or not the position of the unseen loudspeakers. Data on pupil dilation and verbal response time were acquired simultaneously for each experimental trial.

Although the correlation between subjective and behavioral/physiological measures has been analyzed in recent works, the relation between pupil dilation and verbal response time remained unexplored, due to the different requirements of the two measurement procedures (presence vs. absence of a retention period after the end of the auditory stimulus). No retention period was included in this study, so that a mean pupil dilation was calculated only over the listening phase. Despite this methodological constraint when measuring dilation, the results indicated that all three proxies of LE were sensitive to SNR. In addition, a significant interaction between SNR and vision emerged for response time, and a significant interaction between SNR and attention was found for mean pupil dilation. This indicates that the two quantities were sensitive to the cognitive manipulations included in the study. No significant correlation was found between the three proxy measures, supporting the hypothesis that they analyze different aspects of LE.

38 The effects of working memory load and working memory capacity on online spoken word recognition: evidence from eye movements

Gal Nitsan[1,2], Karen Banai[1], Boaz M. Ben-David[3]
1. University of Haifa, Israel | 2. Interdisciplinary Center (IDC) Herzliya, Israel | 3. The Interdisciplinary Center (IDC) Herzliya, Israel

Difficulties in speech perception, especially in adverse noisy conditions, are highly prevalent among older adults. However, the degree of the deficit is highly variable across older listeners. Previous research in our lab suggests that this variability could be related to individual differences in cognitive capacity (Hadar et al., 2016, Front. Neurosci. 10:221; Nitsan et al., 2019, Tr. Hear. 23:2331216519839624). Using the eye-tracking 'visual world' paradigm, listeners were asked to follow spoken instructions, while retaining either a low load (single-digit) or high (four-digits) load for later recall. In critical trials, instructions (e.g., "point at the candle") directed listeners' gaze to pictures of objects whose names shared either onset or offset sounds with the name of a competitor that was displayed on the screen at the same time (e.g., candy or sandal). We reported that for young listeners, high-load delayed real-time spoken word recognition both in quiet and in noise. Importantly, the interference effect of a concurrent memory load was greater for individuals with a smaller memory span than for those with a larger one. In an ongoing study, we administrated the paradigm with a group of 30 older listeners with clinically normal hearing. We examine whether older adults are similarly affected by a working memory load as younger adults, and how individual differences in older adults' working memory capacity affect the timeline for spoken-word recognition. Preliminary data pointing to age-related similarities and differences in these effects will be discussed.

39 The effects of noise and poor voice quality on spoken language processing in school-aged children: A systematic review

Isabel S. Schiller[1], Angélique Remacle[2], Nancy Durieux[3], Dominique Morsomme[1]
1. Faculty of Psychology, Speech and Language Therapy, and Education, University of Liège, Belgium | 2. Fund for Scientific Research - F.R.S. - FNRS, Brussels, Belgium | 3. ULiege Library, Université de Liège, Belgium

At school, children often face challenging listening conditions due to high noise levels or because they are exposed to dysphonic speakers. To date, no comprehensive review has evaluated how this might affect spoken language processing (SLP). Our aim was to systematically review the literature on the effects of noise and/or impaired voice quality on regular school-aged children's SLP. Eligibility was restricted to studies that assessed 6-18 year-old children's performance and response times in listening tasks presented

in noise and/or an impaired voice quality. We searched Medline/Ovid, PsycINFO/Ovid, Eric/Ovid, and Scopus up to August 2018. Risk of bias was determined using an adapted version of the NIH Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies. We classified and discussed findings in the light of three SLP components: speech perception, listening comprehension, and auditory working memory. We identified 24 eligible studies on the effect of noise (n=14), impaired voice (n=8), and the combination of noise and impaired voice (n=2). Sixteen of these studies were evaluated to be of good quality, eight of fair quality. For each SLP component, there was evidence for the disruptive effect of either noise or impaired voice on task performance or response times. However, there was no indication of an interaction between noise and impaired voice. Results from our systematic review suggest that acoustic degradations may impede children's speech perception, comprehension of spoken language, and ability to retain speech-encoded information. This has important implications for the educational setting and highlights the need for improved listening conditions in learning spaces.

40 Speech-on-speech perception of musicians and non-musicians: the role of prosodic cues

Elif C. Kaplan, Deniz Başkent, Anita E. Wagner
UMCG, Netherlands | University of Groningen, Netherlands

In the current study, we investigated the role of prosodic cues in speech-on-speech perception in musicians and non-musicians. Earlier studies have shown that musicians may have an advantage in speech-on-speech performance in behavioral tasks [1,2]. Previously, we have also shown in an eye-tracking study that musical experience has an effect on the timing of resolution of lexical competition when processing quiet vs. masked speech [3]. In particular, musicians were faster in lexical decision-making when a two-talker masker was added to target speech. However, the source of the difference observed between groups remained unclear. Here, by employing a visual world paradigm, we aimed to clarify whether musicians and non-musicians differ in their use of durational cues that contribute to prosodic boundaries in Dutch in resolving lexical competition, when processing quiet vs two-talker masked speech.

The materials consisted of Dutch bisyllabic words, presented in two conditions: matching vs mismatching duration and quiet vs masked. In the matching condition, the duration of the first syllable within a target word (e.g., painter) was matching the duration of that syllable. In the mismatching condition, the duration of the first syllable was mismatched by embedding the recording of a monosyllabic word (e.g., pain), which was longer as it preceded a phrase boundary. Listeners were presented with four pictures that contained the target ("painter"), the competitor ("pain"), and two distractor pictures. If musical training preserves listeners' sensitivity to the acoustic correlates of prosodic boundaries when processing speech, we expected to observe more fixations towards the competitor in the mismatching condition.

We compared gaze-fixations of both groups across conditions (matching vs mismatching duration) and masking (two-talker masker vs quiet). Our results showed that both in quiet and masked speech, musicians exhibited more lexical competition and a delay in resolving the lexical ambiguity in the mismatching duration condition. This indicated that musicians pick up more of the durational cues both in quiet and in masked speech.

References:

1. Başkent D, Gaudrain E. Musician advantage for speech-on-speech perception. J Acoust Soc Am. 2016;139(3):EL51–6.
2. Swaminathan J, Mason CR, Streeter TM, Best V, Kidd G, Patel AD. Musical training, individual differences and the cocktail party problem. Sci Rep. 2015;1–10.
3. Kaplan EC, Wagner AE, Başkent D. Are musicians at an advantage when processing speech on speech? In: Parncutt R, Sattmann S, editors. Proceedings of ICMPC15/ESCOM10. Graz, Austria: Centre for Systematic Musicology, University of Graz; 2018. p. 233–6.

41 Spectrotemporal prediction errors support perception of degraded speech

Matthew Davis
University of Cambridge, UK

Ediz Sohoglu
University of Sussex, Brighton, UK

Speech perception depends not only on signal quality but also on supportive contextual cues or prior knowledge (Sohoglu et al., 2014, JEP:HPP 40:186). Predictive coding (PC) theories provide a common framework to explain the neural impact of these two changes to speech perception. According to PC accounts, neural representations of expected sounds are subtracted from bottom-up signals, such that only the unexpected parts (i.e. 'prediction error') are passed up the cortical hierarchy to update higher-level representations (Rao and Ballard, 1999, Nat. Neurosci. 2:79). Previous multivariate fMRI data (Blank and Davis, 2016, PLoS Biol. 14: e1002577) show that when listeners' predictions are weak or absent, neural representation are enhanced for higher-fidelity speech sounds. However, when listeners make accurate predictions (e.g. after matching text), higher-fidelity speech leads to suppressed neural representations despite better perceptual outcomes. Computational simulations reported by Blank and Davis (2016) demonstrate that these observations are uniquely consistent with prediction error computations, and challenge alternative accounts in which all forms of perceptual improvement should enhance neural representation (Aitchison and Lengyel, 2017, Curr. Opin. Neurobiol. 46:219). In the current work we applied forward encoding models (Crosse et al., 2016, Front. Hum. Neurosci. 10:604) to MEG data and test whether cross-over interactions between signal quality and prior knowledge on neural representations are evident at early stages of processing (within 200 ms of speech input).

We analysed data from a previous MEG study (N=21, English speakers) which measured evoked responses to degraded spoken words (Sohoglu and Davis, 2016, PNAS 13:E1747). Listeners heard noise-vocoded speech with varying signal quality (spectral channels), preceded by matching or mismatching written text (prior knowledge). Consistent with previous findings (Sohoglu et al., 2014, JEP:HPP 40:186), ratings of speech clarity were enhanced by greater spectral detail and matching text.

We report two main MEG findings: (1) MEG responses to speech were best predicted using spectrotemporal modulations (outperforming envelope, spectrogram and phonetic feature representations). (2) We observed a cross-over interaction between clarity and prior knowledge, consistent with prediction error representations; if matching text preceded speech then greater spectral detail was associated with reduced forward encoding accuracy whereas increased encoding accuracy was observed with greater spectral detail following mismatching text. This interaction emerged in MEG responses within 200 ms of speech input, consistent with early computations of prediction error proposed by PC theories. These findings contribute towards the detailed specification of a computational model of speech perception based on PC principles.

42 Practice listening and understanding speech (PLUS): Two novel auditory-cognitive training programs for hearing-impaired listeners

Antje Heinrich
Manchester Centre for Audiology and Deafness, University of Manchester, UK

Helen Henshaw
National Institute for Health Research (NIHR) Nottingham Biomedical Research Centre, Nottingham, UK

Melanie Ferguson
National Acoustic Laboratories, Macquarie, Australia

Our research suggests that auditory training, using an adaptive phoneme discrimination task, results in significant improvements in speech perception and cognition for people with hearing loss (PHL) and for hearing aid (HA) users, and that these improvements are driven by refinements in higher order cognitive control. Furthermore, a recent meta-analysis shows the largest benefits to cognition for PHL may be achieved by combined auditory-cognitive training approaches.

Based on these previous findings we have developed two bespoke auditory-cognitive training programmes that target bottom-up refinement of sensory and cognitive skills (phoneme discrimination n-back training) and the top-down development of cognitive control for speech perception (2-talker competing speech training). Phoneme stimu-

li are those reported by Ferguson et al. (2014, Ear Hear. 35:e110), presented within an n-back odd-one-out paradigm. Novel stimuli for competing speech training are based on challenging listening situations HA users encountered regularly as identified using the qualitative method Photovoice.

The training programs will be provided to first-time HA users across two UK National Health Service (NHS) audiology services to assess the feasibility of conducting a full-scale NHS multicentre RCT of intervention effectiveness and cost-effectiveness.

This is a summary of independent research funded by the National Institute for Health Research (NIHR) RfPB Programme (PB-PG-0816-20044).

43 Good auditory ecology for active and healthy aging

Elisabeth Ingo[1], Valerie Hazan[2], Inga Holube[3], Joerg Bitzer[3], Mary Rudner[1]
1. Linnaeus Centre HEAD, Swedish Institute for Disability Research, Department of Behavioural Sciences and Learning, Linköping University, Linköping, Sweden | 2. Department of Speech Hearing and Phonetic Sciences, UCL, UK | 3. Institute of Hearing Technology and Audiology, Jade University of Applied Sciences, Oldenburg, Germany

Background: One-third of people over the age of 65 have a hearing impairment that affects everyday conversation. Hearing impairment is also associated with several physical, cognitive, and psychosocial health problems. Hearing aids have limited value in noisy environments. Ergonomic listening environments or good auditory ecology is therefore needed. However, we have limited knowledge of the auditory ecology that older people encounter, and how the auditory ecology affects communication for this group. Purpose: The purpose of the project is twofold: Firstly, to investigate the auditory ecology encountered by older people in their daily lives, and secondly to investigate how auditory ecology affects their communication.

Methods: The project includes three studies. In Study 1, we will investigate auditory ecology in real-time using our novel version of Ecological Momentary Assessment (EMA). Our EMA technique combines objective (yet integrity-protected) recording of the acoustic environment, simultaneously with subjective assessment (listening activity, listening effort, speech understanding, motivation, and acoustic environment). In Studies 2 and 3 we will investigate how the auditory ecology identified in Study 1 affects speech recognition (Study 2) and dialogue (Study 3) in older people, under controlled laboratory conditions. In Study 2, we will use the Swedish hearing in noise test (HINT) to study the effect of auditory ecology identified using EMA on speech recognition. In Study 3, we will use Diapix (a 'spot the difference' picture task performed in pairs) to study the effect of auditory ecology identified using EMA on communicative efficiency. In Study 3, we will also study associations between communicative efficiency and listening effort, fatigue, cognitive abilities, motivation, quality of life and psychosocial health. For the project, we will recruit 72 participants aged 65 – 80 years (with equal numbers of men and women). This is required to achieve 95% power for ANOVA's and 80% power for

correlations. Half (36) will take part in all three studies. The other 36 (matched on age and gender to the first group) will take part as dialogue partners in study 3. We will test hearing, screen for cognitive decline and assess quality of life, psychosocial health, and hearing disability for all 72 participants. Implications: The results can be used to facilitate active and healthy aging by promoting ergonomic listening and increased participation. Results will also form a knowledge base for good auditory ecology and better hearing rehabilitation for older people.

44 The effects of linguistic variability and CI processing on voice cue perception

Thomas Koelewijn[1], Floor Arts[1], Etienne Gaudrain[2,1], Terrin N. Tamati[3], Deniz Başkent[1]
1. University Medical Center Groningen, University of Groningen, Department of Otolaryngology, Netherlands | 2. CNRS, Lyon Neuroscience Research Center, France | 3. The Ohio State University, United States

For cochlear-implant (CI) users, speech perception can be challenging, especially in adverse listening conditions (e.g., cocktail party situations). Talkers' voices play an important role in speech perception, since identifying individual talkers can facilitate speech communication in challenging conditions. While previous research has suggested that several linguistic factors broadly influence talker perception, how these factors influence perception of the individual voice cues, and how this is affected by CI processing remains unclear.

The current study investigated the role of linguistic variability in voice cue perception, specifically fundamental frequency (F0) and vocal-tract length (VTL). For normal hearing adults, Just-Noticeable-Differences (JNDs) were obtained using a 3AFC adaptive paradigm. Effects of word status (words, nonwords), token identity (whether the same word/nonword was repeated in the three intervals [fixed], or different items were used [variable]), word characteristics (lexical frequency, neighborhood density), and nonword characteristics (phonotactic probability, neighborhood density) were examined.

Results demonstrated that voice cue perception in normal hearing participants was primarily affected by token identity. JNDs for F0 and VTL were significantly lower for fixed than for variable words and nonwords. This effect was largest for words. For word characteristics, F0 and VTL JNDs were affected by phonological information in nonwords, i.e., phonotactic probability, but not by lexical information, i.e., lexical frequency and neighborhood density. However, VTL JNDs varied less across linguistic conditions than F0 JNDs, suggesting different processing mechanisms for these voice cues. These findings show that linguistic variation interferes with the perception of voice cues, and that the perception of individual voice cues may be closely related to phonological processing.

The observed interactions may be different when listening to degraded speech. How the effect of linguistic variation on F0 and VTL voice cue perception is affected by CI processing was addressed in a follow-up experiment using a similar design, again with normal hearing participants, but this time including CI simulated speech. Preliminary results will be discussed. These outcomes will provide better insight into the interaction of voice cues and linguistic information, which may also improve our understanding of speech perception processes in populations with limitations in voice perception, such as CI users.

45 Simulating sensorineural pathologies to help refine their diagnosis

Jacques Grange, John Culling
Cardiff University, Cardiff, United Kingdom

Sensorineural hearing loss (SNHL) can express itself in many ways, but underlying pathologies cannot currently be finely diagnosed. In order to establish the psychophysical signature(s) of a given pathology, the performance of young normally-hearing listeners is measured as they attend to a SNHL simulator based on a version of the physiological Model of the Auditory Periphery (MAP, Meddis et al. 1986~2018) impaired accordingly. The acoustic stimulus is passed through MAP, then reconstructed from the predicted auditory-nerve (AN) firing response. Identifying psychophysical tests that can help discriminate between pathologies is essential to refining the diagnosis of SNHL, such that targeted treatment may become conceivable.

Early work (SpIN 2019) demonstrated that medial olivo-cochlear (MOCR) and acoustic (AR) efferent reflexes are essential to the faithful coding of the temporal modulations that carry speech information. Efferent reflexes play a key role in the AN dynamic-range adaptation to context level that prevents information loss (e.g. through AN saturation). The simulator was also validated for its normal-hearing version by obtaining speech reception thresholds (SRT) only 1 dB higher than those obtained with unprocessed stimuli.

A first study measured the impact of simulated pathologies on speech intelligibility in noise. While deactivating 70% of ANs or halving the endocochlear potential did not lead to any appreciable SRT inflation, total loss of outer haircells led to a significantly smaller SRT inflation than the 3-4 dB found earlier, when both MOCR and AR were knocked out. Making use of different maskers (broad-band speech-modulated noise, or speech, instead of steady speech-shaped noise) may enable improved pathology discrimination.

A second study aimed to measure the amount of simulated deafferentiation that leads to significant SRT elevation. 90-95% deafferentiation was found to be required to significantly elevate SRTs in steady noise. This outcome is consistent with the effect of stochastic under-sampling predicted by Lopez-Poveda and Barrios (2013, Front. Neurosci. 7:124).

A third study established changes in binaural masking level difference (BMLD) for the pathologies simulated in the first study. This was done by comparing diotic (N0S0) to dichotic (N0Sπ) thresholds of audibility of a 250-Hz tone in white noise, for which a 13-dB BMLD is found with normal hearing (Hirsh and Burgeat, 1958, JASA 30:827). Only a modest BMLD drop was found with simulated pathologies, despite individual thresholds being inflated.

46 Machine Learning Challenges to Revolutionise Hearing Device Processing

Simone Graetzer, Trevor Cox
Acoustics Research Centre, University of Salford

Jon Barker
University of Sheffield

Michael Akeroyd
University of Nottingham

John Culling
Cardiff University

Graham Naylor
University of Nottingham

In this project, we will run a series of machine learning challenges to revolutionise speech processing for hearing devices. Over five years, there will be three paired challenges. Each pair will consist of a challenge focussed on hearing-device processing and another focussed on speech perception modelling. The series of processing challenges will help to develop new and improved approaches for hearing device signal processing for speech. The parallel series of perception challenges will develop and improve methods for predicting speech intelligibility and quality for hearing impaired listeners.

To facilitate the challenges, we will generate open-access datasets, models and infrastructure. These will include: (1) open-source tools for generating realistic test/training materials for different listening scenarios; (2) baseline models of hearing impairment; (3) baseline models of hearing-device speech processing; (4) baseline models of speech perception and (5) databases of speech perception in noise. The databases will include the results of listening tests that characterise how real people, including those who are

hearing impaired, perceive speech in noise, along with a comprehensive characterisation of each test subject's hearing ability. This will allow us to improve on existing knowledge about how best to characterise listeners individually for the purpose of predicting their speech perception in noise.

The data, models and tools we generate will form a test-bed to allow other researchers to develop their own algorithms for speech and hearing aid processing in different listening scenarios. Providing open access to these resources will lower barriers that prevent researchers from considering hearing impairment. Through this, we aim to increase the number of researchers including hearing impairment in their work.

In round one, speech will occur in the context of a 'living room', i.e., a person speaking in a moderately reverberant room with minimal background noise. Entries can be submitted to either the processing or perception challenge, or both. We expect to open round one in October 2020 for a closing date in June 2021 and results in October 2021.

This project involves researchers from the Universities of Sheffield, Salford, Nottingham and Cardiff in conjunction with the Hearing Industry Research Consortium, Action on Hearing Loss, Amazon, and Honda. It is funded by EPSRC. For more information, go to www.claritychallenge.org.

47 Listening effort and oesophageal speech: An EEG study

Sneha Raman[1], Axel Winneke[2], Inma Hernaez[1], Eva Navas[1]
1. AHOLAB Signal Processing Laboratory, University of the Basque Country (UPV/EHU), Bilbao, Spain | 2. Fraunhofer Institute for Digital Media Technology, Branch Hearing, Speech and Audio Technology, Oldenburg, Germany

Listening effort (LE) is being increasingly investigated in several ways as it helps us understand not just how much of the information was understood correctly, but also how taxing it was to listen to the message. Electroencephalography (EEG) has been used to investigate brain activity related to LE. However, this has been done mostly for speech in noise (SPiN) and/or in the context of hearing impairment. EEG experiments for pathological speech — which is intrinsically noisy — have not been explored extensively.

We performed an EEG experiment with participants with normal hearing (N=12) to look at the differences in brain activity when listening to healthy and impaired speech. The impaired speech we have used is oesophageal speech (OS), which is one of the speech production mechanisms adopted following a laryngectomy (removal of larynx). OS — generated using the vibrations of the oesophagus — lacks fundamental frequency and contains unwanted artefacts such as swallowing sounds, which affects the intelligibility, rhythm and prosody of speech. These alterations make OS more effortful to listen to compared to healthy speech. Participants listened to sentences in healthy and impaired speech and were asked to subjectively rate the perceived LE on a 14-point scale ranging

from ‘no effort’ to ‘extreme effort’. A multichannel continuous EEG was recorded while the task was performed. We hypothesised differences in LE when listening to healthy and impaired speech and that these differences should also be reflected in the brain activity. As expected, the subjective data revealed significantly more LE associated with impaired as compared to healthy speech.

The focus for the EEG data analysis was placed on activity in the alpha and delta frequency band. Alpha (8-12 Hz) power has been found to be related to LE for SPiN and suggested to reflect the suppression of task-irrelevant information. Increased delta activity (0.5-3 Hz) in the frontal region has been found to increase with increasing levels of concentration in tasks such as mental calculations. In our EEG data, no significant differences between the two conditions were found with respect to alpha power in the parietal region. However, there was a significant increase in the delta band power when listening to impaired as compared to healthy speech, particularly in frontal regions. These results suggest that the difficulty in listening to impaired speech may not be attributed to requiring suppression of distractors but with needing increased concentration to perceive it.

48 Listening effort of natural speaking styles

Maria Koutsogiannaki[1], **Olympia Simantiraki**[2], Martin Cooke[3], Marie Lallier[1]
1. Basque center on Cognition Brain and Language, Donostia, Spain | 2. University of the Basque Country, Vitoria-Gasteiz, Spain | 3. Ikerbasque (Basque Science Foundation), Vitoria-Gasteiz, Spain

According to the hyper-hypo model of speech communication two principles govern human speech production; efficacy and economy. From the speaker’s side, there is a constant negotiation between maximizing clarity and minimizing production effort which leads to natural speech adaptations from effortless to effortful and vice versa that aim to maintain intelligibility in dynamically changing listening environments. From the listener’s side, speech in noise studies have shown that speaker’s adaptations are beneficial for the listener, reducing listening effort.

In this work, we aim to measure objectively listeners’ effort on processing natural speaking styles. Unlike other studies that vary intelligibility levels by degrading (e.g. vocoded speech) or modifying speech properties (spectral boosting, duration transformations), our main focus is on the cognitive demands of processing natural speech. Physiological responses to clear, casual and Lombard speaking styles in quiet, reverberation and cafeteria noise were measured using pupil dilation metrics. A total of 40 normal-hearing Spanish natives were tested using a combination of the above speaking styles and the masker types at a specific signal-to-noise ratio level for restaurant and reverberation estimated to induce 70% of intelligibility on the style of speech considered most effortful, namely casual speech. To evaluate the performance of pupil dilation metrics, subjective evaluations have also been collected on speech intelligibility through oral responses on the attended stimuli and on listening effort in the form of questionnaires.

A mixed effects model with subjective listening effort and intelligibility scores as fixed factors and participant as random factor, revealed significant effects and interactions on peak pupil dilation. Both masker type and speaking style contributed to this significance. Interestingly, the peak pupil dilation metric follows the subjective listening effort pattern of the speaking style which may suggest that it reveals the cognitive effort of processing speech.

Last, we introduce the phonological awareness theory to explain participants' variability on listening effort and physiological responses. Phonological awareness capacity is participants' ability to manipulate and process speech sounds. It has been found that for participants with impeded phonological processing, listening effort increases significantly in noisy conditions. To estimate participants' phonological awareness, we designed a battery of phonological tasks for adults. Surprisingly enough, correlation analysis on the phonological scores and listening effort evaluations and on the phonological scores and physiological metrics did not show any significant relationships.

49 Selecting laboratory test scenarios

Karolina Smeds, Florian Wolters, Sarah Gotowiec, Petra Herrlin, Josefina Larsson, Martin Dahlquist
ORCA Europe, WS Audiology, Sweden

When performing hearing-related laboratory tests, a selection of test scenarios is needed. Traditionally, various speech situations (in quiet or in noise) have been implemented, with varying degree of ecological validity. Some research groups suggest a set of "prototype listening situations" that can be used for laboratory testing. None of these sets have been widely adopted. Other research investigates the listening situations encountered in everyday life in order to learn more about people's auditory reality. This information could potentially be used when selecting laboratory test scenarios. Ecological Momentary Assessments (EMA) is one method that can be used to investigate everyday listening situations as they occur. EMA methodology increases ecological validity by studying individuals' real time experiences rather than retrospective reports. After a summary of previous work on prototype listening situations, this presentation will show how recently collected EMA data from our laboratory can be used to learn more about common, important, and difficult everyday listening situations.

The results of our study showed that around one third of EMA reports were related to speech communication, and one quarter to focused listening (mainly to TV or radio). Of note, there were also a considerable amount of non-speech, non-active listening reports. When focusing on the very important situations (constituting 24% of the reports), more than half of the situations were of the speech communication type. When investigating only very important situations that occur "almost daily" (13% of the reports), the focused listening situations, mainly listening to TV or radio, constituted half of the situations. When focusing on the very difficult situations (constituting 8% of the reports),

again around half of the reported situations were of the speech communication type. However, among the very difficult situations, non-active listening situations constituted almost one quarter of the reports. These non-active listening situations mainly took place in noise.

Based on prior literature and our current data, we will suggest a small set of relevant laboratory test scenarios.

50 Interaction of acoustic and semantic context information on phonetic identification

Loriane Leprieur
LLING - Laboratoire de linguistique de Nantes, FR

Olivier Crouzet
LLING - Laboratoire de Linguistique de Nantes, FR

Etienne Gaudrain
CNRS, Lyon Neuroscience Research Center, FR

Background : Speech comprehension depends on both phonetic and semantic processes. Phonetic mechanisms aim at categorizing phonemes from the acoustic input. Various sources of information are available among which some are intrinsically related to the current segment (formant frequencies, formation ratios) while others relate to extrinsic information relating to acoustic context (e.g. speaker-specific properties associated with neighbouring acoustic information). These processes would interact with a semantic analysis of the message in order to provide an optimal interpretation of the signal. The influence of semantic information on phonetic identification is also viewed as a context effect. Therefore, two different sources of context information are available to listeners : phonetic context is associated with how listeners adapt to acoustic information neighbouring the target segment, while semantic context is associated with how the identification of neighbouring words along with sentence interpretation affect speech perception. Both context mechanisms may interact with bottom-up acoustic analysis in order to reach the final categorisation.

Our aim is to investigate how these two sources of contextual information interact with each other in speech perception. More precisely, we want to assess how the acoustic information is processed and related to contextual information when acoustic and semantic contexts interact together.

Method : In order to address this issue, we designed French linguistic material in which a target word is preceded by a subject-verb structure which is expected to favour one interpretation over the other. Target words were selected in order to constitute CV or CVC word pairs exhibiting vowel alternations associated with distinct but acoustically close vowels (e.g. « balle » (/bal/) / « belle » (/bɛl/)). Three semantic contexts were designed for each word pair (Context 1 favours Word 1, Context 2 favours Word 2, Context

0, neutral, favours none). These semantic contexts are combined with synthetic modifications of the formant frequencies in the first part of the carrier sentence (excluding the target word). The final word is selected among a set of words resynthesized along a 7-step formant continuum (F1/F2/F3/F4) in order to provide measurements to compute psychometric curves. 2-AFC classification frequencies and psychometric modelling are computed in order to investigate how acoustic and semantic information interact. Our predictions are that acoustic contextual information will compete with semantic information. The results will then provide the basis for modelling uncertainty processing in speech perception. Data collection is still in preparation and preliminary results will be presented during the conference.

51 Perceptually trained end-to-end FFTNet neural model for single channel speech enhancement

Muhammed P.V. Shifas
PhD Research Scholar, Speech Signal Processing Lab (SSPL), University of Crete (UoC), Greece

Nagaraj Adiga, Vassilis Tsiaras
Researcher, SSPL, UoC, Greece

Yannis Stylianou
Professor, SSPL, UoC, Greece

Single channel speech enhancement is challenging task. Recent advancements in machine learning (ML) show that the combination of linear and non-linear operators are able to model the complex characteristics of signals, including that of speech and noise. However, the modeling power of ML models highly depends on the layer-wise design of their architecture as well as the loss function on which the parameters are optimized. Though different neural approaches have been suggested to suppress the noise artifacts for speech enhancement, not many have tried to explore the statistical differences between speech and noise signals in a mixture. Further, most existing models performance were optimized on the waveform domain of speech, ignoring the frequency selective aspects of human auditory perception.

To address these constraints, recently, we had suggested a parallel, non-causal, waveform domain end-to-end FFTNet neural architecture. In this work, we suggest an extension of the FFTNet model optimized on the perceptually salient spectral domain of the enhanced signal for single channel speech enhancement. The proposed model has a dilation pattern which resembles to the classical FFT coefficient computing Butterfly structure. In contrast to other waveform based approaches like WaveNet, FFTNet uses an initial wide dilation pattern. Such an architecture better represents the long term correlated structure of speech in the time domain. On the contrary noise is usually highly non-correlated in such a wide dilation pattern. To further strengthen this feature of FFTNet, we suggest a non-causal FFTNet architecture, where the present sample in

each layer is estimated from the past and future samples of the previous layer. By optimizing the parameters on the spectral domain objective, the suggested model can better learn the features which are perceptually more significant than the temporal training approach. We refer to that as SE-FFTNet.

The suggested SE-FFTNet model has shown considerable improvement over existing models like WaveNet or SEGAN while having far-lesser parameters. Perceptually, it improves PESQ by up to 14% over WaveNet and SEGAN. In terms of reconstructed spectral components, the Log-spectral-distortion (LSD) has been reduced by 7.6% over SEGAN and 16.1% over WaveNet. Informal listening tests and objective metrics confirm that the suggested model optimized on spectral objective produces better enhancement than the same model trained by minimizing sample objective function (improving PESQ 10% and LSD 6.9%). Formal listening tests are ongoing.

52 Can people with hearing loss benefit from speech training at home?

Maja Serman, Kaja Kallisch
Sivantos GmbH, Erlangen, Germany

Speech training is known to improve speech understanding in noise for some normal hearing, hearing aid and cochlear implant users. Different speech training methods are currently available, depending on the speech components used for the training (phoneme, word, sentence or a mixture of these). Previously, we have shown that phoneme-based training in the lab improves speech understanding of untrained material in noise for normal hearing listeners, hearing aid and cochlear implant users (Serman, 2012; Schumann et al., 2015; Schumann et al., 2017). Here, we report on the benefits of app-based, self-performed phoneme training at home for fifteen subjects with mild to moderate hearing loss.

Objective measures showed that, as expected, subjects improved on trained material (phoneme recognition in quiet and in noise), whereas the performance on untrained monosyllabic words in quiet (fixed level) did not change before and after training. The most interesting finding was that phoneme training at home improved speech reception thresholds in noise, for untrained material.

References:
Serman, M., 2012, Paper presented at the 15th Annual meeting of the German Society of Audiology, 2012; Erlangen, Germany
Schumann et al., 2015, International Journal of Audiology, 54(3):190-198
Schumann et al., 2017, Laryngorhinootology 96(2): 98-102

53 What are some of the challenges in dynamic cocktail party listening?

Moritz Wächtler[1], Fabian Wenzel[1], Josef Kessler[2], Martin Walger[3], Hartmut Meister[1]
1. Jean Uhrmacher Institute, University of Cologne, Germany | 2. Department of Neurology, University Hospital Cologne, Germany | 3. Clinic of Otorhinolaryngology, Head and Neck Surgery, University of Cologne, Germany

In everyday life, listeners are often confronted with situations in which multiple talkers speak simultaneously. Those so-called cocktail-party situations can be static, that is, the target talker always remains the same, or they can be dynamic, meaning the target talker changes in an unpredictable manner. Previous studies have shown that, in dynamic situations, the listener's speech intelligibility often decreases after a transition from one target talker to the other. Afterwards it takes some time until the speech intelligibility is back to its initial state again. This study examined which factors contribute to the drop in speech intelligibility after a transition. To this end, a setup involving three competing talkers uttering matrix sentences was used (cf. Meister et al., this conference). The talkers differed in terms of their voices (medium male, deep female and high female) and spatial positions (-60°, 0° and +60° azimuth angle). The target talker was unknown to the listeners in advance and was indicated by the first word of the sentence. Depending on the condition, target talkers changed either after each trial (transition probability = 100%) or only in 20% of the trials. Two different transition types were used: 1) The talkers either remained at the same positions and the target voice changed; 2) the talkers switched positions but the target voice remained the same. Fourteen younger normal-hearing and 18 older normal-hearing listeners participated in the study.

This poster will present an analysis of the influence of different factors on speech intelligibility after transitions. These factors include the target talker position, the transition probability and the individual hearing thresholds. Furthermore, we will address the question of how fast listeners were able to switch attention from one talker to the other.

Funding: Supported by Deutsche Forschungsgemeinschaft (ME2751/3-1).

54 Pupillary correlates of auditory emotion recognition in older hearing-aid users

Julie Kirwan[1], Anita E. Wagner[1], Peter Derleth[2], Julia Rehmann[2], Deniz Başkent[1]
1. University Medical Center Groningen, Netherlands | 2. Sonova AG, Switzerland

Hearing-impaired (HI) individuals are known to have difficulties in auditory emotion recognition tasks compared to normal hearing individuals. It is still unclear if this is due to difficulties in the lower levels of auditory processing or to the categorisation of emotions that is involved in the experimental task (Picou et al., 2018, Tr. Hear. 22:4). An objective index of emotion recognition can be observed in pupil dilations, which have recently been shown to dilate more for emotionally meaningful speech in comparison to emotionally neutral speech (Jürgens, Fischer and Schacht, 2018, Front. Psychol 9:13). In this study, we investigated pupil dilations as a measure of emotion recognition in an older HI population, all users of hearing aids, and by correlating this measure with behavioural responses we aimed to gain insight into the potential difficulties that this population faces with an emotion recognition task. These difficulties may arise from this population being unable to hear the signal clearly, or due to the effect of cognitive involvement in the experimental task. We further hypothesised that high-frequency information is important for emotion recognition, so we tested our participants both with and without a frequency lowering enabled hearing aid.

We fitted 8 older HI participants, who had moderate to severe sloping high-frequency hearing loss, with frequency lowering enabled hearing aids for an acclimatisation period of 3-6 weeks. We recorded their pupil dilations in response to emotional speech with and without frequency lowering, during both a passive and active-listening condition. The active condition included a behavioural emotion identification task, where participants were given a forced-choice task after each stimulus was presented.

Preliminary results indicate that the pupillary response of the passive-listening condition reveals that emotional arousal can be elicited in an older HI population, and even when only passively listening to emotional speech. Functional differences are seen between the passive and active listening conditions, indicating that the cognitive involvement elicited by the emotion categorisation task is reflected by the pupillary responses. Frequency lowering does not provide benefits on emotion categorisation at the group level, but may have benefits for individual participants.

55 Static and dynamic cocktail party listening – Effects of age-related hearing loss

Hartmut Meister, Moritz Wächtler, Fabian Wenzel
Jean Uhrmacher Institute, University of Cologne, Germany

Josef Kessler
Department of Neurology

Martin Walger
Clinic of Otorhinolaryngology, Head and Neck Surgery

Verbal communication often involves situations with several talkers speaking simultaneously. These "cocktail party" situations are typically "dynamic", since the talker of interest may change in a possibly unpredictable manner. However, clinical assessments and research mainly consider "static" cocktail party listening with one fixed target talker and the competing talker(s) serving primarily as masker(s).

Recently, it has been shown that different types of attention – such as focusing, "dividing" and switching attention - are required in static and dynamic cocktail party listening (Meister et al., 2019). Especially the need to divide and switch attention is associated with cognitive "costs" reflected in decreasing performance (Brungart and Simpson 2007, Lin & Carlile 2015, Meister et al., 2019) and increasing reaction times (Oberem et al., 2017).

Older people appear to be at a particular disadvantage in cocktail party listening due to age related hearing loss and decline in several cognitive abilities, such as attention, working memory, and executive functions. Notably, however, older listeners with near-normal hearing and good cognitive skills did not perform significantly worse in static and dynamic cocktail party situations than younger subjects (Meister et al., 2019): Especially the ability to switch attention was largely preserved whereas older listeners tended to show more problems with dividing attention.

The present study shows results of an ongoing investigation into the possible effects of hearing impairment on static and dynamic cocktail party listening. Based on a paradigm with three competing talkers and different types and probabilities of switching the talker of interest (cf. Wächtler et al., this conference), we consider older listeners with near-normal hearing to moderate impairment. We hypothesize that the combination of hearing loss and attentional load is especially detrimental for the demanding dynamic condition resulting in higher costs compared to the static condition.

Supported by Deutsche Forschungsgemeinschaft (ME2751/3-1).

References:
Brungart, D.S., Simpson, B.D., 2007. Cocktail party listening in a dynamic multitalker environment. Percept. Psychophys. Jan;69(1):79-91.
Lin, G., Carlile, S., 2015. Costs of switching auditory spatial attention in following conversational turn-taking. Front. Neurosci. Apr. 20;9:124. doi: 10.3389/fnins.2015.00124.
Meister, H., Wenzel, F., Gehlen, A., Kessler, J., Walger, M. Attentional mechanisms in static and dynamic cocktail-party listening. Proceedings of the 23rd International Congress on Acoustics, September 9 to 13, 2019, Aachen, Germany
Oberem, J., Koch, I., Fels, J., 2017. Intentional switching in auditory selective attention: Exploring age-related effects in a spatial setup requiring speech perception. Acta Psychol. (Amst). Jun;177:36-43. doi: 10.1016/j.actpsy.2017.04.008.

56 Evaluation of the performance of a model-based adaptive beamformer

Alastair H. Moore, Rebecca R. Vos, Patrick A. Naylor, Mike Brookes
Imperial College, London, UK

Adaptive beamforming has great potential to improve the performance of hearing aids, provided that the characteristics of the signal required by the design procedure are estimated to a sufficient degree of accuracy. Continuously updating the properties of the interfering noise field, such as the estimated noise covariance matrix, allows an immediate response to changes in the acoustic scene. This has the potential to not only improve noise cancellation itself, but key parameters such as the response time of the model.

A recent conference presentation by Naylor et al [1] proposed a method for improving the robustness of adaptive beamformers, using a straightforward model for the sound field. Simulation experiments were conducted using measured reverberant impulse responses and challenging levels of realistic noise (including distributed babble and white noise, with interfering male and female speakers), and the results showed that the proposed adaptive beamforming method outperformed a fixed beamformer by ≥ 1 dB over a range of acoustic scenarios.

In the current study, the performance of the method proposed in [1] is re-evaluated using real measurements taken from the Oldenburg database [2], an eight-channel database of head-related impulse responses (HRIRs) and binaural room impulse responses (BRIRs), including BRIRs for multiple, realistic head and sound-source positions in four natural environments reflecting daily-life communication situations.

As in [1], the performance of the proposed method is compared to a baseline method assuming cylindrically isotropic noise. The metrics used to assess the performance are (1) the noise reduction, (described by the change in frequency-weighted SNR), (2) the speech intelligibility (using the short-time objective intelligibility (STOI) algorithm), (3) the speech quality (using the perceptual evaluation of speech quality (PESQ)), and (4) the robustness of the method.

References:

[1] PA. Naylor, AH. Moore, M. Brookes, Improving Robustness of Adaptive Beamforming for Hearing Devices, 2019, International Symposium on Auditory and Audiological Research (ISAAR), Topic: Auditory Learning in Biological and Artificial Systems, August 21-23, Nyborg, Denmark.

[2] H. Kayser, SD. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, Database of Multichannel In-ear and Behind-the-ear Head-related and Binaural Room Impulse Responses, EURASIP J. Adv. Signal Process, 2009, 6:1--6:10.

57 Comparison of ideal mask-based enhancement methods for highly degraded speech

Simone Graetzer
University of Salford, UK

Carl Hopkins
University of Liverpool, UK

This paper compares the performance of three ideal mask-based enhancement methods for speech mixed with white Gaussian noise at very low signal-to-noise ratios (SNRs). Ideal masks, which require a priori knowledge of both the target signal and the masker, set the upper limit of what can be achieved if the instantaneous or 'local' SNR estimator is accurate and reliable. The standard ideal binary mask (IBM) is constructed by means of a binary classification of sound sources as target or interferer in the time-frequency domain. In each time-frequency bin, the local SNR is compared with a threshold referred to as a 'Local Criterion' (LC). When the LC is exceeded, a value of one is assigned to the mask. Otherwise, a value of zero is assigned. Subsequently, regions of the mixture signal in which the mask contains zeros are removed. In this study, the IBM was compared with an alternative binary mask comprising [0.1,1] rather than [0,1] gains, and an ideal ratio mask (IRM), which can take any value between 0 and 1. Speech produced by twelve speakers of British English was mixed with white Gaussian noise at SNRs between -29 and -5 dB before enhancement. The results demonstrate that IRMs can be used to obtain near maximal speech intelligibility even at very low mixture SNRs. Some benefits were found of raising the lower gain of [0,1] masks to 0.1 for LC = 0 when the SNR was greater than or equal to -20 dB. These are likely to be due to a reduction in the crudeness of the binary switching and therefore the audibility of artefacts. The results indicate the importance of mask density when mixture SNRs are low, where mask density is defined as the number of ones in the mask.

Unfortunately this author could not make it to present their poster. We have left the abstract for your information.

57 Studying the effects of background noise on preschool children's novel word learning using a multi-session paradigm.

Meital Avivi-Reich[1,2], Tina M. Grieco-Calub[2]

1. Communication Arts, Sciences and Disorders, Brooklyn College of City University of New-York (CUNY), NY, United States | 2. The Roxelyn & Richard Pepper Department of Communication Sciences and Disorders, Northwestern University, IL, United States

Common daily environments often contain background noise that disrupts access to speech. Thus, it is reasonable to suspect that background noise imposes challenges for individuals who are attempting to learn something from the speech input, such as new words. Extant data on the effects of background noise on novel word learning, however, have shown mixed results. One possible reason for this observation is the variability in methodology across studies, including the age of participants, the types of background noise used, and the way that word learning has been quantified. In addition, most studies assess the effect of background noise in a single session, which limits the ability to test the effect of background noise over multiple exposures of the word. The present line of studies aims to add to this body of work by comparing novel word learning of preschool-age children between a multi-session paradigm and a one-day paradigm while controlling for the number of exposures to the novel word. Preschool-age children were chosen for this study because speech input is the sole way that they learn about words in their native language and because they are more susceptible to the negative effects of background noise due to their immature language and cognitive abilities. In this study, children were exposed to two stories presented through a movie, with each story containing four novel CVC words. Children were exposed to both stories, one in quiet and one in the presence of four-talker babble presented at 0 dB signal-to-noise ratio. After each story, receptive word learning was quantified with a four-alternative-forced-choice task, and expressive word learning was quantified by the number of novel labels correctly produced when their corresponding objects were shown to the children. In the first study, eight children were exposed to the two stories once during each experimental session and were required to complete five sessions over the course of two weeks. In the second study, eight children were exposed to each of the stories five times during the first experimental session and once during the second experimental session a few days later. Results suggested that children's receptive and expressive word learning improved by session for children who participated in the multi-session paradigm versus for children who participated in the one-day paradigm. Greater improvement was observed for the words exposed in quiet. The results and their implications will be further discussed.

58 Measuring the benefit of NELE algorithms for hearing aid users in realistic scenarios with the AFC-MHA platform

Carol Chermaz
The Centre for Speech Technology Research, University of Edinburgh, United Kingdom

Matthias Vormann, Kirsten Wagener
Hörzentrum Oldenburg GmbH, Oldenburg, Germany

Volker Hohmann
Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, Oldenburg, Germany

When determining the hearing profile of a listener, it is commonplace to measure the SRT (Speech Reception Threshold) with non-reverberant speech against artificially created speech-shaped noise. However, such conditions do not reflect real-world acoustic environments in which speech communication is actually experienced.

In this study we measured the psychometric functions for 20 hearing aid users in two realistic acoustic scenes, which are representative of everyday scenarios: a living room and a busy cafeteria. All the subjects were native German listeners (mean age: 73 years, hearing profiles N3/N4). The German Matrix test was used as speech corpus; binaural recordings of real-world noise and impulse responses were used in order to recreate the acoustic scenarios.

In order to preserve an accurate representation of spatial cues while providing an adequate compensation for hearing loss, we presented the stimuli via headphones using the openMHA [1] as a simulation of hearing aids. The audio output of the AFC [2] test platform was routed to the openMHA in real time via the Jack Audio Connection Kit. Stimuli were played back at realistic presentation levels, i.e. 65 dBA for the living room and 75 dBA for the cafeteria, while speech was scaled to match the desired SNR. Given the intensity of the stimuli and the increased loudness sensitivity of the HI (Hearing Impaired) subjects, we used the compressive CR2-NALRP fitting rule, which is based on [3]. Individual SRT50 and slope of the psychometric functions were estimated concurrently with an adaptive procedure.

The results of this study will be used as a starting point for an evaluation of NELE (Near-End Listening Enhancement) algorithms for HI subjects in realistic noise, which follows in the footsteps of a 2019 study with NH native English listeners [4].

Funding: This project has received funding from the EU's H2020 research and innovation programme under the MSCA GA 67532 (the ENRICH network: www.enrich-etn.eu).

References:
1. Herzke, Tobias, et al. “Open signal processing software platform for hearing aid research (openMHA).” Proceedings of the Linux Audio Conference. 2017.
2. Ewert, Stephan D. “AFC—A modular framework for running psychoacoustic experiments and computational perception models.” Proceedings of the international conference on acoustics AIA-DAGA. 2013.
3. Grimm, Giso, et al. “Implementation and evaluation of an experimental hearing aid dynamic range compressor.” Threshold 80.90 (2015): 100.
4. Chermaz, Carol, et al. “Evaluating Near End Listening Enhancement Algorithms in Realistic Environments.” Proc. Interspeech 2019 (2019): 1373-1377.

Unfortunately this author could not make it to present their poster. We have left the abstract for your information.

59 The effect of sound source diffuseness on speech perception in young and older adults

Meital Avivi-Reich
Communication Arts, Sciences and Disorders, Brooklyn College of City University of New-York (CUNY), NY, United States

Bruce A. Schneider
Psychology Department, University of Toronto Mississauga, Canada

The variety and nature of auditory scenes have changed significantly over the years due to electronic amplification. It is important to understand how these changes affect speech perception across the lifespan. When amplification is used, each sound source is often presented over multiple loudspeakers, which can alter its timbre, and introduce comb-filtering effects. Increasing the diffuseness of a sound by presenting it over several spatially-separated loudspeakers might affect the listeners’ ability to form a coherent auditory image of it, alter its perceived location, and may even affect the extent to which it competes for the listener’s attention. In addition, it can lead to comb filtering effects that can alter the spectral profiles of sounds arriving at the ears. The current study aims to systematically study the effects of different amplified acoustic scenes on speech perception in young and older adults.

In this study, 24 young adults and 24 older adults were asked to repeat nonsense sentences presented in either noise, babble or competing speech maskers. Participants were divided into two experimental groups; 1) A Compact-Target Timbre group where the target sentences were presented over a single central loudspeaker (compact target), while the masker was either presented over three loudspeakers (diffuse) or over the same single loudspeaker (compact); 2) A Diffuse-Target Timbre group, where the target sentences were diffuse while the masker was either compact or diffuse. The sentences presented under each of the four timbre conditions were played in 4 different SNRs for each type of masker (Noise, Babble, Speech). The correct repetition of the three target words in each sentence was recorded and the 50% correct corresponding SNR thresholds and the slopes of the psychometric functions were calculated and analyzed.

The current results show that in the absence of a timbre-contrast, when both the masker and the target were either compact or diffuse, the results in both conditions were similar. However, when there was a timbre-contrast, the signal-to-noise ratios needed for 50% correct recognition of the target speech were higher (worse) when the masker was compact, and lower (better) when the target was compact. In addition, older adults had higher SNR thresholds than the young adults under the different timbre conditions but demonstrated similar pattern of results. The possible implications of amplification, and how they may differ depending on the listener's age, will be discussed.

60 Using automatic speech recognition to predict aided speech-in-noise intelligibility

Lionel Fontan, Maxime Le Coz
Archean LABS, Montauban, France

Jérôme Farinas
IRIT - Université de Toulouse, France

Bertrand Segura
Ecole d'Audioprothèse de Cahors, Université Paul Sabatier - Toulouse III, France

Michael Stone
Manchester Centre for Audiology and Deafness, School of Health Sciences, UK

Christian Füllgrabe
School of Sport, Exercise and Health Sciences, Loughborough University, UK

As the main complaint of people with age-related hearing loss (ARHL) is difficulty understanding speech, the success of rehabilitation through hearing aids (HAs) is often measured through speech intelligibility tests. These tests can be fairly lengthy and therefore cannot be conducted for all HA settings that might yield optimal speech intelligibility to the hearing-impaired listener.

Recent studies showed that automatic speech recognition (ASR) can be used as an objective measure for the prediction of unaided speech intelligibility in quiet in people with real or simulated ARHL (Fontan et al., 2017; Fontan et al., in revision). The aim of the present study was to assess the applicability of ASR to a wider range of listening conditions, involving unaided and aided speech-in-noise perception in older hearing-impaired (OHI) listeners.

Twenty-eight OHI participants (mean age = 73.3 years) were recruited for this study. They completed several speech-identification tasks, involving logatoms, words, and sentences. All speech materials were mixed with a background noise with the long-term average speech spectrum (LTASS) and presented monaurally through headphones at 60 dB SPL. The signal-to-noise ratio was -1.5 dB. Participants completed the identification tasks unaided and aided using a HA simulator implementing individual gains prescribed by the CAM2b fitting rule.

A speech-intelligibility prediction system was set up, consisting of: (1) the HA simulator used for the OHI participants (Moore et al., 2010), (2) an age-related-hearing-loss simulator implementing the algorithms described by Nejime and Moore (1997), and (3) an HMM-GMM-based ASR system using the Julius decoder software (Nagoya Institute of Technology, Japan), with acoustic models trained on speech in LTASS noise, and a different language model for each of the speech materials. Human and machine intelligibility scores were calculated as the percentage of logatoms or words that were correctly identified.

The results show that, on average, the implementation of CAM2b gains significantly improved speech-in-noise intelligibility performances both in OHI listeners and the ASR system.

61 Near-end listening enhancement in cars

Enguerrand Gentet
Groupe PSA, France

Gaël Richard
LTCI, Téléecom Paris, Institut Polytechnique de Paris, France

Bertrand David
LTCI, Téléecom Paris, Institut Mines-Télécom, France

Sébastien Denjean, Vincent Roussarie
Groupe PSA, France

Near-end listening enhancement is a growing field of research that aims at increasing intelligibility of speech signals in noisy environments. Voice transformation techniques are usually used under the constraint of keeping the Signal to Noise Ratio (SNR) unchanged. In our work we propose a slightly different approach where different near-end listening methods applied to in-car noisy environments are studied under the constraint of a fixed perceived loudness.

The first method consists of an adaptive equalizer which reallocates the energy of frequency bands to maximize the Speech Intelligibility Index (SII). Perceptual tests have been carried out and demonstrate a variable performance of the algorithm depending on the shape of the noise spectrum. We also highlighted limitations of perceptual tests based on the Speech Reception Threshold (SRT) as it does not reflect real-life situations.

The second method is based on deep parallel learning models which automatically learn the voice transformations from speech datasets. We introduced a novel duration modification feature and we studied the use of recurrent architectures combined with wavelet description of features. Objective results and preliminary listening tests show the merit of this approach.

62 Speech in noise perception and sound localization, relationship with pure tone audiometry in unilateral hearing loss

Mariam Alzaher, Mathieu Marx, Pascal Barone
CerCo-CNRS, Toulouse, France

Binaural hearing yields different cues necessary for sound localization and speech recognition performances. These cues are altered in the case of monaural hearing due to the alteration of neural mechanisms supporting interaural level differences and interaural time differences. This alteration may vary according to the severity of deafness. In this study, we investigated sound localization and speech in noise performance for 21 unilateral hearing loss (UHL) patients compared to a population of 20 normal hearing subjects (NHS) subjects in binaural condition and fitted with an earplug for the simulation of monaural condition. UHL patients were assigned into two groups according to their pure tone averages (PTA; mild<72.5 dB and severe >72.5 dB). Our findings demonstrated a deficit in speech in noise and in spatial performances after unilateral hearing loss compared to NHS. The results of speech in noise evaluation reported better sppech reception thresholds (SRT) for moderate hearing loss compared to the sever population only in dichotic condition, suggesting a possible link between PTA and SRT. Furthermore, moderate UHL showed better localization accuracies than severe UHL in addition to a strong correlation between PTA and root mean square errors (RMS). However, we did not find a correlation between RMS errors and SRT levels. In addition, when compared to NHS, speech in noise deficit was present for moderate and severe UHL; however, moderate UHL demonstrated higher localization accuracies compared to NHS with monaural plug. These results suggest that pure tone audiometry can be a good predictor of sound localization and possibly, in speech in noise perception in the dichotic condition; however, spatial abilities and speech in noise recognition are two different processes that are not directly related at the behavioral level. UHL patients seem to rely more on monaural spectral cues for spatial location rather that speech in noise segregation.

63 Relating the speech-derived frequency-following response to speech intelligibility in noise

Tijmen Wartenberg[1], Markus Garrett[2], Sarah Verhulst[1]
1. Hearing Technology @ WAVES, dept. of Information Technology, Ghent University, Ghent, Belgium | 2. Medizinische Physik and Cluster of Excellence "Hearing 4 All", Dept. of Medical Physics & Acoustics, University of Oldenburg, Oldenburg, Germany

The envelope following response (EFR) is a brainstem auditory evoked potential (AEP) to modulated sound and its strength was shown to relate to the individual ability to detect amplitude-modulation perceptually. Recent work from our research group has shown that the EFR to a modulated square-wave stimulus is sensitive to age-related temporal-envelope processing deficits associated with cochlear synaptopathy This raised our interest in using AEPs as an objective tool to detect problems in speech perception.

Here, we study whether the AEP extracted from a speech token can also predict individual speech recognition performance. Brainstem AEPs contains phase-locked information to the harmonics of the sound, including the fundamental frequency (F0) of the speaker. Because simulations using an auditory model of the auditory periphery suggest that phase-locking to F0 is reduced with cochlear synaptopathy, we hypothesized that the phase-locked brainstem response to the harmonic content of the presented speech token might relate to speech intelligibility. We further investigated whether adding stationary background noise would affect the relationship.

Speech AEPs were analyzed in a group of young normal hearing (yNH), elderly normal hearing (oNH) and elderly hearing-impaired (oHI) participants to 3000 iterations of the CV /da/ in quiet and in speech-weighted noise, spoken by a male speaker. A linear regression analysis was performed with the EFR (adding + and - polarities) and spectral FFR (subtracting + and - polarities) as predictor variables for the speech-reception threshold to broadband and filtered (<1.5 kHz; > 1.65 kHz) sentences from the German OLSA test.

Our results confirm that the amplitude of the speech-derived EFR matched the EFR to amplitude-modulated stimuli in the same listeners well. The EFR was strongest in the yNH group, followed by the oNH group and oHI group. The group ranking of the EFR explained group differences in the speech-reception threshold for high-pass filtered speech (> 1.65 kHz) in noise. Significant within group correlations were found using either the EFR or spectral FFR to predict speech performance. However, no major differences were found between the responses recorded in the quiet and noise condition. Our results suggest that the brainstem AEP to a short CV can to some extent inform about the speech reception threshold.

64 Using fNIRS to explore emotional prosody perception

Ryssa Moffat[1,2], David McAlpine[3], Deniz Başkent[4], Robert Luke[3], Lindsey van Yper[3]
1. International Doctorate of Experimental Approaches to Language and Brain (IDEALAB), University of Potsdam, Germany; University of Groningen, Netherlands; Newcastle University, UK; and Macquarie University, Australia | 2. Department of Cognitive Science, The Australian Hearing Hub, Macquarie University, Sydney, Australia | 3. Department of Linguistics, The Australian Hearing Hub, Macquarie University, Sydney, Australia | 4. University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

Recognising emotional prosodies in speech is a key element in verbal communication. Recipients of cochlear implants (CIs) with good speech recognition perform below normal hearing (NH) peers on emotional prosody recognition tasks. Evidence from behavioural studies indicates that CI recipients rely on temporal and intensity cues to compensate for the device's poor transmission of spectral cues. Little is known about the brain mechanisms underlying emotional prosody recognition in CI hearing. We employed the brain-imaging tool functional near-infrared spectroscopy (fNIRS) to examine cortical processing of emotional prosody in normal-hearing (NH) listeners and to prepare an appropriate paradigm for CI recipients.

Forty NH adults participated in a behavioural forced-choice listening task and a fNIRS passive listening task. Six-syllable sentences with pseudo content words and real function words were used in both tasks (e.g., "the larfle is himber"). Stimuli were recorded with prosodic features adjudged neutral, happy, sad, fearful and angry. To examine each group's reliance on acoustic cues, stimuli were manipulated in four separate ways: 1) pitch cues equalised, 2) pitch and intensity cues equalised, 3) pitch and rate cues equalised, and 4) rate and intensity cues equalised. In the forced-choice listening task, participants identified the emotion conveyed in each stimulus (N=100). During the passive listening fNIRS task, participants heard 10 blocks of each emotion-condition pair (N=20). Comparisons will be made between prosodies, as well as within and between acoustic conditions. Correlations between metabolic and behavioural responses for each prosody-condition pair will be reported. This method offers insight into the relative importance of pitch in emotional prosody recognition and provides the groundwork for understanding emotional prosody processing in CI hearing.

The Speech in Noise Workshop is generously supported by: