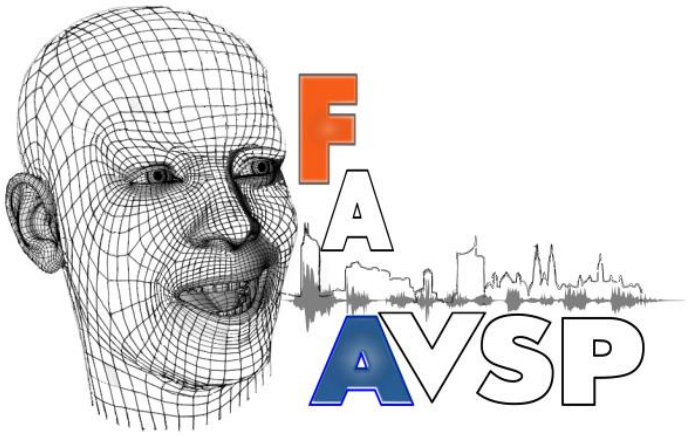
FAAVSP 2015

The 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing

Kardinal König Haus, 11 – 13 September 2015, Vienna, Austria



Foreword

Welcome to the 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing which brings together the 13th International Conference on Auditory-Visual Speech Processing (AVSP 2015) and the 4th International Symposium on Facial Analysis and Animation (FAA 2015).

Both the AVSP and FAA conferences have a common focus on facial communication research. The AVSP conference concentrates on how auditory and visual speech information plays a role in human perception, machine recognition, and human-machine interaction. FAA focuses on facial animation analysis and synthesis addressed in the fields of computer graphics, computer vision and psychology.

The two conferences attract researchers from diverse fields, such as speech processing, computer graphics and computer vision, psychology, neuroscience, linguistics, robotics and electrical engineering.

The aim of this first joint conference is to bring together, from both academia and industry, the two communities of auditory-visual speech processing (AVSP) and facial animation (FAA) to discuss research and exchange ideas, data and experiences.

FAAVSP 2015 Sponsors

We would like to thank our conference sponsors for their support.

The Telecommunications Research Center Vienna (FTW)

ftw Creating Communication Technologies

The Signal Processing and Speech Communication Laboratory (SPSC Lab) of Graz University of Technology

TU Graz

The International Speech Communication Association (ISCA)

isca

Disney Research

Disney Research

The Centre for Digital Entertainment.

CDE Centre for Digital Entertainment

FAAVSP 2015 Proceedings and Invited Speakers

Full 4-6 page AVSP papers will be published in the ISCA archive (http://www.isca-speech.org/iscaweb/index.php/archive/online-archive). FAA abstracts will be published by ACM (http://www.acm.org/publications/icp_series). We would like thank the members of the scientific committee and all other reviewers for their valuable contribution to the paper selection process.

For this meeting, we are fortunate to have keynotes from five excellent speakers

- Volker Helzle (Filmakademie Baden-Württemberg, Institute of Animation)
- Veronica Orvalho (University of Porto, Department of Computer Science)
- Jean-Luc Schwartz (GIPSA Lab, Grenoble)
- Frank Soong & Lijuan Wang (Microsoft Research Asia).

The goal of keynote speeches is to get a glance at fields related to AVSP/FAA, and their current problems and innovative methods.

Committees

Organising Committee:

Michael Pucher (Telecommunications Research Center Vienna)
Gernot Kubin (Technical University Graz)
Darren Cosker (University of Bath)
Chris Davis (University of Western Sydney)
Slim Ouni (University of Lorraine)
William Smith (University of York)
Eva Krumhuber (University College London)

Scientific Committee:

AVSP

Gerard Bailly (GIPSA-lab)
Jonas Beskow (KTH)
Nick Campbell (Trinity College Dublin)
Vincent Colotte (University of Lorraine)
Piero Cosi (Institute of Cognitive Sciences and Technologies)
Chris Davis (University of Western Sydney)
Olov Engwall (KTH)
Sasha Fagel (zoobe)
Maëva Garnier (GIPSA-Lab)
Beatrice de Gelder (Masstricht University)
Asif A. Ghazanfar (Princeton University)
Björn Granström (KTH)
Alexandra Jesse (University of Massachusetts Amherst)
Jeesun Kim (University of Western Sydney)
Sonja Kotz (University of Manchester)
Emiel Krahmer (Tilburg University)
Dominic W. Massaro (University of California, Santa Cruz)
Slim Ouni (University of Lorraine)
Gerasimos Potamianos (University of Thessaly)
Marc Swerts (Tilburg University)
Lawrence D. Rosenblum (University of California Riverside)
Jean-Luc Schwartz (GIPSA-lab)
Kaoru Sekiyama (Kumamoto University)
Jean Vroomen (Tilburg University)
Takeshi Saitoh (Kyushu Institute of Technology)

FAA

Matthew Aylett (University of Edinburgh / Cereproc)
Thabo Beeler (Disney Research Zürich)
Michael Berger (University of Edinburgh / Speech Graphics)
Bernd Bickel (Institute of Science and Technology Austria)
Volker Blanz (University of Siegen)
Darren Cosker (University of Bath)
Adrian Hilton (University of Surrey)
Eva Krumhuber (University College London)
Gernot Kubin (Technical University Graz)
Jacqueline Leybaert (Université Libre de Bruxelles)
Stephen Maddock (University of Sheffield)
Rachel McDonnell (Trinity College Dublin)
Michael Pucher (Telecommunications Research Center Vienna)
Dietmar Schabus (Telecommunications Research Center Vienna)
William Smith (University of York)
Ingmar Steiner (Saarland University)
Barry Theobald (University of East Anglia)
Markus Toman (Telecommunications Research Center Vienna)

Call for papers:

Submission of papers are invited in all areas of auditory-visual speech processing and facial animation and including but not limited to:

- Acquisition of Facial Shape, Motion and Texture
- Facial animation and rendering techniques
- Facial Model Based Coding and Compression
- Facial Analysis and Animation for Mobile Applications
- Embodied Virtual Agents
- Visual and Audiovisual Speech Synthesis
- Human and machine recognition of audio-visual speech
- Human and machine models of multimodal integration
- Multimodal and perceptual processing of facial animation and audiovisual events
- Cross-linguistic studies of audio-visual speech processing
- Developmental studies of audio-visual speech processing
- Audio-visual prosody
- Emotion and Expressivity modeling
- Gestures accompanying speech and non-linguistic behavior
- Neuropsychology and neurophysiology of audio-visual speech processing
- Scene analysis using audio and visual speech information
- Data collection and corpora for audio-visual speech processing

The conference will be held in Vienna, Austria, 11.-13. September 2015. The session on September 11 will be devoted to FAA topics and those on September 12-13 to AVSP topics. The keynotes will present topics relevant to both communities.

Two types of submission are possible: Abstracts (1 pages) for FAA and AVSP topics, and full papers (4 to 6 pages) for AVSP topics. Abstracts and full papers will be peer reviewed. Full papers will be published in the ISCA archive.

The conference is a satellite event of INTERSPEECH 2015 in Dresden

PROGRAM

Friday, 11th September

08:30 **Registration Opens**

09:15 **Introduction**

09:30 – 10:30 **Keynote 1: An Artistic & Tool Driven Approach For Believable Digital Characters**
(Chair: Cosker)
Volker Helzle, *Filmakademie Baden-Württemberg, Institute of Animation*

10:30 – 11:00 **Coffee Break**

11:00 – 12:20 **Oral Session 1: Facial Analysis and Synthesis**
(Chair: Cosker)

- **Towards Synthesis of Novel, Photorealistic 3D Faces**
  Arnaud Dessein, *Université de Bordeaux;* Aleksejs Makejevs & William Smith*, *University of York*
- **Personalization of Statistical Face Models for Tracking and Animation**
  Markus Kettern*, Anna Hilsmann, & Peter Eisert, *Fraunhofer HHI*
- **Image-Based Expressive Speech Animation Based on the OCC Model of Emotions**
  Paula Costa* & José Mario De Martino, *University of Campinas*
- **Threshold-Based Lip Segmentation Using Feedback of Shape Information**
  Ashley Gritzman*, Vered Aharonson, David Rubin, & Adam Pantanowitz, *University of the Witwatersrand*

12:20 – 13:10 **Lunch Break**

13:10 – 14:40 **Poster Session 1**
(Chair: Smith)

- **Interface for Monitoring of Engagement from Audio-Visual Cues**
  João Cabral*, Yuyun Huang, Christy Elias, Ketong Su, & Nick Campbell, *Trinity College Dublin*
- **Boxing the Face: A Comparison of Dynamic Facial Databases Used in Facial Analysis and Animation**
  Pasquale Dente* & Dennis Küster, *Jacobs University Bremen;* Eva Krumhuber, *University College London*
- **Visio-Articulatory to Acoustic Conversion of Speech**
  Michael Pucher* & Dietmar Schabus, *Telecommunications Research Center Vienna*
- **Perceived Emotionality of Linear and Non-Linear AUs Synthesised Using a 3D Dynamic Morphable Facial Model**
  Darren Cosker*, *University of Bath;* Eva Krumhuber, *University College London;* Adrian Hilton, *University of Surrey*

- **4D Cardiff Conversation Database (4D CCDb): A 4D Database of Natural, Dyadic Conversations**
  Jason Vandeventer*, Andrew Aubrey, Paul Rosin, & David Marshall, *Cardiff University*
- **Improved Visual Speech Synthesis Using Dynamic Viseme k-means Clustering and Decision Trees**
  Christiaan Rademan* & Thomas Niesler, *University of Stellenbosch*
- **Scattering vs. Discrete Cosine Transform Features in Visual Speech Processing**
  Etienne Marcheret, *IBM Thomas J. Watson Research Center;* Gerasimos Potamianos*, *University of Thessaly;* Josef Vopicka, *IBM Czech Republic;* Vaibhava Goel, *IBM Thomas J. Watson Research Center*
- **Speaker-Independent Machine Lip-Reading With Speaker-Dependent Viseme Classifiers**
  Helen Bear*, Stephen Cox, & Richard Harvey, *University of East Anglia*
- **Stream Weight Estimation using Higher Order Statistics in Multi-modal Speech Recognition**
  Kazuto Ukai*, Satoshi Tamura, & Satoru Hayamizu, *Gifu University*

14:40 – 15:40 (Chair: Cosker) **Keynote 2: How To Create A Look-A-Like Avatar Pipeline Using Low-Cost Equipment**
Veronica Orvalho, *University of Porto*

15:40 – 16:10 **Coffee Break**

16:10 – 17:30 (Chair: Krumhuber) **Oral Session 2: Perception, Emotion, and Corpora**

- **The University of Edinburgh Speaker Personality and MoCap Dataset**
  Kathrin Haag* & Hiroshi Shimodaira, *University of Edinburgh*
- **Is Your Body Lying? Exploring Bodily Cues for Deception Using an Automated Movement Analysis**
  Mariana Serras Pereira*, Suleman Shahid, Eric Postma, & Marc Swerts, *Tilburg University*
- **Speech Independent Emotion Transfer for Virtual Faces**
  Timothy Costigan* & Rachel McDonnell, *Trinity College Dublin*
- **The Effect of Animation Realism on Face Ownership and Engagement**
  Elena Kokkinara* & Rachel McDonnell, *Trinity College Dublin*

18:00 **Social Event: Wine Tavern "Zur Wildsau" (30 minutes walk)**
Meet in front of Kardinal-Koenig Haus!

Saturday, 12th September

09:00 – 10:00 **Keynote 3: Audiovisual Binding In Speech Perception.**
(Chair: Davis) Jean-Luc Schwartz, *GIPSA Lab, Grenoble*

10:00 – 10:20 **Coffee Break**

10:20 – 12:00 **Oral Session 3: Life Span**
(Chair: Krumhuber)

- **Children's Spontaneous Emotional Expressions While Receiving (Un)wanted Prizes in the Presence of Peers**
  Mandy Visser*, Emiel Krahmer, & Marc Swerts, *Tilburg University*
- **You Can Raise Your Eyebrows, I Don't Mind: Are Monolingual and Bilingual Infants Equally Good at Learning from the Eyes Region of a Talking Face?**
  Mathilde Fort*, Anira Escrichs, Alba Ayneto-Gimeno, & Núria Sebastián-Gallés, *Universitat Pompeu Fabra*
- **Comparison of Visual Speech Perception of Sampled-Based Talking Heads: Adults and Children With and Without Developmental Dyslexia**
  Paula Costa*, Daniella Batista, Mayara Toffoli, & Keila Knobel, *University of Campinas;* Cintia Alves Salgado, *Federal University of Rio Grande do Norte;* José Mario De Martino, *University of Campinas*
- **Cross-Modality Matching of Linguistic Prosody in Older and Younger Adults**
  Simone Simonetti*, Jeesun Kim, & Christopher Davis, *University of Western Sydney*
- **"I Do Not See What You Are Saying": Reduced Visual Influence on Mulimodal Speech Integration in Children With SLI**
  Aurélie Huyse, *Université Libre de Bruxelles;* Frédéric Berthommier, *Gipsa Lab;* Jacqueline Leybaert*, *Université Libre de Bruxelles*

12:00 – 13:40 **Lunch Break**

13:40 – 15:20 **Oral Session 4: Emotion, Personality, and Dialogue**
(Chair: Davis)

- **Message vs. Messenger Effects on Cross-Modal Matching for Spoken Phrases**
  Catherine Best*, *University of Western Sydney;* Christian Kroos, *Curtin University;* Karen Mulak, *University of Western Sydney;* Shaun Halovic, *Westmead Hospital;* Mathilde Fort, *Universitat Pompeu Fabra;* Christine Kitamura, *University of Western Sydney*
- **Audiovisual Generation of Social Attitudes from Neutral Stimuli**
  Adela Barbulescu*, *GIPSA Lab, INRIA Grenoble;* Gérard Bailly, *GIPSA Lab;* Rémi Ronfard, *INRIA Grenoble;* Mael Pouget, *GIPSA Lab*
- **Classification of Auditory-Visual Attitudes in German**
  Angelika Hönemann*, *University of Bielefeld;* Hansjoerg Mixdorff, *Beuth University Berlin;* Albert Rilliard, *LIMSI-CNRS*

- **Delayed Auditory Feedback with Static and Dynamic Visual Feedback**
  Elizabeth Stelle*, *University of British Columbia;* Caroline Smith, *University of New Mexico;* Eric Vatikiotis-Bateson, *University of British Columbia*
- **Visual vs. Auditory Emotion Information: How language and Culture Affect Our Bias Towards the Different Modalities**
  Chee Seng Chong*, Jeesun Kim & Christopher Davis, *University of Western Sydney*

15:20 – 15:40 **Coffee Break**

15:40 – 16:40 **Poster Session 2**
(Chair: Ouni)

- **Environmental, Linguistic, and Developmental Influences on Mothers' Speech to Children: An Examination of Audible and Visible Properties**
  Nicholas Smith* & Timothy Vallier, *Boys Town National Research Hospital;* Bob McMurray, *University of Iowa;* Christine Hammans, *Boys Town National Research Hospital,* Julia Garrick, *Boys Town National Research Hospital, University of Cincinnati*
- **Dynamics of Audiovisual Binding in Elderly Population**
  Attigodu Ganesh*, Frédéric Berthommier, & Jean-Luc Schwartz, *GIPSA Lab*
- **Combining Acoustic and Visual Features to Detect Laughter in Adults' Speech**
  Hrishikesh Rao*, Zhefan Ye, Yin Li, Mark Clements, Agata Rozga, & James Rehg, *Georgia Institute of Technology*
- **Auditory-Visual Perception of VCVs Produced by People with Down Syndrome: a Preliminary Study**
  Alexandre Hennequin, Amélie Rochet-Capellan, & Marion Dohen*, *GIPSA-Lab*
- **The Perceived Sequence of Consonants in McGurk Combination Illusions Depends on Syllabic Stress**
  Bo Holm-Rasmussen* & Tobias Andersen, *Technical University of Denmark*
- **An Answer to a Naïve Question to the McGurk Effect: Why Does Audio /b/ Give More /d/ Percepts With Visual /g/ Than With Visual /d/?**
  Tobias Andersen*, *Technical University of Denmark*
- **Optimal Timing of Audio-Visual Text Presentation: The Role of Attention**
  Maiko Takahashi* & Akihiro Tanaka, *Tokyo Woman's Christian University*
- **Anticipation of Turn-Switching in Auditory-Visual Dialogs**
  Hansjörg Mixdorff*, *Beuth University Berlin;* Angelika Hönemann, *University of Bielefeld;* Jeesun Kim & Christopher Davis, *University of Western Sydney*

16:40 – 18:00 **Oral Session 5: Culture and Language**
(Chair: Krumhuber)

- **Comparison of Multisensory Display Rules in Expressing Complex Emotions between Cultures**
  Sachiko Takagi*, *Tokyo Woman's Christian University;* Shiho Miyazawa, *JEED;* Elisabeth Huis In 't Veld, *Tilburg University;* Beatrice de Gelder, *Maastricht University;* Akihiro Tanaka, *Tokyo Woman's Christian University*
- **Towards the Development of Facial and Vocal Expression Database in East Asian and Western Cultures**
  Akihiro Tanaka* & Sachiko Takagi, *Tokyo Woman's Christian University, Waseda University;* Saori Hiramatsu, *Waseda University;* Elisabeth Huis In 't Veld, *Tilburg University;* Beatrice de Gelder, *Maastricht University*
- **The Effect of Modality and Speaking Style on the Discrimination of Non-Native Phonological and Phonetic Contrasts in Noise**
  Sarah Fenwick*, Christopher Davis, Catherine Best, & Michael Tyler, *University of Western Sydney*
- **Audio-Visual Perception of Mandarin Lexical Tones in AX Same-Different Judgment Task**
  Rui Wang*, Biao Zeng, & Simon Thompson, *Bournemouth University*

19:00 **Social Event: Wambacher Restaurant (opposite the conference venue)**
Meet inside the restaurant!

Sunday, 13th September

09:00 – 10:00 (Chair: Pucher) **Keynote 4: From Text-to-Speech (TTS) to Talking Head - A Machine Learning Approach to A/V Speech Modeling and Rendering.**
Frank Soong & Lijuan Wang, *Microsoft Research Asia*

10:00 – 10:20 **Coffee Break**

10:20 – 12:00 (Chair: Ouni) **Oral Session 6: Visual Speech Synthesis**

- **Lip Animation Synthesis: a Unified Framework for Speaking and Laughing Virtual Agent**
  Yu Ding* & Catherine Pelachaud, *Télécom-Paris Tech*
- **Comparison of Dialect Models and Phone Mappings in HSMM-Based Visual Dialect Speech Synthesis**
  Dietmar Schabus* & Michael Pucher, *Telecommunications Research Center Vienna*
- **HMM-Based Visual Speech Synthesis Using Dynamic Visemes**
  Ausdang Thangthai* & Barry-John Theobald, *University of East Anglia*
- **Investigating the Impact of Artificial Enhancement of Lip Visibility on the Intelligibility of Spectrally-Distorted Speech**
  Najwa Alghamdi*, Stephen Maddock, Guy Brown, & Jon Barker, *University of Sheffield*
- **The Stability of Mouth Movements for Multiple Talkers over Multiple Sessions**
  Christopher Davis*, Jeesun Kim, Vincent Aubanel, Greg Zelic & Yatin Mahajan, *University of Western Sydney*

12:00 – 13:40 **Lunch Break**

13:40 – 15:20 (Chair: Pucher) **Oral Session 7: Audio-Visual Speech Recognition**

- **Voicing Classification of Visual Speech Using Convolutional Neural Networks**
  Thomas Le Cornu* & Ben Milner, *University of East Anglia*
- **Comparison of Single-model and Multiple-model Prediction-based Audiovisual Fusion**
  Stavros Petridis*, Varun Rajgarhia, & Maja Pantic, *Imperial College London*
- **Finding Phonemes: Improving Machine Lip-Reading**
  Helen Bear*, Richard Harvey, & Yuxuan Lan, *University of East Anglia*
- **Discovering Patterns in Visual Speech**
  Stephen Cox*, *University of East Anglia*
- **Improving Lip-reading Performance for Robust Audiovisual Speech Recognition using DNNs**
  Kwanchiva Thangthai*, Richard Harvey, Stephen Cox, & Barry-John Theobald, *University of East Anglia*

15:20 – 15:40 **Coffee Break**

15:40 – 16:25 **Poster Session 3**
(Chair: Pucher)

- **Visual Cues to Phrase Segmentation and the Acquisition of Word Order**
  Irene de la Cruz-Pavía*, *Université Paris Descartes, University of British Columbia;* Michael McAuliffe & Janet Werker, *University of British Columbia;* Judit Gervain, *Université Paris Descartes;* Eric Vatikiotis-Bateson, *University of British Columbia*
- **Head Movements, Eyebrows, and Phonological Prosodic Prominence Levels in Stockholm Swedish News Broadcasts**
  Gilbert Ambrazaitis* & Malin Svensson Lundmark, *Lund University;* David House, *KTH*
- **Visual Lip Information Supports Auditory Word Segmentation**
  Antje Strauss*, Christophe Savariaux, Sonia Kandel, & Jean-Luc Schwartz, *GIPSA-lab*
- **The Multi-Modal Nature of Trustworthiness Perception**
  Elena Tsankova, *Jacobs University Bremen;* Eva Krumhuber*, *University College London;* Andrew Aubrey, *Cardiff University;* Arvid Kappas & Guido Möllering, *Jacobs University Bremen;* David Marshall & Paul Rosin, *Cardiff University*
- **Face-Speech Sensor Fusion for Non-Invasive Stress Detection**
  Vasudev Bethamcherla*, William Paul, Cecilia Alm, Reynold Bailey, Joe Geigel, & Linwei Wang, *Rochester Institute of Technology*
- **The Development of Patterns of Gaze to a Speaking Face**
  Julia Irwin* & Lawrence Brancazio, *Yale University, Southern Connecticut State University*

16:25 – 17:45 **Oral Session 8: Visual Speech Perception**
(Chair: Davis)

- **Integration of Auditory, Labial and Manual Signals in Cued Speech Perception by Deaf Adults: An Adaptation of the McGurk Paradigm**
  Clémence Bayard, *CNRS Grenoble;* Jacqueline Leybart & Cécile Colin*, *Université Libre de Bruxelles*
- **Explaining the Visual and Masked-Visual Advantage in Speech Perception in Noise: The Role of Visual Phonetic Cues**
  Vincent Aubanel*, Christopher Davis & Jeesun Kim, *University of Western Sidney*
- **Analysing the Importance of Different Visual Feature Coefficients**
  Danny Websdale* & Ben Milner, *University of East Anglia*
- **Auditory and Audiovisual Close-Shadowing in Normal and Cochlear-Implanted Hearing Impaired Subjects**
  Lucie Scarbel*, Denis Beautemps, & Jean-Luc Schwartz, *GIPSA-lab;* Marc Sato, *Laboratoire Parole et Langage*

17:45 – 18:00 **Closing**

Social Events

Friday, Sept 11th
Wine Tavern "Zur Wildsau"

Address: Slatingasse 22, 1130 Wien
Tel: +4301 8764653
Homepage: http://www.wildsau.at/

The rustic aspects and unique location of this tavern are enchanting and the ideal place to admire the beautiful views over Vienna while enjoying a true Viennese cuisine and regional wine. For all attendees who would like to join this event, we will meet at **18:00 o'clock in front of Kardinal-Koenig Haus**, the site of the conference. We will then walk together to the wine tavern which will be a 30 minutes walk. Please note that you will have to pay for food and drinks yourself!

For those who would like to take public transport, you can take the **local buses 54B or 55B from Jagdschlossgasse** which is a side street to Kardinal-Koenig Platz. **Alight at Ghelengasse** and then **walk towards Slatingasse** where you will find directions to the tavern.

Saturday, Sept 12th
Wambacher Restaurant

Address: Lainzer Strasse 123, 1130 Wien
Tel: +4301 8048366
Homepage: http://wambacher.co.at/

For 170 years this has been a well-known and popular wine tavern and restaurant. Known for the legendary Wambacher Schnitzel and for the lovely courtyard garden, which is nicely shaded by the old trees. Come enjoy some wonderful Viennese cuisine with a fine selection of foods and drinks.

For all attendees who would like to join this event, we will meet at **19:00 o'clock inside the Wambacher restaurant.**
Please note that you will have to pay for food and drinks yourself!

Tourist Sights

Wambacher Restaurant on *Lainzer Straße 123*

- A beautiful wine tavern with a terrace that serves Viennese food and drinks.
  Tel: +431 8048366
  Open daily: 11:00 – 24:00

Café Dommayer on *Dommayergasse 1*

- A lovely café with a garden where you can enjoy a refreshing cup of coffee, a nice hot cup of tea, along with delicious cake or pastries.
  Tel: +431 87754650
  Open daily: 07:00 – 22:00

Zoo Schönbrunn on *Maxingstraße 13b*

- A big zoo where you can see diverse species of animals. Take a relaxing walk while surrounded by the most exotic specimens of animals in their nicely built habitats!
  Open daily: 09:00 – 18:30

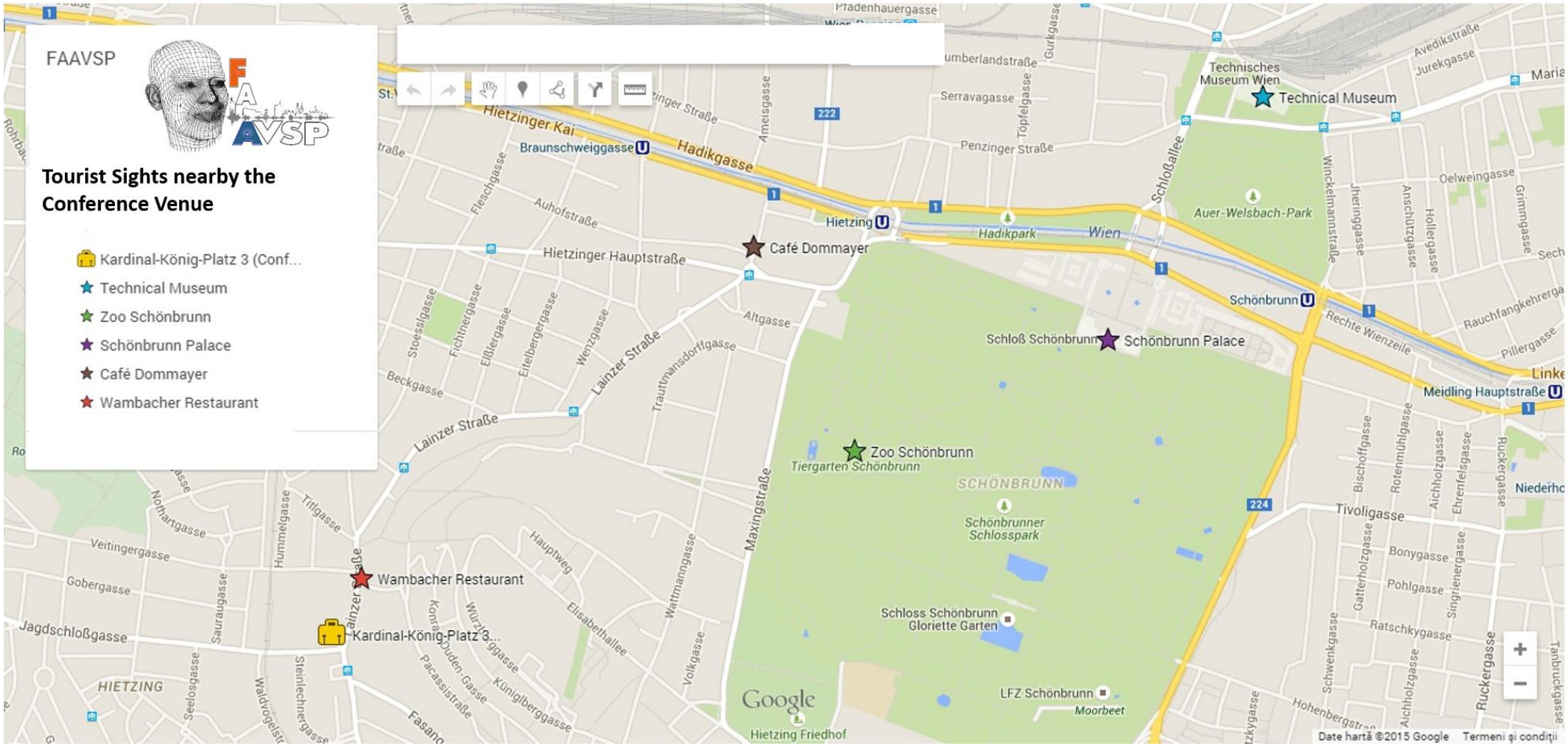
Schönbrunn Palace on *Schönbrunner Schloßstraße 47*

- A grandiose, unique palace to visit and admire its stunning architectural details.
  Tel: +431 81113239
  Open daily: 08:30 – 17:30

Technical Museum on *Mariahilfer Straße 212*

- A well-suited museum for those with a technical mind! See how techniques and machinery are applied.
  Tel: +431 899980
  Fri – Sun: 10:00 – 18:00

Local Area Map with Tourist Sights



Abstracts

(in order of appearance)

Keynote 1

An Artistic & Tool Driven Approach For Believable Digital Characters

Volker Helzle
Filmakademie Baden-Württemberg, Institute of Animation

This talk focuses on the practical tools developed at the Filmakademie for the creation of believable digital characters. We will discuss solutions that have been implemented to achieve realistic and physically plausible facial deformations during a short setup time. We will also look into new applications that made use of these characters like the cloud based animated messaging service for mobile devices (Emote), an interactive installation where animated characters recite poetry in an emotional way, or the approach we are taking to use stylized animated faces in the research on Autism.

Volker Helzle is in charge of Research and Development at the Institute of Animation at Filmakademie Baden-Württemberg. After graduating from HDM Stuttgart Media University (Dipl. Ing. AV-Medien) in 2000 he moved to California and for three years worked at Eyematic Interfaces (later acquired by Google) where his team pioneered facial performance capture substantially contributing to the engineering and development of the Eyematic Facestation. In 2003 he joined Filmakademie where he supervises the research and development department at the Institute of Animation. The primary focus of his first few years at Filmakademie has been the development of facial animation tools. This led to one of the first plausible technology tests, realizing a virtual actor in an exemplary VFX production. In addition to the technical research Volker is supervising the curriculum for the postgraduate Technical Director (TD) course at Filmakademie. TDs tackle the technological challenges of Animation, VFX and Transmedia productions at Filmakademie. The close relation to the research group allows students to engage in multidisciplinary projects. As a program consultant he contributes to the organization of the annual FMX conference. Being a C-64 kid of the 80ties, Volker's life was strongly influenced by video games and early computer graphics. To this day he is a passionate gamer but also finds interest in completely analogical activities like mountain hiking, gardening or yoga.

Oral Session 1

Towards Synthesis of Novel, Photorealistic 3D Faces

Arnaud Dessein, *Université de Bordeaux;* Aleksejs Makejevs & William Smith, *University of York*

In this abstract we describe a method for synthesising novel, photorealistic 3D faces. Our approach is based on selecting patches from real faces that are consistent with a texture provided by a global statistical model. Since the texture patches come from real faces, they are guaranteed to contain plausible levels of detail. We also describe an extension where the patch database is built using intrinsic textures captured in a lightstage. This makes the resulting models relightable.

Personalization of Statistical Face Models for Tracking and Animation

Markus Kettern, Anna Hilsmann, & Peter Eisert, *Fraunhofer HHI*

Linear and multilinear geometric models of human faces (e.g. blend shapes) are one of the prime representations for facial action in conputer vision and graphics. Well-crafted, person-specific models involve lots of manual labour but enable photorealistic animations. On the other hand, simple and generic models are well established for tracking facial action in video since they are robust and allow to estimate facial geometry even from monocular data. Between those extremes, statistical models created from semantically aligned sets of 3D face scans can cover a part of the variety of shapes and expressions that human faces may take on while still looking rather realistic. However, the shapes they can resemble are limited and often lack detail since they only represent a subset of the variations in the data used to create them and in most cases are restrained to lower order moments of these data. We present an approach to adapt statistical geometry models to a specific person via one or more 3D face scans which are not semantically aligned. In this way, the flexibility of these models can be exploited for tracking a 3D face model in video or for animation with increased level of detail and many facial characteristics of the target person.

Image-Based Expressive Speech Animation Based on the OCC Model of Emotions

Paula Costa & José Mario De Martino, *University of Campinas*

Expressive speech animation is a key aspect of the implementation of embodied conversational agents (ECA) capable of inspiring user thrust and empathy. Categorical models of emotions, like the "Big Six" emotions of Ekman, provide a small and clear vocabulary of emotions which have been adopted in many initiatives to model the speech accompanied by the expression of emotions. However, such a small set of stereotyped emotions seems not be appropriate to reproduce the complex combinations of facial expressions observed in everyday dialogues. The present work proposes an image-based expressive speech animation synthesis methodology that is based on the Ortony, Clore and Collins (OCC) appraisal model of emotions.

Threshold-based Lip Segmentation using Feedback of Shape Information

Ashley Gritzman, Vered Aharonson, David Rubin, & Adam Pantanowitz, *University of the Witwatersrand*

This research addresses the challenge of automatic threshold computation for lip segmentation. Solving this challenge will facilitate accurate and efficient lip segmentation using colour-based methods, while avoiding the drawbacks of more complex techniques. The proposed technique uses feedback of shape information to select the threshold value.

Poster Session 1

Interface for Monitoring of Engagement from Audio-Visual Cues

João Cabral, Yuyun Huang, Christy Elias, Ketong Su, & Nick Campbell, *Trinity College Dublin*

Our research goal is to develop an interactive multi-modal platform which can take multi-modal input to infer about user-level states, in particular engagement of the user with the system. This type of information is very important because it enables the platform to adapt its output modalities and interaction strategies in order to optimize the user experience and success of communication. For example, in the call centre scenario the automated agent could switch its operation mode or even redirect the user to a human operator upon a nonengaged situation. In this work we present the interface component of the platform that permits to visualise in real-time audio-visual features estimated from input video and the result of user engagement detection. The use case is the interaction with a Lego robot equipped with a web cam and the necessary software for conversing with a person. Currently, the output interface can be used to monitor and test the performance of the engagement detector by a human. Moreover, this component is also useful to manually adapt the feedback of the system by the person, called the "wizard", before the automated response implementation.

Boxing the Face: Comparison of Dynamic Facial Databases Used in Facial Analysis and Animation

Pasquale Dente & Dennis Kuester, *Jacobs University Bremen*; Eva Krumhuber, *University College London*

Facial animation is a difficult task that is based on an approximation of subtle facial movements [Trutoiu et al. 2014], and that needs to be well grounded in real life dynamic facial behaviour to be convincing. Yet while the endpoints of expressions in still images can be defined relatively precisely using the Facial Action Coding System FACS [Ekman et al. 2002], the design of facial dynamics requires additional high-resolution data (e.g., [Trutoiu et al. 2014]). However, for animation designers who do not have the resources to elicit, record, and validate such expressions, the question arises which of the extant and freely available dynamic facial databases might best serve this purpose. In this work, we examine the technical quality of 16 databases of dynamic facial expressions, including 7 databases showing spontaneous expressions. The relative proportion of the estimated visible facial area showed a lot of variation, ranging from about 7% (BINED) to up to 57% (STOIC) of the total available area of the image. More important than these proportional values, however, is arguably the effective amount of pixels² upon which facial animation modelling can be based. For some otherwise promising databases, the estimated visible facial area was only slightly above 100 pixels². In combination with sometimes high compression rates, this suggests that some of these databases may not be suitable for animation modelling despite their conceptual relevance. However, a few databases are download-able at a considerably higher resolution (up to 1280x1024), yielding substantially better results. We suggest that databases should provide the possibility to enlarge the face, as well as to provide profile views or 3D models for facial animation design. Overall, spontaneous databases appeared to be somewhat less advanced in respect to these technical parameters.

Visio-Articulatory to Acoustic Conversion of Speech

Michael Pucher & Dietmar Schabus, *Telecommunications Research Center Vienna*

In this paper we evaluate the performance of combined visual and articulatory features for the conversion to acoustic speech. Such a conversion has possible applications in silent speech interfaces, which are based on the processing of non-acoustic speech signals. With an intelligibility test we show that the usage of joint visual and articulatory features can improve the reconstruction of acoustic speech compared to using only articulatory or visual data. An improvement can be achieved when using the original or using no voicing information.

Perceived Emotionality of Linear and Non-Linear AUs Synthesised Using a 3D Dynamic Morphable Facial Model

Darren Cosker, *University of Bath;* Eva Krumhuber, *University College London;* Adrian Hilton, *University of Surrey*

Research using dynamic facial expressions in computer science and psychology is largely focused on facial models with control param- eters based on the Facial Action Coding System (FACS) [Ekman et al. 2002]. Facial models used in research and production are typically linear in nature, whereas real expressions are non-linear. Using a 3D Dynamic Morphable Model [Cosker et al. 2010], in this work we explore the effect of linear and non-linear facial movement on expression recognition. We believe that this has implications in the validity of using linear or non-linear models in facial experiments, and also impacts on the design of facial models in general.

4D Cardiff Conversation Database (4D CCDb): A 4D Database of Natural, Dyadic Conversations

Jason Vandeventer, Andrew Aubrey, Paul Rosin, & David Marshall, *Cardiff University*

The 4D Cardiff Conversation Database (4D CCDb) is the first 4D (3D Video) audio-visual database containing natural conversations between pairs of people. This publicly available database contains 17 conversations which have been fully annotated for speaker and listener activity: conversational facial expressions, head motion, and verbal/non-verbal utterances. It can be accessed at http://www.cs.cf.ac.uk/CCDb. In this paper we describe the data collection and annotation process. We also provide results of a baseline classification experiment distinguishing frontchannel from backchannel smiles, using 3D Active Appearance Models for feature extraction, polynomial fitting for representing the data as 4D sequences, and Support Vector Machines for classification. We believe this expression-rich, audio-visual database of natural conversations will make a useful contribution to the computer vision, affective computing, and cognitive science communities by providing raw data, features, annotations, and baseline comparisons.

Improved Visual Speech Synthesis Using Dynamic Viseme k-means Clustering and Decision Trees

Christiaan Rademan & Thomas Niesler, *University of Stellenbosch*

We present a decision tree-based viseme clustering technique that allows visual speech synthesis after training on a small dataset of phonetically-annotated audiovisual speech. The decision trees allow improved viseme grouping by incorporating k-means clustering into the training algorithm.
The use of overlapping dynamic visemes, defined by tri-phone time-varying oral pose boundaries, allows improved modelling of coarticulation effects. We show that our approach leads to a clear improvement over a comparable baseline in perceptual tests. The avatar is based on the freely available MakeHuman and Blender software components.

Scattering vs. Discrete Cosine Transform Features in Visual Speech Processing

Etienne Marcheret, *IBM Thomas J. Watson Research Center;* Gerasimos Potamianos, *University of Thessaly;* Josef Vopicka, *IBM Czech Republic;* Vaibhava Goel, *IBM Thomas J. Watson Research Center*

Appearance-based feature extraction constitutes the dominant approach for visual speech representation in a variety of problems, such as automatic speechreading, visual speech detection, and others. To obtain the necessary visual features, typically a rectangular region-of-interest (ROI) containing the speaker's mouth is first extracted, followed, most commonly, by a discrete cosine transform (DCT) of the ROI pixel values and a feature selection step. The approach, although algorithmically simple and computationally efficient, suffers from lack of DCT invariance to typical ROI deformations, stemming, primarily, from speaker's head pose variability and small tracking inaccuracies. To address the problem, in this paper, the recently introduced scattering transform is investigated as an alternative to DCT within the appearance-based framework for ROI representation, suitable for visual speech applications. A number of such tasks are considered, namely, visual-only speech activity detection, visual-only and audio-visual sub-phonetic classification, as well as audio-visual speech synchrony detection, all employing deep neural network classifiers with either DCT or scattering-based visual features. Comparative experiments of the resulting systems are conducted on a large audio-visual corpus of frontal face videos, demonstrating, in all cases, the scattering transform superiority over the DCT.

Speaker-Independent Machine Lip-Reading With Speaker-Dependent Viseme Classifiers

Helen Bear*, Stephen Cox & Richard Harvey, *University of East Anglia*

In machine lip-reading, which is identification of speech from visual-only information, there is evidence to show that visual speech is highly dependent upon the speaker [1]. Here, we use a phoneme-clustering method to form new phoneme-to-viseme maps for both individual and multiple speakers. We use these maps to examine how similarly speakers talk visually. We conclude that broadly speaking, speakers have the same repertoire of mouth gestures, where they differ is in the use of the gestures.

Stream Weight Estimation using Higher Order Statistics in Multi-modal Speech Recognition

Kazuto Ukai, Satoshi Tamura, & Satoru Hayamizu, *Gifu University*

In this paper, stream weight optimization for multi-modalspeech recognition using audio information and visual infor-mation is examined. In a conventional multi-stream Hidden Markov Model (HMM) used in multi-modal speech recognition, a constraint in which the summation of audio and visual weight factors should be one is employed. This means balance between transition and observation probabilities of HMM is fixed. We study an effective weight estimation indicator when releasing the constraint. Recognition experiments were conducted using an audio-visual corpus CENSREC-1-AV [1]. In noisy environments, effectiveness of deactivating the constraint is clarified for improving recognition accuracy. Subsequently higher-order statistical parameter (kurtosis) based stream weights were proposed and tested. Through recognition experiments, it is found proposed stream weights are successful.

Keynote 2

How To Create A Look-A-Like Avatar Pipeline Using Low-Cost Equipment

Veronica Orvalho, *University of Porto*

Creating a 3D avatar that looks like a specific person is time-consuming, requires expert artists, expensive equipment and a complex pipeline. In this talk I will walk you through the avatar animation pipeline created at PIC (Porto Interactive Center, www.portointeractivecenter.org) for the VERE (Virtual Embodiment and Robotic re-Embodiment, http://www.vereproject.eu/) European Project. This new pipeline does not require the user to have artistic knowledge, uses regular cameras to create the 3D avatar and a web cam to generate the animation. In this talk i will explain how we designed and created the look-a-like system at each stage: modelling, rigging and animation. I will also describe the challenge we had to overcome and the current status of the system. I will show some of our current avatar results, which could be used for example in games, interactive applications and virtual reality. I look forward to see you at the talk!

Verónica Costa Orvalho holds a Ph.D in Software Development (Computer Graphics) from Universitat Politécnica de Catalunya (2007), where her research centred on "Facial Animation for CG Films and Videogames". She has been working in IT companies for the past 15 years, such as IBM and Ericsson, and Film companies, including Patagonik Film Argentina. She has given many workshops and has international publications related to game design and character animation in conferences such as SIGGRAPH. She has received international awards for several projects: "Photorealistic facial animation and recognition", "Face Puppet" and "Face In Motion". She has received the 2010 IBM Scientific Award for her work of facial rig retargeting. Now, she is a full time professor of Porto University. In 2010 she founded Porto Interactive Center (www.portointeractivecenter.org) at Porto University, which is the host of several International and national projects as project coordinator or participant. She has strong connections with the film and game companies and provided consulting and participated in several productions like Fable 2, The Simpsons Ride. She has current and past close collaboration with film and game companies such as: Blur Studios, Electronic Arts and Microsoft. Her main research interests are in developing new methods related to motion capture, geometric modeling and deformation, facial emotion synthesis and analysis, real time animation for virtual environments and the study of intelligent avatars.

Oral Session 2

The University of Edinburgh Speaker Personality and MoCap Dataset

Kathrin Haag & Hiroshi Shimodaira, *University of Edinburgh*

We announce the release of a dialogue dataset with motion capture for the head and body which includes introverted and extroverted speaker personality styles. The dataset will be used to synthesize personality based non-verbal behaviour from speech. Initial copy synthesis experiments show that the recorded head motion clearly reflects characteristics of introverted and extroverted speakers.

Is Your Body Lying? Exploring Bodily Cues for Deception Using an Automated Movement Analysis

Mariana Serras Pereira, Suleman Shahid, Eric Postma, & Marc Swerts, *Tilburg University*

The present study explores the cue validity of children's body movement for deception detection. To achieve this, we use an automated method to look at children's nonverbal behaviour, in particular to body movement as a possible cue for deception detection. Results based on videos from truthful and deceptive children reveal that children tend to exhibit more movement during a deceptive situation when compared with a truthful situation.

Speech Independent Emotion Transfer for Virtual Faces

Timothy Costigan & Rachel McDonnell, *Trinity College Dublin*

Animating a virtual face is a difficult and expensive undertaking. Whether using a motion capture system or hand animating, it will likely require considerable skill and perhaps the services of a good actor. It is therefore important that mistakes are minimised to avoid having to recapture large sequences. Methods do exist however, that can alter existing clips to take on new properties such as emotion. We aim to produce an automatic system capable of transferring emotion to previously unseen facial motions. Results showed that our system is capable of applying and transferring between emotions while retaining the same speech content.

The Effect of Animation Realism on Face Ownership and Engagement

Elena Kokkinara & Rachel McDonnell, *Trinity College Dublin*

Recent advances in facial tracking technologies have allowed us to create realistic animations of virtual faces that would function even in real-time. A number of systems have recently been developed for gaming and VR platforms, mainly aiming to track actors' ex- pressions and using them for off-line editing. Animation realism of virtual characters' faces has been considered highly important for conveying emotions and intent [Hyde et al. 2013]. However, to our knowledge, no perceptual experiments have assessed the way that participants engage with their own ani- mated virtual face and what are the influencing factors, when they see a real-time mirrored representation of their facial expressions mapped on the virtual face. Studies in immersive virtual environments have shown that it is possible to feel an illusory ownership over a virtual body, when the body is seen from a fist person perspective and when partic- ipants receive synchronous tapping on the virtual body and their hidden real body [Slater et al. 2009]. Similarly, when participants see their collocated virtual body animating in synchrony with their tracked real body, they can feel a sense of ownership and control over their virtual representation [Kokkinara and Slater 2014]. Here, we consider the possibility to perceive ownership and control over a mirrored virtual face with synchronous animated expressions to the tracked real face.

Keynote 3

Audiovisual Binding In Speech Perception

Jean-Luc Schwartz, *GIPSA Lab*

We have been elaborating in the last years in Grenoble a series of experimental works in which we attempt to show that audiovisual speech perception comprises an "audiovisual binding" stage before fusion and decision. This stage would be in charge to extract and associate the auditory and visual cues corresponding to a given speech source, before further categorisation processes could take place at a higher stage. We developed paradigms to characterize audiovisual binding in terms of both "streaming" and "chunking" adequate pieces of information. This can lead to elements of a possible computational model, in relation with a larger theoretical perceptuo-motor framework for speech perception, the "Perception-for-Action-Control" Theory.

Jean-Luc Schwartz, Research Director at CNRS, has been leading ICP (Institut de la Communication Parlée, Grenoble France) from 2003 to 2006. His main areas of research involve perceptual processing, perceptuo-motor interactions, audiovisual speech perception, phonetic bases of phonological systems and the emergence of language, with publications in cognitive psychology (e.g. Cognition, Perception & Psychophysics, Behavioral & Brain Sciences, Hearing Research), neurosciences (e.g. Neuroimage or Human Brain Mapping), signal processing and computational modelling (e.g. IEEE Trans. Speech and Audio Processing, JASA, Computer Speech and Language, Language and Cognitive Processes), and phonetics in relation with phonology (e.g. Journal of Phonetics or Phonology Laboratory). He has been involved in many national and European projects, and responsible of some of them. He coordinated a number of special issues of journals such as Speech Communication, Primatology, Philosophical Transactions of the Royal Society B, Frontiers in Psychology, Journal of Phonetics. He organized several international workshops on Audiovisual Speech Processing, Language Emergence or Face-to-Face Communication.

Oral Session 3

Children's Spontaneous Emotional Expressions While Receiving (Un)wanted Prizes in the Presence of Peers

Mandy Visser, Emiel Krahmer, & Marc Swerts, *Tilburg University*

In this research, we studied the course of emotional expressions of 8- and 11-year-old children after winning a (large) first prize or a (substantially smaller) consolation prize, while playing a game competing the computer or a physically co-present peer. We analyzed their emotional reactions by conducting two perception tests in which participants rated children's level of happiness. Results showed that co-presence positively affected children's happiness only when receiving the first prize. Moreover, for children who were in the presence of a peer, we found that eye contact affected expressions of happiness of 8-year-old children negatively and that of 11-year-old children positively. Finally, this study showed that having eye contact with their co-present peer affected children's emotional expressions. Overall, we can conclude that, as children grow older and their social awareness increases, the presence of a peer affects their nonverbal expressions, regardless of their appreciation of their prize.

You Can Raise Your Eyebrows, I Don't Mind: Are Monolingual and Bilingual Infants Equally Good at Learning from the Eyes Region of a Talking Face?

Mathilde Fort, Anira Escrichs, Alba Ayneto-Gimeno, & Núria Sebastián-Gallés, *Universitat Pompeu Fabra*

In this study we investigate whether paying attention to a speaker's mouth impacts 15- and 18-month-old infants' ability to process visual information displayed in the talker's eyes or mouth region. Our results showed that both monolingual and bilingual 15 month-olds could *detect* the apparition of visual information appearing in the eyes/mouth region but only 15-month-old monolinguals and 18-month-old bilinguals could learn to *anticipate* its appearance in the eyes region. Overall, we demonstrate that specific language constrains (i.e., bilingualism) not only influences how infants selectively deploy their attention to different region of human faces, but also impact their ability to learn from them.

Comparison of Visual Speech Perception of Sampled-Based Talking Heads: Adults and Children With and Without Developmental Dyslexia

Paula Costa, Daniella Batista, Mayara Toffoli, & Keila Knobel, *University of Campinas;* Cintia Alves Salgado, *Federal University of Rio Grande do Norte;* José Mario De Martino, *University of Campinas*

Among other applications, videorealistic talking heads are envisioned as a programmable tool to train skills which involve the observation of human face. This work presents partial results of a study conducted with adults and children with and without developmental dyslexia to evaluate the level of speech intelligibility provided by a sample-based talking head model in comparison with unimodal auditory and real video stimuli. The results obtained so far indicate that dyslexic children, when compared to control group, are less capable of dealing with the imperfections of a synthetic facial animation visual speech model. The present study is motivated by the hypothesis that a training program using facial animation stimuli could improve the phoneme awareness of dyslexic children, a skill that is considered significantly related to success in the early stages of reading and spelling.

Cross-Modality Matching of Linguistic Prosody in Older and Younger Adults

Simone Simonetti, Jeesun Kim, & Christopher Davis, *University of Western Sydney*

Older adults perform worse than younger adults in recognising auditory linguistic prosody. Such problems may result from deterioration at the sensory level (e.g., hearing loss). The current study used a novel approach to examine this, by determining older adult's performance on a visual prosody task. If older adults are able to process visual prosody this suggests that any difficulty they show when processing auditory linguistic expressions could be related to hearing loss. The current study presented 18 younger and 11 older adults with pairs of sentences spoken by the same talker. They decided whether the pair contained the same or different prosody in a simple AX matching task. Sentence pairs were presented in four different ways; auditory-auditory (AA), visual-visual (VV), audio-visual (AV), and visual-audio (VA). Compared to older adults, younger adults exhibited similar performance for focused statements but showed better performance for questions. We suggest that problems processing questioning expressions might result from hearing loss, problems perceiving pitch, or problems at the cognitive level.

"I Do Not See What You Are Saying": Reduced Visual Influence on Mulimodal Speech Integration in Children With SLI

Aurélie Huyse, *Université Libre de Bruxelles;* Frédéric Berthommier, *Gipsa Lab;* Jacqueline Leybaert, *Université Libre de Bruxelles*

The impact of language impairment on audio-visual integration of speech in noise is examined here by testing the influence of the degradation of the auditory and the visual speech cue. Fourteen children with specific language impairment (SLI) and 14 age-matched children with typical language development (TLD) had to identify /aCa/ syllables presented in auditory only (AO), visual only (VO) and audiovisual (AV) congruent and incongruent (McGurk stimuli) conditions, embedded either in stationary noise (ST) or amplitude modulated noise (AM), in a masking release paradigm. Visual cues were either reduced (VR) or clear (VCL). In the AO modality, children with SLI had poorer performance than TLD children in AM noise but not in ST noise, leading to a weaker masking release effect. In VO modality, children with SLI had weaker performance both in VCL and VR conditions. Analyses revealed reduced AV gains in children with SLI compared to control children. In the McGurk trials, SLI children showed a decreased influence of visual cues on AV perception in the SLI group compared to the TLD group. Data analysis in the framework of the Fuzzy-Logical Model of Perception suggested that children with SLI had preserved integration abilities; the differences with TLD children were rather due to differences in the unisensory modalities. An increased weight of audition in the VR condition compared to the VCL condition was observed in both groups, suggesting that participants awarded more weight to audition when the visual input was degraded.

Oral Session 4

Message vs. Messenger Effects on Cross-Modal Matching for Spoken Phrases

Catherine Best, *University of Western Sydney;* Christian Kroos, *Curtin University;* Karen Mulak, *University of Western Sydney;* Shaun Halovic, *Westmead Hospital;* Mathilde Fort, *Universitat Pompeu Fabra;* Christine Kitamura, *University of Western Sydney*

A core issue in speech perception and word recognition research is the nature of information perceivers use to identify spoken utterances across indexical variations in their phonetic details, such as talker and accent differences. Separately, a crucial question in audio-visual research is the nature of information perceivers use to recognize phonetic congruency between the audio and visual (talking face) signals that arise from speaking. We combined these issues in a study examining how differences between connected speech utterances (messages) versus between talkers and accents (messenger characteristics) contribute to recognition of crossmodal articulatory congruence between audio-only (AO) and video-only (VO) components of spoken utterances. Participants heard AO phrases in their native regional English accent or another English accent, and then saw two synchronous VO displays of point-light talking faces from which they had to select the one that corresponded to the audio target. The incorrect video in each pair was either the same or a different phrase as the audio target, produced by the same or a different talker, who spoke in either the same or a different English accent. Results indicate that cross-modal articulatory correspondence is more accurately and quickly detected for message content than for messenger details, suggesting that recognising the linguistic message is more fundamental than messenger features is to cross-modal detection of audio-visual articulatory congruency. Nonetheless, messenger characteristics, especially accent, affected performance to some degree, analogous to recent findings in AO speech research.

Audiovisual Generation of Social Attitudes from Neutral Stimuli

Adela Barbulescu*, *GIPSA Lab, INRIA Grenoble;* Gérard Bailly, *GIPSA Lab;* Rémi Ronfard, *INRIA Grenoble;* Mael Pouget, *GIPSA Lab*

The focus of this study is the generation of expressive audiovisual speech from neutral utterances for 3D virtual actors. Taking into account the segmental and suprasegmental aspects of audiovisual speech, we propose and compare several computational frameworks for the generation of expressive speech and face animation. We notably evaluate a standard framebased conversion approach with two other methods that postulate the existence of global prosodic audiovisual patterns that are characteristic of social attitudes. The proposed approaches are tested on a database of ”Exercises in Style” [1] performed by two semi-professional actors and results are evaluated using crowdsourced perceptual tests. The first test performs a qualitative validation of the animation platform while the second is a comparative study between several expressive speech generation methods. We evaluate how the expressiveness of our audiovisual performances is perceived in comparison to resynthesized original utterances and the outputs of a purely framebased conversion system

Classification of Auditory-Visual Attitudes in German

Angelika Hönemann, *University of Bielefeld;* Hansjörg Mixdorff, *Beuth University Berlin;* Albert Rilliard, *LIMSI-CNRS*

This paper presents results from an auditory-visual recognition experiment employing short utterances of German produced with varying attitudinal expressions. It is based on 16 different kinds of social and/or propositional attitudes which place speakers in various social interactions with a partner of inferior, equal or superior status, and having a communication aim with a positive, neutral or negative, valence. Data from ten German subjects were classified by native perceivers regarding the attitude portrayed. Participants were given five choices: The intended attitude, two closely related attitudes, and two randomly chosen ones. Higher recognition scores were obtained in audio-visual presentations (45%), over 36% with audio-only stimuli. The best recognized attitudes were doubt, (neutral) statement, surprise and irritation which all yielded audio-visual recognition scores over 50%. Lowest recognition scores were obtained for irony, 'walking-on-eggs' and politeness. A hierarchical clustering based on correspondence analysis showed that groupings of stimuli in one cluster are consistent with their original labels - these consistent stimuli yield better recognition scores. Conversely, clusters with heterogeneous populations simply aggregate bad performances.

Delayed Auditory Feedback with Static and Dynamic Visual Feedback

Elizabeth Stelle, *University of British Columbia;* Caroline Smith, *University of New Mexico;* Eric Vatikiotis-Bateson, *University of British Columbia*

Visual speech information influences the accuracy and content of auditory speech perception, with both the static and dynamic components of the visual signal playing a role. However, less is known about the effect of visual information when it is presented as a source of speech production feedback. This paper presents results from a delayed auditory feedback paradigm which contrasted the presentation of static and dynamic visual feedback. Participants repeated short sentences in six conditions, and speech rate and speech errors were measured. Speech rate was reliably reduced when dynamic visual feedback was paired with delayed auditory feedback, and there was a weak effect for speech errors, which were reduced when dynamic visual feedback was paired with normal auditory feedback. Within condition trends raise questions about adaptation to different types of feedback. Our results suggest that dynamic, but not static, visual feedback affects speech production. Even if the auditory and visual feedback are misaligned, it is important that temporal dynamics be present in both signals.

Visual vs. Auditory Emotion Information: How language and Culture Affect Our Bias Towards the Different Modalities

Chee Seng Chong, Jeesun Kim, & Christopher Davis, *University of Western Sydney*

This study investigated if familiarity with a language that an emotion is expressed in, affects how information from the different sensory modalities are weighed in auditory-visual (AV) processing. The rationale for this study is that visual information may drive multisensory perception of emotion when a person is unfamiliar with a language, and this visual dominance effect may be reduced when a person is able to understand and extract emotion information from the language. To test this, Cantonese, English and Malay speakers were presented spoken
Cantonese and English emotion expressions (angry, happy, sad, disgust and surprise) in AO, VO or AV conditions. Response matrices were examined to see if patterns of responses changed as a function of whether the expressions were produced in their native or non-native language. Our results show that the visual dominance effect for Cantonese and Malay participants changed depending on the language an emotion was expressed in, while the English participants showed a strong visual dominance effect regardless of the language of expression.

Poster Session 2

Environmental, Linguistic, and Developmental Influences on Mothers' Speech to Children: An Examination of Audible and Visible Properties

Nicholas Smith & Timothy Vallier, *Boys Town National Research Hospital;* Bob McMurray, *University of Iowa;* Christine Hammans, *Boys Town National Research Hospital;* Julia Garrick, *Boys Town National Research Hospital, University of Cincinnati*

Mothers adapt their speech in various ways when talking to infants and young children. These adaptations have both audible and visible properties, and are thought to serve important functions that promote speech, language, and cognitive development in children. We examined mothers' speech to their preschool children in the context of an interactive speech perception task in which mothers produced target words in order to direct their child (or another adult) to select the correct matching picture on a touch screen. All interaction took place via real-time video under different levels (56 and 76 dB SPL) of background noise, presented to mothers and children through headphones in order to permit uncontaminated recordings of mothers' speech. Acoustic-phonetic analyses of target words, together with video-based analyses of mothers' speech-related facial movements, provide coordinated measures of how mothers modify their speech as a function of listener age (child or adult) as well as perceptual challenges related to background noise level, target word properties, and context.

Dynamics of Audiovisual Binding in Elderly Population

Attigodu Ganesh, Frédéric Berthommier, & Jean-Luc Schwartz, *GIPSA Lab*

In a previous set of experiments, we showed that if a McGurk stimulus is preceded by an incongruent Audio-Visual (AV) context (composed of incongruent auditory and visual speech materials) the amount of McGurk fusion is largely decreased. We interpreted this result in the framework of a two-stage "binding and fusion" model of AV speech perception. The aim of the present study is to measure the dynamics of the AV binding mechanism in elder compared with younger adults. The results are in line with previous studies and suggest that elder subjects might have more modulations due to context – and hence a larger binding/unbinding dynamics - than younger ones.

Combining Acoustic and Visual Features to Detect Laughter in Adults' Speech

Hrishikesh Rao, Zhefan Ye, Yin Li, Mark Clements, Agata Rozga, & James Rehg, *Georgia Institute of Technology*

Laughter can not only convey the affective state of the speaker but also be perceived differently based on the context in which it is used. In this paper, we focus on detecting laughter in adults' speech using the MAHNOB laughter database. The paper explores the use of novel long-term acoustic features to capture the periodic nature of laughter and the use of computer vision-based smile features to analyze laughter. The classification accuracy of the leave-one-speaker-out cross-validation using a cost-sensitive learning approach with a random forest classifier with 100 trees for detecting laughter in adults' speech was 93.06% using acoustic features alone. Using only the visual features, the accuracy was 89.48%. Early fusion of audio and visual features resulted in an absolute improvement in the accuracy, compared to using only acoustic features, by 3.79% to 96.85%. The results indicate that the novel acoustic features do capture the repetitive characteristics of laughter, and the vision-based smile features can provide complementary visual cues to discriminate between speech and laughter. The significant finding of the study is the improvement of not only the accuracy, but a reduction in the false positives using the fusion of audio-visual features.

Auditory-Visual Perception of VCVs Produced by People with Down Syndrome: A Preliminary Study

Alexandre Hennequin, Amélie Rochet-Capellan, & Marion Dohen, *GIPSA-LAB*

Down syndrome (DS) is the most frequent genetic disorder in humans and is present throughout society. When questioned about their child's speech, all parents of a child with DS report issues in speech intelligibility [1]. People with DS actually have better receptive than expressive speech abilities [1]. Improving speech production of people with DS is an important aspect of their quality of life. Understanding how perception of speech produced by people with DS could be improved could also have positive effects on their social integration. Speech difficulties in people with DS originate from anatomical and physiological specificities as well as motor impairments and appear in early childhood. For example, people with DS have a smaller vocal tract and their tongue is bigger relatively to the size of their oral cavity. Other anatomical and perceptual specificities affect their ability to produce speech (see [2] for a review). All these have acoustical consequences. To our knowledge no study has explored auditory-visual perception of speech produced by people with DS whereas it is well known that speech perception benefits from vision especially in disturbed conditions (e.g. in noise: [3]). This study aims at exploring if and how vision can improve the perception, by "ordinary" people, of speech produced by people with DS.

The Perceived Sequence of Consonants in McGurk Combination Illusions Depends on Syllabic Stress

Bo Holm-Rasmussen & Tobias Andersen, *Technical University of Denmark*

The McGurk illusion is a striking proof that humans integrate auditory and visual speech. The illusion is created by presenting a video of a face producing one sound with a dubbed audio track of another sound and the stimuli together then creates either a fusion or combination of the two modalities; or the visual stimuli dominates. Exactly what defines the sequence of consonants in the combination illusion is not well known. In this work we have conducted a listening test that indicates that syllabic stress has an influence on the order of consonants in the McGurk combination illusions.

An Answer to a Naïve Question to the McGurk Effect: Why Does Audio /b/ Give More /d/ Percepts With Visual /g/ Than With Visual /d/?

Tobias Andersen, *Technical University of Denmark*

When the sound of /aba/ is dubbed onto a video of a talking face saying /ada/ it is often heard as /aga/ due to the McGurk fusion illusion. Sometimes, it is heard as /ada/. When /aba/ is dubbed onto a talking face saying /aga/ it can cause the same two percepts. Naively, one would expect more /ada/ percepts when the visual stimulus is /ada/ than when it is /aga/ but Cathiard et al. (2001) noted that the opposite is often the case. They isolated this effect in a data set containing responses to auditory /aba/, visual /ada/ and /aga/, and the two corresponding audiovisual stimuli (in their Table 4). Recently, Andersen (2015) applied an early model of integration based on the Maximum Likelihood Estimation (MLE) principle to audiovisual speech perception. Here, we include a geometry where the representation of /d/ is located between /b/ and /g/ into this early MLE model and fit it to the data set described by Cathiard et al. The result shows that the early MLE model is able to account for Cathiard et al.'s observations.

Optimal Timing of Audio-Visual Text Presentation: The Role of Attention

Maiko Takahashi & Akihiro Tanaka, *Tokyo Woman's Christian University*

This study investigates the optimal timing of audio-visual presentation for text comprehension. In Experiment 1, participants were asked to read and/or hear the expository passages in three conditions; visually presented, auditorily presented, and audio-visually presented condition. Comprehension performance in audio-visual presentation condition was not different from that in visual or auditory condition, raising the possibility that cognitive load on processing both visual and auditory information negated the positive effect of multi-modal presentation. To reduce the load of processing dual information simultaneously, we proposed to delay the timing of one of the two modality presentation and to direct participants' attention to one of the two modality information during audio-visual text presentation. In Experiment 2 to 4, passages were presented audio-visually in three conditions; auditorily preceding, simultaneous, and visually preceding conditions. Participants were instructed to direct their attention to whole information (Exp. 2), visual information (Exp. 3), and auditory information (Exp. 4). The results showed that comprehension performance was higher in the visually preceding condition when their attention was directed whole or visual information. Based on the results, the integration process of audio-visually presented text information was discussed.

Anticipation of Turn-Switching in Auditory-Visual Dialogs

Hansjörg Mixdorff, *Beuth University Berlin;* Angelika Hönemann, *University of Bielefeld;* Jeesun Kim & Christopher Davis, *University of Western Sydney*

This paper presents an experiment in which we examined whether German and Australian English perceivers were able to predict imminent turn-switching in Australian English auditory-visual dialogs. Subjects were presented excerpts of one and four second duration either preceding a switch or taken from inside a turn and had to decide which condition they saw. Stimuli were either A/V, video-only or audio-only. Results on the one second excerpts were close to random. In general we found a preference for non-switching. Australian subjects outperformed the German subjects in the audio-only condition, but outcomes were almost equal on the A/V stimuli. Analysis regarding the syntactic and prosodic properties of the stimuli showed that phrase-final statement as well as question intonation facilitated recognition presumably due to these acting as markers of turn-switch preparation; whereas incomplete sentences and non-terminal intonation were indicative of turn-internal excerpts. As to visual cues signaling a following switch results were rather varied. An open mouth on the part of the listener more often preceded switches than not.

Oral Session 5

Comparison of Multisensory Display Rules in Expressing Complex Emotions between Cultures

Sachiko Takagi, *Tokyo Woman's Christian University;* Shiho Miyazawa, *JEED;* Elisabeth Huis In 't Veld, *Tilburg University;* Beatrice de Gelder, *Maastricht University;* Akihiro Tanaka, *Tokyo Woman's Christian University*

Previous studies have suggested that there are cultural differences in display rules and decoding rules of the emotions. In this study, we examined the cultural differences in the multisensory display rules of the six basic emotions (happiness, anger, disgust, sadness, fear and surprise) and the six complex emotions (interest, contempt, embarrassment, shame, guilt, and envy) between Japanese and Dutch. In the experiment, we used the six kinds of faces and voices showing the six basic emotions. Japanese and Dutch participants were asked to create audiovisual movies expressing each of twelve emotions by combining one of the six faces and one of the six voices. Our results showed cultural differences in expressing complex emotions. Specifically, there are two ways to express complex emotion by combining a face and voice showing the basic emotions: connection or substitution. Japanese participants tend to use the way of connection and combine the face and voice showing different emotions. On the other hand, Dutch participants tend to use the way of substitution and combine the face and voice showing the same basic emotions. Our findings indicate differential display rules and decoding rules between cultures in expressing complex emotions.

Towards the Development of Facial and Vocal Expression Database in East Asian and Western Cultures

Akihiro Tanaka & Sachiko Takagi, *Tokyo Woman's Christian University, Waseda University;* Saori Hiramatsu, *Waseda University;* Elisabeth Huis In 't Veld, *Tilburg University;* Beatrice de Gelder, *Maastricht University*

The purpose of this study is to develop a stimulus set, in which facial and vocal expressions by East Asian and Western speakers are recorded. In the recording session, facial and vocal expressions, in which the six basic emotions were expressed, were recorded from Japanese and Dutch models, using equivalent linguistic phrases and identical recording settings and procedures. After selection, facial and vocal expressions from eight Japanese and eight Dutch models were evaluated by Japanese and Dutch students, respectively. Results showed that accuracies for facial expressions were above 60% except for fear in both Japanese and Dutch groups. Although the facial and vocal expressions were recorded simultaneously, accuracies for vocal expressions were not so high as facial expressions. This study is the first step towards the development of a new stimulus set of facial and vocal expressions from East Asian and Western models. We expect our stimulus set to be used in future studies comparing between facial and vocal emotion perception, and those comparing unisensory and/or multisensory emotion perception between cultures.

The Effect of Modality and Speaking Style on the Discrimination of Non-Native Phonological and Phonetic Contrasts in Noise

Sarah Fenwick, Christopher Davis, Catherine Best, & Michael Tyler, *University of Western Sydney*

Auditory speech is difficult to discern in degraded listening conditions, however the addition of visual speech can improve perception. The Perceptual Assimilation Model [1] suggests that non-native contrasts involving a native phonological difference (two-category assimilation) should be discriminated more accurately than those involving a phonetic goodness-offit difference (category-goodness assimilation), but it is not known whether auditory-visual (AV) benefit is greater for phonological than phonetic differences when the acoustic signal is degraded by speech-shaped-noise. In auditory-only (AO) and AV conditions, monolingual Australian English participants completed AXB discrimination tasks on twocategory versus category-goodness Sindhi contrasts. We also examined the relative influences of phonetic feature difference (laryngeal vs. place-of-articulation [POA]) and speaking style (clear vs. citation speech) on discrimination accuracy. AV benefit was found for POA contrasts, but no effect of speaking style, and AV benefit was larger for two-category than category-goodness contrasts. For laryngeal contrasts, AV benefit was found for the two-category contrasts (across speaking style), but for the category-goodness contrast only when it was clearly articulated. These results indicate that non-native perceivers use visual speech to their advantage, and to a greater extent for phonological contrasts, but speaking style contributes in AV conditions only for a less salient phonetic contrast.

Audio-Visual Perception of Mandarin Lexical Tones in AX Same-Different Judgment Task

Rui Wang, Biao Zeng, & Simon Thompson, *Bournemouth University*

Two same-different discrimination tasks were conducted to test whether Mandarin and English native speakers use visual cues to facilitate Mandarin lexical tone perception. In the experiments, the stimuli were presented in 3 modes: audio-only (AO), audio-video (AV) and video-only (VO) under the clear and two levels of signal-to-noise ratio (SNR) -6dB and -9dB noise condition. If the speakers' perception of AV is better than that of AO, the extra visual information of lexical tones contributes tone perception. In Experiment 1 and 2, we found that Mandarin speakers had no visual augmentation under clear and noise conditions. For English speakers, on the other hand, extra visual information hindered their tone perception (visual reduction) under SNR -9dB noise. This suggests that English speakers rely more on auditory information to perceive lexical tones. Tone pairs analysis in both experiments found that visual reduction in tone pair T2-T3 and visual augmentation in tone pair T3-T4. It indicates that acoustic tone features (e.g. duration, contour) can be seen and be involved in the process of audiovisual perception. Visual cues facilitate or inhibit tone perception depends on whether the presented visual features of the tone pairs are distinctively recognised or highly confusing to each other.

Keynote 4

From Text-to-Speech (TTS) to Talking Head - A Machine Learning Approach to A/V Speech Modeling and Rendering

Frank Soong & Lijuan Wang, *Microsoft Research Asia*

In this talk, we will present our research results in A/V speech modeling and rendering via a statistical, machine learning approach. A Gaussian Mixture Model (GMM) based Hidden Markov Model (HMM) will be reviewed first in speech modeling where GMM is for modeling the stochastic nature of speech production while HMM, for characterizing the Markovian nature of speech parameter trajectories. All speech parametric models are estimated via an EM algorithm based maximum likelihood procedure and the resultant models are used to generate speech parameter trajectories for a given text input, say a sentence, in the maximum probability sense. Thus generated parameters is then used to synthesize corresponding speech waveforms via a vocoder or to render high quality output speech by our "trajectory tiling algorithm" where appropriate segments of the training speech database are used to "tile" the generated trajectory optimally. Similarly, the lips movement of a talking head, along with the jointly moving articulatory parts like jaw, tongue and teeth, can also be trained and rendered according to the optimization procedure. The visual parameters of a talking head can be collected via 2D- or 3D-video(via stereo, multi-camera recording equipment or consumer grade, capturing devices like Microsoft Kinect)and the corresponding visual trajectories of intensity, color and spatial coordinates are modeled and synthesized similarly. Recently, feedforward Deep Neural Net (DNN) and Recurrent Neural Net machine learning algorithms have been applied to speech modeling for both recognition and synthesis applications. We have deployed both forms of neural nets in TTS training successfully. The RNN, particularly, with a longer memory can model speech prosody of longer contexts in speech, say in a sentence, better. We will also cover the topics of cross-lingual TTS and talking head modeling, where audio and visual data collected in one source language can be used to train a TTS or talking head in a different target language. The mouth shapes of a mono-lingual speaker have also been found adequate for rendering synced lips movement of talking heads in different languages. Various demos of TTS and talking head will be shown to illustrate our research findings.

Frank K. Soong is a Principal Researcher and Research Manager, Speech Group, Microsoft Research Asia (MSRA), Beijing, China, where he works on fundamental research on speech and its practical applications. His professional research career spans over 30 years, first with Bell Labs, US, then with ATR, Japan, before joining MSRA in 2004. At Bell Labs, he worked on stochastic modeling of speech signals, optimal decoder algorithm, speech analysis and coding, speech and speaker recognition. He was responsible for developing the recognition algorithm which was developed into voice-activated mobile phone products rated by the Mobile Office Magazine (Apr. 1993) as the "outstandingly the best". He is a co-recipient of the Bell Labs President Gold Award for developing the Bell Labs Automatic Speech Recognition (BLASR) software package. He has served as a member of the Speech and Language Technical Committee, IEEE Signal Processing Society and other society functions, including Associate Editor of the IEEE Speech and Audio Transactions and chairing IEEE Workshop. He published extensively with more than 200 papers and co-edited a widely used reference book, Automatic Speech and Speech Recognition- Advanced Topics, Kluwer, 1996. He is a visiting professor of the Chinese University of Hong Kong (CUHK) and a few other top-rated universities in China. He is also the co-Director of the National MSRA-CUHK Joint Research Lab. He got his BS, MS and PhD from National Taiwan Univ., Univ. of Rhode Island, and Stanford Univ, all in Electrical Eng. He is an IEEE Fellow "for contributions to digital processing of speech".

Lijuan Wang received B.E. from Huazhong Univ. of Science and Technology and Ph.D. from Tsinghua Univ., China in 2001 and 2006 respectively. In 2006, she joined the speech group of Microsoft Research Asia, where she is currently a lead researcher. Her research areas include audio-visual speech synthesis, deep learning (feedforward and recurrent neural networks), and speech synthesis (TTS)/recognition. She has published more than 25 papers on top conferences and journals and she is the inventor/co-inventor of more than 10 granted/pending USA patents. She is a senior member of IEEE and a member of ISCA.

Oral Session 6

Lip Animation Synthesis: a Unified Framework for Speaking and Laughing Virtual Agent

Yu Ding & Catherine Pelachaud, *Télécom-Paris Tech*

This paper proposes a unified statistical framework to synthesize speaking and laughing lip animations for virtual agents in real time. Our lip animation synthesis model takes as input the decomposition of a spoken text into phonemes as well as their duration. Our model can be used with synthesized speech. First, Gaussian mixture models (GMMs), called lip shape GMMs, are used to model the relationship between phoneme duration and lip shape from human motion capture data; then an interpolation function is learnt from human motion capture data, which is based on hidden Markov models (HMMs), called HMMs interpolation. In the synthesis step, lip shape GMMs are used to infer a first lip shape stream from the inputs; then this lip shape stream is smoothed by the learnt HMMs interpolation, to obtain the synthesized lip animation. The effectiveness of the proposed framework is confirmed in the objective evaluation.

Comparison of Dialect Models and Phone Mappings in HSMM-Based Visual Dialect Speech Synthesis

Dietmar Schabus & Michael Pucher, *Telecommunications Research Center Vienna*

In this paper we evaluate two different methods for the visual synthesis of Austrian German dialects with parametric Hidden-Semi-Markov-Model (HSMM) based speech synthesis. One method uses visual dialect data, i.e. visual dialect recordings that are annotated with dialect phonetic labels, the other methods uses a standard visual model and maps dialect phones to standard phones. This second method is more easily applicable since most often visual dialect data is not available. Both methods employ contextual information via decision tree based visual clustering of dialect or standard visual data. We show that both models achieve a similar performance on a subjective pair-wise comparison test. This shows that visual dialect data is not necessarily needed for visual modeling of dialects if a dialect to standard mapping can be used that exploits the contextual information of the standard language.

HMM-Based Visual Speech Synthesis Using Dynamic Visemes

Ausdang Thangthai & Barry-John Theobald, *University of East Anglia*

In this paper we incorporate dynamic visemes into hidden Markov model (HMM)-based visual speech synthesis. Dynamic visemes represent intuitive visual gestures identified automatically by clustering purely visual speech parameters. They have the advantage of spanning multiple phones and so they capture the effects of visual coarticulation explicitly within the unit. The previous application of dynamic visemes to synthesis used a sample-based approach, where cluster centroids were concatenated to form parameter trajectories corresponding to novel visual speech. In this paper we generalize the use of these units to create more flexible and dynamic animation using a HMM-based synthesis framework. We show using objective and subjective testing that a HMM synthesizer trained using dynamic visemes can generate better visual speech than HMM synthesizers trained using either phone or traditional viseme units.

Investigating the Impact of Artificial Enhancement of Lip Visibility on the Intelligibility of Spectrally-Distorted Speech

Najwa Alghamdi, Stephen Maddock, Guy Brown, & Jon Barker, *University of Sheffield*

The intelligibility of visual speech can be affected by a number of facial visual signals, e.g. lip emphasis, teeth and tongue visibility, and facial hair. This paper focuses on lip visibility.
In the study presented in this paper, we use spectrally-distorted speech to train groups of non-native, English-speaking Saudi listeners using three different forms of speech: audio-only, audiovisual, and enhanced audiovisual, which is achieved by artificially colouring the lips of the speaker to improve lip visibility. The reason for using spectrally-distorted speech is that the longer term aim of our work is to employ these ideas in a training system for hearing-impaired users, in particular cochlear-implant users. Our initial work uses non-native Saudi listeners based on the assumption that their reduced processing abilities for native speech can be compared to the reduced processing abilities of cochlear implant users as a result of the inherent noise in the processing of sound by a cochlear implant. The results suggest that using enhanced audiovisual speech during auditory training improves the training gain when subsequently listening to audio-only spectrally-distorted speech. The results also suggest that spectrally-distorted speech intelligibility during training is improved when an enhanced visual signal is used.

The Stability of Mouth Movements for Multiple Talkers over Multiple Sessions

Christopher Davis, Jeesun Kim, Vincent Aubanel, Greg Zelic & Yatin Mahajan, *University of Western Sydney*

To examine the stability of visible speech articulation (a potentially useful biometric) we examined the degree of similarity of a speaker's mouth movements when uttering the same sentence on six different occasions. We tested four speakers of differing language background and compared within- and across speaker variability. We obtained mouth motion data using an inexpensive 3D close range sensor and commercial face motion capture software. These data were exported as c3d files and the analysis was based on guided principal components derived from custom Matlab scripts. We showed that within-speaker repetitions were more similar than between speaker ones; that language background did not affect the stability of the utterances and that the patterns of articulation from different speakers were relatively distinctive.

Oral Session 7

Voicing Classification of Visual Speech Using Convolutional Neural Networks

Thomas Le Cornu & Ben Milner, *University of East Anglia*

The application of neural network and convolutional neural network (CNN) architectures is explored for the tasks of voicing classification (classifying frames as being either non-speech, unvoiced, or voiced) and voice activity detection (VAD) of visual speech. Experiments are conducted for both speaker dependent and speaker independent scenarios. A Gaussian mixture model (GMM) baseline system is developed using standard image-based two-dimensional discrete cosine transform (2D-DCT) visual speech features, achieving speaker dependent accuracies of 79% and 94 %, for voicing classification and VAD respectively. Additionally, a single layer neural network system trained using the same visual features achieves accuracies of 86% and 97 %. A novel technique using convolutional neural networks for visual speech feature extraction and classification is presented. The voicing classification and VAD results using the system are further improved to 88% and 98% respectively. The speaker independent results show the neural network system to outperform both the GMM and CNN systems, achieving accuracies of 63% for voicing classification, and 79% for voice activity detection.

Comparison of Single-Model and Multiple-Model Prediction-based Audiovisual Fusion

Stavros Petridis, Varun Rajgarhia, & Maja Pantic, *Imperial College London*

Prediction-based fusion is a recently proposed audiovisual fusion approach which outperforms feature-level fusion on laughter-vs-speech discrimination. One set of predictive models is trained per class which learns the audio-to-visual and visual-to-audio feature mapping together with the time evolution of audio and visual features. Classification of a new input is performed via prediction. All the class predictors produce a prediction of the expected audio / visual features and their prediction errors are combined for each class. The model which best describes the audiovisual feature relationship, i.e., results in the lowest prediction error, provides its label to the input. In all the previous works, a single set of predictors was trained on the entire training set for each class. In this work, we investigate the use of multiple sets of predictors per class. The main idea is that since models are trained on clusters of data, they will be more specialised and they will produce lower prediction errors which can in turn enhance the classification performance. We experimented with subject-based clustering and clustering based on different types of laughter, voiced and unvoiced. Results are presented on laughter-vs-speech discrimination on a cross-database experiment using the AMI and MAHNOB databases. The use of multiple sets of models results in a significant performance increase with the latter clustering approach achieving the best performance. Overall, an increase of over 4% and 10% is observed for F1 speech and laughter, respectively, for both datasets.

Finding Phonemes: Improving Machine Lip-Reading

Helen Bear, Richard Harvey, & Yuxuan Lan, *University of East Anglia*

In machine lip-reading there is continued debate and research around the correct classes to be used for recognition. In this paper we use a structured approach for devising speaker-dependent viseme classes, which enables the creation of a set of phoneme-to-viseme maps where each has a different quantity of visemes ranging from two to 45. Viseme classes are based upon the mapping of articulated phonemes, which have been confused during phoneme recognition, into viseme groups. Using these maps, with the LiLIR dataset, we show the effect of changing the viseme map size in speaker-dependent machine lip-reading, measured by word recognition correctness and so demonstrate that word recognition with phoneme classifiers is not just possible, but often better than word recognition with viseme classifiers. Furthermore, there are intermediate units between visemes and phonemes which are better still

Discovering Patterns in Visual Speech

Stephen Cox, *University of East Anglia*

We know that an audio speech signal can be unambiguously decoded by any native speaker of the language it is uttered in, provided that it meets some quality conditions. But we do not know if this is the case with visual speech, because the process of lipreading is rather mysterious and seems to rely heavily on the use of context and non-speech cues. How much information about the speech content is there in a visual speech signal? We make some attempt to provide an answer to this question by 'discovering' matching segments of phoneme sequences that represent recurring words and phrases in audio and visual representations of the same speech. We use a modified version of the technique of segmental dynamic programming that was introduced by Park and Glass. Comparison of the results shows that visual speech displays rather less matching content than the audio, and reveals some interesting differences in the phonetic content of the information recovered by the two modalities.

Improving Lip-reading Performance for Robust Audiovisual Speech Recognition using DNNs

Kwanchiva Thangthai, Richard Harvey, Stephen Cox, & Barry-John Theobald, *University of East Anglia*

This paper presents preliminary experiments using the Kaldi toolkit [1] to investigate audiovisual speech recognition (AVSR) in noisy environments using deep neural networks (DNNs). In particular we use a single-speaker large vocabulary, continuous audiovisual speech corpus to compare the performance of visual-only, audio-only and audiovisual speech recognition. The models trained using the Kaldi toolkit are compared with the performance of models trained using conventional hidden Markov models (HMMs). In addition, we compare the performance of a speech recognizer both with and without visual features over nine different SNR levels of babble noise ranging from 20dB down to -20dB. The results show that the DNN outperforms conventional HMMs in all experimental conditions, especially for the lip-reading only system, which achieves a gain of 37.19% accuracy (84.67% absolute word accuracy). Moreover, the DNN provides an effective improvement of 10 and 12dB SNR respectively for both the single modal and bimodal speech recognition systems. However, integrating the visual features using simple feature fusion is only effective in SNRs at 5dB and above. Below this the degradion in accuracy of an audiovisual system is similar to the audio only recognizer.

Poster Session 3

Visual Cues to Phrase Segmentation and the Acquisition of Word Order

Irene de la Cruz-Pavía, *Université Paris Descartes, University of British Columbia;* Michael McAuliffe & Janet Werker, *University of British Columbia;* Judit Gervain, *Université Paris Descartes;* Eric Vatikiotis-Bateson, *University of British Columbia*

The present investigation aims to determine the visible facial events that accompany a specific type of prosodic information – phrasal prominence –, which has been proposed as potentially allowing prelexical infants to discover basic word order. The acoustic realization of phrasal prominence correlates systematically with the natural languages' basic word order: prominence is realized as a durational contrast in V(erb)-O(bject) languages, and as a pitch/intensity contrast in O(bject)-V(erb) languages. In a production experiment with adult native talkers of Japanese (OV) and English (VO) we examined the visible facial events that accompany phrasal prominence in order to determine potential differences in: (a) OV versus VO languages, and (b) Infant- versus Adult-Directed Speech. Participants were videotaped producing target phrases consistent with two words – one of them carrying phrasal prominence –, which were embedded in a carrier sentence. Analysis of head and face motion was conducted using Optical Flow (OF) analysis. OF is a technique that is easily implemented and detects motion very accurately. Eyebrow movements were manually coded. Analysis of the target IDS productions is complete, and nearly complete for the ADS productions. Cross-linguistic differences were found in the movements of Japanese and English talkers in the IDS productions. Only the English talkers produced greater vertical movement in the word containing phrasal prominence, which presumably corresponds to the motion derived from head nods. Both the English and Japanese talkers raised their eyebrows to mark the beginning of the target phrase. A preliminary analysis of the ADS productions suggests a lower frequency of eyebrow movements in ADS than IDS. The presence of multimodal events in IDS might help infants locate the boundaries of phrases and detect their prominent element, which might in turn allow them to discover the basic word order of the language(s) under acquisition.

Head Movements, Eyebrows, and Phonological Prosodic Prominence Levels in Stockholm Swedish News Broadcasts

Gilbert Ambrazaitis & Malin Svensson Lundmark, *Lund University;* David House, *KTH*

This contribution presents a first analysis of the distribution of head and eyebrow movements as a function of phonological prominence levels in Swedish news broadcasts. Swedish has a binary lexical pitch accent distinction (Accent 1, 2). In addition, words can be highlighted at the sentence level. For Stockholm Swedish, a phonological distinction is generally assumed between the non-focal, accented realization of a word and a focal realization of a word. Our hypothesis was that focally accented words would coincide with head or eyebrow movements more often than non-focal words, while the lexical accent category should have no effect of the distribution of movements.

A corpus of news readings from Swedish Television was annotated for focal accents, head and eyebrow movements. The results revealed a dependency of the distribution of movements on the one hand and focal accents on the other, confirming our hypothesis. Also, no consistent effects of the word accent type on eyebrow and head movements were found. However, there was an effect of the word accent type on the annotations of 'double' head beats, which might be explained as follows: Only nine words in the entire corpus (of 986 words) were annotated with a double head beat (to be compared with 220 annotations of simple head movements), of which seven were Accent 2 compounds. Compounds are characterized by a main and a secondary stress. This and the fact that simple head beats occurred frequently in the corpus suggest an association of a head beat with lower-level prominence and phonological-prosodic structure. Eyebrow movements, on the other hand, might be more restricted to higher-level prominence and information-structure coding, as they occurred sparsely in the corpus. To conclude, this study suggest that head and eyebrow movements can represent two quite different modalities of prominence cuing, both from a formal and a functional point of view, rather than just being cumulative prominence markers.

Visual Lip Information Supports Auditory Word Segmentation

Antje Strauss, Christophe Savariaux, Sonia Kandel, & Jean-Luc Schwartz, *GIPSA-lab*

Word segmentation is one of the initial processes that needs to be solved when acquiring the first or learning a second language. Acoustic cues like the fundamental frequency and segment durations have been shown to facilitate the detection of word boundaries. The role of visual speech and in particular of lip movements in word segmentation is still rather unknown. In French, liaisons, e.g. between determiner and noun, often pose a problem of several segmentation possibilities (e.g., the sequence /lafiS/ with liaison ("l'affiche") means the poster whereas without liaison ("la fiche") it means the file.). Here, we use 17 ambiguous French sequences with and without liaison. They were presented in carrier sentences either with clear acoustic cues for the first or the second segmentation possibility or with ambiguous acoustic cues. The three audio conditions were combined with lip movements hyper-articulating either the first or the second segmentation possibility in order to observe the influence of visual information on segmentation. The participants had to indicate as quickly as possible which of the two versions they understood (e.g., "l'affiche" or "la fiche"?). Results show that lip information indeed biases the word segmentation decision. These data provide important implications for audiovisual integration processes.

The Multi-Modal Nature of Trustworthiness Perception

Elena Tsankova, *Jacobs University Bremen;* Eva Krumhuber, *University College London;* Andrew Aubrey, *Cardiff University;* Arvid Kappas & Guido Möllering, *Jacobs University Bremen;* David Marshall & Paul Rosin, *Cardiff University*

Most past work on trustworthiness perception has focused on the structural features of the human face. The present study investigates the interplay of dynamic information from two channels – the face and the voice. By systematically varying the level of trustworthiness in each channel, 49 participants were presented with either facial or vocal information, or the combination of both, and made explicit judgements with respect to trustworthiness, dominance, and emotional valence. For most measures results revealed a primacy effect of facial over vocal cues. In examining the exact nature of the trustworthiness - emotion link we further found that emotional valence functioned as a significant mediator in impressions of trustworthiness. The findings extend previous correlational evidence and provide important knowledge of how trustworthiness in its dynamic and multi-modal form is decoded by the human perceiver.

Face-Speech Sensor Fusion for Non-Invasive Stress Detection

Vasudev Prasad Bethamcherla, William Paul, Cecilia Alm, Reynold Bailey, Joe Geigel, & Linwei Wang, *Rochester Institute of Technology*

We describe a human-centered multimodal framework for automatically measuring cognitive changes. As a proof-ofconcept, we test our approach on the use case of stress detection. We contribute a method that combines non-intrusive behavioural analysis of facial expressions with speech data, enabling detection without the use of wearable devices. We compare these modalities' effectiveness against galvanic skin response (GSR) collected simultaneously from the subject group using a wristband sensor. Data was collected with a modified version of the Stroop test, in which subjects perform the test both with and without the inclusion of stressors. Our study attempts to distinguish stressed and unstressed behaviors during constant cognitive load. The best improvement in accuracy over the majority class baseline was 38%, which was only 5% behind the best GSR result on the same data. This suggests that reliable markers of cognitive changes can be captured by behavioral data that are more suitable for group settings than wearable devices, and that combining modalities is beneficial.

The Development of Patterns of Gaze to a Speaking Face

Julia Irwin & Lawrence Brancazio, *Yale University, Southern Connecticut State University*

Pattern of gaze to a speaking face was examined in typically developing children ranging from 5-10 years of age and in adults. Children viewed the speaking face in a visual only (speechreading) condition, in the presence of auditory noise and in an audiovisual mismatch (McGurk) condition. Amount of gaze to the face and fixation on the mouth of the speaker were examined over the course of a trial where a consonant vowel /ma/ or /na/ was spoken. Results indicate an increase in gaze on the face, and more specifically in fixations on the mouth of a speaker, between the ages of 5 and 10. These results reveal changes in pattern of gaze to the face with development. Further, pattern of gaze to the face of the speaker may help account for previous findings in the literature showing that visual influence on heard speech increases with development.

Oral Session 8

Integration of Auditory, Labial and Manual Signals in Cued Speech Perception by Deaf Adults: An Adaptation of the McGurk Paradigm

Clémence Bayard, *CNRS Grenoble*; Jacqueline Leybart & Cécile Colin, *Université Libre de Bruxelles;*

Among deaf individuals fitted with a cochlear implant, some use Cued Speech (CS; a system in which each syllable is uttered with a complementary manual gesture) and therefore, have to combine auditory, labial and manual information to perceive speech. We examined how audio-visual (AV) speech integration is affected by the presence of manual cues and on which form of information (auditory, labial or manual) the CS receptors primarily rely depending on labial ambiguity. To address this issue, deaf CS users (N=36) and deaf CS naïve (N=35) participants were submitted to an identification task of two AV McGurk stimuli (either with a plosive or with a fricative consonant). Manual cues were congruent with either auditory information, lip information or the expected fusion. Results revealed that deaf individuals can merge audio and labial information into a single unified percept. Without manual cues, participants gave a high proportion of fusion response (particularly with ambiguous plosive McGurk stimuli). Results also suggested that manual cues can modify the AV integration and that their impact differs between plosive and fricative McGurk stimuli.

Explaining the Visual and Masked-Visual Advantage in Speech Perception in Noise: The Role of Visual Phonetic Cues

Vincent Aubanel, Christopher Davis & Jeesun Kim, *University of Western Sydney*

Visual enhancement of speech intelligibility, although clearly established, still resists a clear description. We attempt to contribute to solving that problem by proposing a simple account based on phonetically motivated visual cues. This work extends a previous study quantifying the visual advantage in sentence intelligibility across three conditions with varying degrees of visual information available: auditory only, auditory visual orally masked and auditory-visual. We explore the role of lexical as well as visual factors, the latter derived from groupings in visemes. While lexical factors play an undiscriminative role across modality conditions, some measure of viseme confusability seems to capture part of the performance results. A simple characterisation of the phonetic content of sentences in terms of visual information occurring exclusively inside the mask region was found to be the strongest predictor for the auditory-visual masked condition only, demonstrating a direct link between localised visual information and auditory-visual speech processing performance.

Analysing the Importance of Different Visual Feature Coefficients

Danny Websdale & Ben Milner, *University of East Anglia*

A study is presented to determine the relative importance of different visual features for speech recognition which includes pixel-based, model-based, contour-based and physical features. Analysis to determine the discriminability of features is performed through F-ratio and J-measures for both static and temporal derivatives, the results of which were found to correlate highly with speech recognition accuracy (r = 0:97). Principal component analysis is then used to combine all visual features into a single feature vector, of which further analysis is performed on the resulting basis functions. An optimal feature vector is obtained which outperforms the best individual feature (AAM) with 93:5% word accuracy.

Auditory and Audiovisual Close-Shadowing in Normal and Cochlear-Implanted Hearing Impaired Subjects

Lucie Scarbel, Denis Beautemps, & Jean-Luc Schwartz, *GIPSA-lab;* Marc Sato, *Laboratoire Parole et Langage*

This study takes place in the theoretical background of perceptuo-motor linkage in speech perception. A close-shadowing experiment has been carried out on post-lingual cochlear implanted (CI) and normal-hearing (NH) adults in order to evaluate sensory-motor interactions during joint perception and production of speech. To this aim, participants have to categorize audio (A) and audiovisual (AV) syllables as quickly as possible, with two modes of responses, oral or manual. Overall, responses from NH were globally faster and more precise than those of CI, although adding the visual modality led to a gain in performance in CI. Critically, oral responses were faster but less precise than manual responses in the two groups, with a stronger difference observed for CI than for NH. Despite auditory deprivation, these results suggest the involvement of sensory-motor interactions during speech perception in CI, albeit possibly less efficient than in NH.