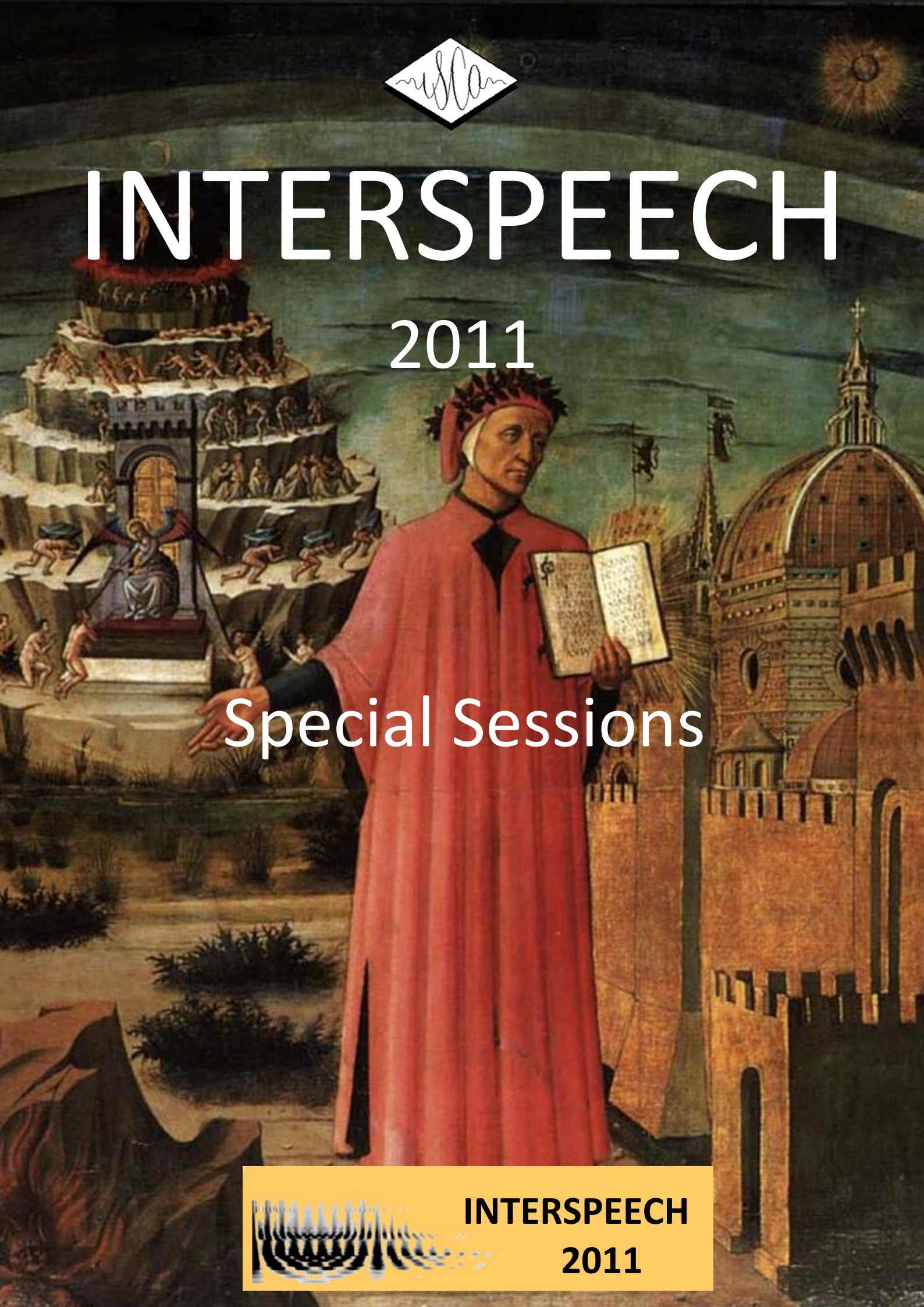

INTERSPEECH

2011

Special Sessions

INTERSPEECH
2011

Special Sessions

The Interspeech 2011 Organisation Committee is pleased to announce acceptance of the following Special Sessions at Interspeech 2011.

Speech and Language Processing Based Assistive Technologies and Health Applications

Sun-Ses2-S1-O (oral): *Sunday 13:30 - 14:30*
Sun-Ses2-S1-P (poster): *Sunday 14:30 - 15:30*
Place: Pala Affari, Caravaggio (Adua 1)

The goal of this special session is to bring together different communities working towards speech-based assistive technologies and health applications. In this special session, we hope to provide a venue not only to share technical approaches towards building speech-based health informatics and assistive technologies but also to attract the speech and language processing community to this exciting emerging field. With the advances in speech and spoken language processing in the last decade, many speech-based assistive and enabling applications have emerged towards easing the daily life and work, continuously monitor, assess, and understand activities using speech, help people in need for various reasons using either speech-based solutions or using features from their speech, or even just to keep company when needed. These include a diverse area of research efforts:

Speech technology can be utilized for interactive speech-based personal assistant technologies in smart environments or mobile devices. Some applications can be command/control systems (e.g., entertainment center control in a car), question answering systems (either factoid, e.g., where is the nearest gas station?, or personal, e.g., do I have a meeting tomorrow?), or information access or distillation systems using a document collection, such as news or web (e.g., anything new about Prop 8 in California?). Designing tailored spoken language interfaces targeting the needs of a range of populations, from the elderly to young children, especially those with specific mental of physical health issues is a great need, and an open technical challenge. Speech and language technology can monitor the well-being of people, such as elderly using cues from their speech, assist people in everyday life, such as reading web pages or books for blind people or children who do not know how to read. The latter may include educational applications such as tutoring systems for children or adults or helping users to learn a new language for pronunciation or grammar, or even new socio-communication behavior (such as in children with Autism).

Speech technology opens a plethora of powerful health applications. Different developmental, mental, or anatomic defects or diseases can lead to voice, speech, language and reading disorders. Examples are cognitive impairment, depression and PTSD, dyslexia and Alzheimer's, Autism, oral and laryngeal cancer, dysarthria, cleft lip and palate, Sigmatism, and stuttering. Monitoring technologies using acoustic, prosodic, and lexical cues can be utilized to detect disorders or specific articulation difficulties or to assess and quantize voice, speech and language (how "good" is a person's voice, speech, or language). A reliable assessment allows to monitor the progress of diseases or therapies and a comparison between different possible therapies, different hospitals and doctors. It provides the basis for interactive training tools with direct feedback for home usage. Examples are training tools where patients can train their speech and voices after total or partial laryngectomy in order to improve their intelligibility, practicing tools in order to improve certain phonemes specific issues like Sigmatism, or training tools to avoid stuttering. Voice conversion technologies allow the transformation of handicapped voices to more natural voices and would improve the communication affected persons.

It is clear that significant technical improvements are needed to enable such assistive technologies in diverse areas of research, such as machine learning, speech and language processing, and multi-modal human/machine interaction, and in most cases, collaboration with social science fields, such as gerontology, pediatrics, preventive medicine or education. However, most underlying technologies are now at a level to make these applications possible in everyday use.

Organisers: Gokhan Tur (gokhan.tur@ieee.org), Microsoft; Elmar Nöth (noeth@informatik.uni-erlangen.de), FAU Erlangen-Nuremberg, Germany; Shrikanth Narayanan (shri@sipi.usc.edu), USC-University of Southern California, LA, CA, US; Tobias Bocklet

(tobias.bocklet@informatik.uni-erlangen.de), FAU Erlangen-Nuremberg, German.

Crowdsourcing for speech processing

Sun-Ses3-S1-O (oral): *Sunday 16:00 - 17:00*

Sun-Ses3-S1-P (poster): *Sunday 17:00 - 18:00*
Place: Pala Affari, Caravaggio (Adua 1)

As the amount of speech data that is used in research and commercial applications has risen, so has the issue of how to obtain this data, how to transcribe it and how to assess the quality of systems that use speech. Until recently the solution to this has been expensive and lengthy. Experts have, for example, either annotated data by themselves or trained groups of people to do the task. It is costly to pay them and the combination of the training process and the subsequent throughput has added a considerable delay to system development. Recently, in answer to this, several researchers have turned to crowdsourcing, where non-experts perform tasks in exchange for a certain incentive. The literature in this area (http://sites.google.com/site/amtworkshop2010/) shows that, when used properly, this approach can produce results that are comparable to what is obtained from experts in less time and more inexpensively. Typically, crowdsourcing involves asking several workers to perform very small tasks (like labelling one sentence) and remunerating them with small amounts (like $.05 per task) for this. With access to a multitude of workers, the tasks are accomplished very quickly. Typical sources of workers, although certainly not the only ones, are Amazon Mechanical Turk and CrowdFlower. This session will examine the breadth of work in this area and include papers about the use of crowdsourcing for speech processing. The following is a list of possible areas, although papers in other related areas are welcome: data acquisition, speech labeling, assessment and evaluation, user studies involving speech.

While the sources of workers are not yet available in some countries, several efforts have been started to create local groups of workers and we welcome researchers who can describe their efforts to set up such a source.

We expect that this session will not only interest those who have used crowdsourcing and want to show their findings, but also those who are curious about how crowdsourcing can be useful for their speech processing needs..

Organisers: Maxine Eskenazi (max@cmu.edu), Carnegie Mellon University, Pittsburgh, PA, US; Helen Meng (hmmeng@se.cuhk.edu.hk), The Chinese University of Hong Kong; David Suendermann (david@speechcycle.com), SpeechCycle, Inc., NY, US; Gina Levow (levow@u.washington.edu) , University of Washington, US.

Spoken Language Processing of Human-Human Conversations

Tue-Ses2-S1-O (oral): *Tuesday 13:30 - 14:30*
Tue-Ses2-S1-P(poster): *Sunday 14:30 - 15:30*
Place: Pala Affari, Caravaggio (Adua 1)

Conversations provide an efficient way of human interaction, and create unique knowledge sharing opportunities between people with different areas of expertise. Although conversations are so common, there is still no globally adopted automated or semi-automated mechanism for processing these conversations, saving and analyzing their content for later use, or mining them.

The availability of human-human phone conversations, such as the Switchboard corpus, and multi-party meeting recordings such as the ICSI and AMI corpora, publicly broadcast conversations such as talk shows, and shared task evaluations such as the ones performed by NIST, have facilitated research on automatic processing of conversational speech. There have been active research efforts on processing of human conversations, covering a broad range of tasks from speech recognition, speaker diarization, to speech understanding and social signal analysis. The purpose of this special session is to offer the opportunity to the researchers working on speech and language processing of conversations to share ideas and have constructive discussions and provide all others an opportunity to find out about the latest developments in this field, in a single, focused, special conference session. The focus areas in the proposed special session include: Dialog act tagging of conversations, Automatic summarization of conversations, Speech analytics on human-human dialogs, Social signal analysis in conversations, Extraction of conversation structure, Annotation and evaluation issues, Information extraction from human conversations, Multimodal analysis of human conversations.

Organisers: Dilek Hakkani-Tur (dilek@ieee.org), Microsoft Speech Labs, US; Yang Liu (email:yang.liu@utdallas.edu), University of Texas at Dallas, US.

Speech and audio processing for human-robot interaction

Tue-Ses3-S1-O (oral): *Tuesday 16:00 - 17:00*
Tue-Ses3-S1-P(poster): *Sunday 17:00 - 18:00*
Place: Pala Affari, Caravaggio (Adua 1)

The field of human-robot interaction is attracting an increasing amount of interest from researchers making this an ideal time to highlight the work being done in human-robot spoken language interaction. Social interaction is

characterized by a continuous and dynamic exchange of informationcarrying signals. Producing and understanding these signals allow humans to communicate simultaneously on multiple levels. Such signals include: speech and non-speech sounds, gesture, facial expression and pose. Among these channels, vocal expression is best suited for communicating a rich variety of information; it is also the most natural modality for communicating meaning, emotion and personality. Vocal expression is characterized by a verbal component (language) and by a non-verbal component (prosody, intonation, hesitation).

Our current ability to model vocal communication is quite limited; spoken language systems, robots in our case, are able to communicate concrete meaning through language but their ability to detect (or for that matter generate) non-linguistic information streams is quite primitive. The ability to understand this information, and for that matter adapt generation to the goal of the communication and the characteristic of particular interlocutors, constitutes a significant aspect of natural interaction. The purpose of this special session is to bring together researchers who are exploring vocal expression from different perspectives, including detection, modeling and generation. The focus of the session is on audio verbal and non-verbal cues required for the design of natural interaction between a human and a robot. Special session topics may include, but are not limited to: Speech recognition systems for HRI, Dialog systems for HRI, Automatic emotion detection from verbal and non-verbal cues, Automatic recognition of user personality in dialog, Multimodal speech/audio expression generation in robots, Perception-action loops in robots, Back-channel generation and understanding, Interpretation of prosodic information, Timing in discourse, Integrated models of vocal communication, Natural Human-Robot Interaction (HRI).

Organiser: Laurence Devillers (devil@limsi.fr), LIMSI-CNRS, FR; Agnes Delaborde (agnes.delaborde@limsi.fr), LIMSI-CNRS, FR; Alexander Rudnicky (Alex.Rudnicky@cs.cmu.edu) , Carnegie Mellon University, Pittsburgh, PA, US.

Speech Technology for Under-Resourced Languages

Wed-Ses1-S2-O Wednesday 10:00 - 11:00
Wed-Ses1-S2-P Wednesday 11:00 - 12:00
Place: Pala Affari, Caravaggio (Adua 1)

Most research in the field of speech technology has traditionally been conducted in very few languages such as English, French, Spanish or Chinese. The given special poster session is focused on research and development of new speech technologies for less-resourced national languages, mainly, used in the following large geographical regions: Eastern Europe, South and Southeast Asia, West Asia, North Africa, Sub-Saharan Africa. Usually these languages are under-presented at the Interspeech conferences and other ISCA-originated programs.

The special session is open to discuss problems and peculiarities of targeted languages in application to spoken language technologies, including but not limited to automatic speech recognition, text-to-speech, speech-to-speech translation, spoken dialogue systems in an internationalized context. When developing speech-based technologies researchers are faced with many new problems from lack of audio databases and linguistic resources (lexicons, grammars, text collections), to inefficiency of existing methods for language and acoustical modeling, and limited infrastructure for the creation of relevant resources. They often have to deal with novel linguistic phenomena that are poorly studied or researched (for instance, clicks in southern African languages, tone in many languages of the world, language switching in multilingual systems, etc). There are also lots of common problems and features in many groups of under-resourced languages and the aim of this special session is to provide a forum for the analysis of discussion of issues related to development of speech technologies/models/methods for these languages as well as to find some common approaches and solutions.

Well-written papers on targeted multi-language speech technologies are encouraged, and papers describing original results obtained for under-resourced languages, but important for well-elaborated languages too, are invited as well. Good papers from any countries and any authors may be accepted if they present new speech studies concerning the languages of interest of the special session. Submissions from countries suffering from under-resource language problems are high-priority for the given special session. All papers are expected to be submitted following the same schedule, procedure and format as for regular Interspeech papers, and will undergo the same blind review process by anonymous and independent reviewers as regular Interspeech-2011 papers. Authors shall also declare that their contributions are original and not being submitted for publication elsewhere (e.g., another conference, workshop, or journal). 4-page paper submission deadline is 31 March 2011 via the on-line paper submission system www.interspeech2011.org/ICMS/submissions. Accepted papers will appear in the Interspeech-2011 Proceedings and indexed in relevant databases (ISI), at least one author of each successful paper must register and attend the Interspeech-2011 in Florence and present own research in the poster form.

Organizers:

The Interspeech-2011 special poster session on Speech Technology for Under-Resourced Languages is organized by a consortium of ISCA International Affairs Committee, SLTU International Workshop & Google Research. The alphabetical list of organizers is:
Etienne Barnard, North-West University, South Africa (etienne.barnard@gmail.com)
Laurent Besacier, Laboratoire d'Informatique de Grenoble, France (laurent.besacier@imag.fr)
Alexey Karpov, St. Petersburg Institute for Informatics and Automation, Russia (karpov_a@mail.ru)
Chafic Mokbel, University of Balamand, Lebanon (chafic.mokbel@balamand.edu.lb)
Pedro J. Moreno, Google Research, USA (pedro@google.com)
Yoshinori Sagisaka, Waseda University, Tokyo, Japan (ysagisaka@gmail.com)
Thippur Sreenivas, Indian Institute of Science, India (tvsree@ece.iisc.ernet.in)
Yun-Hsuan Sung, Google Research, USA (yhsung@google.com)

Special Events

The Interspeech 2011 Organisation Committee is pleased to announce acceptance of the following Special Events at Interspeech 2011.

Speech Processing Tools

Wed-Ses1-S3 (poster): Wednesday 10:00 - 12:00
Place: Pala Affari, Ghirlandaio & Masaccio (room 226 & 336)

The proposed special session aims to establish a forum for developers of off-the shelf software for speech processing. Such a forum will improve the scientific visibility of tool development and allow developers to obtain academic recognition for their highly interdisciplinary work. For Interspeech 2011, the focus of the special session should be on the interoperability of tools. The special session on Speech Processing Tools is about off-the-shelf software for speech technology - software that one can download or license which supports tasks relevant to the processing of speech data and which is intended to serve the linguist, phonetician, technologist or other researcher in his or her daily work. Examples of such software are SFS, Praat, ELAN, SpeechRecorder, Emu, Transcriber, NXT, etc. This session is not about clever algorithms (which are always useful and needed), but focuses much more on the joy of use, stability, appropriateness, availability, and interoperability. The motivation for such a special session is that currently magnificent tools exist which are used in many labs. However, the people who developed these tools have problems getting proper recognition for their work (outside of the satisfaction of seeing one's own tools on other people's laptops during a conference). In the scientific application area of software, tool development is considered less interesting than publishing results obtained with these tools. In computer science or software engineering, developing tools for a given application is considered "just application work" that only sometimes has to do with the frontiers of research. Most speech technology or speech researchers will have had the experience that obtaining funding for tool development is next to impossible - one often has to hide development costs somewhere in the budget. With tool development receiving academic recognition this might slowly change for the better.

Organisers: Christoph Draxler (draxler@phonetik.uni-muenchen.de), Institute of Phonetics and Speech Processing, Munich DE; Toomas Altosaar (Toomas.Altosaar@hut.fi), Aalto University School of Science and Technology, Helsinki University of Technology, SF.; Sadaoki Furui (furui@cs.titech.ac.jp), Tokyo Institute of Technology, Tokyo, Japan; Mark Liberman, Linguistic Data Consortium, Philadelphia, PA, US.

Speaker State Challenge - Intoxication and Sleepiness

Wed-Ses1-S1(oral) Wednesday 10:00 - 12:00
Wed-Ses2-S1 (oral) Wednesday 13:30 - 15:30
Place: Pala Affari, Raffaello - (3rd floor)

While the first open comparative challenges in the field of paralinguistics targeted more "conventional" phenomena such as emotion, age, and gender, there still exists a multiplicity of not yet covered, but highly relevant speaker states and traits. Thus, the INTERSPEECH 2011 Speaker State Challenge broadens the scope by addressing two less researched speaker states while focusing on the crucial application domain of security and safety: the computational analysis of intoxication and sleepiness in speech. Apart from intelligent and socially competent future agents and robots, main applications are found in the medical domain and surveillance in high-risk environments such as driving, steering or controlling. The INTERSPEECH 2011's theme "Speech science and technology for real life" is not only generally reflected in these every-day application scenarios, but also in particular by the conditions of the Challenge

such as naturalistic paralinguistic phenomena and no pre-selection of instances. For these Challenge tasks, the ALCOHOL LANGUAGE CORPUS (ALC) and the SLEEPY LANGUAGE CORPUS (SLC) with genuine intoxicated and sleepy speech will be provided by the organisers. The first consists of 39 hours of speech, stemming from 154 speakers in gender balance, and will serve to evaluate features and algorithms for the estimation of speaker intoxication in gradual blood alcohol percentage. The second features 10 hours of speech recordings of 50 subjects, annotated in 10 different levels of sleepiness. The verbal material consists of different complexity reaching from sustained vowel phonation to natural communication. The corpora further feature detailed speaker meta data, orthographic transcript, phonemic transcript, and segmentation and multiple annotation tracks. Both are given with distinct definitions of test, development, and training partitions, incorporating speaker independence as needed in most real-life settings. Benchmark results of the most popular approaches will be provided.

Two Sub-Challenges are addressed: the Intoxication Sub-Challenge, the degree of speakers' intoxication by alcohol consumption has to be determined by regression, covering blood alcohol concentration from 0 to 1.6 per mill. The measures of competition will thus be cross-correlation and mean linear error; the Sleepiness Sub-Challenge, the sleepiness of a speaker in an ordinal scale from 1 to 10 has to be determined by a suited regression algorithm and

Transcription of the train and development sets will be known. Contextual knowledge may be used, as the sequence of chunks will be given. All Sub-Challenges allow contributors to find their own features with their own machine learning algorithm. However, a standard feature set will be provided per corpus that may be used. Participants will have to stick to the definition of training, development, and test sets. They may report on results obtained on the development set, but have only three trials to upload their results on the test sets, whose labels are unknown to them. Each participation will be accompanied by a paper presenting the results that undergoes peer-review and has to be accepted for the conference in order to participate in the Challenge.

Organisers: Björn Schuller (email: schuller@IEEE.org), Technische Universität, Munich, DE; Stefan Steidl (steidl@icsi.berkeley.edu), ICSI Berkeley, CA, US; Anton Batliner (batliner@informatik.uni-erlangen.de), Friedrich-Alexander-University, Erlangen-Nuremberg, DE; Florian Schiel (schiel@phonetik.uni-muenchen.de), University of Munich / Bavarian Archive for Speech Signals Services, Munich, DE; Jarek Krajewski (krajewsk@uni-wuppertal.de), Bergische Universität Wuppertal, Wuppertal, DE.

Sponsors: HUMAINE Association (www.emotion-research.net); Bavarian Archive for Speech Signals (BAS)

Show and Tell (DEMO session)

Mon-Ses2-S1 (poster) *Monday 13:30 - 15:30*
Place: Pala Congressi - Donatello -(Onice) - ground floor
[Pala Affari, Masaccio (room 226)]

Tue-Ses1-S1 (poster) *Tuesday 10:00 - 12:00*
Place: Pala Congressi - Donatello -(Onice) - ground floor
[Pala Affari, Masaccio (room 226)]

We are delighted to host the first Show and Tell event at INTERSPEECH 2011. Show&Tell will provide researchers, technologists and practitioners from academia, industry and government the opportunity to demonstrate their latest research systems and interact with the attendees in an informal setting. Demonstrations need to be based on innovations and fundamental research in areas of human speech production, perception, communication, and speech and language technology and systems.

Demonstrations has been peer-reviewed by members of the Interspeech Program Committee, who judged the originality, significance, quality, and clarity of each submission.

Organiser: Mazin Gilbert (AT&T Labs Research, USA), Dimitrios Dimitriadis (AT&T Labs Research, USA)