SCHEDULE

	Monday 1	Tuesday 2	Wednesday 3	Thursday 4	Friday 5
9:30-10:00	Registration- Opening	Audio Segmentation Part I	Speaker Recognition Part I	Vicomtech presentation	Poster Session: Student PhD-Thesis Work
10:30 – 11:00	Speech Synthesis Part I			Language Recognition Part I	
11:00-11:30	Coffee Break	Coffee Break	Coffee Break	Coffee Break	Coffee Break
11:30 -13:00	Speech Synthesis Part II	Audio Segmentation Part II	Speaker Recognition Part II	Language Recognition Part II	David van Leeuwen Keynote Speech
13:00 -14:30	Lunch at "EET Cafeteria"	Lunch at "EET Cafeteria"	Lunch at "EET Cafeteria"	Lunch at "EET Cafeteria"	Closing ceremony - Lunch
14:30 – 16:45	Speech Synthesis Part III	Audio Segmentation Part III	Speaker Recognition Part III	Language Recognition- Part III	
16:45 – 17:10	Speech Synthesis - Closing & Summary	Audio Segmentation - Closing & Summary	Speaker Recognition - Closing & Summary	Language Recognition - Closing & Summary	
17:10 – 17:30		Spotlights of Student PhD-Thesis Work S1	Spotlights of Student PhD-Thesis Work S2		
17:30 – 23:55				Social Gathering : Trip to Pazo Baion, an outsanting Albariño Winery and Gala Dinner	
21:00 – 23:00	Opening Dinner ("Hotel Bahia")				

PROGRAM DESCRIPTION

Opening Ceremony:

- Doña Asunción Longo, Vicerrectora de Investigación de la Universidad de Vigo
- Doña Edita de Lorenzo, Directora de la Escuela de Ingeniería de Telecomunicación de la Universidad de Vigo
- Doña Nuria González Prelcic, Directora de AtlantTIC Research Center
- Doña Carmen García Mateo, Chair of 2013 RTTH Summer School

Technical Sessions

There will four sessions, each devoted to one single speech technology evaluation. The program completes with a keynote lecture from an expert on tecnhnoloy evaluation.

Date	Topic	Lecturers
Monday, July 1st	Speech Synthesis	**Juan Manuel Montero Martínez** Speech Technology Group (GTH) Department of Electronic Engineering (IEL) Technical University of Madrid (UPM)
Tuesday, July 2nd	Audio Segmentation	**Laura Docio Fernández** Multimedia Technologies Group (GTM) AtlantTIC Research Center University of Vigo
		Paula López Otero Multimedia Technologies Group (GTM) AtlantTIC Research Center University of Vigo
		Alfonso Ortega Vivolab, Aragon Institute for Engineering Resarch (I3A), University of Zaragoza
Wednesday, July 3rd	Speaker Recognition	**Joaquin González-Rodríguez** Biometric Recognition Group – ATVS Escuela Politecnica Superior Universidad Autonoma de Madrid
		Javier González-Dominguez Biometric Recognition Group – ATVS Escuela Politecnica Superior Universidad Autonoma de Madrid

Date	Topic	Lecturers
Thursday, July 4th	Language Recognition	**Luis Javier Rodríguez Fuentes** Software Technologies Working Group (GTTS) Department of Electricity and Electronics (ZTF-FCT) University of the Basque Country (UPV/EHU)
		Mikel Penagarikano Software Technologies Working Group (GTTS) Department of Electricity and Electronics (ZTF-FCT) University of the Basque Country (UPV/EHU)
Friday, July 5th	Keynote Speech: **The importance of evaluation in speech engineering**	**David van Leeuwen** Netherlands Forensic Institute and Radboud Univ. Nijmegen (Netherland)

Keynote Speech:

Title: **The importance of evaluation in speech engineering**

Abstract

Speech technology is an engineering discipline, and as such has to meet a clear requirement: it must work. This apparently should set clear criteria for what is acceptable for speech technological implementations. However, speech is also an expression of language, and is therefore inherently diverse and variable and the true answer may not alway be well defined. The acceptability of a technology is therefore not purely a binary decision, but should be measured on a continuous scale of usefulness. Speech technology can only move forward if its performance can be evaluated in a way that better performance corresponds to a more useful application. This allows the researcher to measure small improvements in utility, and cumulative improvements will gradually move the technology forward to be more useful.

This lecture reviews the general requirements for the specification of a clear task, metric and methodology in speech technologies. It will do this by analysing some existing evaluation paradigms, and looking at what these have meant to the direction of the research field as a whole.

Session on Emotional Text-To-Speech Evaluation

There will be two sessions: a theoretical introduction on Emotional TTS and evaluation (about two hours long) and a practical session on building an emotional TTS system (about 4 hours long). The theoretical lecture will describe current trends in speech synthesis, especially the techniques used by the systems of the Albayzin 2012 Speech Synthesis evaluation: HMM-based systems, hybrid systems, grapheme-based systems...
The practical session will comprise two parts. The first one will deal with processing the text in order to automatically predict the appropriate emotional polarity (happiness, sadness or neutral) and the appropriate emotional strength to synthesize each sentence of the text, using open-source machine learning tools. The second part will fine-tune the synthesis component the system, using emotion interpolation techniques and an HMM-based emotional TTS engine.

Session on Audio Segmentation Evaluation

This session will consist of two parts: an introductory talk (about two hours long) and a practice session (about four hours long).

In the introductory talk, a background on audio segmentation will be given to the students: description of the task, applications, main issues to deal with, and the evaluation framework as defined in Albayzin evaluations. After this, the most habitual audio features (MFCCs and its derivatives, LPCs, etc) and the main state-of-the-art approaches for audio segmentation and classification will be introduced (BIC algorithm, HMM segmentation, SVMs). Right after, fusion techniques for combining different audio segmentation systems will be presented. Lastly, the Albayzin Audio Segmentation Evaluation carried out in 2010 and 2012 will be described: description of the task, datasets, evaluation metrics and performance achieved by the proposed systems. A brief summary of the main issues found by the participants will be given.

In the practice session we will evaluate the performance of some of the previously mentioned techniques for audio segmentation. To that end, we will use some well-known toolkits. Specifically, we will use HTK for HMM-based segmentation and MatLab for BIC-based segmentation and system fusion. The data we will work with are the datasets proposed for the Albayzin 2012 Audio Segmentation Evaluation.

Session on Speaker Recognition Evaluation

The Speaker Recognition tutorial will consist in two lectures (1 hour each), and two laboratory sessions (2 hours each). The first lecture will give a chronological overview on speaker recognition technologies synchronized with a history of NIST Speaker Recognition evaluations, with their data sets and scoring procedures, analyzing the new challenges progressively introduced in each new eval and how the speaker recognition community have addressed the new problems introduced every new eval. Once the basic technologies have been introduced, in the second lecture we will present in detail the different elements to build the PLDA i-vector state-of-the-art speaker recognition system which will be implemented by the students in the laboratory sessions.

The practical part of this tutorial will guide students to build their own state-of-art speaker recognition (SR) system. Specifically, a SR system based on Probabilistic Linear Discriminant Analysis (PLDA) over i-vectors will be step-by-step developed from acoustic features. This practice session will be divided in two main blocks. The first block will be focused on extracting i-vectors from MFCC acoustic features; while the second block will be concerned with session variability compensation of the i-vectors through classical LDA and its probabilistic version: PLDA. Developed systems will be assessed following a standard evaluation protocol similar to those used in the well-known speaker recognition evaluations conducted by NIST (NIST SRE series). Regarding code and implementation, a standalone framework based on public tools and written in MATLAB will be provided

---

Session on Language Recognition Evaluation

This session will consist of two parts: an introductory talk (about two hours long) and a practice session (about 4 hours long).

In the introductory talk, we first aim to make the students understand the task of spoken language recognition (SLR) as defined in NIST and Albayzin evaluations. Then we will introduce different approaches to SLR ---which are classified either as acoustic or phonotactic, depending on the relying features---, from those applied in the nineties of the last century (GMM-UBM and PPRLM approaches) to the most successful state-of-the-art technologies (Phone-Lattice-SVM and MFCC-SDC-iVector approaches). The third part of the talk will deal with the backend and fusion models typically applied to combine different systems and get improved performance. Finally, we will describe the three Albayzin Language Recognition Evaluations (LRE) carried out in 2008, 2010 and 2012, emphasizing the differences among them and with regard to NIST LRE, and focusing on the conditions, datasets and evaluation measure defined for the Albayzin 2012 LRE, which will be used as benchmark in the practice session.

The practice session will be based on Matlab, so it would be advisable for the students to have basic notions of this language/framework. We will guide students in the development of two acoustic SLR systems, both based on MFCC-SDC features: a GMM-UBM approximation using dot-scoring and a state-of-the-art iVector system with generative (Gaussian) language models. Their performance will be measured and compared on the Albayzin 2010 and 2012 LRE. Finally, both systems will be combined under a discriminative fusion approach to get improved performance on the proposed tasks.

Poster Session: Student PhD-Thesis Work

Session	Name	Surname	Affiliation	Poster Title	Director/es
S1	Jorge	LLombart Gil	Universidad de Zaragoza	Study of the impact of the articulatory features in total variability space on hybrid acoustic models and speaker characterization	Antonio Miguel Artiaga
S1	Domingo	López Oller	Universidad de Granada	Robust Speech Transmission Over Erasure Channels	Ángel Gómez and José Luis Pérez Córdoba
S1	Jaime	Lorenzo Trueba	Speech Technology Group, ETSI Telecomunicacion, Universidad Politecnica de Madrid, Spain	Towards Speaking Style Transplantation in Speech Synthesis	Juan Manuel Montero
S1	Beatriz	Martínez-González	Speech Technology Group, ETSIT, Universidad Politecnica de Madrid	Selection of TDOA Parameters for MDM Speaker Diarization	José Manuel Pardo Muñoz
S1	Ganna	Raboshchuk	Universitat Politècnica de Catalunya	Acoustic event detection in multisource environments using supervised non-negative matrix factorization.	Climent Nadeu
S2	Fernando	de la Calle Silos	Universidad Carlos III de Madrid	Auditory Motivated Structuring Element for Morphological Filtering of Speech Spectrograms Applied to Automatic Speech Recognition	Carmen Peláez Moreno, Ascensión Gallardo Antolín
S2	Javier	Franco Pedroso	Universidad Autónoma de Madrid	Temporal Contours in Linguistic Units for Automatic Text-Independent Speaker Recognition	Joaquín González Rodríguez
S2	Iván	López Espejo	University of Granada	Acoustic Noise Estimation for Robust Speech Recognition	Antonio M. Peinado Herreros y Ángel M. Gómez Garcí
S2	Julia	Olcoz Martínez	Universidad de Zaragoza	Study of Unsupervised Learning Techniques in Automatic Speech Recognition Systems for Unrestricted Domains	Alfonso Ortega Giménez
S2	Dayana	Ribas Gonzalez	Advanced Technologies Application Center (CENATAV)	Noise compensation methods for speaker recognition	Eduardo Lleida Solano, Jose Ramon Calvo de Lara, Antonio Miguel Artiaga