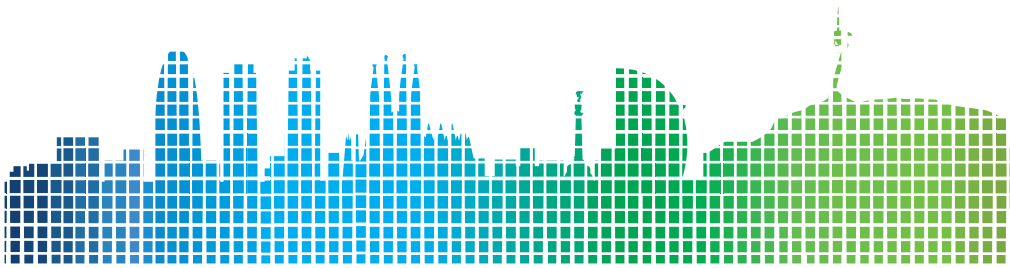
IberSPEECH2018

BARCELONANOVEMBER 21-23



Edited by Antonio Bonafonte, Jordi Luque and Jordi Pons

At the time of release, the proceedings can be downloaded from the website of IberSPEECH 2018: iberspeech2018.talp.cat

Barcelona, November 2018

Contents

Welcome Message

Welcome to IberSPEECH 2018 in Barcelona, November 21-23rd 2018, co-organized by Telefónica Research, the Center for Language and Speech Technologies and Applications (TALP) at the Universitat Politècnica de Catalunya (UPC), and the Research Group on Media Technologies (GTM) at La Salle, Universitat Ramon Llull. The current edition has received inestimable valued support by the Spanish Thematic Network on Speech Technology (RTTH), the Cátedra RTVE and the Voice Input Voice Output Laboratory (Vivolab) at Universidad de Zaragoza, and the ISCA Special Interest Group on Iberian Languages (SIG-IL). In addition, and for the very first time, IberSPEECH becomes an official ISCA Supported Event.

The IberSPEECH 2018 event — the fourth of its kind using this name — brings together the 10th Jornadas en Tecnologías del Habla and the 6th Iberian SLTech Workshop events, aiming to promote interaction and discussion among junior and senior researchers in the field of speech and language processing for Iberian languages.

Barcelona is a modern capital of 1.7 million people and the sixth-most populous urban area in Europe. It is the home of many points of interest declared as World Heritage Sites by UNESCO such as Sagrada Familia, Park Güell, and Palau de la Musica Catalana, and also the birthplace for great minds like Antoni Gaudí, Joan Miró, Montserrat Caballé, Eduardo Mendoza, and many more. Barcelona offers a unique combination of landscapes and weather, coupled with exquisite gastronomical experiences that are the result of a blend of heritage, produce, terroir, tradition, creativity, and innovation.

The venue, the Telefónica Diagonal ZeroZero Tower by Spanish architects EMBA, is located at the very beginning of the famous Diagonal Avenue, which crosses Barcelona diagonally from the sea to the Llobregat river. The position of the Diagonal ZeroZero Tower is exceptional; it is very visible from the city and the coast, and it lays on the border between the consolidated city and the large expanses of public space in the Forum area. In fact, the address of the tower is Diagonal Avenue, number 0 and is just next to the Forum Building. The Forum area is next to the sea and the Besòs River end. It is a new business centre, located in a reformed area. Torre Telefónica is hosting several Telefónica Group companies as well as the Telefónica Research group, a leading industrial research lab following an open research model in collaboration with Universities and other research institutions. It promotes the dissemination of scientific results both through publications in top-tier peer-reviewed international journals, conferences, and technology transfer.

Following with the tradition of previous editions, IberSPEECH 2018 will be a three-day event, bringing together the best researchers and practitioners in speech and language technologies in Iberian languages to promote interaction and discussion. The organizing committee has planned a wide variety of scientific and social activities, including technical paper presen-

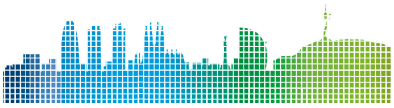
tations, keynote lectures, presentation of projects, laboratories activities, recent PhD thesis, discussion panels, a round table, and awards to the best thesis and papers.

The core of the scientific program of IberSPEECH2018 includes a total of 37 full regular paper contributions that will be presented distributed among 5 oral and 1 poster sessions. To ensure the quality of all the contributions, each submitted paper was reviwed by three members of the scientific review committee. All the papers in the conference will be accessible through the International Speech Communication Association (ISCA) Online Archive. Paper selection was based on the scores and comments provided by the scientific review committee, which includes over 86 researchers from different institutions (mainly from Spain and Portugal, but also from France, Germany, Brazil, Slovakia, Ireland, Greece, Hungary, Slovenia, Austria and United Kingdom).

Furthermore, it is confirmed to publish an extension of selected papers as a special issue of the Journal of Applied Sciences, "IberSPEECH 2018: Speech and Language Technologies for Iberian Languages", published by MDPI with fully open access. In addition to regular paper sessions, the IberSPEECH2018 scientific program features the following activities: the ALBAYZIN evaluation challenge session, a special session including the presentation of demos, research projects and recent PhD thesis, a round table and three keynote lectures.

Following the success of previous ALBAYZIN technology evaluations since 2006, this year ALBAYZIN evaluations have focused around multimedia analysis of TV broadcast content. Under the framework of a newly created Cátedra RTVE at Universidad de Zaragoza, we introduce and report on the results of the IberSPEECH-RTVE 2018 Challenge. The Corporación de Radiotelevisión Española (RTVE) has provided participants with an annotated TV broadcast database and the necessary tools for the evaluations, promoting the fair and transparent comparison of technology in different fields related to speech and language technology. It comprises four different challenge evaluations: Speech to Text Challenge (S2TC), Speaker Diarization Challenge (SDC) and Multimodal Diarization Challenge (MDC), organized by RTVE and Universidad de Zaragoza; and the Search on Speech Challenge (SoSC) jointly organized by Universidad San Pablo-CEU and AuDIaS from Universidad Autónoma de Madrid with the support of the ALBAYZIN Committee. Overall, 7 teams participated in the S2TC challenge, 8 teams in the SDC, 3 teams in the MDC, and 3 more teams in the SoSC challenge, which results in 21 system paper description contributions. Additionally, 11 special session papers are also included in the conference program. These were intended to describe either progress in current or recent research and development projects, demonstration systems, or PhD Thesis extended abstracts to compete in the PhD Award. Furthermore, IberSPEECH2018 features 3 remarkable keynote speakers: Prof. Tanja Schultz (University of Bremen, Germany and Institute of Carnegie Mellon, Pittsburgh, PA USA), Dr. Rob Clark (Google, London, UK) and Dr. Lluis Marquez (Amazon, Barcelona, Spain), to whom we would like to acknowledge for their extremely valuable participation.

Moreover, a round table with recognized experts brought discussion about the role of research and innovation from both academia and industry. Such a symbiosis creates market power by exploring and developing new categories which will eventually become the next blue oceans of our society. However, converting such activities into real businesses, making strong bones and figuring out new products that have a major impact in the real world is a non trivial task. A round table, we expected as an opportunity for both worlds on finding and exploring synergies and collaboration.



The social program of IberSPEECH2018 sets sail with the welcome reception at the Escola d'Enginyeria Barcelona Est (EEBE), at the recently created UPC Diagonal Besòs Campus, next to the Telefonica Tower. EEBE aims to become a top-quality academic centre in the field of engineering for the 21st-century industry that is capable of acting as an agent of transformation at a local and international level. The EEBE was born from the Barcelona College of Industrial Engineering (EUETIB) and from part of the teaching and research activity in chemical and materials engineering hitherto carried out at the Barcelona School of Industrial Engineering (ETSEIB). The gala dinner will be held at Restaurant Marítim, next to the legendary Barcelona Reial Club Marítim, designed by Lázaro Rosa-Violán from Contemporain Studio, to provide the aesthetics and flavours of different Mediterranean paradises.

Finally, we would like to thank all those whose effort made possible this conference, including the members of the organizing committee, the local organizing committee, the ALBAYZIN committee, the scientific reviewer committee, the authors, the conference attendees, the supporting institutions, and so many people who gave their best to achieve a succesful conference.

Barcelona, November 2018
Jordi Luque, General Chair

Organizing Committee

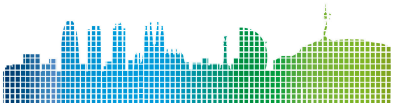
General Chair:
Jordi Luque, Telefónica Research, Spain

General Co-Chairs:
Antonio Bonafonte, Universitat Politècnica de Catalunya, Spain
Francesc Alías Pujol, La Salle – Universitat Ramon LLull, Spain
António Teixeira, Universidade de Aveiro, Portugal

Technical Program Chair:
Javier Hernando Pericás, Universitat Politècnica de Catalunya, Spain

Technical Program Co-Chairs:
Alberto Abad, INESC ID Lisboa / IST Lisboa, Portugal
Xavier Anguera, ELSA Corp., United States
Carlos David Martínez Hinarejos, Universitat Politècnica de València, Spain
Eva Navas, University of the Basque Country, UPV- EHU, Spain
Carlos Segura, Telefónica Research, Spain

Publication Chairs:
Francesc Alías Pujol, La Salle – Universitat Ramon LLull, Spain
Antonio Bonafonte, Universitat Politècnica de Catalunya, Spain



Special Session and Awards Chair:
Ascensión Gallardo-Antolín, Universidad Carlos III, Spain

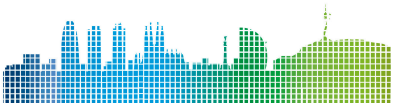
Plenary Talks & Round Table Chairs:
Joan Serrà, Telefónica Research, Spain
Marta Ruiz Costa-Jussà, Universitat Politècnica de Catalunya, Spain

Evaluations Chairs:
Alfonso Ortega, Universidad de Zaragoza, Spain
Eduardo Lleida, Universidad de Zaragoza, Spain
Luis Javier Rodríguez Fuentes, Universidad del País Vasco, Spain

Local Committee:
Francesc Alías Pujol, La Salle – Universitat Ramon LLull, Spain
Antonio Bonafonte, Universitat Politècnica de Catalunya, Spain
Jordi Luque, Telefónica Research, Spain
Jordi Pons, Universitat Pompeu Fabra, Spain
Bardia Rafieian, Universitat Politècnica de Catalunya, Spain
Marta Ruiz Costa-Jussà, Universitat Politècnica de Catalunya, Spain
Carlos Segura, Telefónica Research, Spain
Joan Serrà, Telefónica Research, Spain

Scientific Review Committee

Alberto Abad, INESC-IST, Portugal
Francesc Alías, La Salle – Universitat Ramon Llull, Spain
Aitor Alvarez, Vicomtech-IK4, Spain
Xavier Anguera, Miro ELSA Corp., Portugal
Jorge Baptista, INESC-ID Lisboa, Portugal
Plinio Barbosa, University of Campinas, Brazil
Fernando Batista, INESC-ID & ISCTE-IUL, Portugal
José Miguel Benedí, Universitat Politècnica de València, Spain
Antonio Bonafonte, Universitat Politècnica de Catalunya, Spain
Joao Cabral, Trinity College Dublin, Ireland
Francisco Casacuberta, Universitat Politècnica de València, Spain
María José Castro-Bleda, Universitat Politècnica de València, Spain
Ricardo Córdoba, Universidad Politécnica de Madrid, Spain
Conceicao Cunha, IPS Munich, Germany
Carme de-la-Mota, Universitat Autònoma de Barcelona, Spain
Arantza del Pozo, Vicomtech, Spain
Laura Docío-Fernández, University of Vigo, Spain
Daniel Erro, Cirrus Logic, Madrid, Spain
David Escudero, University of Valladolid, Spain
Nicholas Evans, EURECOM, France
Mireia Farrús, Universitat Pompeu Fabra, Spain
Rubén Fernández, Universidad Politécnica de Madrid, Spain
Javier Ferreiros, Universidad Politécnica de Madrid, Spain
Julian Fierrez, Universidad Autonoma de Madrid, Spain
Ascensión Gallardo, Universidad Carlos III de Madrid, Spain
Fernando García Granada, Universitat Politècnica de València, Spain
Carmen García Mateo, University of Vigo, Spain
Kafentzis George, University of Crete, Greece
Omid Ghahabi, EML European Media Laboratory GmbH, Germany
Juna Ignacio Godino, Llorente Universidad Politécnica de Madrid, Spain



Jon Ander Gómez, Universitat Politècnica de València, Spain
Emilio Granell, Universitat Politècnica de València, Spain
Inma Hernaez, University of the Basque Country (UPV/EHU), Spain
Javier Hernando, Universitat Politècnica de Catalunya, Spain
Lluís-F. Hurtado, Universitat Politècnica de València, Spain
Oliver Jokisch, Leipzig University of Telecommunications (HfTL), Germany
Oscar Koller, Microsoft Germany GmbH, Germany
Eduardo Lleida, University of Zaragoza, Spain
José David Lopes, Heriot Watt University UK
Paula López Otero, Universidade da Coruña, Spain
Jordi Luque, Telefónica Research, Spain
Carlos David Martínez Hinarejos, Universitat Politècnica de València, Spain
Helena Moniz, INESC/FLUL, Portugal
Juan Montero, Universidad Politécnica de Madrid, Spain
Nicolás Morales, Nuance Communications GmbH, Germany
Climent Nadeu, Universitat Politècnica de Catalunya, Spain
Juan L. Navarro-Mesa, Universidad de Las Palmas de Gran Canaria, Spain
Eva Navas, University of the Basque Country, Spain
Géza Németh, Budapest University of Technology & Economics, Hungary
Nelson Neto, Universidade Federal do Pará, Brazil
Hermann Ney, RWTH Aachen University, Germany
Alfonso Ortega, University of Zaragoza, Spain
Yannis Pantazis, Foundations for Research and Technology – Hellas, Spain
Carmen Peláez-Moreno University Carlos III Madrid, Spain
Thomas Pellegrini, Université de Toulouse; IRIT, France
Mikel Penagarikano, University of the Basque Country, Spain
Fernando Perdigao, Institute of Telecommunications (IT), Lisbon, Portugal
José L. Pérez-Córdoba, University of Granada, Spain
Ferran Pla, Universitat Politècnica de València, Spain
Jiri Pribil, Slovak Academy of Sciences Slovakia
Jorge Proenca, IT – Coimbra, Portugal
Michael Pucher, Acoustics Research Institute Austria
Paulo Quaresma, Universidade de Evora, Portugal
Ganna Raboshchuk, ELSA Corp., Portugal
Sam Ribeiro, The University of Edinburgh, UK
Eduardo Rodriguez Banga, University of Vigo, Spain
Marta Ruiz Costa-Jussà, Universitat Politècnica de Catalunya, Spain
Luis Javier Rodríguez-Fuentes, Univ. of the Basque Country UPV/EHU, Spain
Rubén San-Segundo, Universidad Politécnica de Madrid, Spain

Jon Sánchez, Aholab – EHU/UPV, Spain
Joan Andreu, Sanchez Universitat Politècnica de València, Spain
Emilio Sanchis, Universitat Politècnica de València, Spain
Diana Santos, University of Oslo, Norway
Ibon Saratxaga, University of the Basque Country, Spain
Encarna Segarra, Universitat Politècnica de València, Spain
Carlos Segura Perales, Telefónica Research, Spain
Joan Serrà, Telefónica Research, Spain
Alberto Simões, 2Ai Lab – IPCA, Portugal
Rubén Solera-Ureña, INESC-ID Lisboa, Portugal
António Teixeira, University of Aveiro, Portugal
Javier Tejedor, Universidad CEU San Pablo, Spain
Doroteo Toledano, Universidad Autónoma de Madrid, Spain
Isabel Trancoso, INESC ID Lisboa / IST, Portugal
Cassia Valentini-Botinhao, The University of Edinburgh, UK
Amparo Varona, University of the Basque Country, Spain
Andrej Zgank, University of Maribor, Slovenia
Catalin Zorila, Toshiba Cambridge Research Laboratory UK

Organizing Institutions

This conference has been organized by:

with the collaboration of:

IberSPEECH 2018 has been partially funded by the project Red Temática en Tecnologías del Habla 2017 (TEC2017-90829-REDT) founded by Ministerio de Ciencia, Innovación y Universidades.

Awards

Best Paper Award

All regular papers are candidates for this award. The award, given based on the review reports and the presentation at the conference, grants the authors the publicaton of an extended version of their work within the Special Issue of Applied Sciences journal (MDPI) entitled "IberSPEECH 2018: Speech and Language Technologies for Iberian Languages".

Best Albayzin evaluation system

Papers submitted to Albayzin evaluation tasks are candidates for these awards. The awards will be given to the winners of the Albayzin evaluation challenges, in accordance with the evaluation plan and rules defined for each task.

Best PhD Thesis award

Papers submitted to the PhD Thesis special session are candidates for this award. The award is given based on the decision of the committee formed by the members of the General chair, Technical Program chair and Special Session and Awards Chair. The award is given based on different criteria, including the quality of the document, impact of the thesis and clearness of the presentation at the conference.

Professional Career Prize in Speech Technologies

This is an honorary prize awarded by the Spanish Thematic Network on Speech Technology (RTTH) that recognizes experienced individuals who have made outstanding contributions related to speech technology research in Spain.

IberSPEECH 2018 Edition

- José B. Mariño Acebal, Universitat Politècnica de Catalunya
- Antonio Rubio Ayuso, Universidad de Granada.

Venue

Main conference venue

Torre Telefónica

Address:

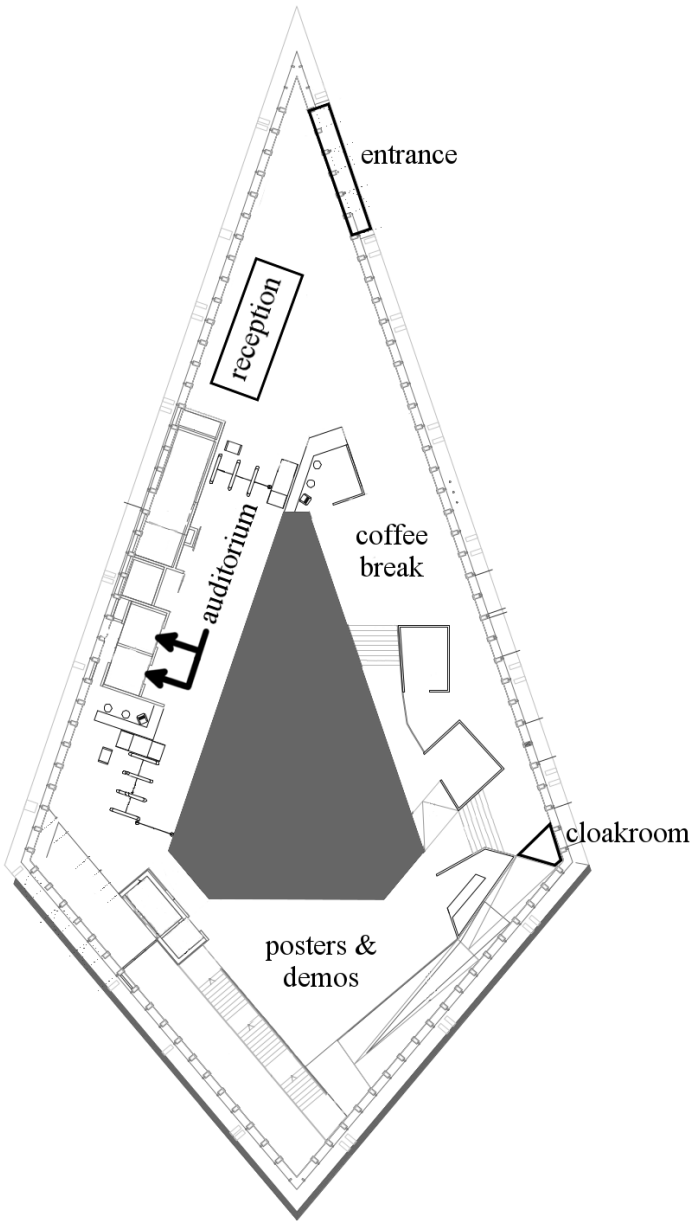
Torre Telefónica – Diagonal 00
Plaza de Ernest Lluch i Martín, 5
08019 Barcelona – Spain

WI-FI access instructions

Connect to **IberSPEECH 2018**'s network with the following password:

- Network: `Iberspeech 2018`
- Password: `*******`

The main body of the conference will be held in Torre Telefónica (floor 0, see figure) and the Auditorium in floor 2 by accessing the elevators depicted in the figure. The following diagram outlines the main conference areas and services:



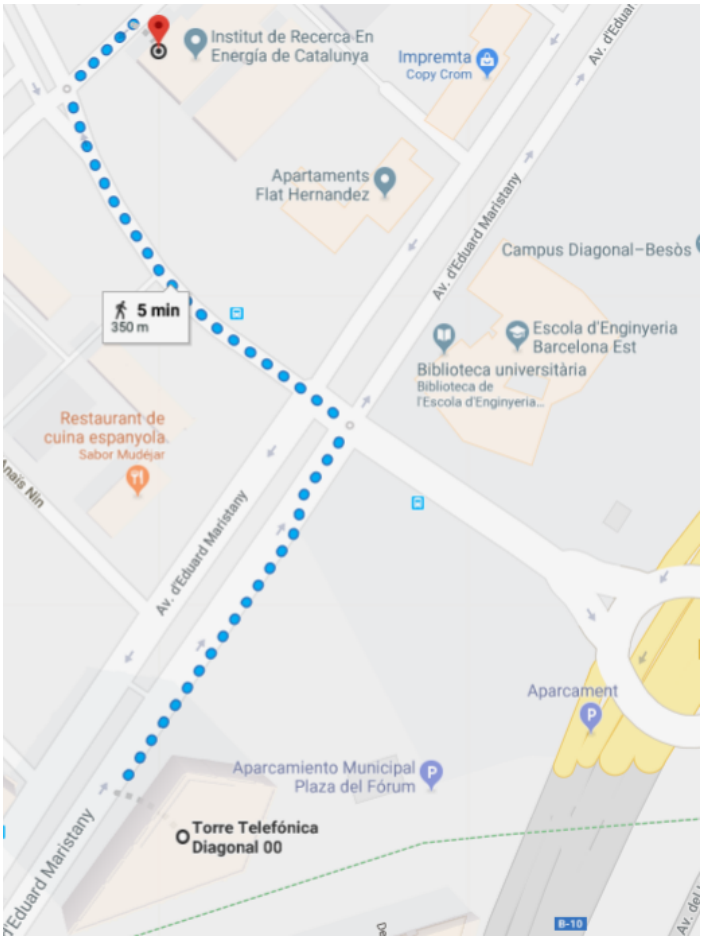
Lunch: 21 and 22 November, 13:30.

D'Ins Escola, Restaurant i Càtering

A gastronomic offer adapted to the occasion and a catering service with an added value that will enrich it: THE SOCIAL VALUE of the PEOPLE that work in this service. People who participate in a training and job placement program developed by the Fundación Formació i Treball.

Address: Carrer de Ramon Llull, 240, 08930, Sant Adrià del Besòs.

Wednesday 21st and Thursday 22nd lunches' will be given very close to Torre Telefónica (350 m).



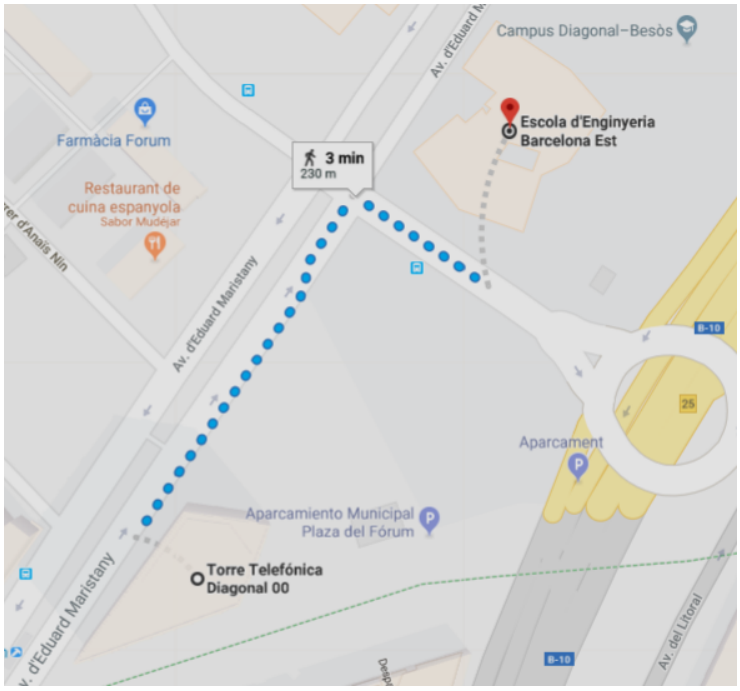
Social Program

Welcome Reception: Wednesday 21 November, 20:00.

Escola d'Enginyeria Barcelona Est (UPC Diagonal Besòs)

Address: Av. Eduard Maristany, 16, 08019, Barcelona

The welcome reception will be given at the **Rambla de Colors** space in UPC Diagonal Besòs (230 m from Torre Telefónica). The space is placed between buildings C and I being possible to access it from both main entrances at each building. Use the number 37400 at the entrance phone to connect with reception and granting access to the building and go down to the **underground level**.



Gala dinner: Thursday 22 November, 20:30.

Restaurant Marítim

Address: Moll d'Espanya, 08039, Barcelona

Telephone: +34 93 221 17 75

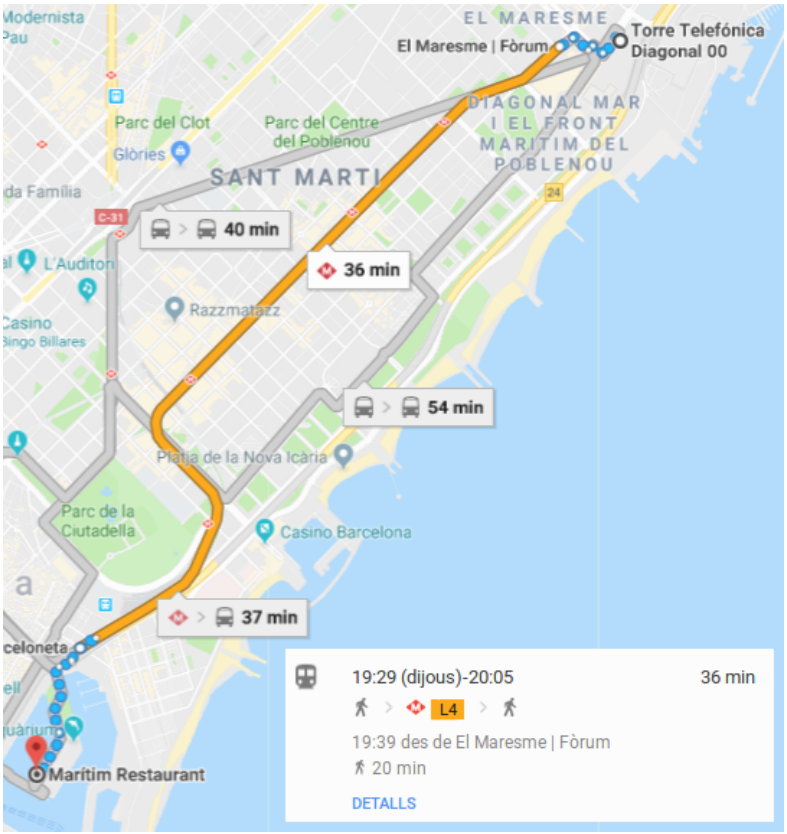
Web: www.maritimrestaurant.es

The gala dinner will be held downtown, next to the sea (close to Cristobal Colon statue).

How do I get there?

Take L4 from El Maresme – Fòrum metro station to Barceloneta.

You can also reach the restaurant by walking the seaside (1h 15 min).



Invited Speakers

Tanja Schultz received her diploma and doctoral degree in Informatics from University of Karlsruhe, Germany, in 1995 and 2000. Prior to these degrees she completed the state exam in Mathematics, Sports, Physical and Educational Science from Heidelberg University, Germany in 1989. She is currently the Professor for Cognitive Systems at the University of Bremen, Germany and adjunct Research Professor at the Language Technologies Institute of Carnegie Mellon, PA USA. Since 2007, she directs the Cognitive Systems Lab, where her research activities include multilingual speech recognition and the processing, recognition, and interpretation of biosignals for human-centered technologies and applications. Prior to joining University of Bremen, she was a Research Scientist at Carnegie Mellon (2000-2007) and a Full Professor at Karlsruhe Institute of Technology in Germany (2007-2015). Dr. Schultz is an Associate Editor of ACM Transactions on Asian Language Information Processing (since 2010), serves on the Editorial Board of Speech Communication (since 2004), and was Associate Editor of IEEE Transactions on Speech and Audio Processing (2002-2004). She was President (2014-2015) and elected Board Member (2006-2013) of ISCA, and a General Co-Chair of Interspeech 2006. She was elevated to Fellow of ISCA (2016) and to member of the European Academy of Sciences and Arts (2017). Dr. Schultz was the recipient of the Otto Haxel Award in 2013, the Alcatel Lucent Award for Technical Communication in 2012, the PLUX Wireless Biosignals Award in 2011, and the Allen Newell Medal for Research Excellence in 2002, and received the ISCA / EURASIP Speech Communication Best paper awards in 2001 and 2015.

Rob Clark received his PhD from the University of Edinburgh in 2003. His primary interest is in producing engaging synthetic speech. Before joining Google Rob was at the University of Edinburgh for many years involved in both teaching and research relating to text-to-speech synthesis. Rob was one of the primary developers and maintainers of the open source Festival text-to-speech synthesis system. In 2015 he joined Google where he is working on text-to-speech synthesis and prosody.

Lluís Màrquez is a Principal Applied Scientist at Amazon Research in Barcelona. From 2013 to 2017 he had a Principal Scientist role at the Arabic Language Technologies group from the Qatar Computing Research Institute (QCRI), and previously, he was Associate Professor at the Technical University of Catalonia (UPC, 2000-2013). He holds a university award-winning PhD in Computer Science from UPC (1999). His research focuses on natural language understanding by using statistical machine learning models. He has 150+ papers in Natural Language Processing and Machine Learning journals and conferences. He has been General and Program Co-chair of major conferences in the area (ACL, EMNLP, EACL, CoNLL, *SEM, EAMT, etc.), and held several organizational roles in ACL and EMNLP too. He was co-organizer of various international evaluation tasks at Senseval/SemEval (2004, 2007, 2010, 2015-2017) and CoNLL shared tasks (2004-2005, 2008-2009). He was Secretary and President of the ACL special interest group on Natural Language Learning (SIGNLL) in the period 2007-2011. More recently, he was President-elect and President of the European Chapter of the ACL (EACL; 2013-2016) and member of the ACL Executive Committee (2015-2016). Luís Màrquez has been Guest Editor of special issues at Computational Linguistics, LRE, JNLE, and JAIR in the period (2007-2015). He has participated in 16 national and EU research projects, and 2 projects with technology transfer to the industry, acting as the principal site researcher in 10 of them, helping companies embed AI in their business.

Technical Program

Speaker Recognition

Wednesday, 21 November 2018, 09:20 – 10:40

Chair: Xavier Anguera, ELSA Corp.

Keynote 1

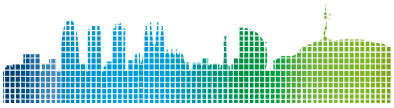
Wednesday, 21 November 2018, 11:00 – 12:00

Chair: Joan Serrà, Telefónica Research

Topics on Speech Technologies

Wednesday, 21 November 2018 **12:00 – 13:30**

Chair: Alberto Abad, INESC-ID/IST



ASR & Speech Applications

Wednesday, 21 November 2018, 15:00 – 16:40

Chair: Carmen García Mateo, University of Vigo

Speech & Language Technologies Applied to Health

Wednesday, 21 November 2018, 17:00 – 18:40

Chair: Mireia Farrús, Universitat Pompeu Fabra

Synthesis, Production & Analysis

Thursday, 22 November 2018, 09:00 – 10:40

Chair: Francesc Alías Pujol, La Salle - Universitat Ramon Llull

Keynote 2

Thursday, 22 November 2018, 11:00 – 12:00

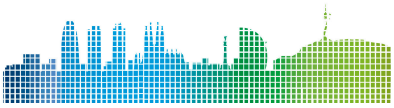
Chair: Antonio Bonafonte, Universitat Politècnica de Catalunya

Special Session

Thursday, 22 November 2018 **12:00 – 13:30**

Chair: Ricardo de Córdoba, Universidad Politécnica de Madrid

Show & Tell



Ongoing Research Projects

PhD Thesis

Albayzin Evaluation

Thursday, 22 November 2018 **15:00 – 16:40**

Chair: Alfonso Ortega & Eduardo Lleida, Universidad de Zaragoza

Multimodal Diarization Challenge

Speaker Diarization Challenge

Search on Speech Challenge

Speech to Text Challenge

Text & NLP Applications

Friday, 23 November 2018, 09:00 – 10:40

Chair: José F. Quesada, Universidad de Sevilla

Keynote 3

Friday, 23 November 2018, 11:00 – 12:00

Chair: Carlos Segura, Telefónica Research

Round Table

Friday, 23 November 2018, 12:00 – 13:00

Chair: Marta R. Costa-Jussà, Universitat Politècnica de Catalunya

Abstracts

Differentiable Supervector Extraction for Encoding Speaker and Phrase Information in Text Dependent Speaker Verification

Victoria Mingote, Antonio Miguel, Alfonso Ortega, Eduardo Lleida
University of Zaragoza, Spain

In this paper, we propose a new differentiable neural network architecture for text dependent speaker verification which uses alignment models to produce a supervector representations of an utterance. Unlike previous works with similar approaches, we do not extract the embedding of an utterance from the mean reduction of the temporal dimension. Our system replaces the mean by a phrase alignment model to keep the temporal structure of each phrase which it is relevant in this application since the phonetic information is part of the identity in the verification task. Moreover, we can apply a convolutional neural network as front-end and thanks to the alignment process being differentiable, we can train the whole network to produce a supervector for each utterance which will be discriminative with respect to the speaker and the phrase simultaneously. As we show, this choice has the advantage that the supervector encodes the phrase and speaker information providing good performance in text-dependent speaker verification tasks. In this work, the process of verification is performed using a basic similarity metric due to simplicity compared to other more elaborate models that are commonly used. The new model using alignment to produce supervectors was tested on the RSR2015-Part I database for text-dependent speaker verification, providing competitive results compared to similar size networks using the mean to extract embeddings.

Phonetic Variability Influence on Short Utterances in Speaker Verification

Ignacio Viñals, Alfonso Ortega, Antonio Miguel, Eduardo Lleida
University of Zaragoza, Spain

This work presents an analysis of i-vectors for speaker recognition working with short utterances and methods to alleviate the loss of performance these utterances imply. Our research reveals that this degradation is strongly influenced by the phonetic mismatch between enrollment and test utterances. However, this mismatch is unused in the standard i-vector PLDA framework. It is proposed a metric to measure this phonetic mismatch and a simple yet effective compensation for the standard i-vector PLDA speaker verification framework. Our results, carried out in NIST SRE10 coreext-coreext female det. 5, evidence relative improvements up to 6.65% in short utterances, and up to 9.84% in long utterances.

Restricted Boltzmann Machine Vectors for Speaker Clustering

Umair Khan, Pooyan Safari, Javier Hernando
Universitat Politècnica de Catalunya, Spain

Restricted Boltzmann Machines (RBMs) have been used both in the front-end and backend of speaker verification systems. In this work, we apply RBMs as a front-end in the context of speaker clustering. Speakers' utterances are transformed into a vector representation by means of RBMs. These vectors, referred to as RBM vectors, have shown to preserve speaker-specific information and are used for the task of speaker clustering. In this work, we perform the traditional bottom-up Agglomerative Hierarchical Clustering (AHC). Using the RBM vector representation of speakers, the performance of speaker clustering is improved. The evaluation has been performed on the audio recordings of Catalan TV Broadcast shows. The experimental results show that our proposed system outperforms the baseline i-vectors system in terms of Equal Impurity (EI). Using cosine scoring, a relative improvement of 11% and 12% are achieved for average and single linkage clustering algorithms respectively. Using PLDA scoring, the RBM vectors achieve a relative improvement of 11% compared to i-vectors for the single linkage algorithm.

Speaker Recognition under Stress Conditions

Esther Rituerto-González, Ascensión Gallardo-Antolín, Carmen Peláez-Moreno
Universidad Carlos III Madrid, Spain

Speaker Recognition systems exhibit a decrease in performance when the input speech is not in optimal circumstances, for example when the user is under emotional or stress conditions. The objective of this paper is measuring the effects of stress on speech to ultimately try to mitigate its consequences on a speaker recognition task. On this paper, we develop a stress-robust speaker identification system using data selection and augmentation by means of the manipulation of the original speech utterances. An extensive experimentation has been carried out for assessing the effectiveness of the proposed techniques. First, we concluded that the best performance is always obtained when naturally stressed samples are included in the training set, and second, when these are not available, their substitution and augmentation with synthetically generated stress-like samples, improves the performance of the system.

Keynote 1 **KN1**
Wednesday, 21 November 2018 **11:00 - 12:00**

Bio signal-based Spoken Communication

Tanja Schultz
University of Bremen, Germany and Institute of Carnegie Mellon, Germany

Speech is a complex process emitting a wide range of biosignals, including, but not limited to, acoustics. These biosignals – stemming from the articulators, the articulator muscle activities, the neural pathways, and the brain itself – can be used to circumvent limitations of conventional speech processing in particular, and to gain insights into the process of speech production in general. In my talk I will present ongoing research at the Cognitive Systems Lab (CSL), where we explore a variety of speech-related muscle and brain activities based on machine learning methods with the goal of creating biosignal-based speech processing devices for communication applications in everyday situations and for speech rehabilitation, as well as gaining a deeper understanding of spoken communication. Several applications will be described such as Silent Speech Interfaces that rely on articulatory muscle movement captured by electromyography to recognize and synthesize silently produced speech, Brain-to-text interfaces that recognize continuously spoken speech from brain activity captured by electrocorticography to transform it into text, and Brain-to-Speech interfaces that directly synthesize audible speech from brain signals.

Bilingual Prosodic Dataset Compilation for Spoken Language Translation

Alp Öktem [1], Mireia Farrús [1], Antonio Bonafonte [2]
[1]Universitat Pompeu Fabra, Spain, [2]Universitat Politècnica de Catalunya, Spain

This paper builds on a previous methodology that exploits dubbed media material to build prosodically annotated bilingual corpora. The almost fully-automatized process serves for building data for training spoken language models without the need for designing and recording bilingual data. The methodology is put into use by compiling an English-Spanish parallel corpus using a recent TV series. The collected corpus contains 7225 parallel utterances totaling to about 10 hours of data annotated with speaker information, word-alignments and word-level acoustic features. Both the extraction scripts and the dataset are distributed open-source for research purposes.

Building an Open Source Automatic Speech Recognition System for Catalan

Baybars Külebi, Alp Öktem
Col·lectivaT SCCL, Spain

Catalan is recognized as the largest stateless language in Europe hence it is a language well studied in the field of speech, and there exists various solutions for Automatic Speech Recognition (ASR) with large vocabulary. However, unlike many of the official languages of Europe, it neither has an open acoustic corpus sufficiently large for training ASR models, nor openly accessible acoustic models for local task execution and personal use. In order to provide the necessary tools and expertise for the resource limited languages, in this work we discuss the development of an ASR system using publicly available data, and CMU Sphinx 5pre-alpha. The resulting models give a WER of 31.95% on an external 4 hour multi-speaker test set. This value was further decreased to 11.68% with language model adaptation.

Multi-Speaker Neural Vocoder

Oriol Barbany Mayor, Antonio Bonafonte, Santiago Pascual de la Puente
Universitat Politècnica de Catalunya, Spain

Statistical Parametric Speech Synthesis (SPSS) offers more flexibility than unit-selection based speech synthesis, which was the dominant commercial technology during the 2000s decade. However, classical SPSS systems generate speech with lower naturalness than unit-selection methods. Deep learning based SPSS, thanks to recurrent architectures, surpasses classical SPSS limits. These architectures offer high quality speech while preserving the desired flexibility in choosing the parameters such as the speaker, the intonation, etc. This paper exposes two proposals conceived to improve deep learning-based text-to-speech systems. First a baseline model, obtained by adapting SampleRNN, making it as a speaker-independent neural vocoder that generates the speech waveform from acoustic parameters. Then two approaches are proposed to improve the quality, applying speaker dependent normalization of the acoustic features, and the look ahead, consisting on feeding acoustic features of future frames to the network with the aim of better modeling the present waveform and avoiding possible discontinuities. Human listeners prefer the system that combines both techniques, which reaches a rate of 4 in the mean opinion score scale (MOS) with the balanced dataset and outperforms the other models.

Improving the Automatic Speech Recognition through the improvement of Laguage Models

Andrés Piñeiro-Martín, Carmen García-Mateo, Laura Docío-Fernández
Multimedia Technologies Group (GTM), Universidade de Vigo, Spain

Language Models are one of the pillars on which the performance of automatic speech recognizer systems is based. Statistical language models based on the probability of word sequence (n-grams) are the most used, although deep neural networks begin to be applied. This is possible due to the increase of computation power along with improvements of algorithms. In this paper, the impact they have on the recognition result is studied in the following situations: 1) when they are adjusted to the work environment of the final application, and 2) when the complexity of these models grows by increasing the order of the n-gram models or applying deep neural networks. Specifically, an automatic speech recognition system with the different language models has been applied to audio recordings corresponding to three experimental frameworks: formal orality, talk on newscasts, and TED talks in Galician. The experimental results showed that improving the language models quality gives an improvement on the recognition performance.

Towards expressive prosody generation in TTS for reading aloud applications

Monica Dominguez [1], Alicia Burga [2], Mireia Farrús [1], Leo Wanner [3]

[1]Universitat Pompeu Fabra, Spain, [2]University Pompeu Fabra, Spain, [3]ICREA and University Pompeu Fabra, Spain

Conversational interfaces involving text-to-speech (TTS) applications have improved expressiveness and overall naturalness to a reasonable extent in the last decades. Conversational features, such as speech acts, affective states and information structure have been instrumental to derive more expressive prosodic contours. However, synthetic speech is still perceived as monotonous, when a text that lacks those conversational features is read aloud in the interface, i.e. it is fed directly to the TTS application. In this paper, we propose a methodology for pre-processing raw texts before they arrive to the TTS application. The aim is to analyze syntactic and information (or communicative) structure, and then use the high-level linguistic features derived from the analysis to generate more expressive prosody in the synthesized speech. The proposed methodology encompasses a pipeline of four modules: (1) a tokenizer, (2) a syntactic parser, (3) a communicative parser, and (3) an SSML prosody tag converter. The implementation has been tested in an experimental setting for German, using web-retrieved articles. Perception tests show a considerable improvement in expressiveness of the synthesized speech when prosody is enriched automatically taking into account the communicative structure.

Performance evaluation of front- and back-end techniques for ASV spoofing detection systems based on deep features

Alejandro Gomez-Alanis [1], Antonio M. Peinado [1], José Andrés González López [2], Angel M. Gomez [1]

[1]University of Granada, Spain, [2]University of Malaga, Spain

As Automatic Speaker Verification (ASV) becomes more popular, so do the ways impostors can use to gain illegal access to speech-based biometric systems. For instance, impostors can use Text-to-Speech (TTS) and Voice Conversion (VC) techniques to generate speech acoustics resembling the voice of a genuine user and, hence, gain fraudulent access to the system. To prevent this, a number of anti-spoofing countermeasures have been developed for detecting these high technology attacks. However, the detection of previously unforeseen spoofing attacks remains challenging. To address this issue, in this work we perform an extensive empirical investigation on the speech features and back-end classifiers providing the best overall performance for an antispoofing system based on a deep learning framework. In this architecture, a deep neural network is used to extract a single identity spoofing vector per utterance from the speech features. Then, the extracted vectors are passed to a classifier in order to make the final detection decision. Experimental evaluation is carried out on the standard ASVSpoof2015 data corpus. The results show that classical FBANK features and Linear Discriminant Analysis (LDA) obtain the best performance for the proposed system.

The observation likelihood of silence: analysis and prospects for VAD applications

Igor Odriozola, Inma Hernaez, Eva Navas, Luis Serrano, Jon Sanchez
Aholab (University of the Basque Country, UPV/EHU), Spain

This paper shows a research on the behaviour of the observation likelihoods generated by the central state of a silence HMM (Hidden Markov Model) trained for Automatic Speech Recognition (ASR) using cepstral mean and variance normalization (CMVN). We have seen that observation likelihood shows a stable behaviour under different recording conditions, and this characteristic can be used to discriminate between speech and silence frames. We present several experiments which prove that the mere use of a decision threshold produces robust results for very different recording channels and noise conditions. The results have also been compared with those obtained by two standard VAD systems, showing promising prospects. All in all, observation likelihood scores could be useful as the basis for the development of future VAD systems, with further research and analysis to refine the results.

On the use of Phone-based Embeddings for Language Recognition

Christian Salamea [1], Ricardo de Córdoba [2], Luis Fernando D'Haro [2], Rubén San-Segundo [2], Javier Ferreiros [2]

[1]Universidad Politécnica Salesiana, Ecuador, [2]Universidad Politécnica de Madrid, Spain

Language Identification (LID) is the process for automatically identifying the language of a given spoken utterance. We have focused in a phonotactic approach in which the system input is the phonemes sequence generated by a speech recognizer (ASR), but instead phonemes we have used phonetic units that contain context information "phone-grams". In this context, we propose the use of Neural Embeddings (NEs) as features for those phone-grams sequences, which are used as entries in a classical i-Vectors framework to train a multi class logistic classifier. These NEs incorporate information from the neighboring phone-grams in the sequence and model implicitly longer-context information. The NEs have been trained using both, Skip-Gram and Glove Model. Experiments have been carried out on the KALAKA-3 database and we have used Cavg as a metric to compare the systems. We propose as baseline the Cavg obtained using the NEs as features in the LID task, 24,69%. Our strategy to incorporate information from the neighboring phone-grams to define the final sequences contributes obtaining up to 24,3% relative improvement over the baseline using Skip-Gram model and up to 32,4% using Glove model. Finally, fusing our best system with an MFCC-based acoustic i-Vectors system provides up to 34,1% improvement.

End-to-End Speech Translation with the Transformer

Laura Cross Vila, Carlos Escolano, José A. R. Fonollosa, Marta R. Costa-Jussà
Universitat Politècnica de Catalunya, Spain

Speech Translation has been traditionally addressed with the concatenation of two tasks: Speech Recognition and Machine Translation. This approach has the main drawback that errors are concatenated. Recently, neural approaches to Speech Recognition and Machine Translation have made possible facing the task by means of an End-to-End Speech Translation architecture. In this paper, we propose to use the architecture of the Transformer which is based solely on attention-based mechanisms to address the End-to-End Speech Translation system. As a contrastive architecture, we use the same Transformer to built the Speech Recognition and Machine Translation systems to perform Speech Translation through concatenation of systems. Results on the Spanish-to-English IWSLT benchmark task show that the end-to-end architecture is able to outperform the concatenated systems by half point BLEU.

Audio event detection on Google's Audio Set database: Preliminary results using different types of DNNs

Javier Darna-Sequeiros, Doroteo T. Toledano
AUDIAS - Universidad Autónoma de Madrid, Spain

This paper focuses on the audio event detection problem, in particular on Google Audio Set, a database published in 2017 whose size and breadth are unprecedented for this problem. In order to explore the possibilities of this dataset, several classifiers based on different types of deep neural networks were designed, implemented and evaluated to check the impact of factors such as the architecture of the network, the number of layers and the codification of the data in the performance of the models. From all the classifiers tested, the LSTM neural network showed the best results with a mean average precision of 0.26652 and a mean recall of 0.30698. This result is particularly relevant since we use the embeddings provided by Google as input to the DNNs, which are sequences of at most 10 feature vectors and therefore limit the sequence modelling capabilities of LSTMs.

Emotion Detection from Speech and Text

Mikel de Velasco, Raquel Justo, Josu Antón, Mikel Carrilero, M. Inés Torres
Universidad del Pais Vasco UPV/EHU, Spain

The main goal of this work is to carry out automatic emotion detection from speech by using both acoustic and textual information. For doing that a set of audios were extracted from a TV show were different guests discuss about topics of current interest. The selected audios were transcribed and annotated in terms of emotional status using a crowdsourcing platform. A 3 dimensional model was used to define an specific emotional status in order to pick up the nuances in what the speaker is expressing instead of being restricted to a predefined set of discrete categories. Different sets of acoustic parameters were considered to obtain the input vectors for a neural network. To represent each sequence of words, a models based on word embeddings was used. Different deep learning architectures were tested providing promising results, although having a corpus of a limited size.

Experimental Framework Design for Sign Language Automatic Recognition

Darío Tilves Santiago [1], Ian Benderitter [2], Carmen García-Mateo [1]
[1]University of Vigo, Spain, [2]Polytech Nantes, France

Automatic sign language recognition (ASLR) is quite a complex task, not only for the intrinsic difficulty of automatic video information retrieval, but also because almost every sign language (SL) can be considered as an under-resourced language when it comes to language technology. Spanish sign language (SSL) is one of those under-resourced languages. Developing technology for SSL implies a number of technical challenges that must be tackled down in a structured and sequential manner. In this paper, the problem of how to design an experimental framework for machine-learning-based ASLR is addressed. In our review of existing datasets, our main conclusion is that there is a need for high-quality data. We therefore propose some guidelines on how to conduct the acquisition and annotation of an SSL dataset. These guidelines were developed after conducting some preliminary ASLR experiments with small and limited subsets of existing datasets.

Baseline Acoustic Models for Brazilian Portuguese Using Kaldi Tools

Cassio Batista [1], Ana Larissa Dias [2], Nelson Sampaio Neto [2]
[1]Federal University of Pará (UFPA), Brazil, [2]Federal University of Pará, Brazil

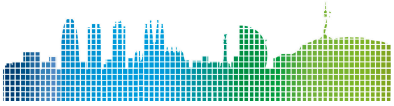
Kaldi has become a very popular toolkit for automatic speech recognition, showing considerable improvements through the combination of hidden Markov models (HMM) and deep neural networks (DNN). However, in spite of its great performance for some languages (e.g. English, Italian, Serbian, etc.), the resources for Brazilian Portuguese (BP) are still quite limited. This work describes what appears to be the first attempt to create Kaldi-based scripts and baseline acoustic models for BP using Kaldi tools. Experiments were carried out for dictation tasks and a comparison to CMU Sphinx toolkit in terms of word error rate (WER) was performed. Results seem promising, since Kaldi achieved the absolute lowest WER of 4.75% with HMM-DNN and outperformed CMU Sphinx even when using Gaussian mixture models only.

Converted Mel-Cepstral Coefficients for Gender Variability Reduction in Query-by-Example Spoken Document Retrieval

Paula López Otero [1], Laura Docío-Fernández [2]

[1]Department of Computer Science, Universidade da Coruña - CITIC, Spain, [2]Department of Signal Theory and Communications, Universidade de Vigo - atlanTTic, Spain

Query-by-example spoken document retrieval (QbESDR) is a task that consists in retrieving those documents where a given spoken query appears. Spoken documents and queries exhibit a huge variability in terms of speaker, gender, accent or recording channel, among others. According to previous work, reducing this variability when following zero-resource QbESDR approaches, where acoustic features are used to represent the documents and queries, leads to improved performance. This work aims at reducing gender variability using voice conversion (VC) techniques. Specifically, a target gender is selected, and those documents and queries spoken by speakers of the opposite gender are converted in order to make them sound like the target gender. VC includes a resynthesis stage that can cause distortions in the resulting speech so, in order to avoid this, the use of the converted Mel-cepstral coefficients obtained from the VC system is proposed for QbESDR instead of extracting acoustic features from the converted utterances. Experiments were run on a QbESDR dataset in Basque language, and the results showed that the proposed gender variability reduction technique led to a relative improvement by 17% with respect to using the original recordings.



A Recurrent Neural Network Approach to Audio Segmentation for Broadcast Domain Data

Pablo Gimeno, Ignacio Viñals, Alfonso Ortega, Antonio Miguel, Eduardo Lleida
University of Zaragoza, Spain

This paper presents a new approach for automatic audio segmentation based on Recurrent Neural Networks. Our system takes advantage of the capability of Bidirectional Long Short Term Memory Networks (BLSTM) for modeling temporal dynamics of the input signals. The DNN is complemented by a resegmentation module, gaining long-term stability by means of the tied-state concept in Hidden Markov Models. Furthermore, feature exploration has been performed to best represent the information in the input data. The acoustic features that have been included are spectral log-filter-bank energies and musical features such as chroma. This new approach has been evaluated with the Albayzín 2010 audio segmentation evaluation dataset. The evaluation requires to differentiate five audio conditions: music, speech, speech with music, speech with noise and others. Competitive results were obtained, achieving a relative improvement of 15.75% compared to the best results found in the literature for this database.

Improving Transcription of Manuscripts with Multimodality and Interaction

Emilio Granell [1], Carlos David Martinez Hinarejos [2], Verónica Romero [3]

[1]Pattern Recognition and Human Language Technology Research Center - Universitat Politècnica de València, Spain, [2]Pattern Recognition and Human Language Technologies Research Center - Universitat Politècnica de València, Spain, [3]Universitat Politècnica de València, Spain

State-of-the-art Natural Language Recognition systems allow transcribers to speed-up the transcription of audio, video or image documents. These systems provide transcribers an initial draft transcription that can be corrected with less effort than transcribing the documents from scratch. However, even the drafts offered by the most advanced systems based on Deep Learning contain errors. Therefore, the supervision of those drafts by a human transcriber is still necessary to obtain the correct transcription. This supervision can be eased by using interactive and assistive transcription systems, where the transcriber and the automatic system cooperate in the amending process. Moreover, the interactive system can combine different sources of information in order to improve their performance, such as text line images and the dictation of their textual contents. In this paper, the performance of a multimodal interactive and assistive transcription system is evaluated on one Spanish historical manuscript. Although the quality of the draft transcriptions provided by a Handwriting Text Recognition system based on Deep Learning is pretty good, the proposed interactive and assistive approach reveals an additional reduction of transcription effort. Besides, this effort reduction is increased when using speech dictations over an Automatic Speech Recognition system, allowing for a faster transcription process.

Improving Pronunciation of Spanish as a Foreign Language for L1 Japanese Speakers with Japañol CAPT Tool

Cristian Tejedor-García [1], Valentín Cardeñoso-Payo [2], María J. Machuca [3], David Escudero-Mancebo [1], Antonio Ríos [3], Takuya Kimura [4]

[1]University of Valladolid, Spain, [2]Universidad de Valladolid, Spain, [3]Autonomous University of Barcelona, Spain, [4]Seisen University, Japan

Availability and usability of mobile smart devices and speech technologies ease the development of language learning applications, although many of them do not include pronunciation practice and improvement. A key to success is to choose the correct methodology and provide a sound experimental validation assessment of their pedagogical effectiveness. In this work we present an empirical evaluation of Japañol, an application designed to improve pronunciation of Spanish as a foreign language targeted to Japanese people. A structured sequence of lessons and a quality assessment of pronunciations before and after completion of the activities provide experimental data about learning dynamics and level of improvement. Explanations have been included as corrective feedback, comprising textual and audiovisual material to explain and illustrate the correct articulation of the sounds. Pre-test and post-test utterances were evaluated and scored by native experts and the ASR, showing a correlation over 0.86 between both predictions. Sounds [s], [fl], [ɾ] and [s], [fɾ], [θ] explain the most frequent failures for discrimination and production, respectively, which can be exploited to plan future versions of the tool, including gamified ones. Final automatic scores provided by the application highly correlate ($r>0.91$) to expert evaluation and a significant pronunciation improvement can be measured.

Exploring E2E speech recognition systems for new languages

Conrad Bernath [1], Aitor Alvarez [1], Haritz Arzelus [1], Carlos David Martínez [2]

[1]Vicomtech, Spain, [2]Pattern Recognition and Human Language Technology Research Center, Universitat Politècnica de València, Spain

Over the last few years, advances in both machine learning algorithms and computer hardware have led to significant improvements in speech recognition technology, mainly through the use of Deep Learning paradigms. As it was amply demonstrated in different studies, Deep Neural Networks (DNNs) have already outperformed traditional Gaussian Mixture Models (GMMs) at acoustic modeling in combination with Hidden Markov Models (HMMs). More recently, new attempts have focused on building end-to-end (E2E) speech recognition architectures, especially in languages with many resources like English and Chinese, with the aim of overcoming the performance of DNN-HMM and more conventional systems. The aim of this work is first to present the different techniques that have been applied to enhance state-of-the-art E2E systems for American English using publicly available datasets. Secondly, we describe the construction of E2E systems for Spanish and Basque, and explain the strategies applied to overcome the problem of the limited availability of training data, especially for Basque as a low-resource language. At the evaluation phase, the three E2E systems are also compared with DNN-HMM based recognition engines built and tested with the same datasets.

Listening to Laryngectomees: A study of Intelligibility and Self-reported Listening Effort of Spanish Oesophageal Speech

Sneha Raman, Inma Hernaez, Eva Navas, Luis Serrano
Aholab (University of the Basque Country, UPV/EHU), Spain

Oesophageal speakers face a multitude of challenges, such as difficulty in basic everyday communication and inability to interact with digital voice assistants. We aim to quantify the difficulty involved in understanding oesophageal speech (in human-human and human-machine interactions) by measuring intelligibility and listening effort. We conducted a web-based listening test to collect these metrics. Participants were asked to transcribe and then rate the sentences for listening effort on a 5-point Likert scale. Intelligibility, calculated as Word Error Rate (WER), showed significant correlation with user rated effort. Speaker type (healthy or oesophageal) had a major effect on intelligibility and effort. Listeners familiar with oesophageal speech did not have any advantage over non familiar listeners in correctly understanding oesophageal speech. However, they reported lesser effort in listening to oesophageal speech compared to non familiar listeners. Additionally, we calculated speaker-wise mean WERs and they were significantly lower when compared to an automatic speech recognition system.

Towards an automatic evaluation of the prosody of people with Down syndrome

Mario Corrales-Astorgano [1], Pastora Martínez-Castilla [2], David Escudero-Mancebo [1], Lourdes Aguilar [3], César González-Ferreras [1], Valentín Cardeñoso-Payo [1]

[1]Universidad de Valladolid, Spain, [2]Universidad Nacional de Educación a Distancia, Spain, [3]Universitat Autònoma de Barcelona, Spain

Prosodic skills may be powerful to improve the communication of individuals with intellectual and developmental disabilities. Yet, the development of technological resources that consider these skills has received little attention. One reason that explains this gap is the difficulty of including an automatic assessment of prosody that considers the high number of variables and heterogeneity of such individuals. In this work, we propose an approach to predict prosodic quality that will serve as a baseline for future work. A therapist and an expert in prosody judged the prosodic appropriateness of individuals with Down syndrome' speech samples collected with a video game. The judgments of the expert were used to train an automatic classifier that predicts the quality by using acoustic information extracted from the corpus. The best results were obtained with an SVM classifier, with a classification rate of 79.30%. The difficulty of the task is evidenced by the high inter-human rater disagreement, justified by the speakers' heterogeneity and the evaluation conditions. Although only 10% of the oral productions judged as correct by the referees were classified as incorrect by the automatic classifier, a specific analysis with bigger corpora and reference recordings of people with typical development is necessary.

Whispered-to-voiced Alaryngeal Speech Conversion with Generative Adversarial Networks

Santiago Pascual de la Puente [1], Antonio Bonafonte [1], Joan Serrà [2], José Andrés González López [3]

[1]Universitat Politècnica de Catalunya, Spain, [2]Telefónica Research, Spain, [3]Universidad de Málaga, Spain

Most methods of voice restoration for patients suffering from aphonia either produce whispered or monotone speech. Apart from intelligibility, this type of speech lacks expressiveness and naturalness due to the absence of pitch (whispered speech) or artificial generation of it (monotone speech). Existing techniques to restore prosodic information typically combine a vocoder, which parameterises the speech signal, with machine learning techniques that predict prosodic information. In contrast, this paper describes an end-to-end neural approach for estimating a fully-voiced speech waveform from whispered alaryngeal speech. By adapting our previous work in speech enhancement with generative adversarial networks, we develop a speaker-dependent model to perform whispered-to-voiced speech conversion. Preliminary qualitative results show effectiveness in re-generating voiced speech, with the creation of realistic pitch contours.

LSTM based voice conversion for laryngectomees

Luis Serrano [1], David Tavarez [2], Xabier Sarasola [1], Sneha Raman [1], Ibon Saratxaga [1], Eva Navas [1], Inma Hernaez [1]

[1]UPV/EHU, Spain, [2]Aholab (UPV/EHU), Spain

This paper describes a voice conversion system designed with the aim of improving the intelligibility and pleasantness of oesophageal voices. Two different systems have been built, one to transform the spectral magnitude and another one for the fundamental frequency, both based on DNNs. Ahocoder has been used to extract the spectral information (mel cepstral coefficients) and a specific pitch extractor has been developed to calculate the fundamental frequency of the oesophageal voices. The cepstral coefficients are converted by means of a LSTM network. The conversion of the intonation curve is implemented through two different LSTM networks, one dedicated to the voiced unvoiced detection and another one for the prediction of F0 from the converted cepstral coefficients. The experiments described here involve conversion from one oesophageal speaker to a specific healthy voice. The intelligibility of the signals has been measured with a Kaldi based ASR system. A preference test has been implemented to evaluate the subjective preference of the obtained converted voices comparing them with the original oesophageal voice. The results show that spectral conversion improves ASR while restoring the intonation is preferred by human listeners.

Sign Language Gesture Classification using Neural Networks

Zuzanna Parcheta [1], Carlos David Martinez Hinarejos [2]
[1]Sciling S.L., Spain, [2]Instituto Tecnológico de Informática - Universitat Politècnica de València, Spain

Recent studies have demonstrated the power of neural networks for different fields of artificial intelligence. In most of fields, as machine translation or speech recognition, neural networks outperform previously used methods (Hidden Markov Models, Statistical Machine Translation, etc.). In this paper we show efficiency of LeNet convolution neural network for sign language recognition. We evaluate different approaches on the Spanish Sign Language dataset where we outperform state-of-the-art results where Hidden Markow Models were applied. As preprocessing step we apply several techniques to get the same size of input matrix containing gesture information.

Influence of tense, modal and lax phonation on the three-dimensional finite element synthesis of vowel [A]

Marc Freixes, Marc Arnela, Joan Claudi Socoró, Francesc Alías Pujol, Oriol Guasch
La Salle - Universitat Ramon Llull, Spain

One-dimensional articulatory speech models have long been used to generate synthetic voice. These models assume plane wave propagation within the vocal tract, which holds for frequencies up to 5kHz. However, higher order modes also propagate beyond this limit, which may be relevant to produce a more natural voice. Such modes could be especially important for phonation types with significant high frequency energy (HFE) content. In this work, we study the influence of tense, modal and lax phonation on the synthesis of vowel [A] through 3D finite element modelling (FEM). The three phonation types are reproduced with an LF (Liljencrants-Fant) model controlled by the Rd glottal shape parameter. The onset of the higher order modes essentially depends on the vocal tract geometry. Two of them are considered, a realistic vocal tract obtained from MRI and a simplified straight duct with varying circular cross-sections. Long-term average spectra are computed from the FEM synthesised [A] vowels, extracting the overall sound pressure level and the HFE level in the 8 kHz octave band. Results indicate that higher order modes may be perceptually relevant for the tense and modal voice qualities, but not for the lax phonation.

Exploring Advances in Real-time MRI for Speech Production Studies of European Portuguese

Conceicao Cunha [1], Samuel Silva [2], António Teixeira [3], Catarina Oliveira [3], Paula Martins [3], Arun Joseph [4], Jens Frahm [4]

[1]IPS Munich, Germany, [2]DETI / IEETA - Universidade de Aveiro, Portugal, [3]University of Aveiro, Portugal, [4]Max Plank Institute for Biophysical Chemistry, Germany

The recent advances in real-time magnetic resonance imaging (RT-MRI) for speech studies, providing a considerable increase in time resolution, potentially improve our ability to study the static and dynamic aspects of speech production. To take advantage of the sheer amount of the resulting data, automated methods need to be used to select, process, and analyze the data, and our work has previously tackled these challenges for an European Portuguese corpus acquired at 14 frames per second (fps). Aiming to further explore RT-MRI in the study of the dynamic characteristics of EP sounds, e.g., nasals, we present a novel 50 fps RT-MRI corpus and assess the applicability, in this new context, of our previous proposals for processing and analysis of these data to extract relevant articulatory information. Importantly, at this stage, we were interested in assessing if and to what extent the new data and the proposed methods provide replicability of the articulatory analysis obtained from the previous corpus. Overall, it was possible to automatically process and analyze the 50 fps data and a comparison of automated analysis performed for the same sounds, for both corpora (i.e., 14fps and 50fps), yields similar results, demonstrating the envisaged replicability.

A postfiltering approach for dual-microphone smartphones

Juan M. Martín-Doñas [1], Iván López-Espejo [2], Angel M. Gomez [1], Antonio M. Peinado [1]
[1]University of Granada, Spain, [2]VeriDas | das-Nano, Spain

Although beamforming is a powerful tool for microphone array speech enhancement, its performance with small arrays, such as the case of a dual-microphone smartphone, is quite limited. The goal of this paper is to study different postfiltering approaches that allow for further noise reduction. These postfilters are applied to our previously proposed extended Kalman filter framework for relative transfer function estimation in the context of minimum variance distortionless response beamforming. We study two different postfilters based on Wiener filtering and non-linear estimation of the speech amplitude. We also propose several estimators of the clean speech power spectral density which exploit the speaker position with respect to the device. The proposals are evaluated when applying speech enhancement on a dual-microphone smartphone in different noisy acoustic environments, in terms of both perceptual quality and speech intelligibility. Experimental results show that our proposals achieve further noise reduction in comparison with other related approaches from the literature.

Speech and monophonic singing segmentation using pitch parameters

Xabier Sarasola [1], Eva Navas [2], David Tavarez [2], Luis Serrano [2], Ibon Saratxaga [2]
[1]UPV/EHU, Spain, [2]University of the Basque Country, Spain

In this paper we present a novel method for automatic segmentation of speech and monophonic singing voice based only on two parameters derived from pitch: proportion of voiced segments and percentage of pitch labelled as a musical note. First, voice is located in audio files using a GMM-HMM based VAD and pitch is calculated. Using the pitch curve, automatic musical note labelling is made applying stable value sequence search. Then pitch features extracted from each voice island are classified with Support Vector Machines. Our corpus consists in recordings of live sung poetry sessions where audio files contain both singing and speech voices. The proposed system has been compared with other speech/singing discrimination systems with good results.

Self-Attention Linguistic-Acoustic Decoder

Santiago Pascual de la Puente [1], Antonio Bonafonte [1], Joan Serrà [2]
[1]Universitat Politècnica de Catalunya, Spain, [2]Telefónica Research, Spain

The conversion from text to speech relies on the accurate mapping from linguistic to acoustic symbol sequences, for which current practice employs recurrent statistical models like recurrent neural networks. Despite the good performance of such models (in terms of low distortion in the generated speech), their recursive structure tends to make them slow to train and to sample from. In this work, we try to overcome the limitations of recursive structure by using a module based on the transformer decoder network, designed without recurrent connections but emulating them with attention and positioning codes. Our results show that the proposed decoder network is competitive in terms of distortion when compared to a recurrent baseline, whilst being significantly faster in terms of CPU inference time. On average, it increases Mel cepstral distortion between 0.1 and 0.3 dB, but it is over an order of magnitude faster on average. Fast inference is important for the deployment of speech synthesis systems on devices with restricted resources, like mobile phones or embedded systems, where speaking virtual assistants are gaining importance.

Synthesizing variation in prosody for Text-to-Speech

Rob Clark
Google, United Kingdom

This talk addresses the issue of producing appropriate and engaging text-to-speech. The quality of speech produced by modern text-to-speech systems is sufficiently intelligible and naturally sounding that we are now seeing it widely used in an increasing number of real world applications. While the speech generated can sound very natural, we are still a long way from ensuring it always sounds appropriate and engaging in the context of a particular discourse or dialogue. We present recent work at Google which begins to address this issue by looking at techniques to generate variation in prosody and speaking style using latent representations and discuss the problems and challenges that we face in going further.

Japañol: a mobile application to help improving Spanish pronunciation by Japanese native speakers

Cristian Tejedor-García, Valentín Cardeñoso-Payo, David Escudero-Mancebo
Departamento de Informática. Universidad de Valladolid, Spain

In this document, we describe the mobile application Japañol, a learning tool which helps pronunciation training of Spanish as a foreign language (L2) at a segmental level. The tool has been specifically designed to be used by native Japanese people, and implies a branch of a previous CAPT gamified tool TipTopTalk!. In this case, a predefined cycle of actions related to exposure, discrimination and production is presented to the user, always under the minimal-pairs approach to pronunciation training. It incorporates freely available ASR and TTS and provides feedback to the user by means of short video tutorials, to reinforce learning progression.

Towards the Application of Global Quality-of-Service Metrics in Biometric Systems

Juan Manuel Espín [1], Roberto Font [1], Juan Francisco Inglés-Romero [1], Cristina Vicente-Chicote [2]

[1]Biometric Vox S.L., Spain, [2]Universidad de Extremadura, Quercus Software Engineering Group, EPCC, Spain, Spain

Performance metrics, such as Equal Error Rate or Detection Cost Function, have been widely used to evaluate and compare biometric systems. However, they seem insufficient when dealing with real-world applications. First, these systems tend to include an increasing number of subsystems, e.g. aimed at spoofing detection or information management. As a result, the aggregation of new capabilities (and their interactions) makes the evaluation of the overall performance more complex. Second, performance metrics only offer a partial view of the system quality in which non-functional properties, such as user experience, efficiency or reliability, are generally ignored. In this paper, we introduce RoQME, an Integrated Technical Project funded by the EU H2020 RobMoSys project. RoQME aims at providing software engineers with methods and tools to deal with system-level non-functional properties, enabling the specification of global Quality-of-Service (QoS) metrics. Although the project is in the context of robotics software, the paper presents potential applications of RoQME to enrich the way in which performance is evaluated in biometric systems, focusing specifically on Automatic Speaker Verification (ASV) systems as a first step.

Special Session
Thursday, 22 November 2018
SP.3
12:20 - 13:30

Incorporation of a Module for Automatic Prediction of Oral Productions Quality in a Learning Video Game

David Escudero-Mancebo [1], Valentín Cardeñoso-Payo [2]
[1]University of Valladolid, Spain, [2]Universidad de Valladolid, Spain

This document presents the research project TIN2017-88858-C2-1-R of the Spanish Government. Antecedents and goals of the project are presented. Current status, recent achievements and collaborations after the first year development are described in the paper.

Silent Speech: Restoring the Power of Speech to People whose Larynx has been Removed

José Andrés González López [1], Phil D. Green [2], Damian Murphy [3], Amelia Gully [3], James M. Gilbert [4]

[1]University of Malaga, Spain, [2]The University of Sheffield, United Kingdom, [3]University of York, United Kingdom, [4]University of Hull, United Kingdom

Every year, some 17,500 people in Europe and North America lose the power of speech after undergoing a laryngectomy, normally as a treatment for throat cancer. Several research groups have recently demonstrated that it is possible to restore speech to these people by using machine learning to learn the transformation from articulator movement to sound. In our project articulator movement is captured by a technique developed by our collaborators at Hull University called Permanent Magnet Articulography (PMA), which senses the changes of magnetic field caused by movements of small magnets attached to the lips and tongue. This solution, however, requires synchronous PMA-and-audio recordings for learning the transformation and, hence, it cannot be applied to people who have already lost their voice. Here we propose to investigate a variant of this technique in which the PMA data are used to drive an articulatory synthesiser, which generates speech acoustics by simulating the airflow through a computational model of the vocal tract. The project goals, participants, current status, and achievements of the project are discussed below.

RESTORE Project: REpair, STOrage and REhabilitation of speech

Inma Hernaez [1], Eva Navas [2], Jose Antonio Municio Martín [3], Javier Gomez Suárez [3]
[1]University of the Basque Country (UPV/EHU), Spain, [2]University of the Basque Country, Spain, [3]Hospital Universitario Cruces - Biocruces, Spain

RESTORE is a project aimed to improve the quality of communication for people with difficulties producing speech, providing them with tools and alternative communication services. At the same time, progress will be made at the research of techniques for restoration and rehabilitation of disordered speech. The ultimate goal of the project is to offer new possibilities in the rehabilitation and reintegration into society of patients with speech pathologies, especially those laryngectomised, by designing new intervention strategies aimed to favour their communication with the environment and ultimately increase their quality of life.

Corpus for Cyberbullying Prevention

Asuncion Moreno, Antonio Bonafonte, Igor Jauk, Laia Tarrés, Victor Pereira
Innovation and Technology Center of the Universitat Politècnica de Catalunya, Spain

Cyberbullying is the use of digital media to harass a person or group of people, through personal attacks, disclosure of confidential or false information, among other means. That is to say, it is considered cyberbullying, or cyber-aggression to everything that is done through electronic communication devices with the intended purpose of harming or attacking a person or a group. In this paper we present a starting project to prevent cyberbullying between kids and teenagers. The idea is to create a prevention system. A system which is installed in the mobile of a kid and, if a harassment is detected, some advice is given to the child. In case of serious or repeated behavior the parents are alerted. The focus of this paper is to describe the characteristics of the database to be used to train the system

Special Session
Thursday, 22 November 2018
SP.7
12:20 - 13:30

EMPATHIC, Expressive, Advanced Virtual Coach to Improve Independent Healthy-Life-Years of the Elderdy

M. Inés Torres [1], Raquel Justo [1], Gérard Chollet [2], César Montenegro [3], Jofre Tenorio-Laranga [4], Olga Gordeeveva [5], Anna Esposito [6], Cornelius Glackin [7], Stephan Schlögl [8], Olivier Deroo [9], Begoña Fernández-Ruanova [10], Riberto Santana [3], Maria S. Kornes [11], Fred Lindner [12], Daria Kyslitska [13], Miriam Reiner [14], Gennaro Cordasco [15], Mari Aksnes [16]

[1]Speech Interactive Group, Universidad del País Vasco UPV/EHU, Spain, [2]Intelligent Voice, United Kingdom, [3]Universidad del Paási Vasco UPV/EHU, Spain, [4]Osatek, Spain, [5]Acapela group, Spain, [6]Universita degli Studi della Campania Luigi Vanvitelli, Italy, [7]Intelligent Voice Ltd, United Kingdom, [8]MCI Management Center Innsbruck, Austria, [9]Acapela Gourp, Belgium, [10]OSATEK, Spain, [11]Oslo Universitetssykehus HF, Norway, [12]Tunstall Nordic, Sweden, [13]eSeniors, France, [14]Technion, Israel, [15]Seconda Università di Napoli (SUN) and IIASS, Italy, [16]Oslo University Hospital, Norway

The EMPATHIC Research & Innovation project researchs, innovates, explores and validates new paradigms and platforms, laying the foundation for future generations of Personalised Virtual Coaches to assist elderly people living independently at and around their home. The project uses remote non-intrusive technologies to extract physiological markers of emotional states in real-time for online adaptive responses of the coach, and advances holistic modelling of behavioural, computational, physical and social aspects of a personalised expressive virtual coach. It develops causal models of coach-user interactional exchanges that engage elders in emotionally believable interactions keeping off loneliness, sustaining health status, enhancing quality of life and simplifying access to future telecare services.

Advances on the Transcription of Historical Manuscripts based on Multimodality, Interactivity and Crowdsourcing

Emilio Granell [1], Carlos David Martinez Hinarejos [2], Verónica Romero [3]

[1]Pattern Recognition and Human Language Technology Research Center - Universitat Politècnica de València, Spain, [2]Pattern Recognition and Human Language Technologies Research Center - Universitat Politècnica de València, Spain, [3]Universitat Politècnica de València, Spain

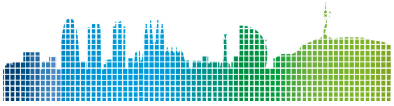
The transcription of digitalised documents is useful to ease the digital access to their contents. Natural language technologies, such as Automatic Speech Recognition (ASR) for speech audio signals and Handwritten Text Recognition (HTR) for text images, have become common tools for assisting transcribers, by providing a draft transcription from the digital document that they may amend. This draft is useful when it presents an error rate low enough to make the amending process more comfortable than a complete transcription from scratch. The work described in this thesis is focused on the improvement of the transcription offered by an HTR system from three scenarios: multimodality, interactivity and crowdsourcing. The image transcription can be obtained by dictating their textual contents to an ASR system. Besides, when both sources of information (image and speech) are available, a multimodal combination is possible, and this can be used to provide assistive systems with additional sources of information. Moreover, speech dictation can be used in a multimodal crowdsourcing platform, where collaborators may provide their speech by using mobile devices. Different solutions for each scenario were tested on two Spanish historical manuscripts, obtaining statistically significant improvements

Special Session
Thursday, 22 November 2018
SP.9
12:20 - 13:30

Bottleneck and Embedding Representation of Speech for DNN-based Language and Speaker Recognition

Alicia Lozano-Diez [1], Joaquin Gonzalez-Rodriguez [1], Javier Gonzalez-Dominguez [2]
[1]Universidad Autonoma de Madrid, Spain, [2]Universidad Autonoma de Madrid / EVO Banco, Spain

In this manuscript, we summarize the findings presented in Alicia Lozano Diez's Ph.D. Thesis, defended on the 22nd of June, 2018 in Universidad Autonoma de Madrid (Spain). In particular, this Ph.D. Thesis explores different approaches to the tasks of language and speaker recognition, focusing on systems where deep neural networks (DNNs) become part of traditional pipelines, replacing some stages or the whole system itself. First, we present a DNN as classifier for the task of language recognition. Second, we analyze the use of DNNs for feature extraction at frame-level, the so-called bottleneck features, for both language and speaker recognition. Finally, utterance-level representation of the speech segments learned by the DNN (known as embedding) is described and presented for the task of language recognition. All these approaches provide alternatives to classical language and speaker recognition systems based on i-vectors (Total Variability modeling) over acoustic features (MFCCs, for instance). Moreover, they usually yield better results in terms of performance.



Deep Learning for i-Vector Speaker and Language Recognition: A Ph.D. Thesis Overview

Omid Ghahabi
EML European Media Laboratory GmbH, Germany

Recent advances in Deep Learning (DL) technology have improved the quality of i-vectors but the DL techniques in use are computationally expensive and need speaker or/and phonetic labels for the background data, which are not easily accessible in practice. On the other hand, the lack of speaker-labeled background data makes a big performance gap, in speaker recognition, between two well-known cosine and PLDA i-vector scoring techniques. This thesis tries to solve the problems above by using the DL technology in different ways, without any need of speaker or phonetic labels. We have proposed an effective DL-based backend for i-vectors which fills 46% of this performance gap, in terms of minDCF, and 79% in combination with a PLDA system with automatically estimated labels. We have also developed an efficient alternative vector representation of speech by keeping the computational cost as low as possible and avoiding phonetic labels. The proposed vectors are referred to as GMM-RBM vectors. Experiments on the core test condition 5 of the NIST SRE 2010 show that comparable results with conventional i-vectors are achieved with a clearly lower computational load in the vector extraction process. Finally, for the LID application, we have proposed a DNN architecture to model effectively the i-vector space of languages in the car environment. It is shown that the proposed DNN architecture outperforms GMM-UBM and i-vector/LDA systems by 37% and 28%, respectively, for short signals 2-3 sec.

Special Session
Thursday, 22 November 2018
SP.11
12:20 - 13:30

Unsupervised Learning for Expressive Speech Synthesis

Igor Jauk
Universitat Politècnica de Catalunya, Spain

This article describes the homonymous PhD thesis realized at the Universitat Politècnica de Catalunya. The main topic and the goal of the thesis was to research unsupervised manners of training expressive voices for tasks such as audiobook reading. The experiments were conducted on acoustic and semantic domains. In the acoustic domain, the goal was to find a feature set which is suitable to represent expressiveness in speech. The basis for such a set were the i-vectors. The proposed feature set outperformed state-of-the-art sets extracted with OpenSmile. Involving the semantic domain, the goal was first to predict acoustic features from semantic embeddings of text for expressive speech and to use the predict vectors as acoustic cluster centroids to adapt voices. The result was a system which automatically reads paragraphs with expressive voice and a second system which can be considered as an expressive search engine and leveraged to train voices with specific expressions. The third experiment evolved to neural network based speech synthesis and the usage of sentiment embeddings. The embeddings were used as an additional input to the synthesis system. The system was evaluated in a preference test showing the success of the approach.

ODESSA/PLUMCOT at Albayzin Multimodal Diarization Challenge 2018

Benjamin Maurice [1], Hervé Bredin [1], Ruiqing Yin [1], Jose Patino [2], Héctor Delgado [2], Claude Barras [1], Nicholas Evans [2], Camille Guinaudeau [1]

[1]LIMSI CNRS, France, [2]EURECOM, France

This paper describes ODESSA and PLUMCOT submissions to Albayzin Multimodal Diarization Challenge 2018. Given a list of people to recognize (alongside image and short video samples of those people), the task consists in jointly answering the two questions "who speaks when?" and "who appears when?". Both consortia submitted 3 runs (1 primary and 2 contrastive) based on the same underlying mono-modal neural technologies : neural speaker segmentation, neural speaker embeddings, neural face embeddings, and neural talking-face detection. Our submissions aim at showing that face clustering and recognition can (hopefully) help to improve speaker diarization.

UPC Multimodal Speaker Diarization System for the 2018 Albayzin Challenge

Miquel Angel India Massana, Itziar Sagastiberri, Ponç Palau, Elisa Sayrol, Josep Ramon Morros, Javier Hernando

Universitat Politècnica de Catalunya, Spain

This paper presents the UPC system proposed for the Multimodal Speaker Diarization task of the 2018 Albayzin Challenge. This approach works by processing individually the speech and the image signal. In the speech domain, speaker diarization is performed using identity embeddings created by a triplet loss DNN that uses i-vectors as input. The triplet DNN is trained with an additional regularization loss that minimizes the variance of both positive and negative distances. A sliding windows is then used to compare speech segments with enrollment speaker targets using cosine distance between the embeddings. To detect identities from the face modality, a face detector followed by a face tracker has been used on the videos. For each cropped face a feature vector is obtained using a Deep Neural Network based on the ResNet 34 architecture, trained using a metric learning triplet loss (available from dlib library). For each track the face feature vector is obtained by averaging the features obtained for each one of the frames of that track. Then, this feature vector is compared with the features extracted from the images of the enrollment identities. The proposed system is evaluated on the RTVE2018 database.

The GTM-UVIGO System for Audiovisual Diarization

Eduardo Ramos-Muguerza, Laura Docío-Fernández, José Luis Alba-Castro
AtlanTTic Research Center, University of Vigo, Spain

This paper explains in detail the Audiovisual system deployed by the Multimedia Technologies Group (GTM) of the atlanTTic research center at the University of Vigo, for the Albayzin Multimodal Diarization Challenge (MDC) organized in the Iberspeech 2018 conference. This system is characterized by the use of state of the art face and speaker verification embeddings trained with publicly available Deep Neural Networks. Video and audio tracks are processed separately to obtain a matrix of confidence values of each time segment that are finally fused to make joint decisions on the speaker diarization result.

The SRI International STAR-LAB System Description for IberSPEECH-RTVE 2018 Speaker Diarization Challenge

Diego Castan [1], Mitchell McLaren [2], Mahesh Kumar Nandwana [1]
[1]SRI International, United States, [2]SRI International, Australia

This document describes the submissions of STAR-LAB (the Speech Technology and Research Laboratory at SRI Interna- tional) to the open-set condition of the IberSPEECH-RTVE 2018 Speaker Diarization Challenge. The core components of the submissions included noise-robust speech activity detec- tion, speaker embeddings for initializing diarization with domain adaptation, and variational Bayes (VB) diarization using a DNN bottleneck i-vector subspaces.

ODESSA at Albayzin Speaker Diarization Challenge 2018

Jose Patino [1], Héctor Delgado [1], Ruiqing Yin [2], Hervé Bredin [2], Claude Barras [2], Nicholas Evans [1]

[1]EURECOM, France, [2]LIMSI, France

This paper describes the ODESSA submissions to the Albayzin Speaker Diarization Challenge 2018. The challenge addresses the diarization of TV shows. This work explores three different techniques to represent speech segments, namely binary key, x-vector and triplet-loss based embeddings. While training-free methods such as the binary key technique can be applied easily to a scenario where training data is limited, the training of robust neural-embedding extractors is considerably more challenging. However, when training data is plentiful (open-set condition), neural embeddings provide more robust segmentations, giving speaker representations which lead to better diarization performance. The paper also reports our efforts to improve speaker diarization performance through system combination. For systems with a common temporal resolution, fusion is performed at segment level during clustering. When the systems under fusion produce segmentations with an arbitrary resolution, they are combined at solution level. Both approaches to fusion are shown to improve diarization performance.

EML Submission to Albayzin 2018 Speaker Diarization Challenge

Omid Ghahabi, Volker Fischer
EML European Media Laboratory GmbH, Germany

Speaker diarization, who is speaking when, is one of the most challenging tasks in speaker recognition, as usually no prior information is available about the identity and the number of the speakers in an audio recording. The task will be more challenging when there is some noise or music on the background and the speakers are changed more frequently. This usually happens in broadcast news conversations. In this paper, we use the EML speaker diarization system as a participation to the recent Albayzin Evaluation challenge. The EML system uses a real-time robust algorithm to make decision about the identity of the speakers approximately every 2 sec. The experimental results on about 16 hours of the developing data provided in the challenge show a reasonable accuracy of the system with a very low computational cost.

In-domain Adaptation Solutions for the RTVE 2018 Diarization Challenge

Ignacio Viñals, Pablo Gimeno, Alfonso Ortega, Antonio Miguel, Eduardo Lleida
Universidad de Zaragoza, Spain

This paper tries to deal with domain mismatch scenarios in the diarization task. This research has been carried out in the context of the Radio Televisión Espa˜nola (RTVE) 2018 Challenge at IberSpeech 2018. This evaluation seeks the improvement of the diarization task in broadcast corpora, known to contain multiple unknown speakers. These speakers are set to contribute in different scenarios, genres, media and languages. The evaluation offers two different conditions: A closed one with restrictions in the resources, both acoustic and further knowledge, to train and develop diarization systems, and an open condition without restrictions to check the latest improvements in the state-of-the-art. Our proposal is centered on the closed condition, specially dealing with two important mismatches: media and language. ViVoLab system for the challenge is based on the i-vector PLDA framework: I-vectors are extracted from the input audio according to a given segmentation, supposing that each segment represents one speaker intervention. The diarization hypotheses are obtained by clustering the estimated i-vectors with a Fully Bayesian PLDA, a generative model with latent variables as speaker labels. The number of speakers is decided by comparing multiple hypotheses according to the Evidence Lower Bound (ELBO) provided by the PLDA.

DNN-based Embeddings for Speaker Diarization in the AuDIaS-UAM System for the Albayzin 2018 IberSPEECH-RTVE Evaluation

Alicia Lozano-Diez, Beltran Labrador, Diego de Benito, Pablo Ramirez, Doroteo T. Toledano
Audias-UAM, Universidad Autonoma de Madrid, Spain

This document describes the three systems submitted by the AuDIaS-UAM team for the Albayzin 2018 IberSPEECH-RTVE speaker diarization evaluation. Two of our systems (primary and contrastive 1 submissions) are based on embeddings which are a fixed length representation of a given audio segment obtained from a deep neural network (DNN) trained for speaker classification. The third system (contrastive 2) uses the classical i-vector as representation of the audio segments. The resulting embeddings or i-vectors are then grouped using Agglomerative Hierarchical Clustering (AHC) in order to obtain the diarization labels. The new DNN-embedding approach for speaker diarization has obtained a remarkable performance over the Albayzin development dataset, similar to the performance achieved with the well-known i-vector approach.

CENATAV Voice-Group Systems for Albayzin 2018 Speaker Diarization Evaluation Campaign

Edward L. Campbell, Gabriel Hernandez, José R. Calvo de Lara
CENATAV, Cuba

Usually, the environment to record a voice signal is not ideal and, in order to improve the representation of the speaker characteristic space, it is necessary to use a robust algorithm, thus making the representation more stable in the presence of noise. A Diarization system that focuses on the use of robust feature extraction techniques is proposed in this paper. The presented features (such as Mean Hilbert Envelope Coefficients, Medium Duration Modulation Coefficients and Power Normalization Cepstral Coefficients) were not used in other Albayzin Challenges. These robust techniques have a common characteristic, which is the use of a Gammatone filter-bank for dividing the voice signal in sub-bands as an alternative option to the classical Triangular filter-bank used in Mel Frequency Cepstral Coefficients. The experiment results show a more stable Diarization Error Rate in robust features than in classic features.

The Intelligent Voice System for the IberSPEECH-RTVE 2018 Speaker Diarization Challenge

Abbas Khosravani, Cornelius Glackin, Nazim Dugan, Gérard Chollet, Nigel Cannings
Intelligent Voice Limited, United Kingdom

This paper describes the Intelligent Voice (IV) speaker diarization system for IberSPEECH-RTVE 2018 speaker diarization challenge. We developed a new speaker diarization built on the success of deep neural network based speaker embeddings in speaker verification systems. In contrary to acoustic features such as MFCCs, deep neural network embeddings are much better at discerning speaker identities especially for speech acquired without constraint on recording equipment and environment. We perform spectral clustering on our proposed CNN-LSTM-based speaker embeddings to find homogeneous segments and generate speaker log likelihood for each frame. A HMM is then used to refine the speaker posterior probabilities through limiting the probability of switching between speakers when changing frames. The proposed system is evaluated on the development set (dev2) provided by the challenge.

JHU Diarization System Description

Zili Huang, L. Paola García-Perera, Jesús Villalba, Daniel Povey, Najim Dehak
Johns Hopkins University, United States

We present the JHU system for Iberspeech-RTVE Speaker Diarization Evaluation. This assessment combines Spanish language and broadcast audio in the same recordings, conditions in which our system has not been tested before. To tackle this problem, the pipeline of our general system, developed entirely in Kaldi, includes an acoustic feature extraction, a SAD, an embedding extractor, a PLDA and a clustering stage. This pipeline was used for both, the open and the closed conditions (described in the evaluation plan). All the proposed solutions use wide-band data (16KHz) and MFCCs as their input. For the closed condition, the system trains a DNN SAD using the Albayzin2016 data. Due to the small amount of data available, the i-vector embedding extraction was the only approach explored for this task. The PLDA training utilizes Albayzin data followed by an Agglomerative Hierarchical Clustering (AHC) to obtain the speaker segmentation. The open condition employs the DNN SAD obtained in the closed condition. Four types of embeddings were extracted, xvector-basic, xvector-factored, i-vector-basic and BNF-i-vector. The x-vector-basic is a TDNN trained on augmented Voxceleb1 and Voxceleb2. The x-vector-factored is a factored-TDNN (F-TDNN) trained on SRE12-micphn, MX6-micphn, VoxCeleb and SITWdev-core. The i-vector-basic was trained on Voxceleb1 and Voxceleb2 data (no augmentation). The BNF-i-vector is a BNF-posterior i-vector trained with the same data as xvector-factored. The PLDA training for the new scenario uses the Albayzin2016 data. The four systems were fused at the score level. Once again, the AHC commputed the final speaker segmentation. We tested our systems in the dev2 data and observed that the SAD is of importance to improve the results. Moreover, we noticed that xvectors were better than i-vectors, as already observed in previous experiments.

GTM-IRLab Systems for Albayzin 2018 Search on Speech Evaluation

Paula López Otero [1], Laura Docío-Fernández [2]
[1]Department of Computer Science, Universidade da Coruña, Spain, [2]University of Vigo, Spain

This paper describes the systems developed by the GTM-IRLab team for the Albayzin 2018 Search on Speech evaluation. The system for the spoken term detection task consists in the fusion of two subsystems: a large vocabulary continuous speech recognition strategy that uses the proxy words approach for out-of-vocabulary terms, and a phonetic search system based on the probabilistic retrieval model for information retrieval. The query-by-example spoken term detection system is the result of fusing four subsystems: three of them are based on dynamic time warping search using different representations of the waveforms, namely Gaussian posteriorgrams, phoneme posteriorgrams and a large set of low-level descriptors; and the other one is the phonetic search system used for spoken term detection with some modifications to manage spoken queries.

AUDIAS-CEU: A Language-independent approach for the Query-by-Example Spoken Term Detection task of the Search on Speech ALBAYZIN 2018 evaluation

Maria Cabello [1], Doroteo T. Toledano [1], Javier Tejedor [2]

[1]AUDIAS - Universidad Autonoma de Madrid, Spain, [2]Universidad CEU - San Pablo, Spain

Query-by-Example Spoken Term Detection is the task of detecting query occurrences within speech data (henceforth utterances). Our submission is based on a language-independent template matching approach. First, queries and utterances are represented as phonetic posteriorgrams computed for English language with the phoneme decoder developed by the Brno University of Technology. Next, the Subsequence Dynamic Time Warping algorithm with a modified Pearson correlation coefficient as cost measure is employed to hipothesize detections. Results on development data showed an ATWV=0.1774 with MAVIR data and an ATWV=0.0365 with RTVE data.

GTTS-EHU Systems for the Albayzin 2018 Search on Speech Evaluation

Luis J. Rodríguez-Fuentes, Mikel Peñagarikano, Amparo Varona, Germán Bordel
Departamento de Electricidad y Electrónica UPV/EHU, Spain

This paper describes the systems developed by GTTS-EHU for the QbE-STD and STD tasks of the Albayzin 2018 Search on Speech Evaluation. Stacked bottleneck features (sBNF) are used as frame-level acoustic representation for both audio documents and spoken queries. In QbE-STD, a flavour of segmental DTW (originally developed for MediaEval 2013) is used to perform the search, which iteratively finds the match that minimizes the average distance between two test-normalized sBNF vectors, until either a maximum number of hits is obtained or the score does not attain a given threshold. The STD task is performed by synthesizing spoken queries (using publicly available TTS APIs), then averaging their sBNF representations and using the average query for QbE-STD. A publicly available toolkit (developed by BUT/Phonexia) has been used to extract three sBNF sets, trained for English monophone and triphone state posteriors (contrastive systems 3 and 4) and for multilingual triphone posteriors (contrastive system 2), respectively. The concatenation of the three sBNF sets has been also tested (contrastive system 1). The primary system consists of a discriminative fusion of the four contrastive systems. Detection scores are normalized on a query-by-query basis (qnorm), calibrated and, if two or more systems re considered, fused with other scores. Calibration and fusion parameters are discriminatively estimated using the ground truth of development data. Finally, due to a lack of robustness in calibration, Yes/No decisions are made by applying the MTWV hresholds obtained for the development sets, except for the COREMAH test set. In this case, calibration is based on the MAVIR corpus, and the 15% highest scores are taken as positive (Yes) detections.

Cenatav Voice Group System for Albayzin 2018 Search on Speech Evaluation

Ana R. Montalvo, Jose M. Ramirez, Alejandro Roble, Jose R. Calvo
Voice Group, Advanced Technologies Application Center, CENATAV, Cuba

This paper presents the system employed in the Albayzin 2018 "Search on Speech" Evaluation by the Voice Group of CENATAV. The system used in the Spoken Term Detection (STD) task consists on an Automatic Speech Recognizer (ASR) and a module to detect the terms. The open source Kaldi toolkit is used to build both modules. ASR acoustic models are based on DNN-HMM, S-GMM or GMM-HMM, trained with audio data provided by the organizers and other obtained from ELDA. The lexicon and trigram language model are obtained from the text associated to the audio. The ASR generates the lattices and the word alignments required to detect the terms. Results with development data shown that DNN-HMM model brings up a behavior better or similar to obtained in previous challenges.

MLLP-UPV and RWTH Aachen Spanish ASR Systems for the IberSpeech-RTVE 2018 Speech-to-Text Transcription Challenge

Javier Jorge [1], Adrià Martínez-Villaronga [1], Pavel Golik [2], Adrià Giménez [1], Joan Albert Silvestre-Cerdà [1], Patrick Doetsch [2], Vicent Andreu Císcar [3], Hermann Ney [2], Alfons Juan [1], Sanchis Albert [1]

[1]Departament de Sistemes Informàtics i Computació, Universitat Politècnica de València, Spain, [2]Human Language Technology and Pattern Recognition, Computer Science Department, RWTH Aachen University, Germany, [3]Escola Tècnica Superior d'Enginyeria Informàtica, Universitat Politècnica de València, Spain

This paper describes the Automatic Speech Recognition systems built by the MLLP research group of Universitat Politècnica de València and the HLTPR research group of RWTH Aachen for the IberSpeech-RTVE 2018 Speech-to-Text Transcription Challenge. We participated in both the closed and the open training conditions. The best system built for the closed condition was an hybrid BLSTM-HMM ASR system using one-pass decoding with a combination of a RNN LM and show-adapted n-gram LMs. It was trained on a set of reliable speech data extracted from the train and dev1 sets using MLLP's transLectures-UPV toolkit (TLK) and TensorFlow. This system achieved 20.0% WER on the dev2 set. For the open condition we used approx. 3800 hours of out-of-domain training data from multiple sources and trained a one-pass hybrid BLSTM-HMM ASR system using open-source tools RASR and RETURNN developed at RWTH Aachen. This system scored 15.6% WER on the dev2 set. The highlights of these systems include robust speech data filtering for acoustic model training and show-specific language modeling.

Limecraft Flow - workflows for story editing, subtitling and archiving

Victor Garcia [1], Nuria Sanchez [1], Angus Knights [2], Maarten Verwaest [3]
[1]VISIONA INGENIERÍA DE PROYECTOS, Spain, [2]SPEECHMATICS, United Kingdom, [3]LIMECRAFT, Belgium

Lots of efforts are being made to progress beyond the state of the art in the area of speech technologies for the purposes of automating indexing and closed captioning. Solutions like Limecraft Flow are currently being used by producers of audio-visual material of various breeds at international level, including TV, corporate and e-learning, to automatically process audio and or video in real-time (e.g. for automating processes like archiving, data-driven journalism, etc.). Properly tuned speech recognition is the corner stone of any possible improvement in automatic indexing and captioning. Limecraft enables its users to automatically transcribe spoken audio in real-time. Limecraft Flow embeds ASR engines from different service providers to address the needs of a specific language or semantic context, being Speechmatics the one that performs best with accuracies well above 95%, indifferent of the speaker, when processing properly recorded European languages. Limecraft Flow also counts on with a patented technology for Natural Language Processing for the purpose of automatically identifying and classifying dialogue fragments. In this paper we will run through the state of the art technologies and highlight some opportunities for innovation. System architecture of this innovative cloud-based application will be presented. The solution consists of a visible and customisable front-end (html5, backbone…) connected with the back-end using node.js. Together, back-end and front-end deliver a user experience that comes close or outperforms the interactivity of native applications, using no more than a modern browser. Finally, system outputs resulting of automatically transcribing different types of TV shows are presented in terms of word error rate (WER) for the Speech-To-Text challenge.

Exploring Open-Source Deep Learning ASR for Speech-to-Text TV program transcription

Juan M. Perero-Codosero [1], Javier Antón-Martín [1], Daniel Tapias Merino [1], Luis A. Hernández-Gómez [2], Eduardo López-Gonzalo [2]

[1]Sigma Technologies S.L., Spain, [2]Universidad Politécnica de Madrid - GAPS, Spain

Deep Neural Networks (DNN) are fundamental part of current ASR. State-of-the-art are "hybrid" models in which acoustic models (AM) are designed using neural networks. However, there is an increasing interest in developing end-to-end Deep Learning solutions where a neural network is trained to predict character/grapheme or sub-word sequences which can be converted directly to words. Though several promising results have been reported for end-to-end ASR systems, it is still not clear if they are capable to unseat hybrid systems. In this contribution, we evaluate open-source state-of-the-art hybrid and end-to-end Deep Learning ASR under the IberSpeech-RTVE Speech to Text Transcription Challenge. The hybrid ASR is based on Kaldi and Wav2Letter will be the end-to-end framework. Experiments were carried out using 6 hours of dev1 and dev2 partitions. The lowest WER on the reference TV show (LM-20171107) was 22.23% for the hybrid system (lowercase format without punctuation). Major limitation for Wav2Letter has been a high training computational demand (between 6 hours and 1 day/epoch, depending on the training set). This forced us to stop the training process to meet the Challenge deadline. But we believe that with more training time it will provide competitive results with the hybrid system.

The Vicomtech-PRHLT Speech Transcription Systems for the IberSPEECH-RTVE 2018 Speech to Text Transcription Challenge

Haritz Arzelus [1], Aitor Alvarez [1], Conrad Bernath [2], Eneritz García [3], Emilio Granell [4], Carlos David Martinez Hinarejos [5]

[1]Vicomtech-IK4, Spain, [2]Vicomtech, Spain, [3]Vicomtec, Spain, [4]Pattern Recognition and Human Language Technology Research Center - Universitat Politècnica de València, Spain, [5]Pattern Recognition and Human Language Technologies Research Center - Universitat Politècnica de València, Spain

This paper describes our joint submission to the IberSPEECH-RTVE Speech to Text Transcription Challenge 2018, which calls automatic speech transcription systems to be evaluated in realistic TV shows. With the aim of building and evaluating systems, RTVE licensed around 569 hours of different TV programs, which were processed, re-aligned and revised in order to discard segments with imperfect transcriptions. This task reduced the corpus to 136 hours that we considered as nearly perfectly aligned audios and that we employed as in domain data to train acoustic models. A total of 6 systems were built and presented to the evaluation challenge, three systems per condition. These recognition engines are different versions, evolution and configurations of two main architectures. The first architecture includes an hybrid LSTM-HMM acoustic model, where bidirectional LSTMs were trained to provide posterior probabilities for the HMM states. The language model corresponds to modified Kneser-Ney smoothed 3-gram and 9-gram models used for decoding and re-scoring of the lattices respectively. The second architecture includes an End-To-End based recognition system, which combines 2D convolutional neural networks as spectral feature extractor from spectrograms with bidirectional Gated Recur- rent Units as RNN acoustic models. A modified Kneser-Ney smoothed 5-gram model was also integrated to re-score the E2E hypothesis. All the systems' outputs were then punctuated using bidirectional RNN models with attention mechanism and capitalized through recasing techniques.

Intelligent Voice ASR system for Iberspeech 2018 Speech to Text Transcription Challenge

Nazim Dugan, Cornelius Glackin, Gérard Chollet, Nigel Cannings
Intelligent Voice Limited, United Kingdom

Provided ground truth transcriptions for training and development are cleaned up using customized clean-up scripts and realigned using a two-step alignment procedure which uses word lattice results coming from a previous ASR system trained for European Spanish. An utterance level selection mechanism is applied on training and development data by calculating word error rate (WER) using the results of previous ASR system. 261 hours of data is selected from train and dev1 subsections of the provided data by applying a selection criterion on the utterance level scoring results. Selected data is merged by 91 hours of training data of previous ASR system and 3-times data augmentation is applied by reverberation using a noise corpus. 1057 hours of final training data is used in the training of a nnet3 chain acoustic model with MFCC's and iVectors as input features using Kaldi framework where GMM iterative phone alignment is used before starting neural network training. Selected text of train and dev1 subsections are also used for new pronunciation additions and language model (LM) adaptation of the LM of the previous ASR System. Generated model is tested using data from dev2 subsection selected with the same procedure as the training data.

The GTM-UVIGO System for Albayzin 2018 Speech-to-Text Evaluation

Laura Docío-Fernández, Carmen García-Mateo
Multimedia Technology Group, AtlanTTic Research Center, University of Vigo, Spain

This paper describes the Speech-to-Text system developed by the Multimedia Technologies Group (GTM) of the atlanTTic research center at the University of Vigo, for the Albayzin Speech-to-Text Challenge (S2T) organized in the Iberspeech 2018 conference. The large vocabulary automatic speech recognition system is built using the Kaldi toolkit. It uses an hybrid Deep Neural Network - Hidden Markov Model (DNN-HMM) for acoustic modeling, and a rescoring of a trigram based word-lattices, obtained in a first decoding stage, with a fourgram language model or a language model based on a recurrent neural network. The system was evaluated only on the open set training condition.

University of the Basque Country (GTTS@EHU) System for IberSPEECH-RTVE 2018 Speech to Text Transcription Challenge

Mikel Penagarikano, Amparo Varona, Luis J. Rodríguez-Fuentes, Germán Bordel
GTTS Group - University of the Basque Country UPV/EHU, Spain

This paper describes the steps to install and use the Google Cloud Platform tools and the Google Cloud Speech-to-Text API to run a state-of-the-art speech-to-text system. The submission of the Software Technology Working Group (http://gtts.ehu.es) of the University of the Basque Country (EHU) for the IberSPEECH-RTVE 2018 Speech to Text Transcription Challenge was just the output of this comercial system. The simplicity of use, economic cost and performance of Google's commercial system makes it suitable to be used as a reference/baseline system by any research group working on Speech Technologies.

Topic coherence analysis for the classification of Alzheimer's disease

Anna Pompili [1], Alberto Abad [1], David Martins de Matos [1], Isabel Pavão Martins [2]
[1]INESC-ID/IST, Portugal, [2]LEL/FMUL, Portugal

Language impairment in Alzheimer's disease is characterized by a decline in the semantic and pragmatic levels of language processing that manifests since the early stages of the disease. While semantic deficits have been widely investigated using linguistic features, pragmatic deficits are still mostly unexplored. In this work, we present an approach to automatically classify Alzheimer's disease using a set of pragmatic features extracted from a discourse production task. Following the clinical practice, we consider an image representing a closed domain as a discourse's elicitation form. Then, we model the elicited speech as a graph that encodes a hierarchy of topics. To do so, the proposed method relies on the integration of various NLP techniques: syntactic parsing for sentence segmentation into clauses, coreference resolution for capturing dependencies among clauses, and word embeddings for identifying semantic relations among topics. According to the experimental results, pragmatic features are able to provide promising results distinguishing individuals with Alzheimer's disease, comparable to solutions based on other types of linguistic features.

Building a global dictionary for semantic technologies

Iklódi Eszter [1], Gábor Recski [1], Gábor Borbély [2], Maria Jose Castro-Bleda [3]

[1]Budapest University of Technology and Economics, Hungary, [2]Department of Algebra, Budapest University of Technology and Economics, Hungary, [3]Departamento de Sistemas Informáticos y Computación, Universitat Politècnica de València, Spain

Computer-driven natural language processing plays an increasingly important role in our everyday life. In the current digital world, using natural language for human-machine communication has become a basic requirement. In order to meet this requirement, it is inevitable to analyze human languages semantically. Nowadays, state-of-the-art systems represent word meaning with high dimensional vectors, known as word embeddings. Within the field of computational semantics a new research direction focuses on finding mappings between embeddings of different languages. This paper proposes a novel method for finding linear mappings among word vectors for various languages. Compared to previous approaches, this method does not learn translation matrices between two specific languages, but between a given language and a shared, universal space. The system was trained in two different modes, first between two languages, and after that applying three languages at the same time. In the first case two different training data were applied; Dinu's English-Italian benchmark data, and English-Italian translation pairs extracted from the PanLex database. In the second case only the PanLex database was used. The system performs on English-Italian languages with the best setting significantly better than the baseline system of Mikolov et al., and it provides a comparable performance with the more sophisticated systems of Faruqui and Dyer and Dinu et al. Exploiting the richness of the PanLex database, the proposed method makes it possible to learn linear mappings among an arbitrary number languages.

TransDic, a public domain tool for the generation of phonetic dictionaries in standard and dialectal Spanish and Catalan

Juan-María Garrido [1], Marta Codina [2], Kimber Fodge [3]

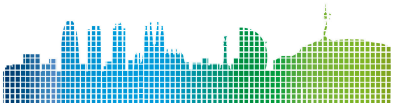
[1]National Distance Education University, Spain, [2]Pompeu Fabra University, Spain, [3]Pompeu Fabra University, United States

This paper presents TransDic, a free distribution tool for the phonetic transcription of word lists in Spanish and Catalan which allows the generation of phonetic transcription variants, a feature that can be useful for some technological applications, such as speech recognition. It allows the transcription in both standard Spanish and Catalan, but also in several dialects of these two languages spoken in Spain. Its general structure, input, output and main functionalities are presented, and the procedure followed to define and implement the transcription rules in the tool is described. Finally, the results of an evaluation carried for both languages are presented, which show that TransDic performs correctly the transcription tasks that it was developed for.

Wide Residual Networks 1D for Automatic Text Punctuation

Jorge Llombart, Antonio Miguel, Alfonso Ortega, Eduardo Lleida
University of Zaragoza, Spain

Documentation and analysis of multimedia resources usually requires a large pipeline with many stages. It is common to obtain texts without punctuation at some point, although later steps might need some accurate punctuation, like the ones related to natural language processing. This paper is focused on the task of recovering pause punctuation from a text without prosodic or acoustic information. We propose the use of Wide Residual Networks to predict which words should have a comma or stop from a text with removed punctuation. Wide Residual Networks are a well-known technique in image processing, but they are not commonly used in other areas as speech or natural language processing. We propose the use of Wide residual networks because they show great stability and the ability to work with long and short contextual dependencies in deep structures. Unlike for image processing, we will use 1-Dimensional convolutions because in text processing we only focus on the temporal dimension. Moreover, this architecture allows us to work with past and future context. This paper compares this architecture with Long-Short Term Memory cells which are used in this task and also combine the two architectures to get better results than each of them separately.



End-to-End Multi-Level Dialog Act Recognition

Eugénio Ribeiro [1], Ricardo Ribeiro [2], David Martins de Matos [1]
[1]Universidade de Lisboa, Portugal, [2]INESC ID Lisboa/ISCTE-IUL, Portugal

The three-level dialog act annotation scheme of the DIHANA corpus poses a multi-level classification problem in which the bottom levels allow multiple or no labels for a single segment. We approach automatic dialog act recognition on the three levels using an end-to-end approach, in order to implicitly capture relations between them. Our deep neural network classifier uses a combination of word- and character-based segment representation approaches, together with a summary of the dialog history and information concerning speaker changes. We show that it is important to specialize the generic segment representation in order to capture the most relevant information for each level. On the other hand, the summary of the dialog history should combine information from the three levels to capture dependencies between them. Furthermore, the labels generated for each level help in the prediction of those of the lower levels. Overall, we achieve results which surpass those of our previous approach using the hierarchical combination of three independent per-level classifiers. Furthermore, the results even surpass the results achieved on the simplified version of the problem approached by previous studies, which neglected the multi-label nature of the bottom levels and only considered the label combinations present in the corpus.

Keynote 3 **KN3**
Friday, 23 November 2018 **11:00 - 12:00**

Automatic Question Answering: Problem Solved?

Lluís Màrquez
Amazon, Spain

Automatic Question Answering (Q&A), i.e., the task of building computer programs that are able to answer question posed in natural language, has a long tradition in the fields of Natural Language Processing and Information Retrieval. In recent years, Q&A applications have had a tremendous impact in industry and they are ubiquitous (e.g., embedded in any of the personal assistants that are in the market, Siri, Alexa, Cortana, Google Assistant, etc.). At the same time, we have witnessed a renewed interest in the scientific community, as Q&A has become one of the paradigmatic tasks for assessing the ability of machines to comprehend text. A plethora of corpora, resources and systems have blossomed and flooded the community in the last three years. These systems can do very impressive things, for instance, finding answers to open ended questions in long text contexts with super-human accuracy, or answering complex questions about images, by mixing the two modalities. As in many other fields, these state-of-the-art systems are implemented using machine learning in the form of neural networks (deep learning). The new AI, of course. But do these Q&A systems really understand what they read? In more simple words, do they provide the right answers for the right reasons? Several recent studies have shown that QA systems are actually very brittle. They generalize badly and they fail miserably when presented with simple adversarial examples. The machine learning algorithms are very good at picking all the biases and artefacts in the corpora, and they learn to find answers based on shallow text properties and pattern matching. But they do not show many understanding or reasoning abilities, after all. Following this serious setback, there is a new push in the community for carefully designing more complex and bias-free datasets, and more robust and explainable systems. Hopefully, this will lead to a new generation of smarter and more useful Q&A engines in the near future. In this talk, I will overview the present and the future of Question Answering by going over all the aforementioned topics.

Panel discussion on Speech technologies: Industry and Academy

Marta R. Costa-Jussà
Universitat Politècnica de Catalunya, Spain

Speech technologies are of increasing interest both at the commercial and scientific level. The recent success of deep learning technologies has provided a boost in these technologies which are highly contributing to outstanding results. Examples of recent advances include achieving human parity for particular tasks in speech recognition, machine translation or natural language understanding. Most of these achievements come from the industry side, which leads to the question of where is the best place to do research nowadays either in industry or academy. Given that deep learning techniques require large amount of computational resources, industry may be more prepared, and if this is the case, how academy could keep the talent. Maybe, restrictions in funding and computational infrastructures can still be compensated with the level of freedom in orienting the research on the own interests. This panel, moderated by Dr Marta R. Costa-jussà researcher at the Universitat Politècnica de Catalunya, will count on experts on both sides: from the academy side, Dr Tanja Schultz, professor at the University of Bremen; and from the industry side, Dr Rob Clark, researcher at Google, and David del Val La Torre, CEO from Telefónica I+D. Panelists will discuss about questions including: how do they envisage speech technologies in 5 years from now?; is deep learning here to stay?; and which is the best place in current days to do research either academy or industry?

Index