# Microsoft Language Development Center's activities in 2008/2009

*Daniela Braga, António Calado, Pedro Silva, Miguel Sales Dias*

Microsoft Language Development Center, Porto Salvo, Portugal
{i-dbraga, i-antonc, i-pedros, Miguel.Dias}@microsoft.com

## 1. Introduction

Microsoft Language Development Center was launched in 2005, integrated in the Portuguese Microsoft (MS) subsidiary. MLDC (http://www.microsoft.com/portugal/mldc) is the first MS R&D center with the mission to bring key language component product development to Europe and neighboring regions. MLDC acts as an expansion branch of the Redmond based product group, responsible for speech R&D in Microsoft and benefits from its experience, technological background and support. The Global Speech Development Group in MS has four locations: Redmond (USA), Mountain View (Silicon Valley, USA), Porto Salvo (Portugal) and Beijing (China). MLDC works closely with the Mountain View, Redmond and Beijing groups in a multicultural and multidisciplinary team dealing with all aspects of ASR, TTS and speech applications for a large number of languages. MLDC staff included 17 researchers (including 3 PhDs), software engineers and computational linguistics in Fiscal Year 2009 and currently includes 11 people in the beginning of Fiscal Year 2010.

## 2. MLDC's action lines and activities

MLDC has based his activity in the following action lines:
1) Performance of R&D in Speech Technology (speech recognition - ASR, speech synthesis - TTS, speech applications), fully integrated in the Microsoft roadmap, applied to a wide range of MS products and platforms (client, server, live, entertainment, automotive, mobility); 2) Establishment of cooperative links with the most innovative universities, institutes, research laboratories and companies in Portugal and Europe, which are active in the speech and HCI areas (Health, Accessibility, Inclusion, Ageing well, Digital Libraries, Robotics), to pursue joint R&D, in natural and multimodal human-computer communication, including speech and natural language; 3) Collection of key multi-language components and resources, such as speech corpora, text corpora and lexica.

MLDC has been collaborating in the entire product R&D life cycle of ASR technology in all major European languages, Brazilian Portuguese and Indian English, from data collection to usability testing and system evaluation: 1) Multilingual telephony and desktop speech data collection in all Western European Languages and American variants and some Eastern Languages for multilingual ASR systems (2006-2007); 2) Several front-end components building for Western European Languages and American variants, such as phonetic lexicons and phone sets (2006-2007), TN (Text Normalization) and ITN (Inverse Text Normalization) (2008-2009), polyphony, morphology modules, POS taggers, word breakers, sentence separators, etc.; 3) Multilingual acoustic models training and grammars building for web search controlled by speech (2008-2009); 4) System evaluation and bug fixing in real scenarios (2009); 5) Usability studies of the ASR system integrated in several Microsoft products (2009);
6) Definition of specifications of features, groups of features, software and systems design.

## 3. Finished projects

In the last 2 years, MLDC was mainly dedicated to develop TTS and ASR technology and components to Microsoft Exchange 2010 in several languages. MLDC developed 4 TTS languages in European Portuguese, Brazilian Portuguese, Catalan and Danish and produced key-language components to 10 ASR European languages (the same 4 languages produced for TTS plus Finnish, Swedish, Italian, Korean, Dutch and Norwegian). MLDC was also dedicated to the Speech International Project (SIP), whose goal was to collect and transcribe large amounts of telephony speech data in several European languages to be used in ASR acoustic model training. This data was also integrated in Microsoft Exchange 2010 product. Smaller finished projects include: Virtual Hélia (Talking head in European Portuguese language), prototype to create personalized TTS voices, User Verification using Compact Signatures of Multiple Face Profiles: Speaker ID + Face ID (a joint project with MS Research India) and Info Service: traffic, news and weather information in MS applications using Speech Technology.

## 4. Ongoing internal projects

We currently have the following ongoing projects:

1) TN for EFIGS (British English, French, Italian, German, Spanish) languages: production of Text Normalization components for TTS and ASR for several Microsoft products;
2) TN and language models for Voice Search for Mobile in 6 languages (British & American English, French, Italian, Spanish, German);
3) Speech Data Collection for Desktop (speech acquisition in 16 kHz and transcription for Russian and Italian).

## 5. Ongoing external projects

MLDC has external collaborative research projects with academia and other industrial partners. The projects on this scope are called "Citizenship" projects. Currently we have one running project, TTS for Galician, a joint project with University of Vigo with the aim of developing a new TTS in Galician using Microsoft technology and University of Vigo's language resources. MLDC has also submitted several European R&D projects and National (FCT and QREN) R&D initiatives with academic and industrial consortia. The most recently approved project in the scope of the 2nd call of the European Ambient Assisted Living Joint Programme was PAE-LIFE (Personal Assistant to Enhance the Social Life of the Seniors), a Personal Life Assistant with the goal of fighting isolation and exclusion of the elderly and allowing them have a more social and fulfilling life through technology controlled by multimodal interfaces. More projects were submitted and are awaiting approval. MLDC belonged to LC-STAR II consortia and is part of the European Center of Excellence in Speech Synthesis network.