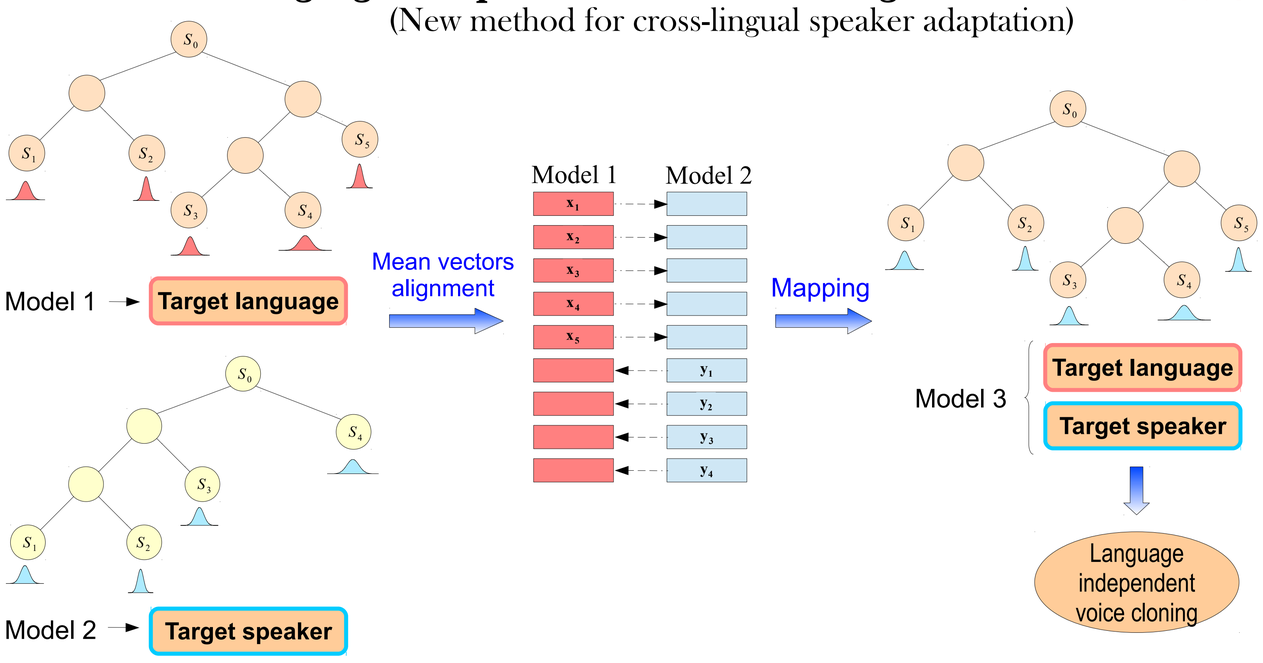
# IMPROVEMENTS IN HMM-BASED AND UNIT-SELECTION SPEECH SYNTHESIS TECHNIQUES

**Carmen Magariños**, Multimedia Technology Group (GTM), University of Vigo
Advisors: Eduardo Rodríguez Banga and Daniel Erro Eslava

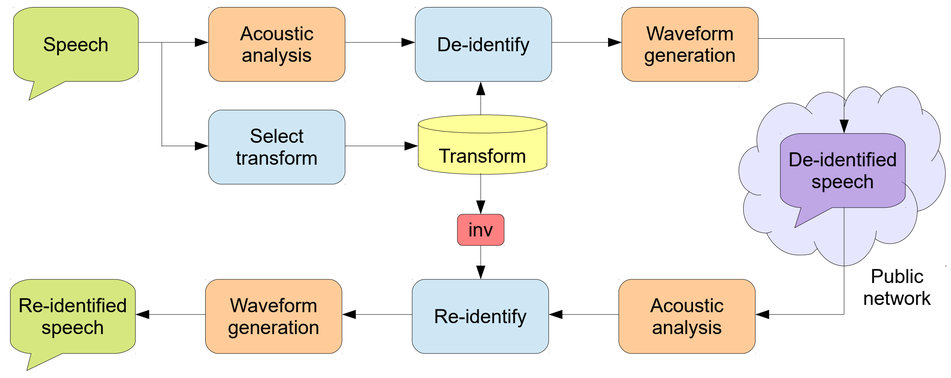
## Motivation of the work

### Language-independent acoustic cloning of HTS[1] voices [1, 2]

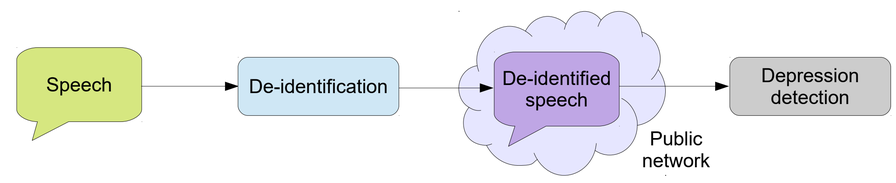
(New method for cross-lingual speaker adaptation)




### Speaker de/re-identification using voice transformation [3, 4]



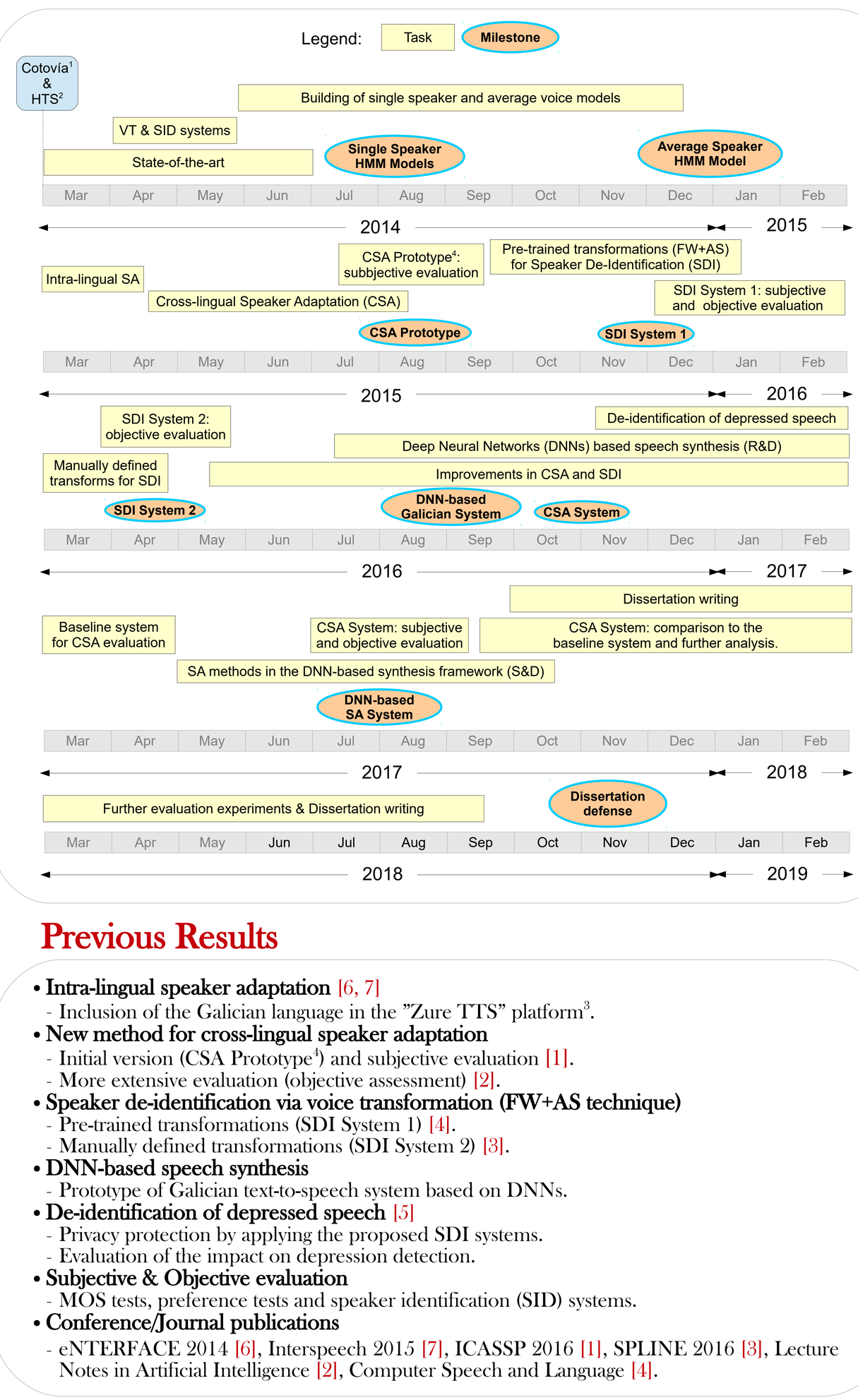

**Application:** De-identification of depressed speech for privacy protection while keeping the information related to the disease [5]




## Thesis objectives

- Analysis of **state-of-the-art techniques** for speech synthesis and speaker adaptation (SA).
- Apply **intra-lingual speaker adaptation techniques** to provide higher flexibility to speech synthesis systems (larger number of speakers, speaking styles and emotions) [6, 7].
- Study and development of **cross-lingual speaker adaptation (CSA) techniques** in order to obtain polyglot speakers (speech-to-speech translation, multilingual speech synthesizers) [1, 2].
- Analysis of different voice transformation (VT) techniques and application in the field of **speaker de-identification (SDI)** [3, 4, 5].

## Research Plan




## Previous Results

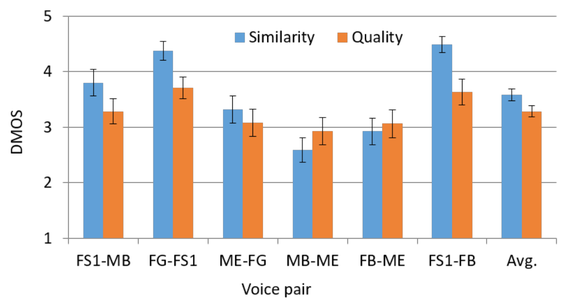
- **Intra-lingual speaker adaptation** [6, 7]
  - Inclusion of the Galician language in the "Zure TTS" platform[3].
- **New method for cross-lingual speaker adaptation**
  - Initial version (CSA Prototype[4]) and subjective evaluation [1].
  - More extensive evaluation (objective assessment) [2].
- **Speaker de-identification via voice transformation (FW+AS technique)**
  - Pre-trained transformations (SDI System 1) [4].
  - Manually defined transformations (SDI System 2) [3].
- **DNN-based speech synthesis**
  - Prototype of Galician text-to-speech system based on DNNs.
- **De-identification of depressed speech** [5]
  - Privacy protection by applying the proposed SDI systems.
  - Evaluation of the impact on depression detection.
- **Subjective & Objective evaluation**
  - MOS tests, preference tests and speaker identification (SID) systems.
- **Conference/Journal publications**
  - eNTERFACE 2014 [6], Interspeech 2015 [7], ICASSP 2016 [1], SPLINE 2016 [3], Lecture Notes in Artificial Intelligence [2], Computer Speech and Language [4].

## New Results & Discussion

- **Cross-lingual speaker adaptation**
  - Extended version of the voice cloning-based method (CSA System) [2].
  - Subjective and objective tests to measure the performance of both versions.
  - Comparison to a baseline method (KLD-based mapping technique) [8].
- **Speaker adaptation for DNN-based speech synthesis**
  - Training of DNN-based average voice models (AVMs) using different techniques.
  - Speaker adaptation experiments: several variants based on the retraining technique.
- **Conference/Journal publications**
  - IET Signal Processing [5], Computer Speech and Language (to be submitted soon) [8].

| | Basque-1 | Basque-2 | Catalan | English | Galician-1 | Galician-2 | Spanish | Original |
|---|---|---|---|---|---|---|---|---|
| VC-1 | 75.13% | 80.00% | 85.13% | 84.87% | 85.13% | 84.47% | | |
| VC-2 | 74.87% | 81.84% | 86.45% | 88.29% | 84.61% | 86.32% | 98.95% | 99.20% |
| KLD | - | - | - | - | 51.18% | 78.02% | | |

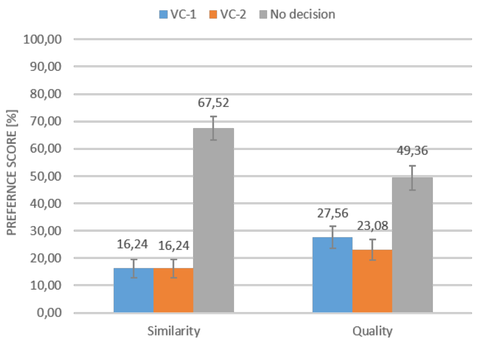
Proposed **CSA System**. Comparison in terms of SID accuracy per language: initial and extended versions (VC-1 and VC-2, respectively) and baseline method (KLD).



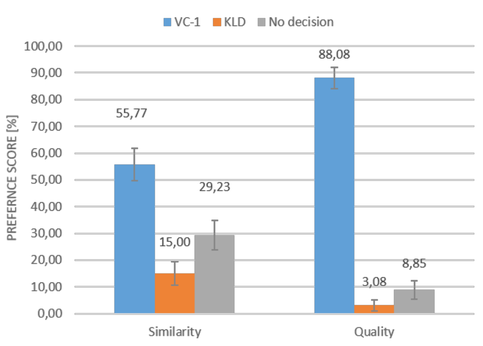

Proposed **CSA System**. DMOS results for the extended version of the proposed voice cloning method. Each pair represents a conversion direction composed of two voices (source and target).

Legend:

F: Female; M: Male
S: Spanish; B: Basque; E: English; G: Galician




Proposed **CSA System**. Results of the preference test when comparing the two versions of the proposed voice cloning method: initial version (VC-1) and extended version (VC-2).




Proposed **CSA System**. Results of the preference test when comparing the initial version of the voice cloning method (VC-1) and the KLD-based mapping technique proposed in [9] (KLD).

## Next Year Planning

- Dissertation writing & Defense

## References

[1] **C. Magariños**, D. Erro, E. R. Banga, "Language-independent acoustic cloning of HTS voices: a preliminary study", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5615-5619, Shanghai, March 2016.
[2] **C. Magariños**, D. Erro, P. Lopez-Otero, E. Rodriguez-Banga, "Language-Independent Acoustic Cloning of HTS Voices: an Objective Evaluation", Lecture Notes in Artificial Intelligence LNCS/LNAI, vol. 10077, pp. 54-63, 2016.
[3] **C. Magariños**, P. Lopez-Otero, L. Docio-Fernandez, D. Erro, E Rodriguez-Banga, C. Garcia-Mateo, "Piecewise Linear Definition of Transformation Functions for Speaker De-Identification", SPLINE, pp. 1-5, Aalborg, July 2016.
[4] **C. Magariños**, P. Lopez-Otero, L. Docio-Fernandez, E. Rodriguez-Banga, D. Erro, C. Garcia-Mateo, "Reversible speaker de-identification using pre-trained transformation functions", Computer Speech & Language, vol. 46, pp. 36-52, 2017.
[5] P. Lopez-Otero, **C. Magariños**, L. Docio-Fernandez, E. Rodriguez-Banga, D. Erro, C. Garcia-Mateo, "On the influence of speaker de-identification in depression detection", IET Signal Processing, vol. 11 (9), pp. 1023-1030, December 2017.
[6] D. Erro, I. Hernaez, E. Navas, A. Alonso, H. Arzelus, I. Jauk, N. Hy, **C. Magariños**, R. Perez-Ramon, M. Sulir, X. Tian, X. Wang, J. Ye, "ZureTTS: online platform for obtaining personalized synthetic voices", Proc. eNTERFACE, pp. 17-25, 2014.
[7] D. Erro, I. Hernaez, A. Alonso, D. Lorenzo, E. Navas, J. Ye, H. Arzelus, I. Jauk, N. Hy, **C. Magariños**, R. Perez-Ramon, M. Sulir, X. Tian and X. Wang, "Personalized Synthetic Voices for Speaking Impaired: Website and App", Interspeech, 2015.
[8] **C. Magariños**, D. Erro, E. Rodriguez-Banga, "Language-independent acoustic cloning of HTS voices", Computer Speech & Language (to be submitted soon).
[9] Y.J. Wu, Y. Nankaku, K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis", Proc. Interspeech, pp. 528-531, 2009.
[10] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, S. King, "A study of speaker adaptation for DNN-based speech synthesis", Proc. Interspeech, pp. 879-883, Dresden, September 2015.

## Acknowledgements

This research was funded by the Spanish Government ('TraceThem' project TEC2015-65345-P and research grant BES-2013-063708), the Xunta de Galicia ('Agrupación Estratexica Consolidada de Galicia' accreditation 2016-2019 and 'Grupos de Referencia Competitiva' GRC2014/024), the European Regional Development Fund (ERDF) and the COST Action IC1206.

[1] http://sourceforge.net/projects/cotovia/ , [2] http://hts.sp.nitech.ac.jp/, [3] http://aholab.ehu.eus/zuretts/, [4] http://goo.gl/FwemL4