Universidade Vigo

GOBIERNO DE ESPAÑA · MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

Unión Europea · Fondo Europeo de Desarrollo Regional · "Una manera de hacer Europa"

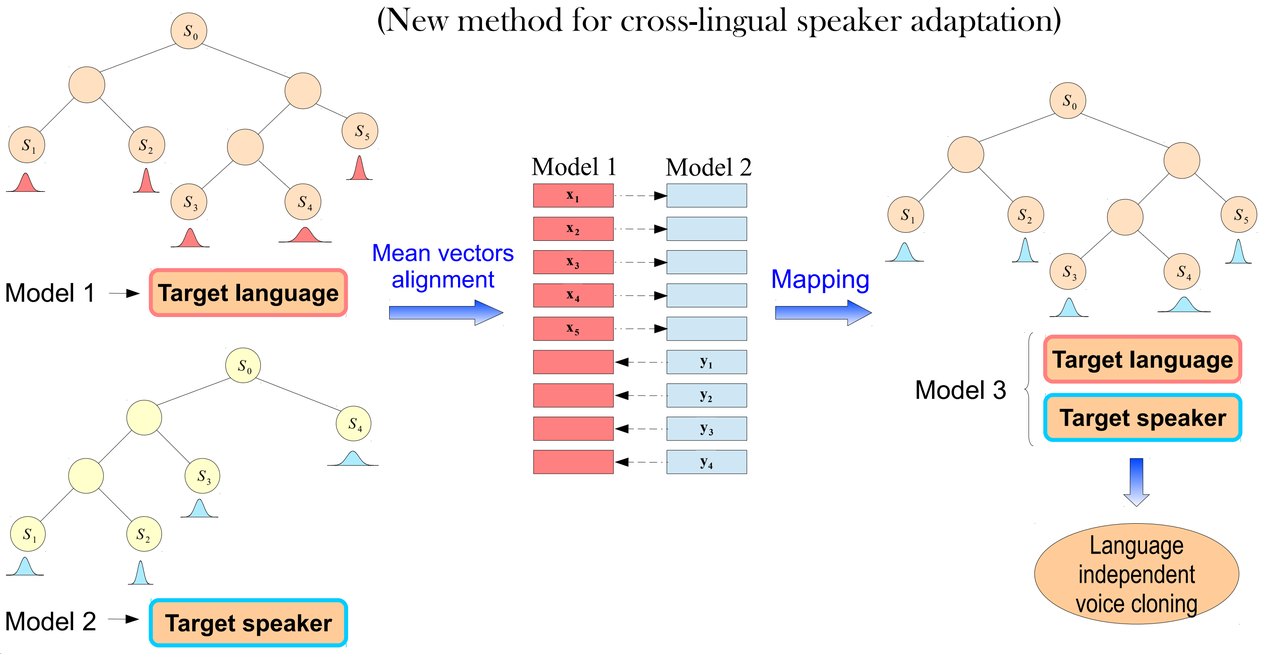
IMPROVEMENTS IN HMM-BASED AND UNIT-SELECTION SPEECH SYNTHESIS TECHNIQUES

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

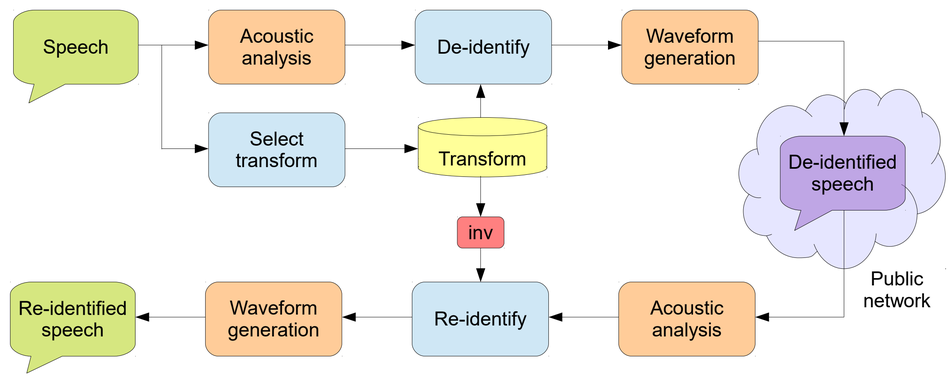
Motivation of the work

Language-independent acoustic cloning of HTS[1] voices [1]

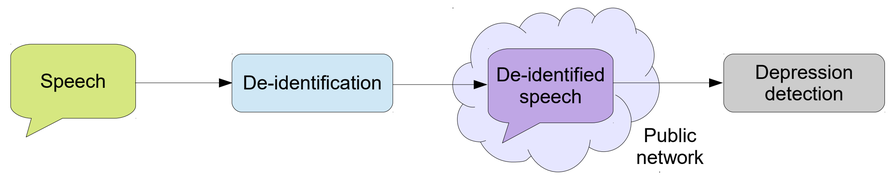
(New method for cross-lingual speaker adaptation)



Speaker de/re-identification using voice transformation [2, 3]



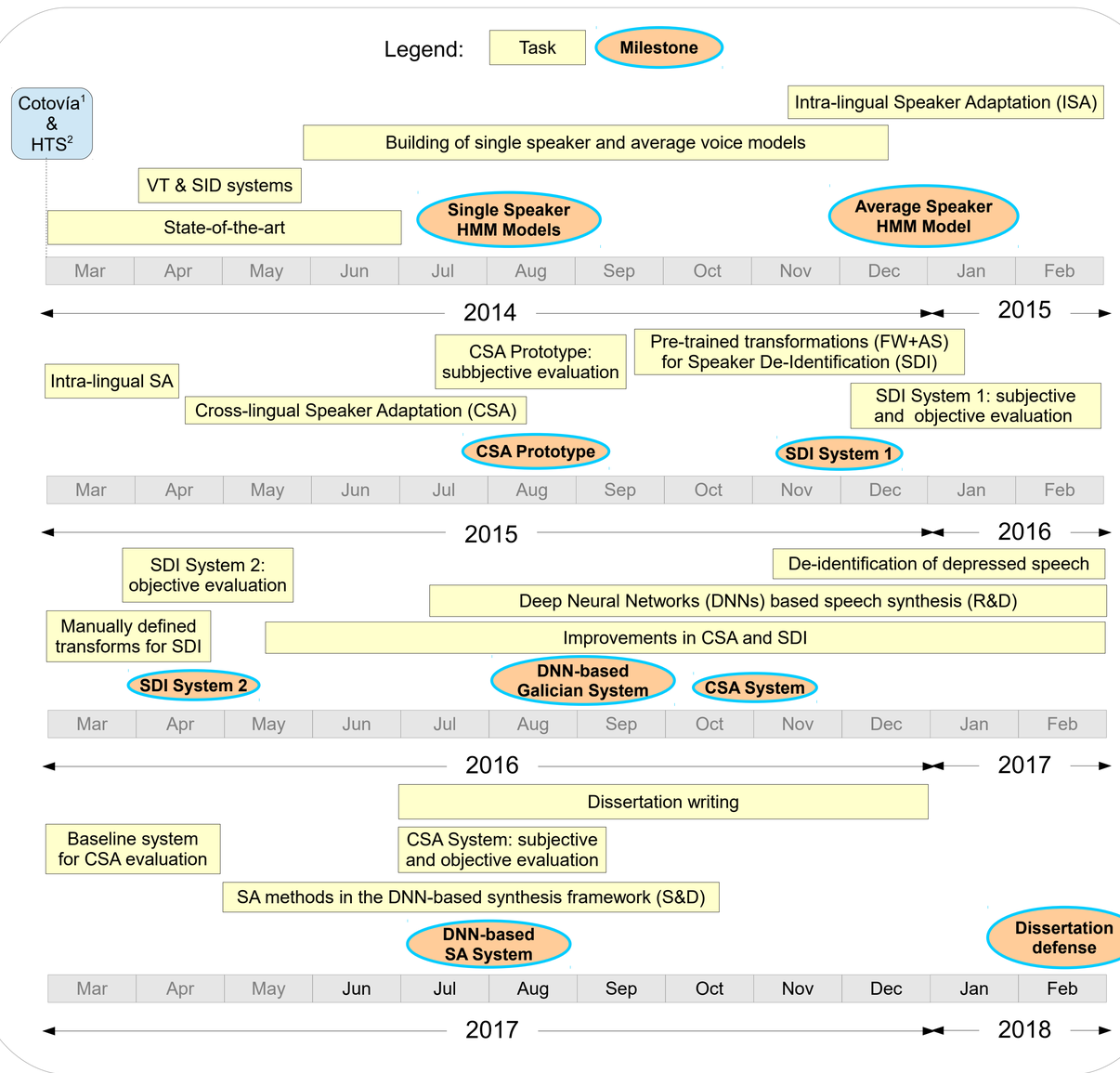
Application: De-identification of depressed speech for privacy-preserving while keeping the information related to the disease [4]



Thesis objectives

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

Research Plan



Previous Results

- **Intra-lingual speaker adaptation** [5, 6]
  - Inclusion of the Galician language in the "Zure TTS" platform[3].
- **New method for cross-lingual speaker adaptation**[4] [1]
- **Speaker de-identification via voice transformation (FW+AS technique)**
  - Initial study of pre-trained transformations (SDI System 1).
  - Manually defined transformations (SDI System 2) [2].
- **Subjective & Objective evaluations: MOS tests and speaker identification (SID) systems**
- **Conference/Journal publications**
  - eNTERFACE 2014 [5], Interspeech 2015 [6], ICASSP 2016 [1], SPLINE 2016 [2]

Acknowledgements

This research was funded by the Spanish Government (project 'TraceThem' TEC2015-65345-P, research grant BES-2013-063708), the European Regional Development Fund (ERDF) and the COST Action IC1206.

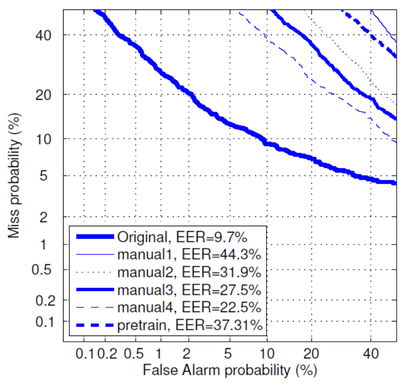
[1] http://sourceforge.net/projects/cotovia/ , [2] http://hts.sp.nitech.ac.jp/, [3] http://aholab.ehu.eus/zuretts/, [4] http://goo.gl/FwemL4, [5] http://www.cstr.ed.ac.uk/projects/merlin/

New Results & Discussion

- **Cross-lingual speaker adaptation**
  - More exhaustive evaluation of the initially proposed method (CSA Prototype) [7].
  - Improvements in the adaptation method (CSA System).
- **Speaker de-identification**
  - Further study of pre-trained transformations functions (SDI System 1) [3].
  - Influence of de-identification in the perception of diseases affecting speech production: de-identification of depressed speech using SDI Systems 1 and 2, and evaluation of the impact on depression detection using automatic tools [4].
- **DNN-based speech synthesis**
  - Initial prototype of Galician text-to-speech system based on DNNs (using The Merlin Toolkit[5]).
- **Conference/Journal publications**
  - Lecture Notes in Artificial Intelligence [7], Computer Speech and Language [3], IET Signal Processing (under review) [4].

	Traditional GMM	FW+AS
converted-target	5.93 ± 0.05	7.79 ± 0.07
converted-source	7.29 ± 0.09	7.72 ± 0.18
reconverted-source	4.28 ± 0.04	0.88 ± 0.03
source-target	8.83 ± 0.09	

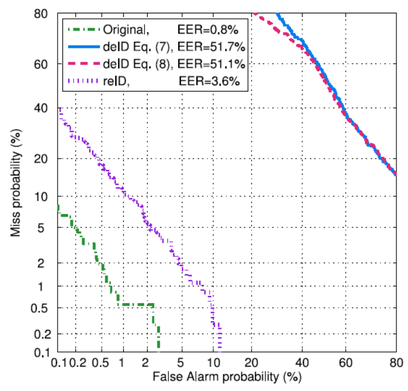
Proposed **SDI System 1**: comparison in terms of average Mel-cepstral distortion with traditional GMM-based voice conversion.



	Matched	Mismatched	Partially matched (known)	Partially matched (unknown)
Transformation	RMSE	RMSE	RMSE	RMSE
original	9.72	n/a	n/a	10.55
manual1	11.85	13.20	11.25	11.71
manual2	11.13	11.26	11.32	11.26
manual3	11.73	10.64	10.89	11.33
manual4	11.39	10.71	10.43	11.02
pretrain	9.95	11.20	9.50	10.03

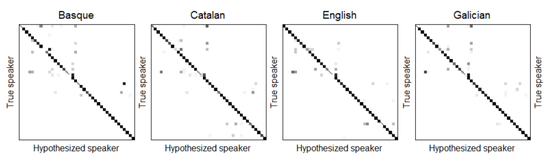
De-identification of depressed speech: DET curves and depression level estimation results for original speech and different transformation functions.

Identification accuracy on original speech	99.2%
De-identification accuracy using Eq. (13)	96.1%
De-identification accuracy using Eq. (14)	95.0%
Re-identification accuracy	97.2%



Proposed **SDI System 1**: results in terms of SID accuracy (i-vectors + PLDA scoring) and DET curves for original, de-identified and re-identified speech.

	Basque	Catalan	English	Galician	Spanish
Accuracy	82.50%	85.39%	86.84%	85.92%	96.97%
Accuracy (2)	90.39%	92.37%	91.05%	93.16%	99.87%



Proposed **CSA Prototype** : SID accuracy (i-vectors + dot-scoring) and confusion matrices per language.

Next Year Planning

- **Cross-lingual speaker adaptation**
  - Evaluation of the CSA System: MOS test and comparison to a baseline system [8].
- **DNN-based speech synthesis (ongoing work at The University of Edinburgh)**
  - Approach to DNN-based speaker adaptation techniques [9].
- **Conference/Journal publications**
  - Coming journal submission: improved cross-lingual adaptation system (CSA System).
- **Dissertation writing**

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**, 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.
[3] **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.
[4] 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 (under review).
[5] 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.
[6] 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.
[7] **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.
[8] 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.
[9] 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.