UniversidadeVigo

GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

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

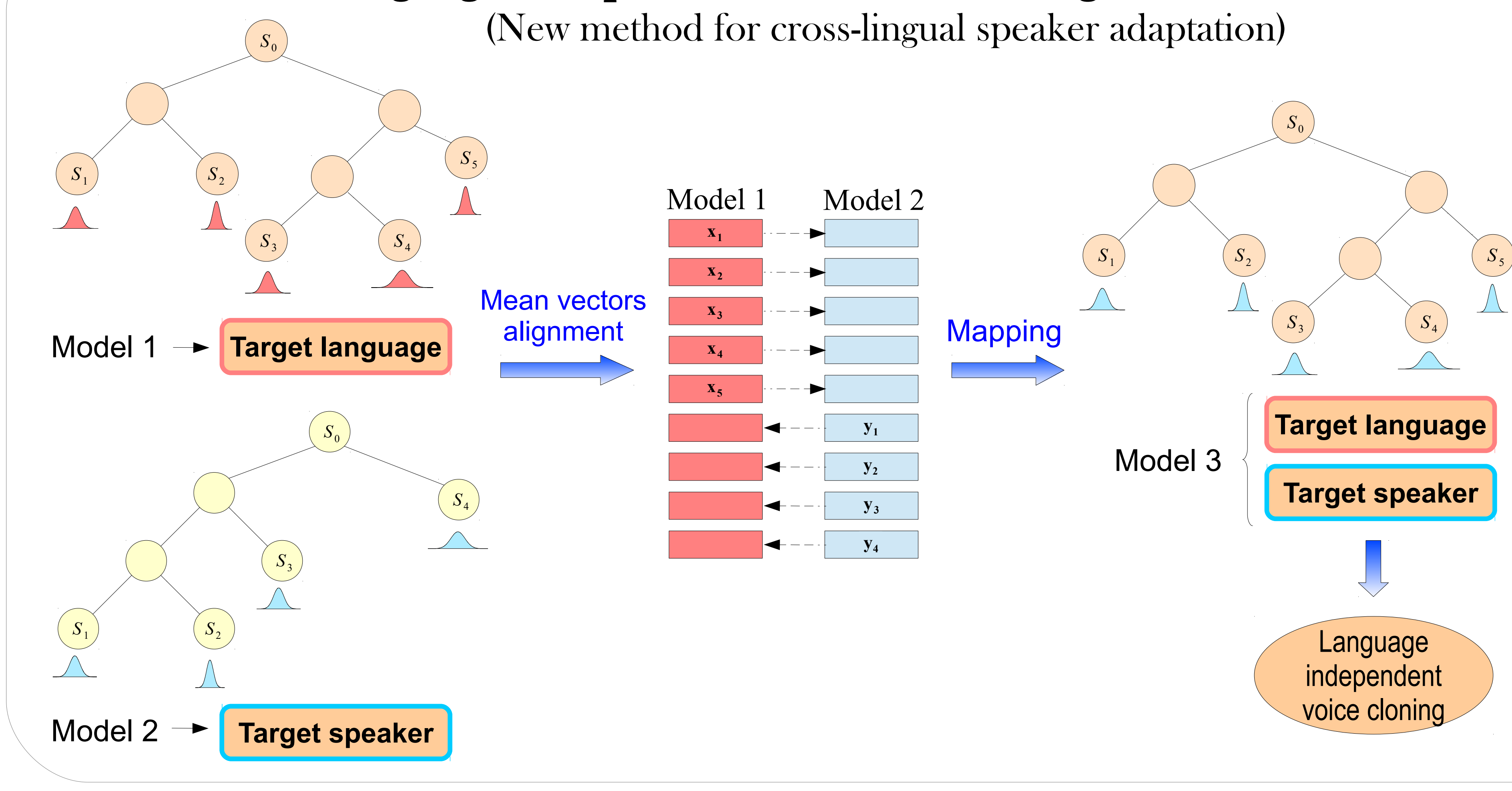
IMPROVEMENTS IN HMM-BASED AND UNIT-SELECTION SPEECH SYNTHESIS TECHNIQUES

Carmen Magariños Iglesias and Eduardo Rodríguez Banga (Advisor)
Multimedia Technology Group (GTM), University of Vigo

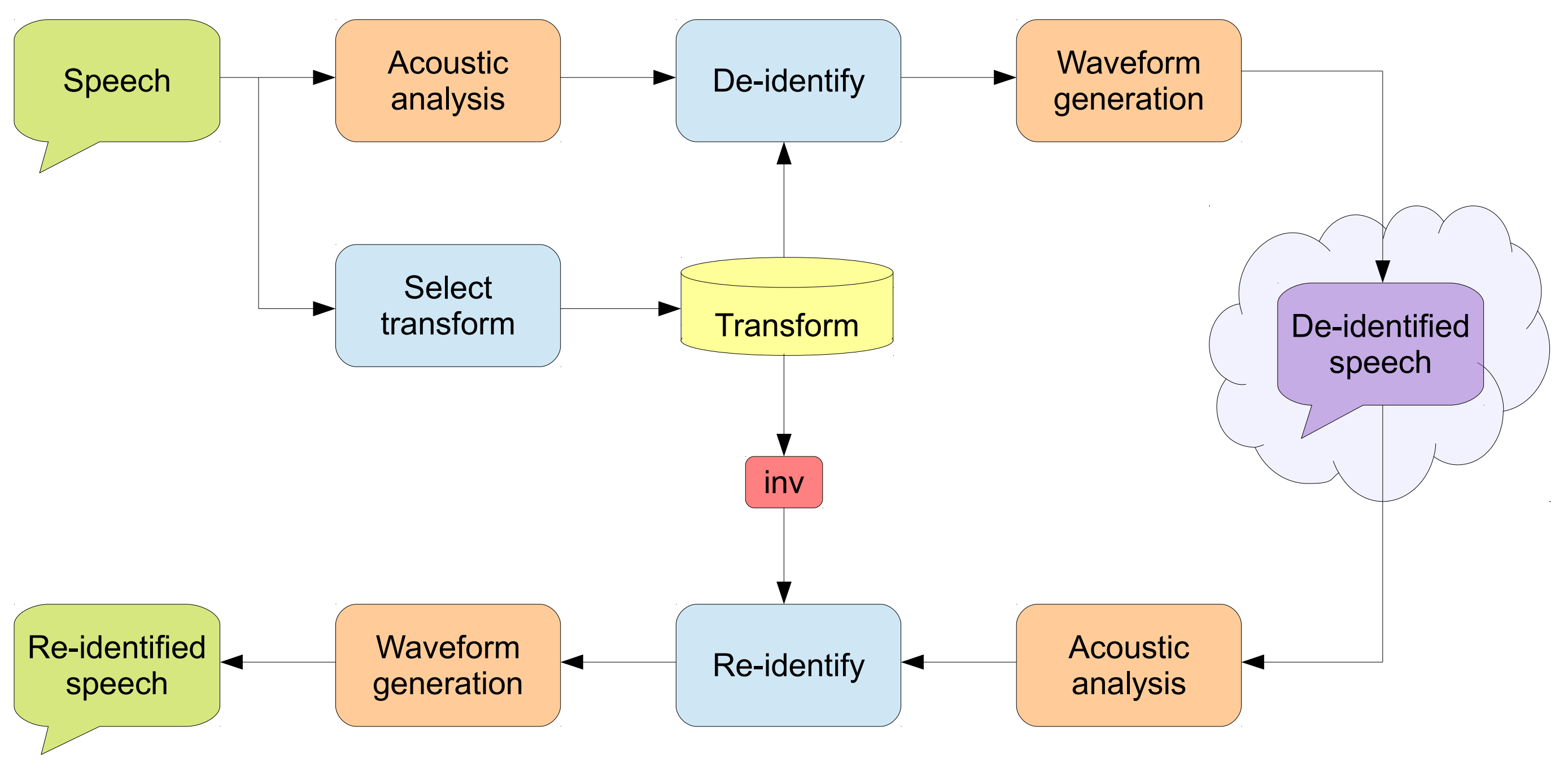
Motivation of the work

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

(New method for cross-lingual speaker adaptation)


Model 1 → Target language
Model 2 → Target speaker
Mean vectors alignment
Model 1 Model 2
Mapping
Model 3
Target language
Target speaker
Language independent voice cloning

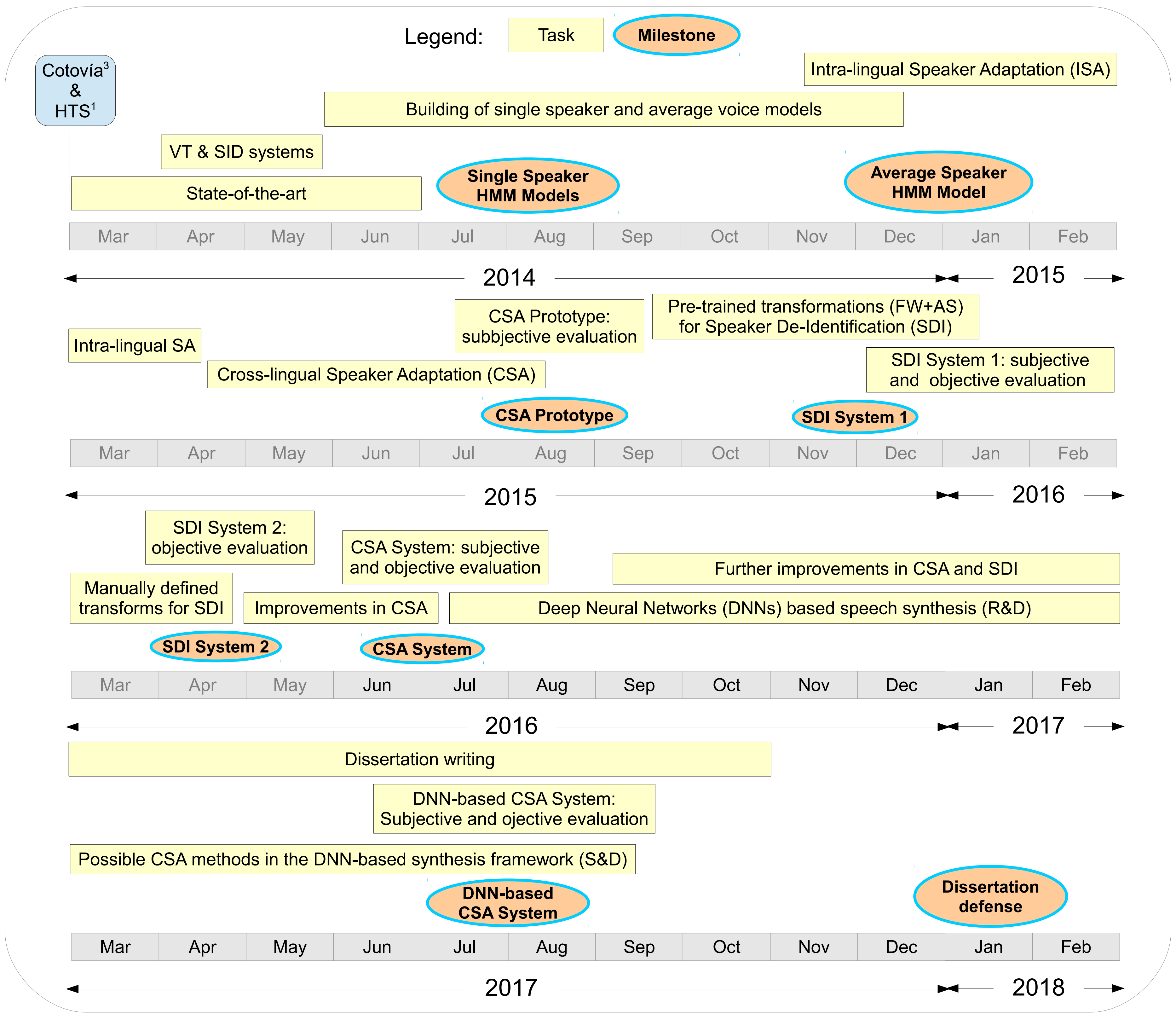
Speaker de/re-identification using voice transformation functions


Speech
Acoustic analysis
De-identify
Waveform generation
Select transform
Transform
De-identified speech
inv
Re-identified speech
Waveform generation
Re-identify
Acoustic analysis

Thesis objectives

- Analysis of **state-of-the-art techniques** for speech synthesis [2], covering speaker adaptation (SA) methods.
- Propose **improvements** to existing techniques, including more **efficient systems** for specific applications with memory or computational load restrictions.
- Apply **intra-lingual speaker adaptation techniques** [3] to increase the flexibility of the speech synthesis systems (larger number of speakers, speaking styles and emotions).
- Study, development and implementation of **cross-lingual speaker adaptation techniques** [4] with the aim of obtaining multilingual speakers (speech-to-speech translation).
- Analysis of different voice transformation (VT) techniques [5] and application in the field of **speaker de-identification**.
- Use of speech synthesis techniques in related applications, such as the **robustness evaluation** of Speaker Identification (SID) Systems. [6]

Research Plan


Legend: Task Milestone
Cotovía[3] & HTS[1]
Intra-lingual Speaker Adaptation (ISA)
Building of single speaker and average voice models
VT & SID systems
State-of-the-art
Single Speaker HMM Models
Average Speaker HMM Model
Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb
2014 2015
CSA Prototype: subjective evaluation
Pre-trained transformations (FW+AS) for Speaker De-Identification (SDI)
Intra-lingual SA
Cross-lingual Speaker Adaptation (CSA)
SDI System 1: subjective and objective evaluation
CSA Prototype
SDI System 1
Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb
2015 2016
SDI System 2: objective evaluation
CSA System: subjective and objective evaluation
Further improvements in CSA and SDI
Manually defined transforms for SDI
Improvements in CSA
Deep Neural Networks (DNNs) based speech synthesis (R&D)
SDI System 2
CSA System
Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb
2016 2017
Dissertation writing
DNN-based CSA System: Subjective and ojective evaluation
Possible CSA methods in the DNN-based synthesis framework (S&D)
DNN-based CSA System
Dissertation defense
Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb
2017 2018

Next Year Planning

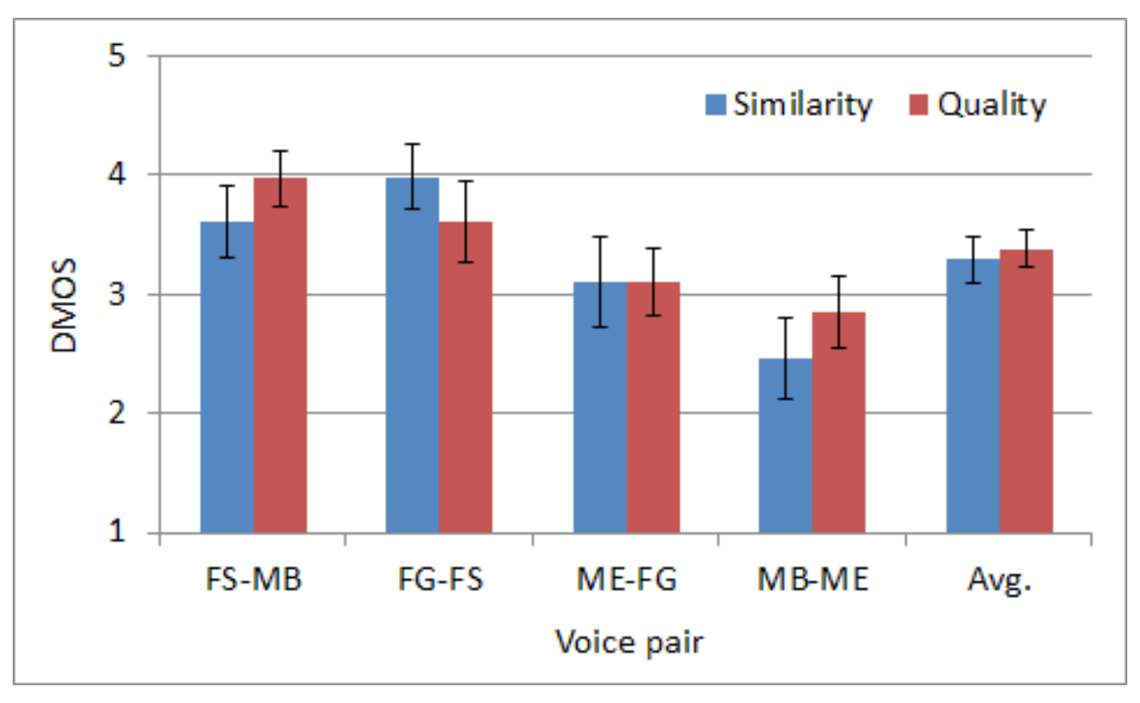
- **Cross-lingual speaker adaptation**
  - Improvements in the adaptation method (final version of the cross-lingual adaptation system).
  - Subjective and objective evaluations (MOS tests and speaker identification system).
- **Speaker de-identification**
  - Further improvements and evaluation.
- **DNN-based speech synthesis**
  - Exploration of DNN-based speech synthesis approaches.
  - Research on possible cross-lingual adaptation methods for DNN-based synthesis.
- **Conference/Journal publications**
  - Coming journal submission: improved cross-lingual adaptation system.

Acknowledgements

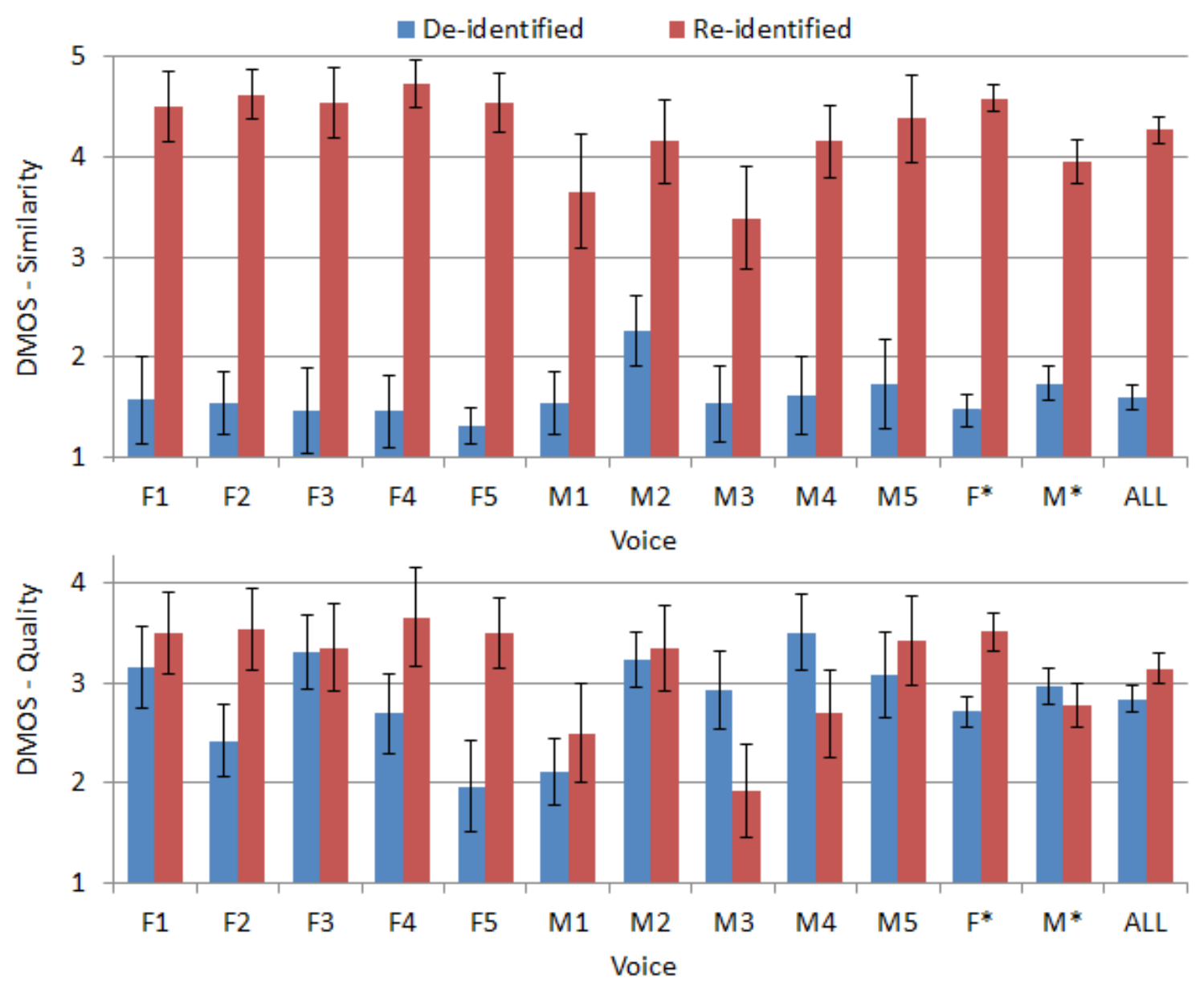
This work has been partially supported by the Spanish Government and the European Regional Development Funds under the project SpeechTech4All (TEC2012-38939-C03-01, BES-2013-063708).

Results & Discussions

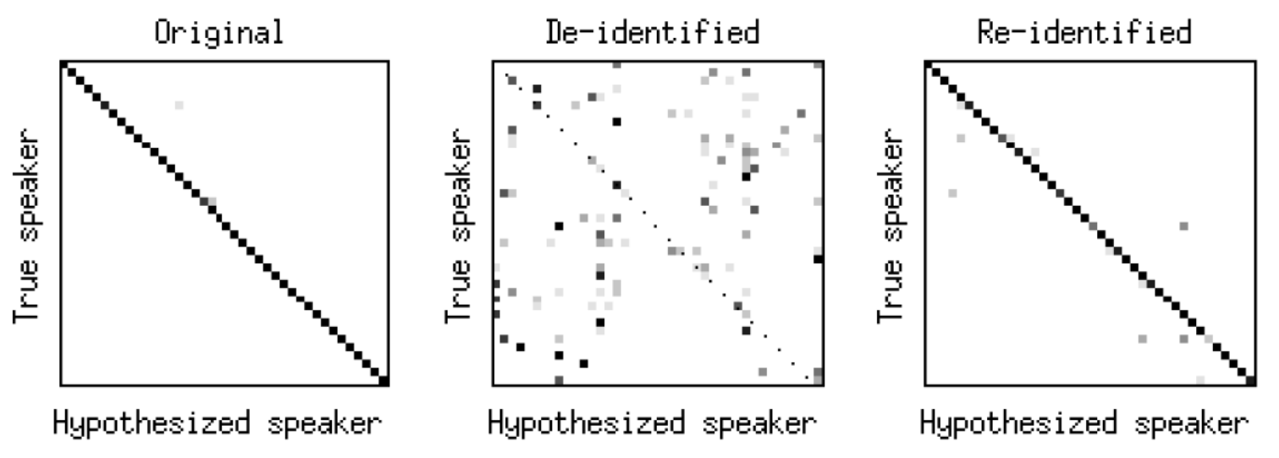
- **Improvements in HMM-based speech synthesis**
  - Ahocoder [7] integration: higher quality than previous vocoder.
- **Intra-lingual speaker adaptation**
  - Average voice model (AVM) for Spanish using Albayzín database.
  - Inclusion of the Galician language in the "Zure TTS" platform[2] [8].
- **New method for cross-lingual speaker adaptation** [1]
  - Language-independent acoustic cloning of HTS[1] voices.
  - Adaptation method based on INCA algorithm [9].
  - Examples at http://goo.gl/FwemL4.
- **Speaker de-identification using voice transformation functions**
  - Pre-trained transformations based on the FW+AS technique [10] (SDI System 1).
  - Manually defined transformations using piecewise linear approximation of FW functions (SDI System 2) [11].
- **Subjective evaluations**
  - Perceptual listening tests.
  - Differential mean opinion score (DMOS).
- **Objective evaluations**
  - Speaker identification system as objective measure.
  - State-of-the-art i-vector approach combined with dot-scoring.
- **Conference/Journal publications**
  - eNTERFACE 2014 [8], Interspeech 2015 [11], ICASSP 2016 [1], SPLINE 2016 (accepted) [12], IEEE Signal Processing Letters (submitted).



Proposed **cross-lingual adaptation** method: DMOS results.



Proposed **SDI System 1**: DMOS results for similarity and quality.



Identification accuracy on original speech	99.2%
De-identification accuracy	89.5%
Re-identification accuracy	94.2%

Proposed **SDI System 1**: confusion matrices and results in terms of accuracy for original, de-identified and re-identified speech.

Transformation	FW	FW+F0	FW+F0+AS
Trans1	82.5%	98.6%	96.9%
Trans2	53.9%	87.2%	88.1%
Trans3	30.6%	64.2%	68.3%
Trans4	4.4%	28.0%	36.7%

Proposed **SDI System 2**: speaker de-identification results in terms of accuracy for the different transformations.

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] Archana Balyan, S. S. Agrawal, Amita Dev, "Speech Synthesis: A Review", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 2- Issue 6, June 2013.

[3] J. Yamagishi, "Average-Voice-Based Speech Synthesis", Ph.D Thesis, Tokyo Institute of Technology, 2006.

[4] K. Oura, J. Yamagishi, M. Wester, S. King, K. Tokuda, "Analysis of unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using KLD-based transform mapping", Speech Communication, vol. 54, pp. 703-714, 2012.

[5] Y. Stylianou, "Voice transformation: A survey," Proc. ICASSP, pp. 3585–3588, 2009.

[6] Q. Jin, A. Toth, A. Black, T. Schultz, "Is voice transformation a threat to speaker identification?", Proc. ICASSP, 2008.

[7] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics Plus Noise Model Based Vocoder for Statistical Parametric Speech Synthesis", IEEE Journal of Selected Topics in Signal Processing, Vol 8, No. 2, 2014.

[8] 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, Bilbao, June 2014.

[9] D. Erro, A. Moreno, A. Bonafonte, "INCA Algorithm for Training Voice Conversion Systems from Nonparallel Corpora", IEEE Transactions on Audio, Speech, and Language Processing (ISSN: 1558-7916), vol. 18(5), pp. 944-953, 2010.

[10] T. C. Zorila, D. Erro, I. Hernaez, "Improving the Quality of Standard GMM-Based Voice Conversion Systems by Considering Physically Motivated Linear Transformations", Communications in Computer and Information Science (ISSN: 1865-0929), vol. 328, pp. 30-39, 2012.

[11] 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.

[12] **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. xxx-xxx, Aalborg, July 2016.

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