Applying Spectral Normalisation and Efficient Envelope Estimation for the VCC 2016

Fernando Villavicencio[1]
Junichi Yamagishi[1]
Jordi Bonada[2]
Felipe Espic[3]

[1]National Institute of Informatics (NII)
[2]Universitat Pompeu Fabra (UPF)
[3]The Centre for Speech Technology Research (CSTR)

Interspeech conference, San Francisco, September 10th 2016

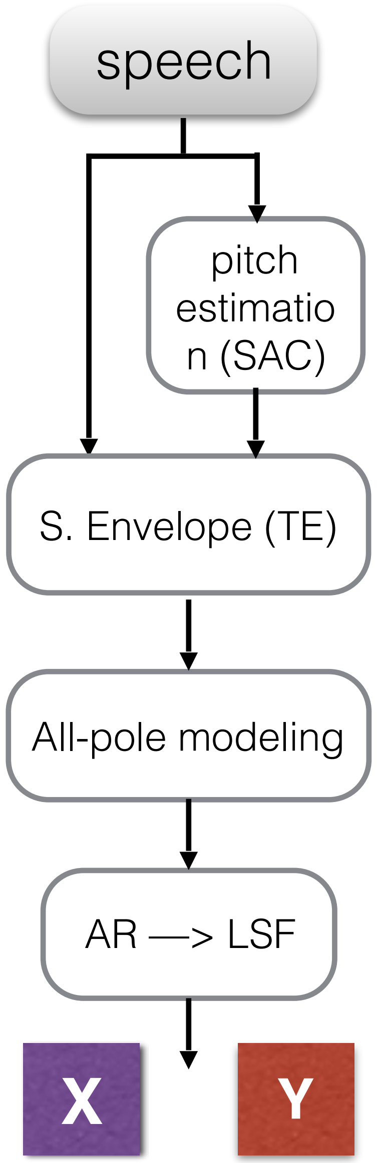
Outline

- Previous work on Voice Conversion (VC)
- Proposed methodology for the VC challenge 2016
- Performance evaluation
- Results at the challenge (perceptual evaluation)

I.I Applying improved spectral modeling to VC

- We introduced the cepstrum-based **"True-Envelope"** (TE) for spectral feature extraction.
- TE allows to **fit closely** the spectral envelope information.
- The cepstral **order** can be **optimized** given f0.
- TE All-Pole (TEAP): **better fitting and stability** conditions than LPC or DAP.
- TEAP applied to GMM-based VC showed **improved** converted **speech quality** [1].

features extraction



[1] **F. Villavicencio, A. Röbel, and X. Rodet, "Applying improved spectral modeling for high-quality voice conversion," in Proc. of ICASSP, 2009.**

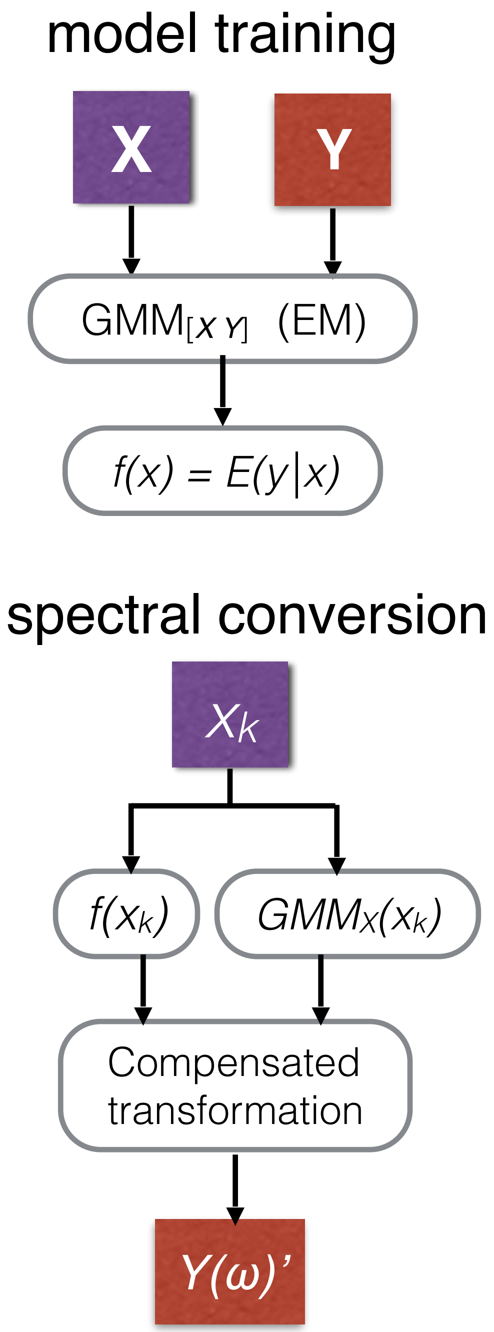
I.II Feature-model error compensation for statistical spectral transformation

- GMM: **error** when representing a feature by the probabilistic model (mixture).

- The difference between source and predicted envelopes can be considered as a **transformation filter**.

- We define/apply this transformation in terms of the **actual** source envelope ***seen*** by the mixture.

- The new transformation **preserves natural features** on poorly modeled spectra: less degradations [2].

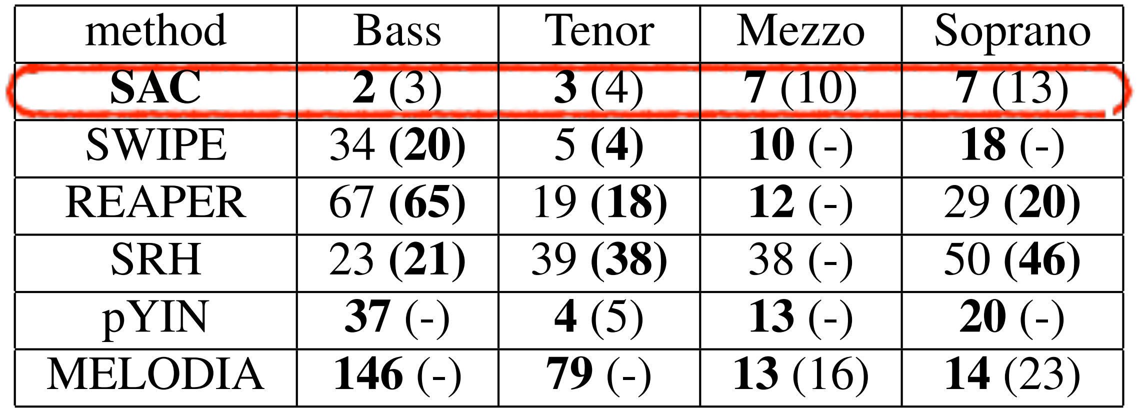


[2] F. Villavicencio, J. Bonada, and Y. Hisaminato, "Observation- model error compensation for enhanced spectral envelope transformation in voice conversion," in Proc. of IEEE-MLSP'15, 2015.

I.III Pitch estimation based on spectral amplitude autocorrelation (SAC)

- Pitch evolution: performance factor on a speech synthesis process.
- Bonada proposed the SAC method for robust and smooth pitch curve estimation.
- The technique takes advantage of the properties of multi-resolution spectra.
- SAC showed lower error rates on large pitch-range singing (opera) [3].

RMSE per voice type (Hertz)



method	Bass	Tenor	Mezzo	Soprano
SAC	**2** (3)	**3** (4)	**7** (10)	**7** (13)
SWIPE	34 **(20)**	5 **(4)**	**10** (-)	**18** (-)
REAPER	67 **(65)**	19 **(18)**	**12** (-)	29 **(20)**
SRH	23 **(21)**	39 **(38)**	38 (-)	50 **(46)**
pYIN	**37** (-)	**4** (5)	**13** (-)	**20** (-)
MELODIA	**146** (-)	**79** (-)	**13** (16)	**14** (23)

[3] F. Villavicenio, J. Bonada, J. Yamagishi, M. Pucher, "Efficient Pitch Estimation on Natural Opera-Singing by a Spectral Correlation based Strategy", IEICE Technical report, 2015.

II.I Observations on the performance of previous work

- The **conversion** effect is **not satisfactory**: alternatives to GMM-based frameworks?.

- **Gender conversion** is not always achieved: a VTLN strategy may be helpful.

- **Average pitch** matching is not sufficient for prosody conversion: additional features should be converted… or broken.

- Spectral amplitude **over-estimations** at the first few harmonics by TE on female voices: necessary to reduce pitch-height **dependency**.

II.II New features incorporated in our entry of the VC challenge 2016

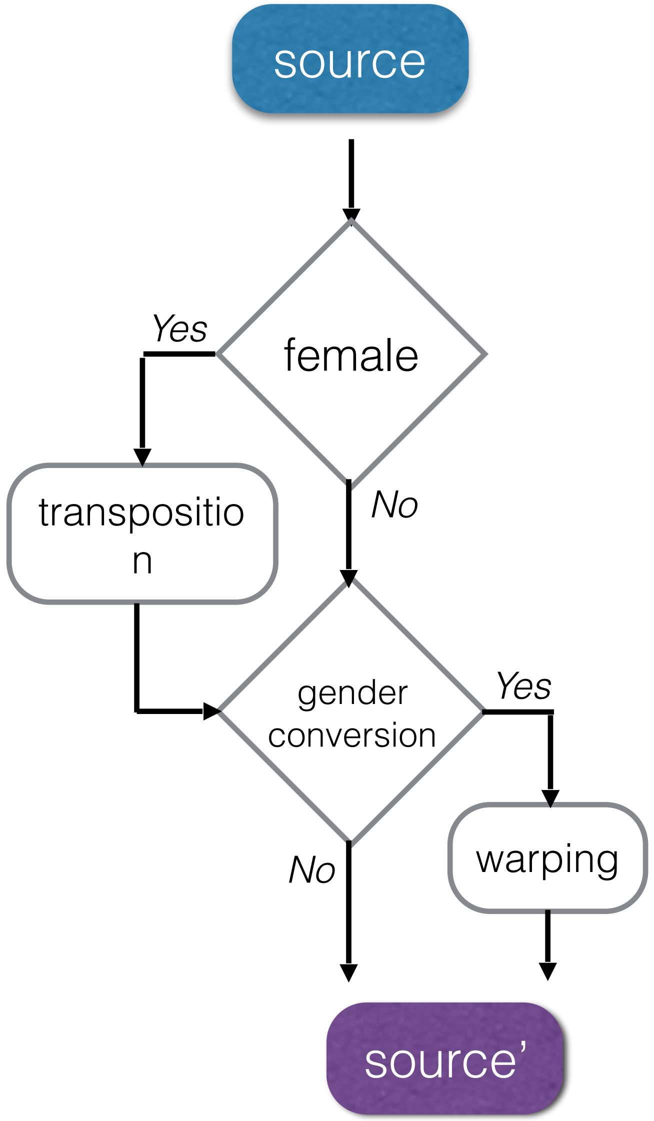
- **Global gender conversion**: A fixed warping factor is applied for inter-gender conversions.

- Female speech: the pitch is transposed to half of its value to artificially **“double” the harmonic structure**.

- The above procedures can be referred to as a **“spectral normalisation”** aiming for preferable processing conditions.

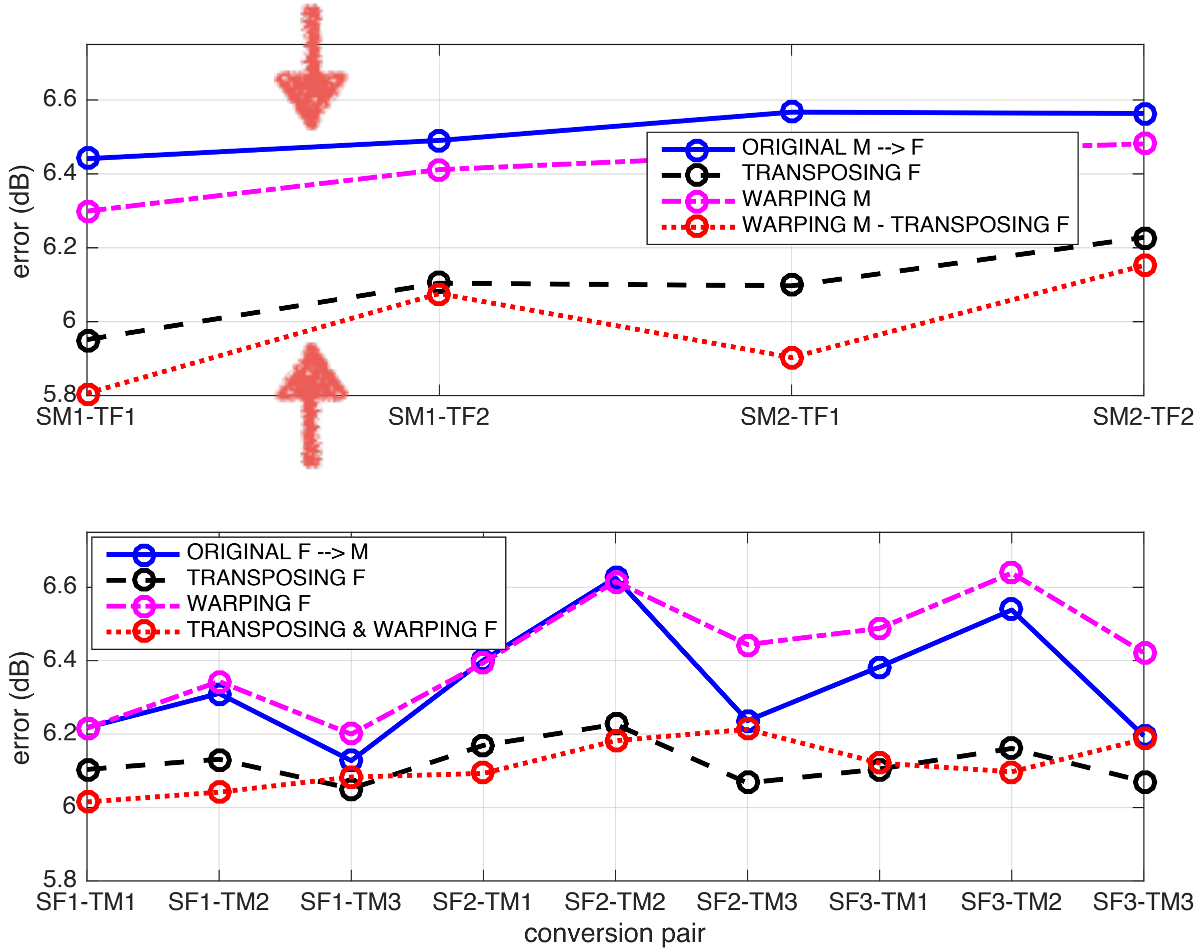
- Global **duration modification** according to the average differences in the training data.

spectral conditions normalisation

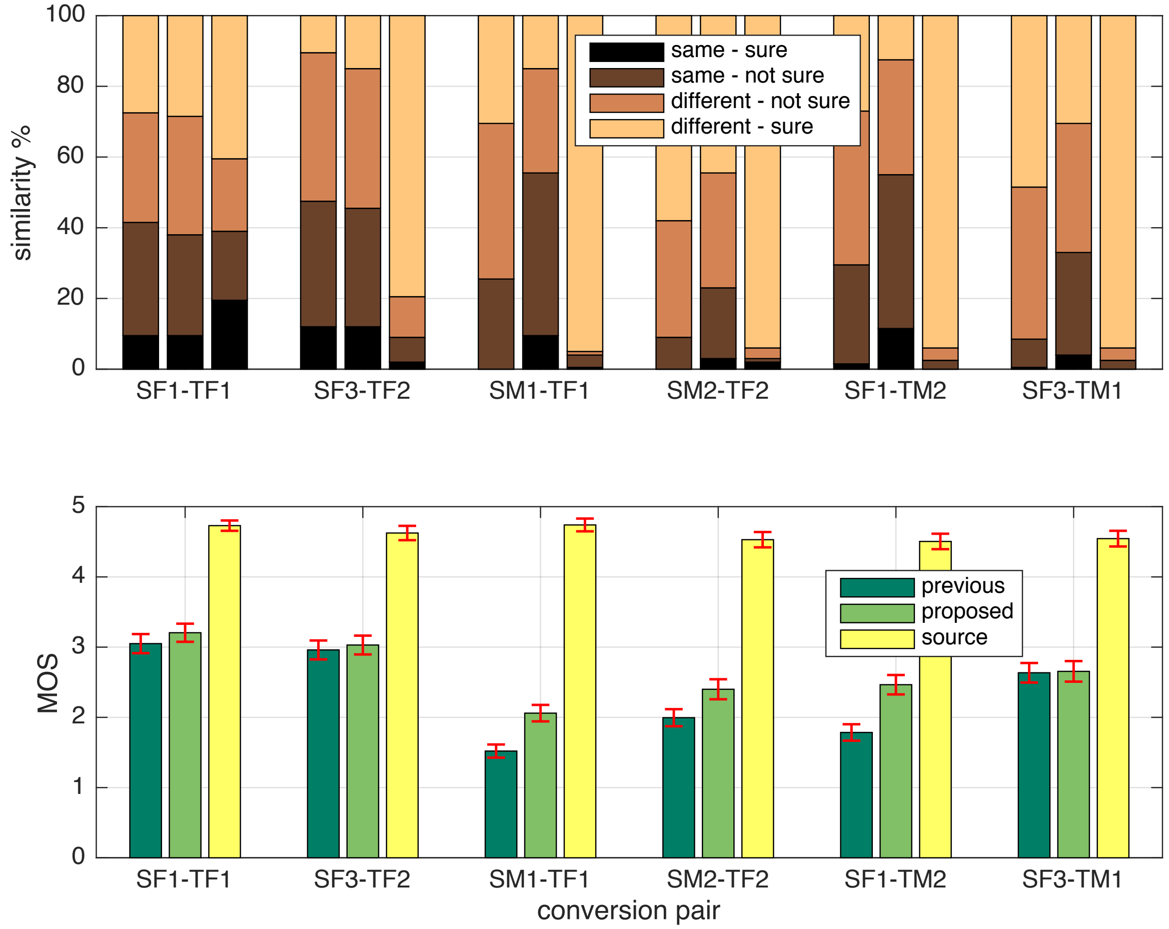


III.I Performance comparison with previous approach

inter-gender spectral distortion

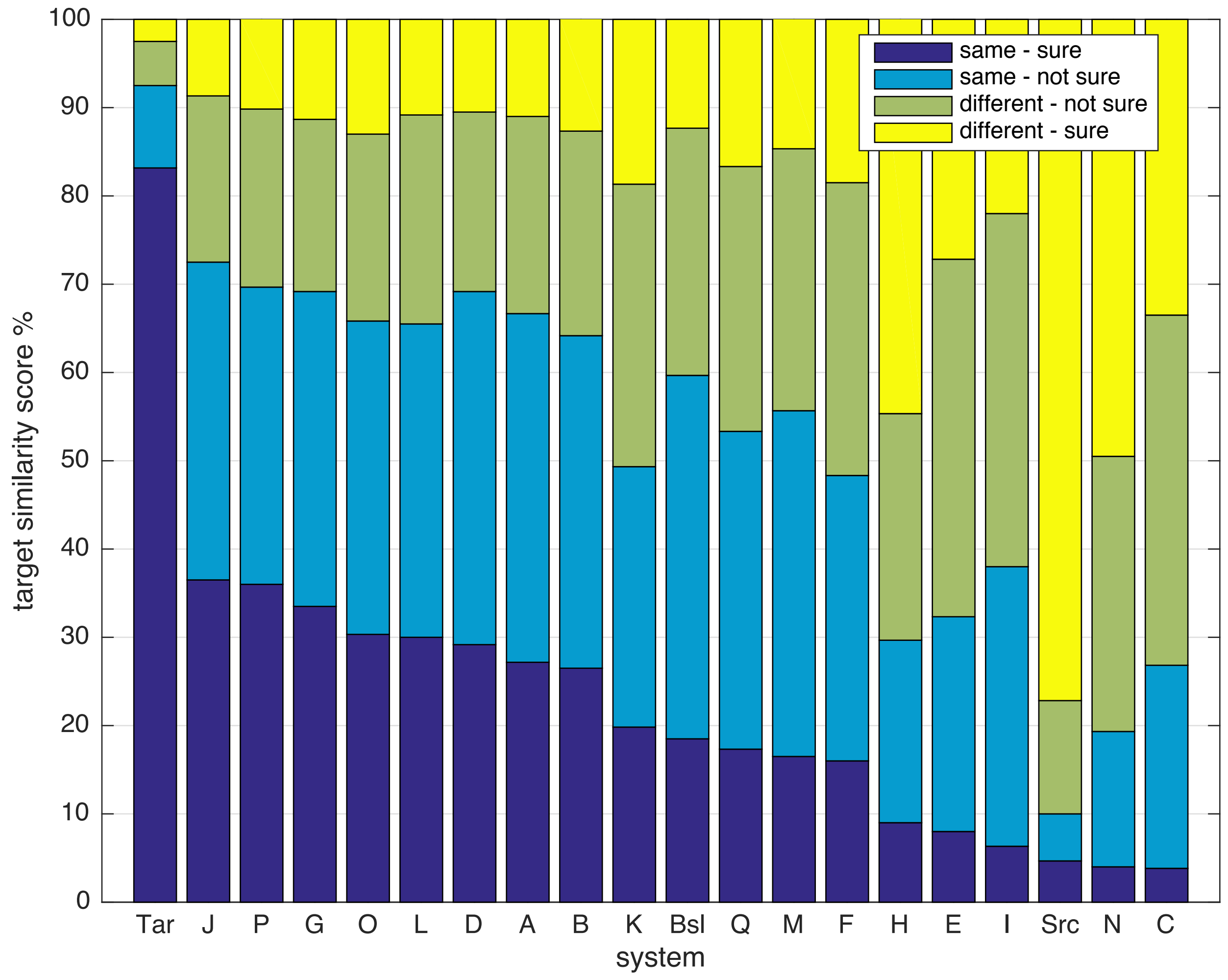


perceptual evaluation

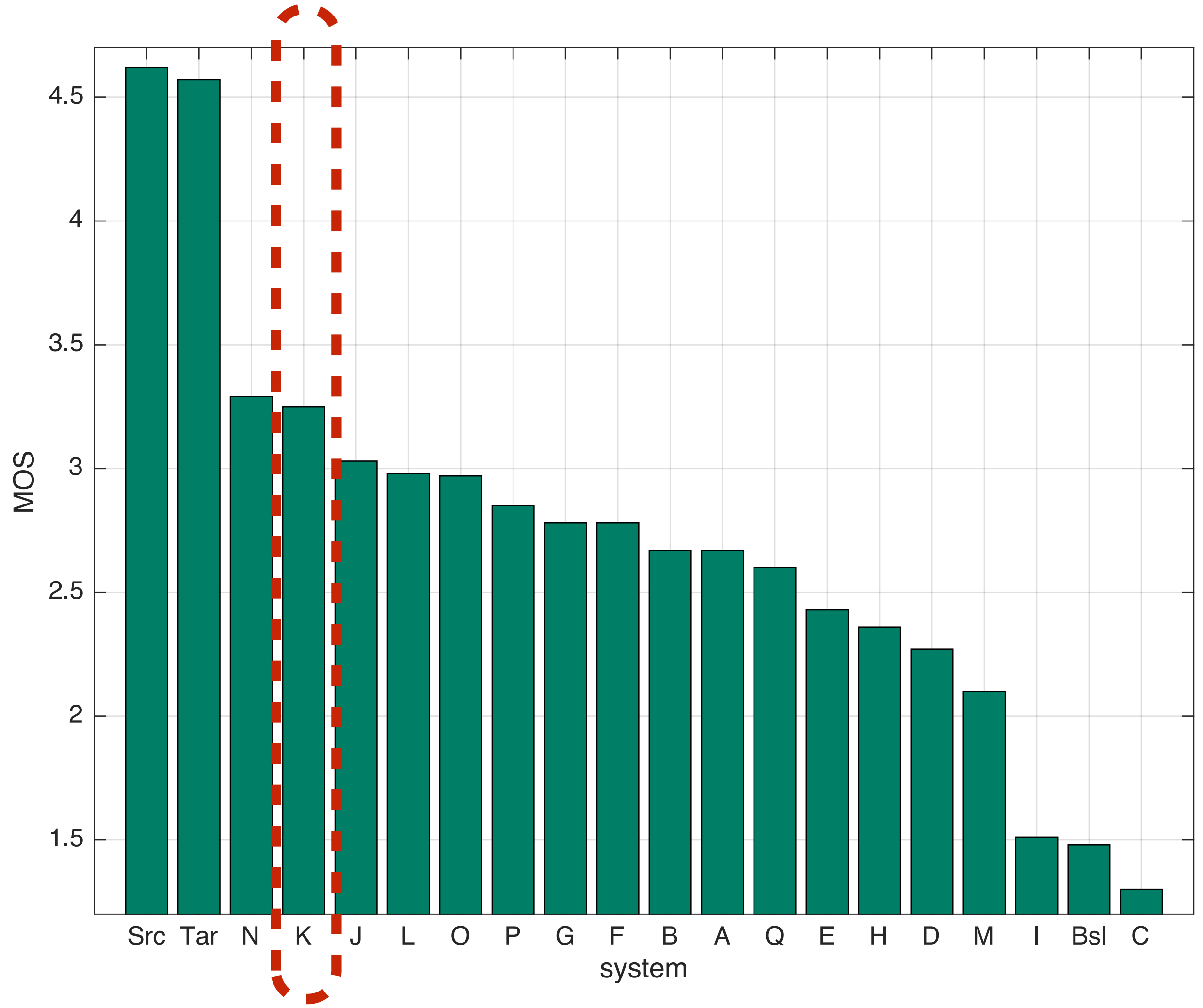


III.II Results at the Voice Conversion Challenge 2016

similarity



quality



M-to-F

F-to-M

Target

Conversion

Some conclusions

- Global warping: **easy** way to **impact** the **gender conversion** performance.
- Necessary to **improve** the spectral modeling of **high-pitched** signals.
- **Recent** conversion **strategies** outperform GMM ones.
- **Similarity** and **quality** are **not yet fully convincing**: incorporating voice quality and prosody features may help.
- The Voice Conversion **Challenge** appears to be a valuable platform for fair evaluation of VC systems.

Thank you.