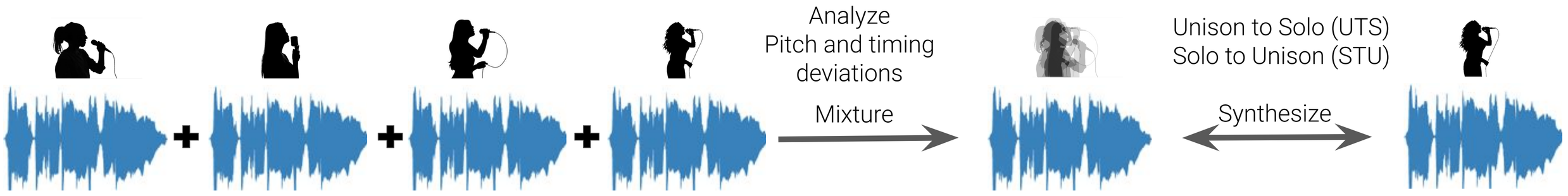
# A Deep Learning based Analysis-Synthesis Framework for Unison Singing

Pritish Chandna[1], Helena Cuesta[1], Emilia Gómez[2,1]

[1] Music Technology Group, Universitat Pompeu Fabra (Barcelona)
[2] Joint Research Centre, European Commission (Sevilla)

## 1 Motivation

Leverage recently developed Deep Learning technologies to analyse real world SATB choral unison singing and facilitate synthesis of Unison from A Capella Input and A Capella from Unison Input.




Multiple singers simultaneously singing the same content.
Natural pitch and timing deviations. Timbral ensemble.

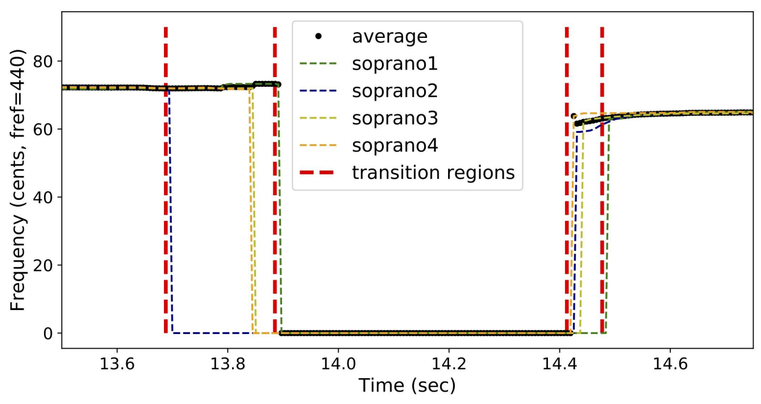
Unison Singing.
Single Perceived Pitch [1]

A Capella Solo Singing

## 2 Unison Singing Analysis

**Choral Singing Dataset (CSD) [2]**

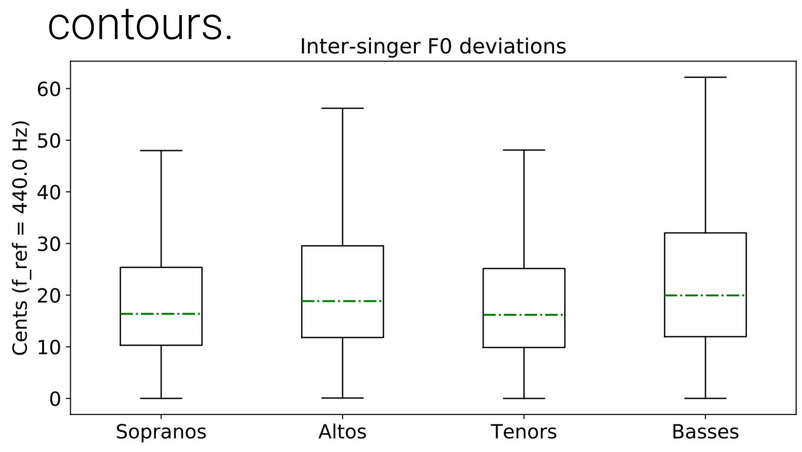
- 3 SATB choir songs
- 16 singers
- 4 singers per section
- Manually corrected F0 annotations



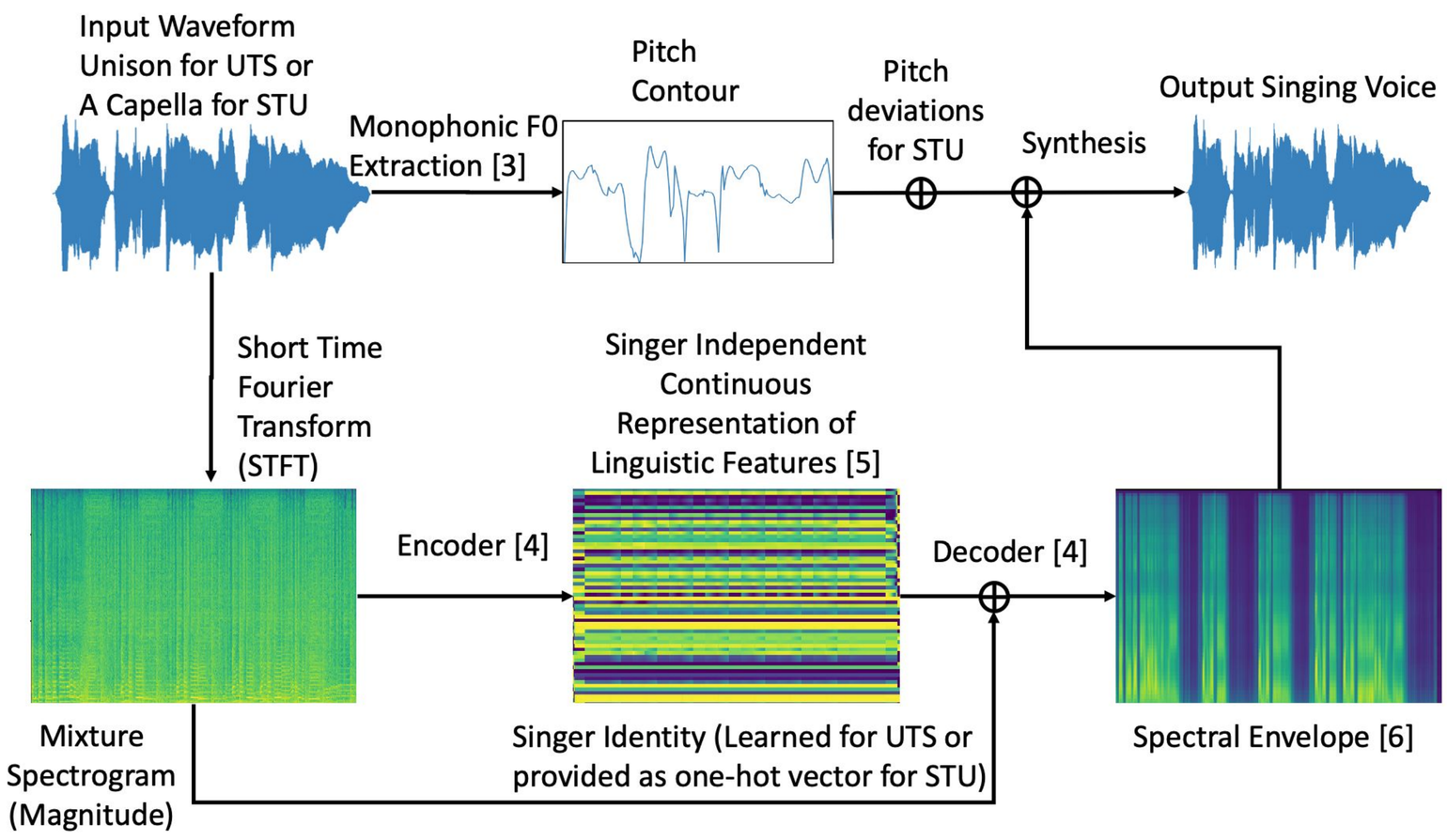

**Timing deviations** are computed at the *transition regions* (voiced ↔ unvoiced), where singers are not entirely in sync.

| Section | Average Timing Deviation ± Standard Deviation |
|---|---|
| Soprano | $0.134 \pm 0.039$ sec |
| Alto | $0.093 \pm 0.0024$ sec |
| Tenor | $0.100 \pm 0.021$ sec |
| Bass | $0.124 \pm 0.021$ sec |

**Inter-singer F0 deviations** (ΔF0s) computed for each pair of singers in the unison as the frame-wise difference between the two F0 contours.




## 3 Synthesis Methodology




- F0 extracted by monophonic F0 extractor [3] used for single pitch for UTS.
- Encoder-Decoder [4] trained on proprietary dataset, no overlap with CSD.
- Singer independent linguistic features as used in Voice Conversion algorithms [5].
- Gender specific timbre changes for STU.
- Pitch deviations sampled from normal distribution.
- Timing deviations implemented using circular shifts between regions of silence.
- WORLD [6] vocoder features used for synthesis.

## 4 Subjective Evaluation

| Test Case | Adherence To Melody | Unison Perception | Audio Quality |
|---|---|---|---|
| UTS | $3.6 \pm 0.93$ | | $2.1 \pm 0.65$ |
| STU_PS | $3.3 \pm 0.83$ | $2.6 \pm 0.85$ | $2.8 \pm 0.45$ |
| STU_PTS | $2.9 \pm 1.14$ | $3.2 \pm 0.96$ | $3.1 \pm 0.63$ |
| STU_TS | $2.3 \pm 1.11$ | | |
| STU_PT | $3.0 \pm 1.23$ | | |

- STU with Pitch (P), Timing (T) and Singer (S) variations.
- **Adherence to melody** shows **F0 extracted by CREPE** [3] can be viewed as a representation of **single perceived pitch of the unison**.
- **Timbre variations do not make significant improvement** to the perception.

## 5 References

[1] S. Ternström, **"Perceptual evaluations of voice scatter in unison choir sounds",** STL-Quarterly Progress and Status Report, vol. 32.

[2] H. Cuesta, et al. **"Analysis of intonation in unison choir singing"**, in Proceedings of the International Conference of Music Perception and Cognition (ICMPC), 2018.

[3] J. W. Kim, et al. **"CREPE: A Convolutional REpresentation for Pitch Estimation"**, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.

[4] P. Chandna, et al. **"Content based singing voice extraction from a musical mixture"**, in Proceedings of the 45th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2020.

[5] K. Qian, et al. **"Autovc: Zero-shot voice style transfer with only autoencoder loss"**, in International Conference on Machine Learning, 2019.

[6] M. Morise, et al. **"World: avocoder-based high-quality speech synthesis system for real-time applications"**, in IEICE TRANSACTIONS on Information and Systems, vol. 99, 2016.

Universitat Pompeu Fabra Barcelona
MTG Music Technology Group
TROMPA
European Commission