DOMAIN ADAPTATION TOWARDS SPEAKER-INDEPENDENT ULTRASOUND TONGUE IMAGING-BASED ARTICULATORY-TO-ACOUSTIC MAPPING

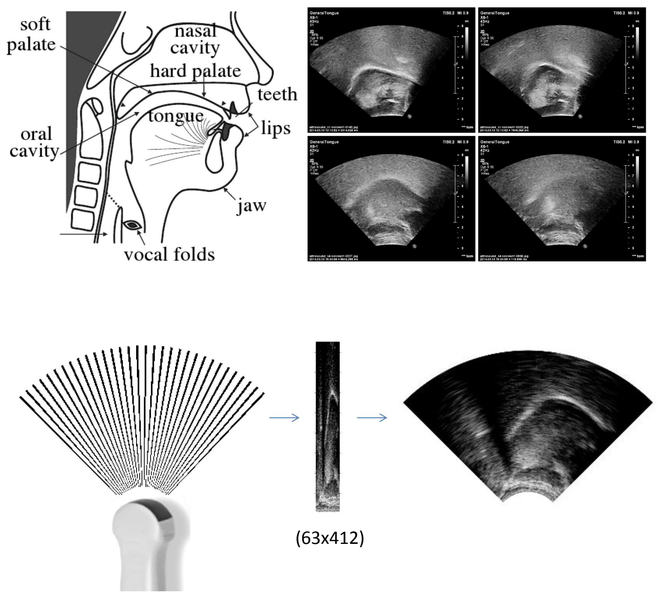
Kang You[1], Kele Xu[2], Ming Feng[1], Jilong Wang[1], Boqing Zhu[2], **Tamás Gábor Csapó**[3], Dawei Feng[2]
[1]Tongji University, College of Computer, Shanghai, China
[2]National University of Defense Technology, College of Computer, China
[3]Budapest University of Technology and Economics, Budapest, Hungary

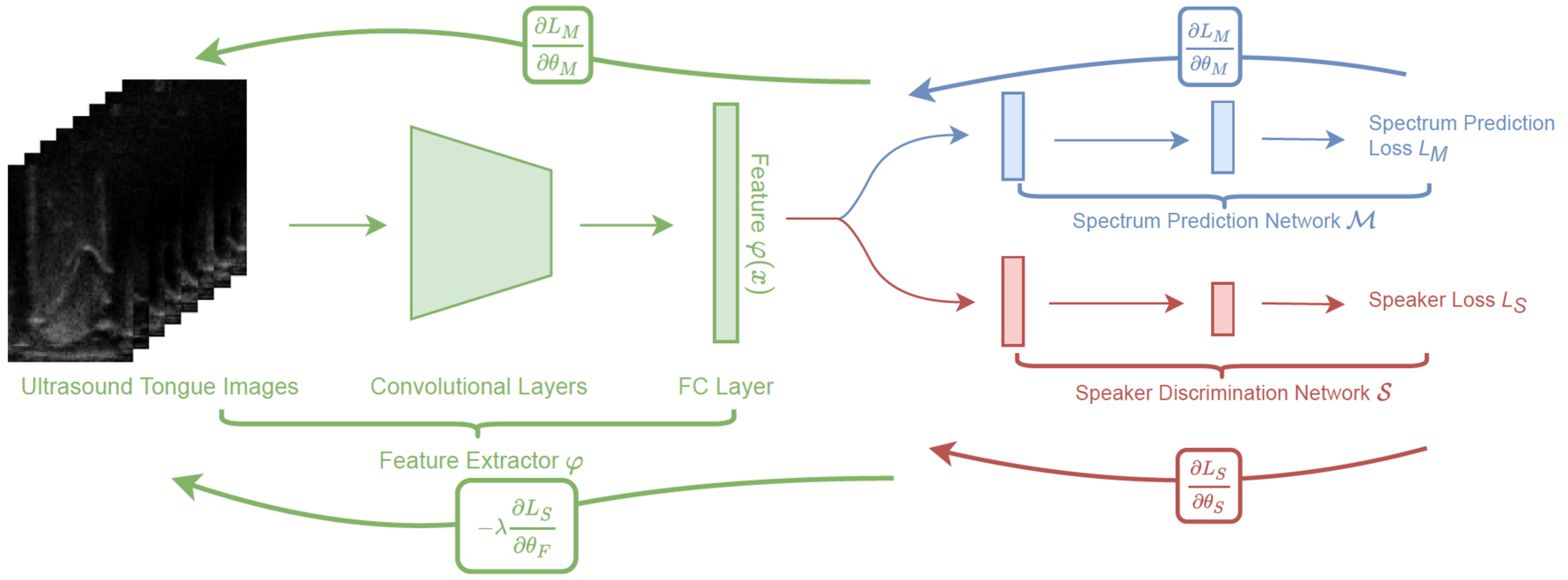
1. Introduction

- articulatory-to-acoustic mapping
- to predict the mel- spectrogram of the audio signals
- from midsagittal ultrasound tongue images (UTI) of the vocal tract
- target application: „Silent Speech Interface”
- unsolved problem before: across speakers

2. Major Contributions

- towards speaker-independent scenario
- domain-adaptation and adversarial learning
- https://github.com/xianyi11/Articulatory-to-Acoustic-with-Domain-Adaptation





3. Methodology

Methodology Overview

- two main parts:
  - left part is a convolutional neural network for the feature extraction from the UTI
  - right part is designed for the estimation task
  - we decouple the latter part into two branches
    - one branch for the regression/generation task
    - another for the speaker discrimination / classification
- joint training
  - novel-designed loss, adversarial learning
    - mel-spectrogram prediction loss (MSE)
    - speaker discrimination loss: shallow speaker discrimination network S which is parameterized by θ_S to recognize the speaker from the ultrasound image
- final loss: weighted average

Database

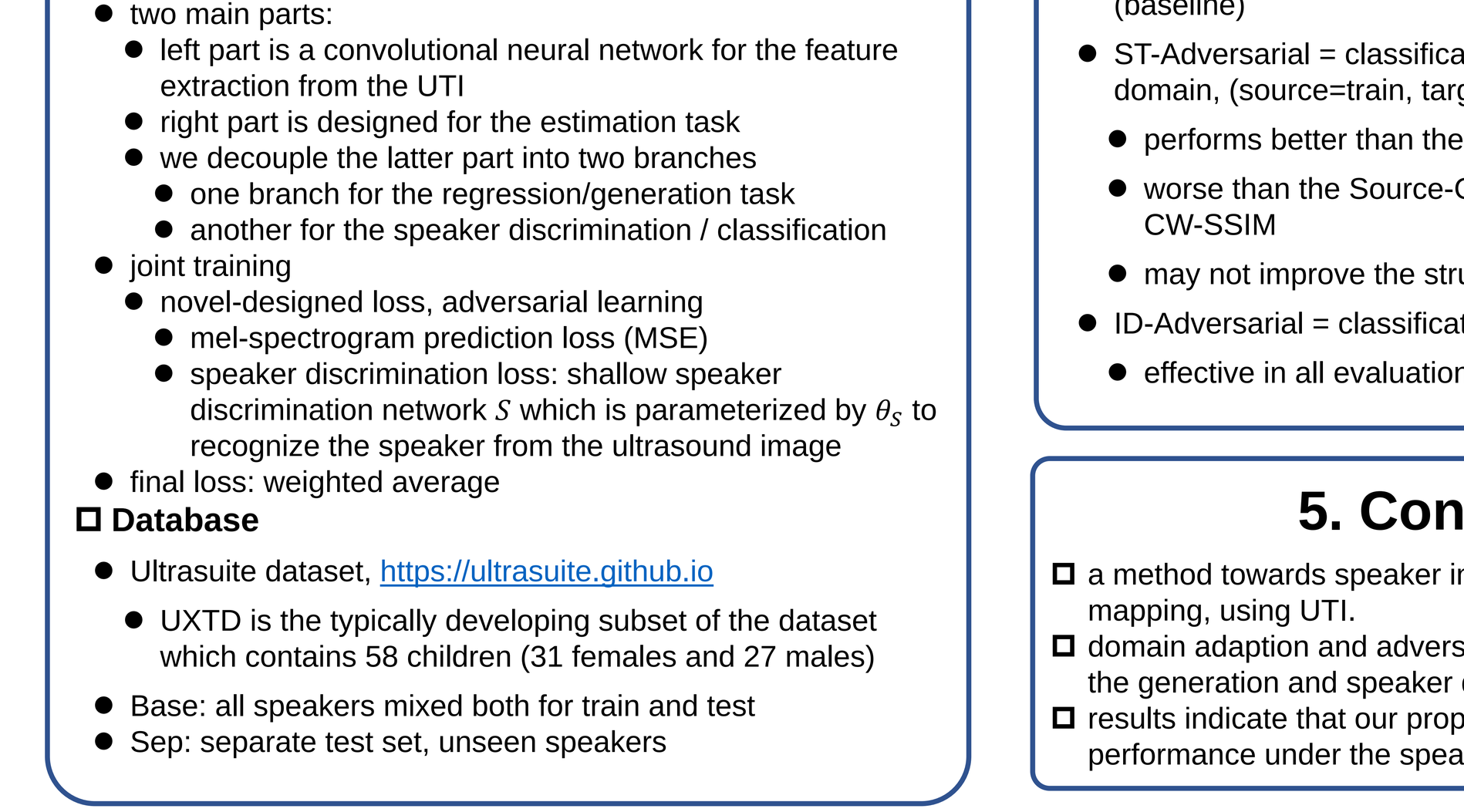
- Ultrasuite dataset, https://ultrasuite.github.io
  - UXTD is the typically developing subset of the dataset which contains 58 children (31 females and 27 males)
- Base: all speakers mixed both for train and test
- Sep: separate test set, unseen speakers

4. Results

- Source-Only = no classification task, speaker dependent (baseline)
- ST-Adversarial = classification task: binary, Source-Target domain, (source=train, target=test)
  - performs better than the Source-Only model with MSE
  - worse than the Source-Only model under the SSIM and CW-SSIM
  - may not improve the structural similarity
- ID-Adversarial = classification task: speaker ID (proposed)
  - effective in all evaluation metrics

5. Conclusions

- a method towards speaker independent articulatory-to-acoustic mapping, using UTI.
- domain adaption and adversarial method, which can decouple the generation and speaker discrimination task
- results indicate that our proposed method can achieve superior performance under the speaker-independent scenario
- future: subjective listening test of synthesized samples
- future: cross-language? (Lukose et al. 2023 ICPhS)
- future: pixel differences? (Giulia, Palo 2023 ICPhS)



		MSE	SSIM	CW-SSIM
Source-Only	Base	1.88	0.73	0.70
	Seq	1.88	0.74	0.73
ST-Adversarial	Base	1.84	0.71	0.68
	Seq	1.78	0.75	0.72
ID-Adversarial	Seq	**1.62**	**0.76**	**0.74**