WINS 2024


TMiT
SmartLab
Intelligent Interactions


Towards Cross-Speaker Articulation-to-Speech Synthesis using Dynamic Time Warping Alignment on Speech Signals

Ibrahim Ibrahimov, Gábor Gosztolya and Tamás Gábor Csapó

05.02.2024

Overview

01

Intro & Dataset

SSI; Ultrasound tongue imagining; UltraSuite-TaL

02

UTI alignment using DTW

DTW distance of speech signals as a tool for UTI alignment

03

Cross-speaker AAM

Articulatory-to-acoustic mapping in a cross-speaker manner

04

Results & Discussion

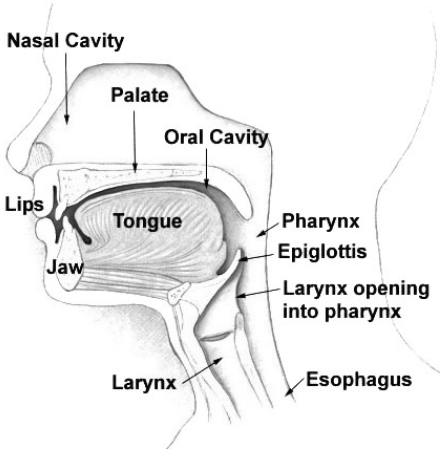
Results of cross-speaker AAM using aligned UTI with DTW distance between speech signals

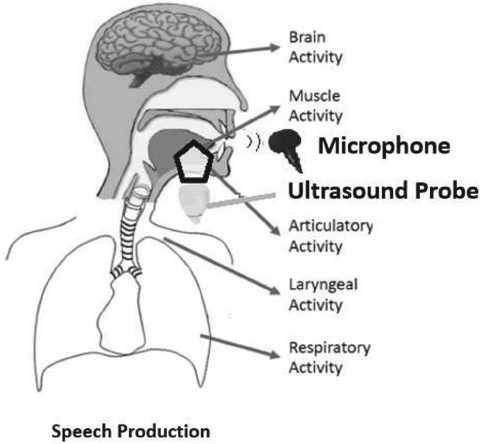
01.a

INTRODUCTION

Silent speech interfaces; articulation-to-speech synthesis; articulatory-to-acoustic mapping; ultrasound tongue imaging

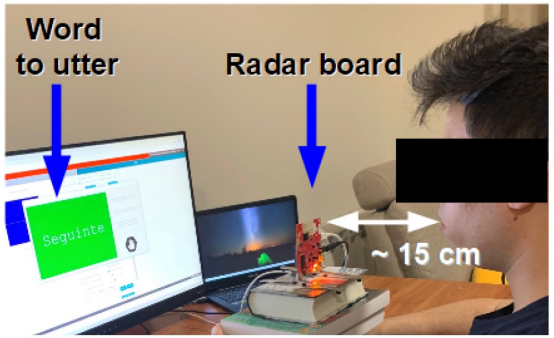
Silent Speech Interfaces (SSI)

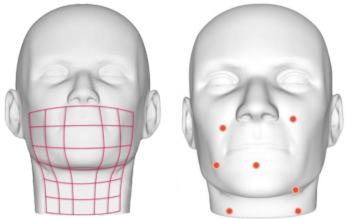
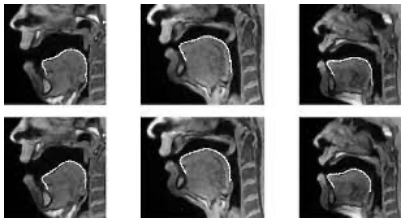
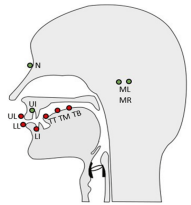


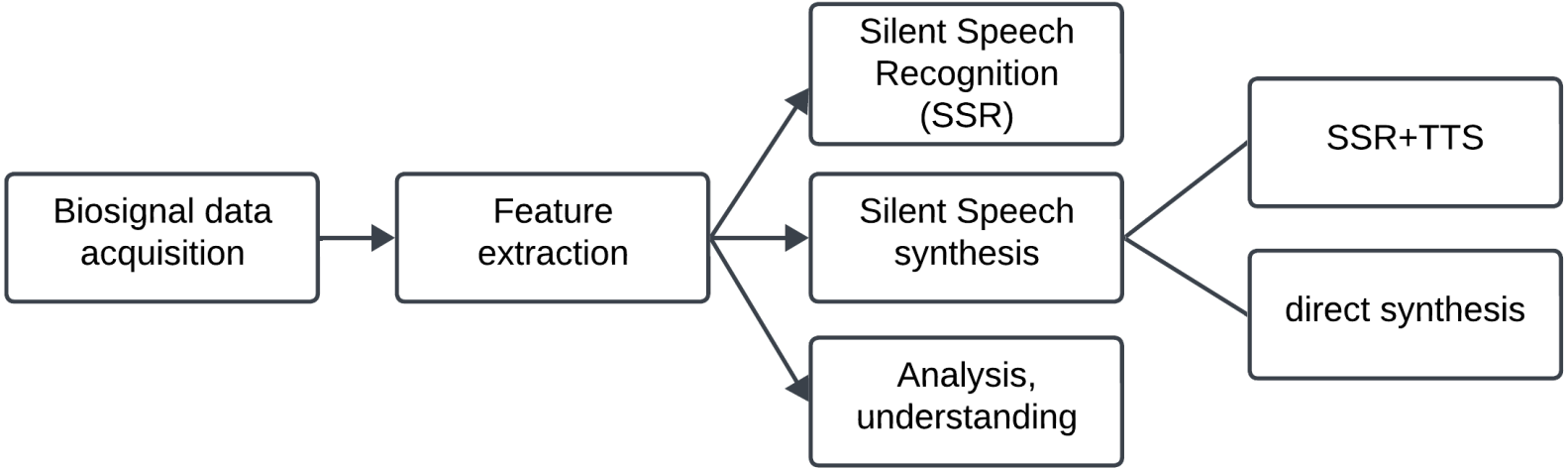


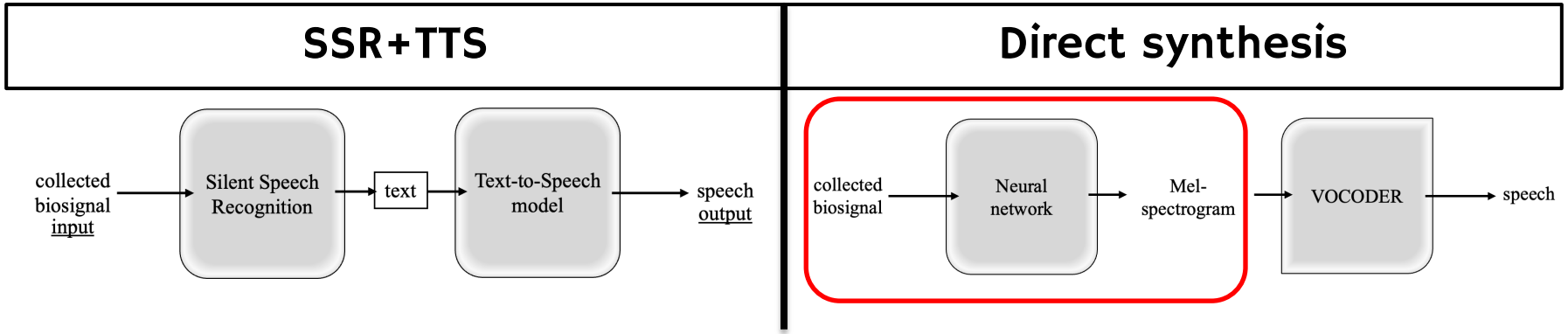
Speech Production

Continuous-Wave Radar

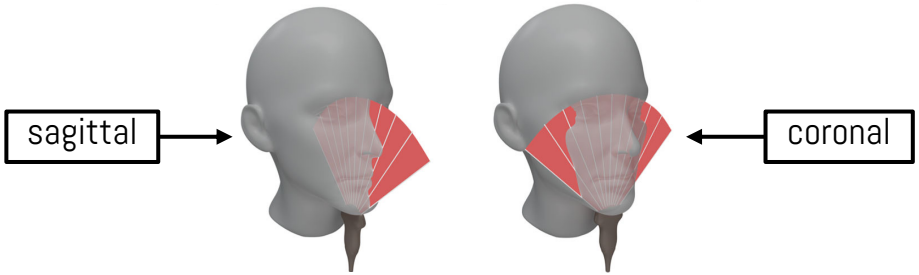


Lip video	surface Electromyagraphy (sEMA)	Real-time MRI	Electromagnetic Articulography (EMA)
			

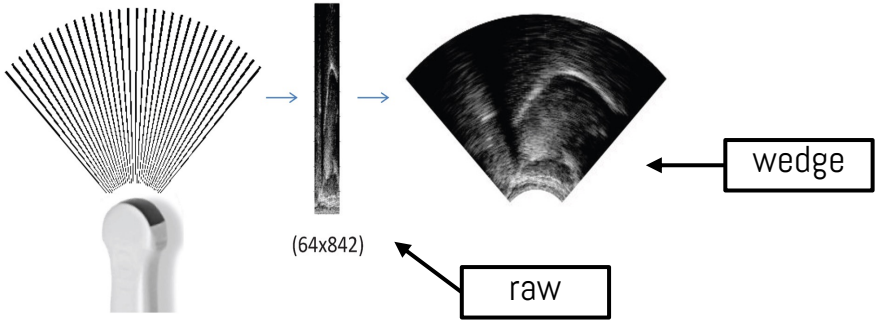

Biosignal data acquisition
Feature extraction
Silent Speech Recognition (SSR)
Silent Speech synthesis
Analysis, understanding
SSR+TTS
direct synthesis


SSR+TTS
Direct synthesis
collected biosignal input
Silent Speech Recognition
text
Text-to-Speech model
speech output
collected biosignal
Neural network
Mel-spectrogram
VOCODER
speech

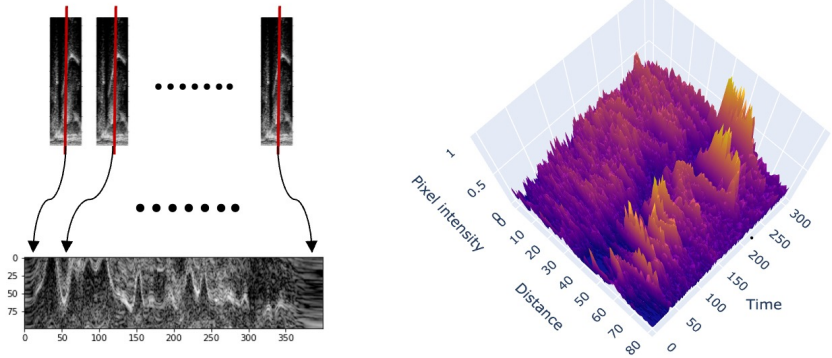
Ultrasound Tongue Imaging (UTI) & Articulatory-to-acoustic Mapping (AAM)



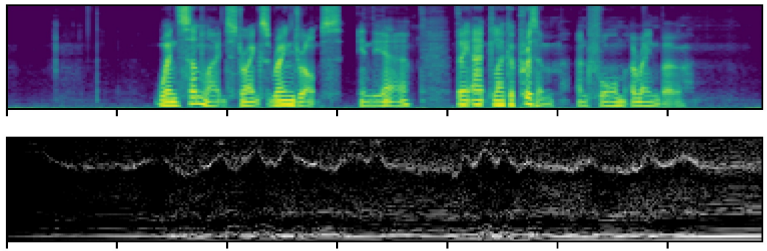
a) Ultrasound probe placements



b) Raw vs wedge representation



c) Ultrasound kymogram visualisation



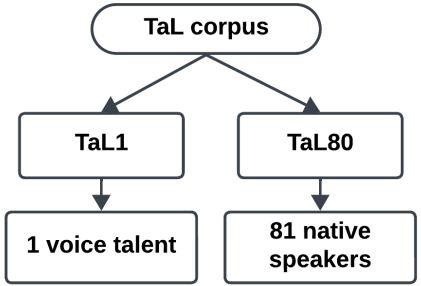
d) Mapping raw ultrasound images to mel-spectral representation of acoustics

01_b

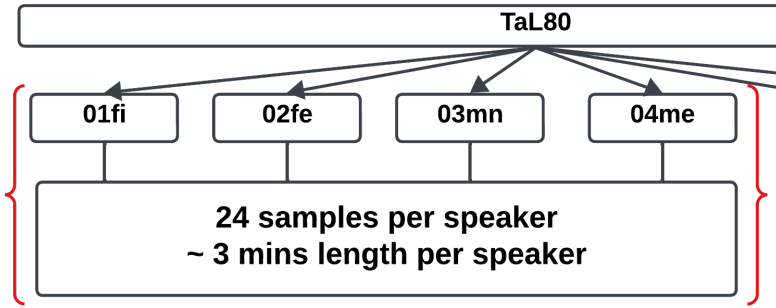
DATASET

UltraSuite-TaL80 – ultrasound tongue and lip video recordings

TaL (Tongue and Lips) corpus



Prompt type	Tag
Read	aud
Silent	sil
Whispered	whi
Read (shared)	xaud
Silent (shared)	xsil
Whispered (shared)	xwhi
Spontaneous	spo
Calibration	cal



More about the dataset ->

Training – Validation
21 - 3

"015" - short - "Other men have tried to explain the phenomenon physically."

"006" - mid-length - "These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon."

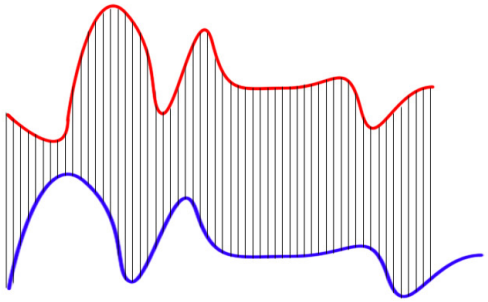
"021" - long sentence - "If the red of the second bow falls upon the green of the first, the result is to give a bow with an abnormally wide yellow band, since red and green light when mixed form yellow."

02

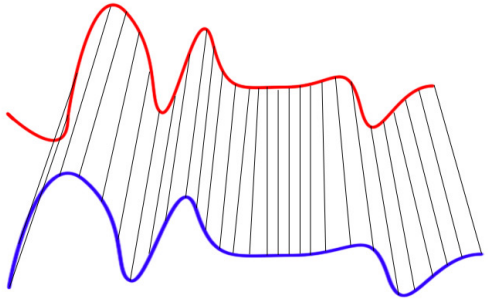
UTI ALIGNMENT USING DTW

Dynamic Time Warping application on speech signals to align ultrasound tongue images

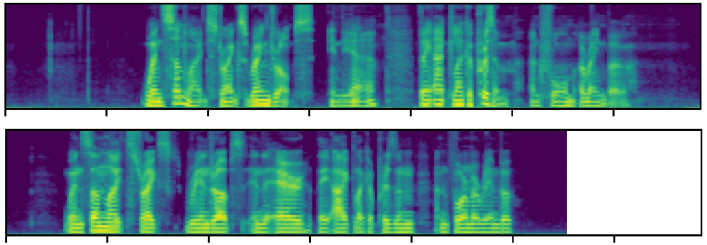
Dynamic Time Warping on Speech Signals



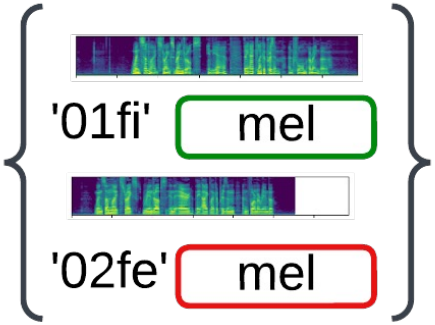
Euclidean Matching



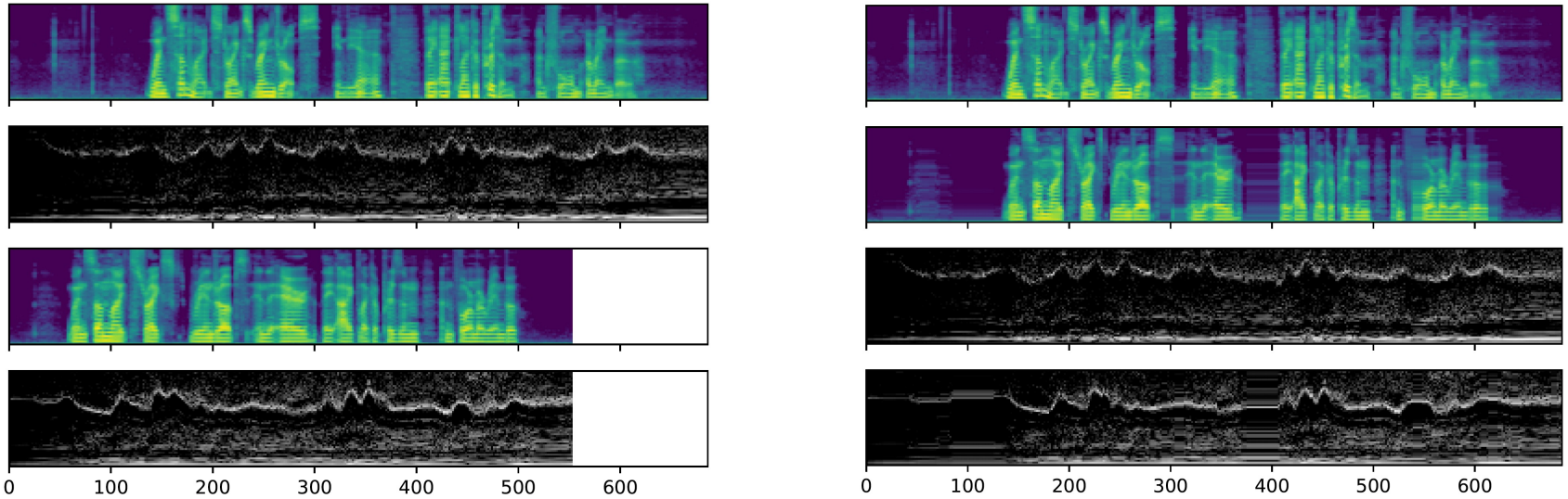
Dynamic Time Warping Matching



APPLY DTW



UTI Alignment using DTW distance

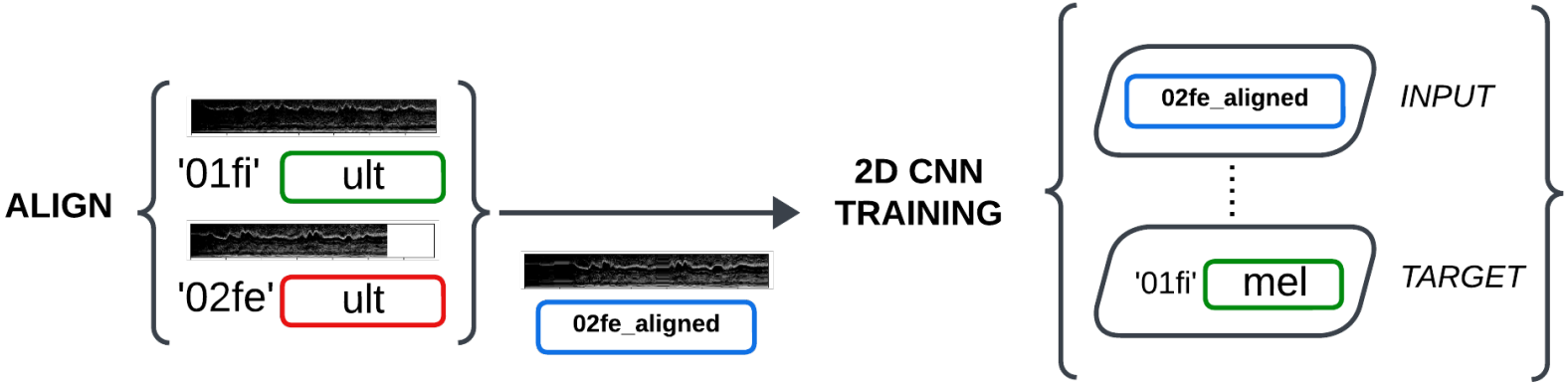


03

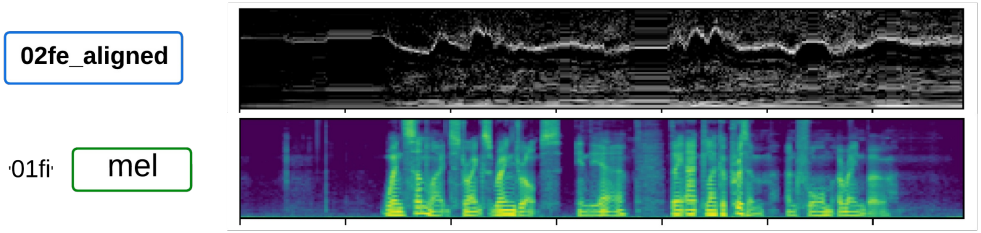
AAM in a cross-speaker manner

Articulatory-to-acoustic mapping using different speaker's input and output values

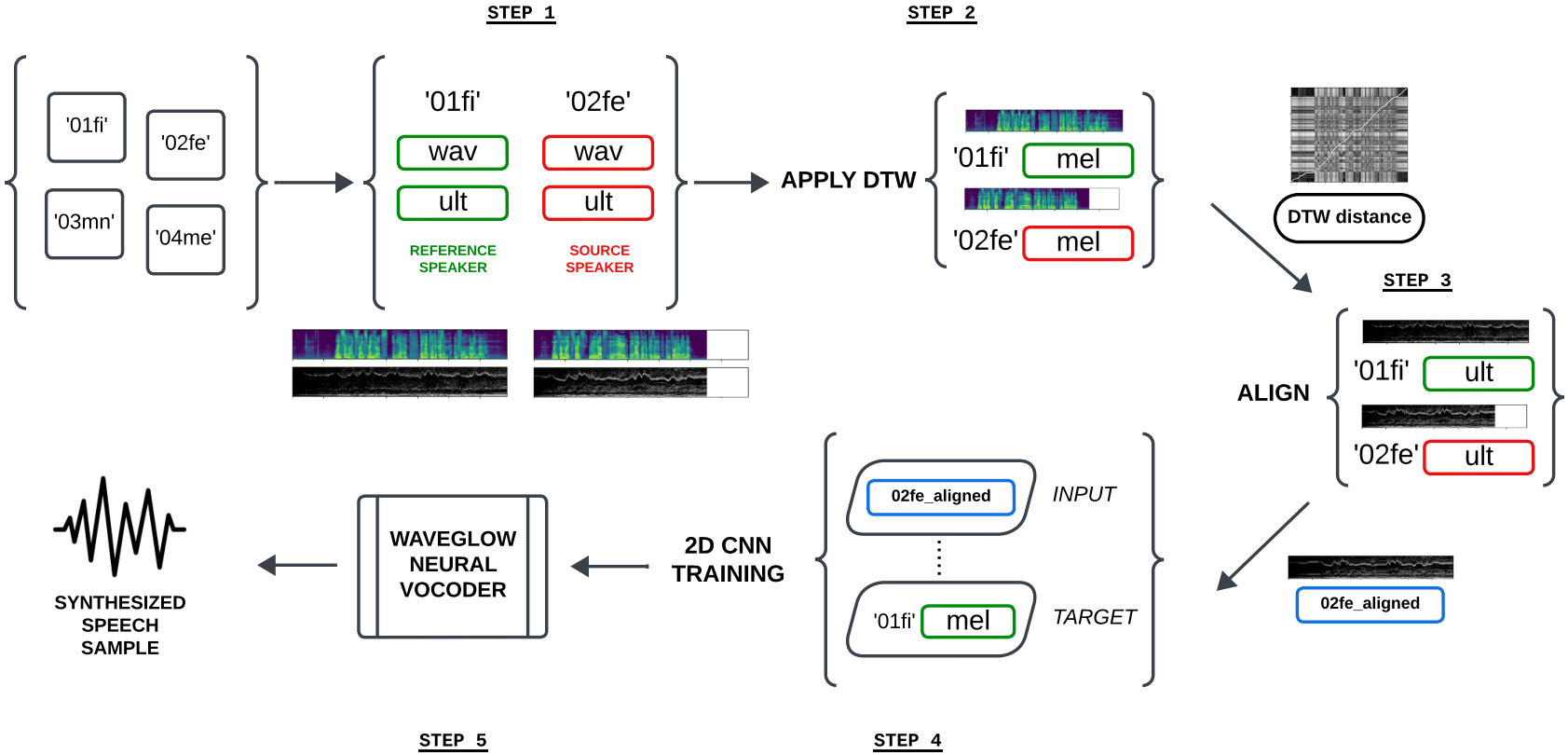
Cross-speaker AAM



Cross speaker articulatory-to-acoustic mapping using 2D CNN



General view to the process followed in this work

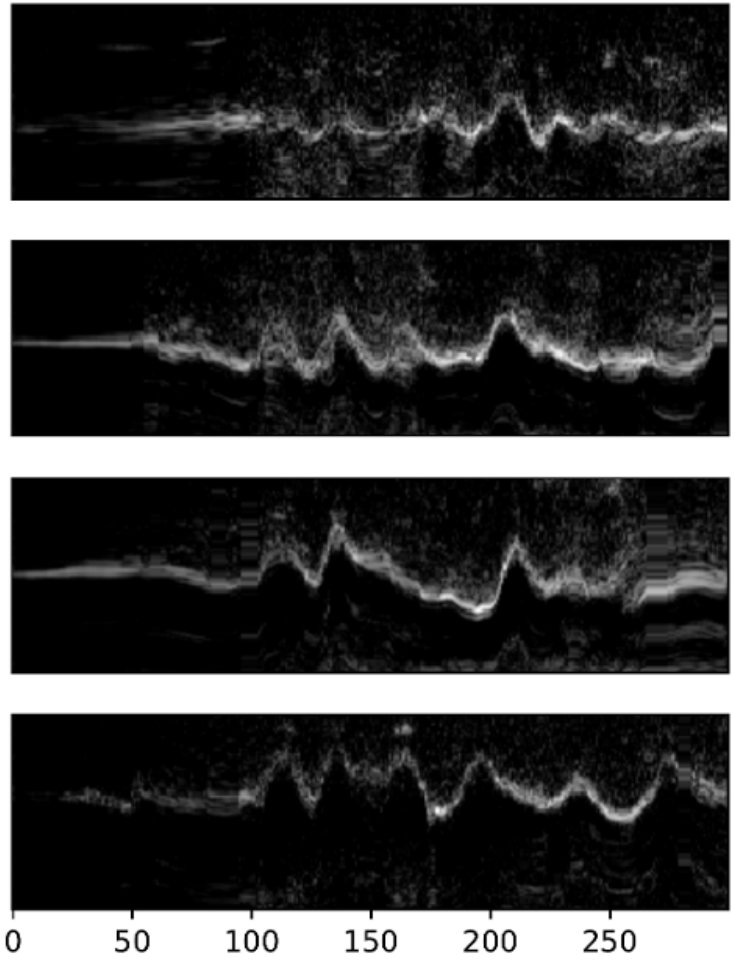


Results and discussion

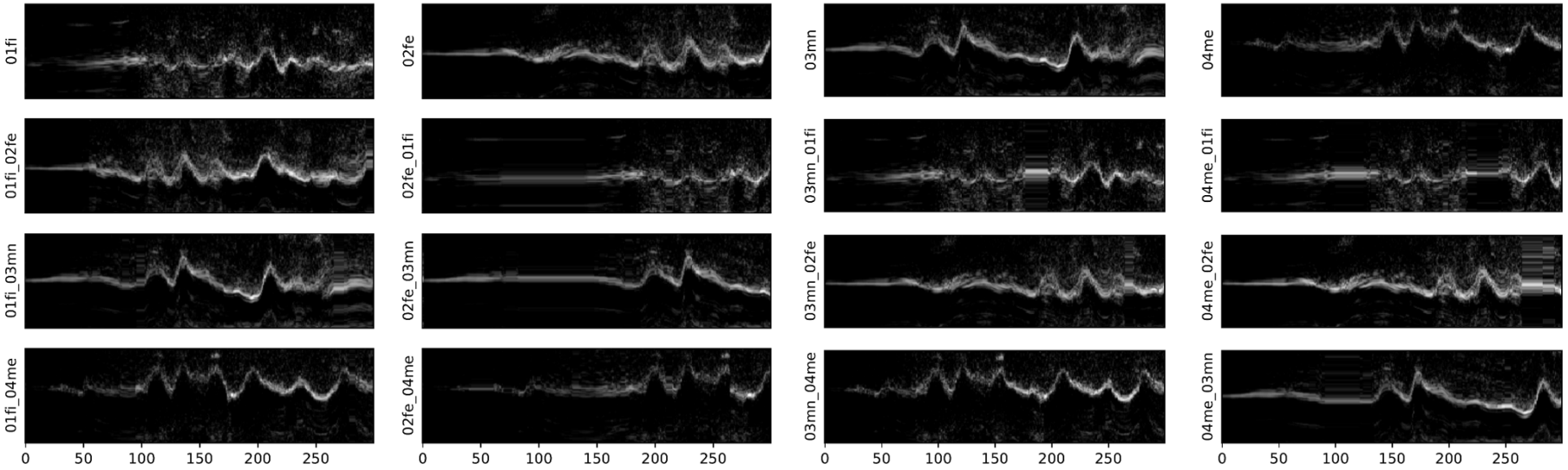
04

“A picture is worth a thousand words”

—Edward Tufte


0
50
100
150
200
250

Kymograms of cross-speaker UTI alignments



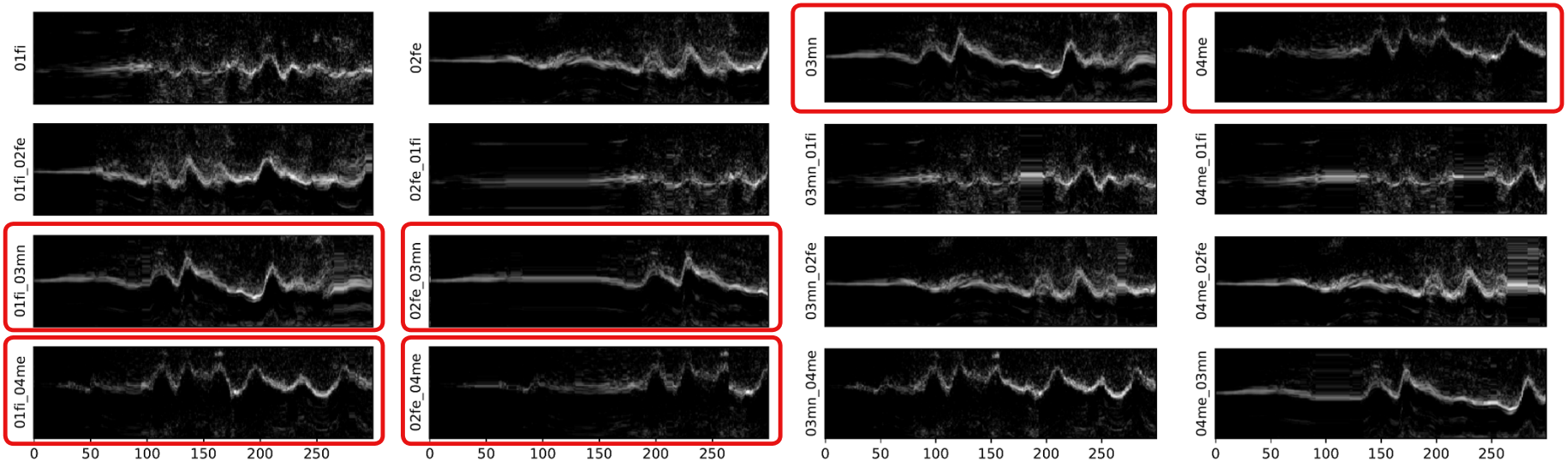
Mean Squared Error values of 2D CNN training

		Reference speaker			
		01fi	02fe	03mn	04me
	Baseline	0.531	0.547	**0.388**	**0.479**
Source spk.	01fi	—	0.655	0.787	0.716
	02fe	0.611	—	0.767	0.754
	03mn	**0.511**	**0.528**	—	0.713
	04me	**0.502**	**0.485**	0.515	—

Mel-cepstral Distortion values of WaveGlow synthesis

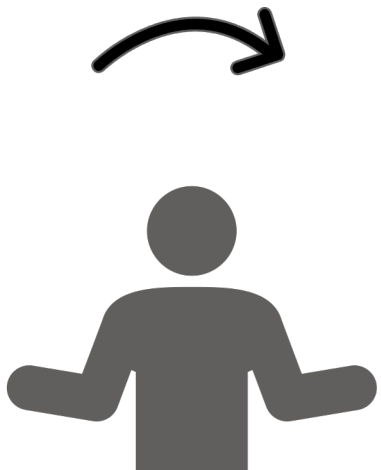
		Reference speaker			
		01fi	02fe	03mn	04me
	Baseline	10.934	11.316	**10.267**	**9.881**
Source spk.	01fi	—	12.726	14.726	12.466
	02fe	11.791	—	14.632	13.304
	03mn	**10.455**	**10.960**	—	12.240
	04me	**9.948**	**10.538**	11.196	—

Kymograms of cross-speaker UTI alignments



THANK YOU FOR YOUR ATTENTION!

New to academia!
Let's connect and discuss ideas!



05.02.2024