Is Dynamic Time Warping of speech signals suitable for articulatory signal comparison using ultrasound tongue images?

Tamás Gábor Csapó, csapot@tmit.bme.hu

Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Hungary

Workshop on Intelligent Infocommunication Networks, Systems and Services (WI^{2}NS^2), Feb 7, 2023

1. Introduction

Ultrasound Tongue Imaging (UTI)

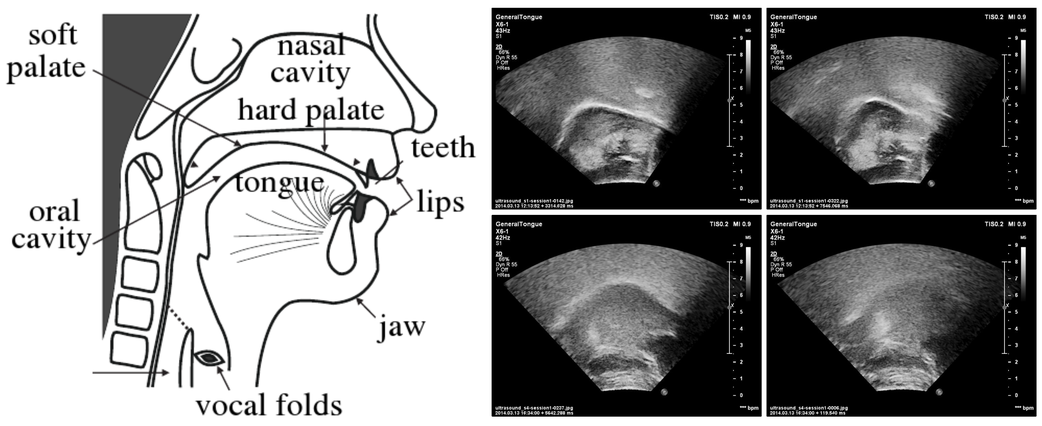

Fig. 1: Vocal tract (left) and sample ultrasound images (right) with the same orientation.

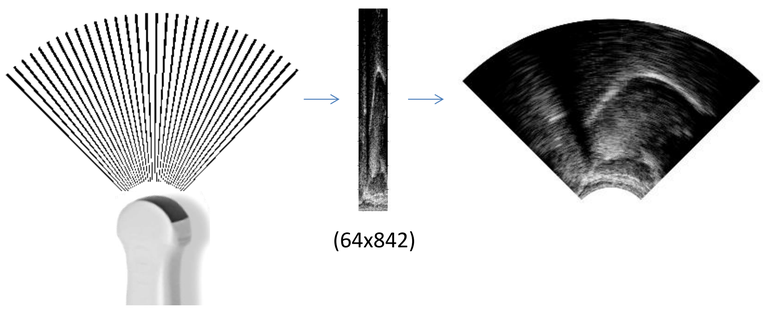

Fig. 2: Ultrasond tongue image representations: raw scanlines / array of raw scanline data / a wedge-formatted image.

Speaker dependence

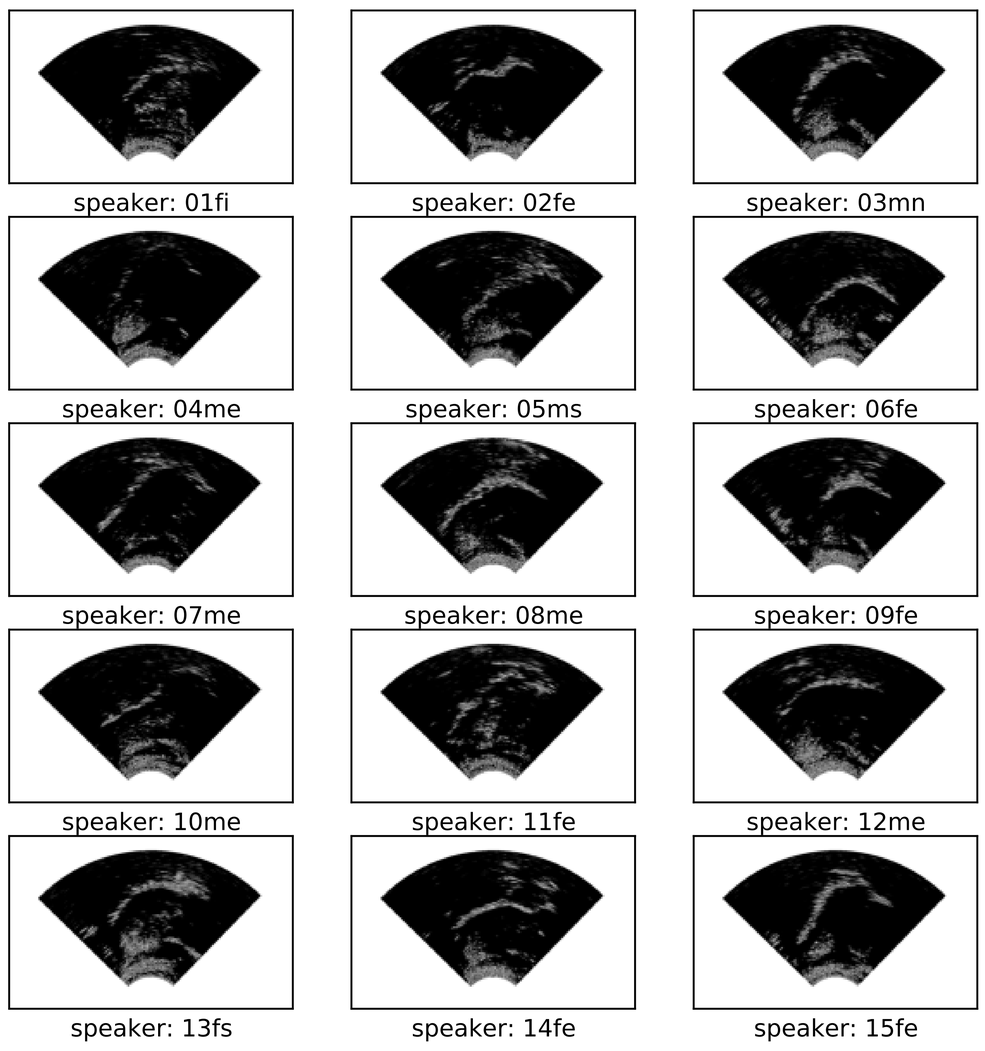

Fig. 3: Examples of the differences in the quality of ultrasound tongue images between speakers from UltraSuite-Tal80.

Dynamic Time Warping (DTW)

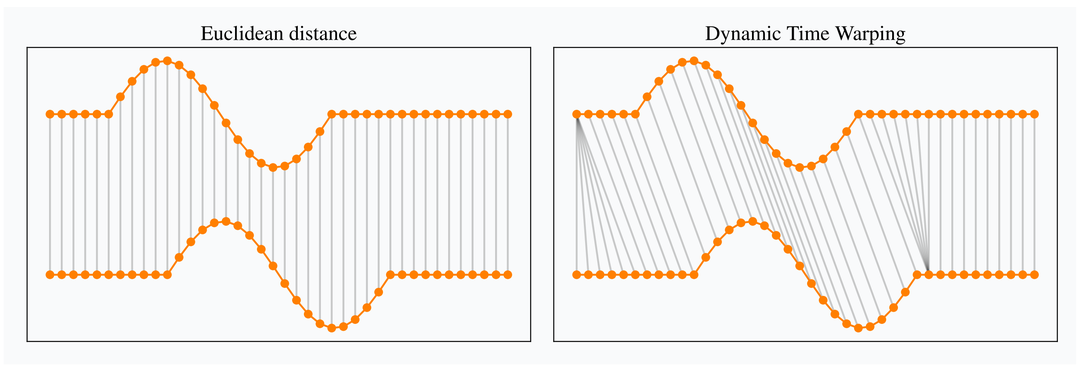

Fig. 4: Comparison between DTW and Euclidean distance. (from `https://rtavenar.github.io/blog/dtw.html`)

3. Experiments and results

Goals of the current study

- analyze the **speaker dependency** of articulatory movement using ultrasound tongue imaging, for future machine learning purposes
- investigate the **applicability of dynamic time warping** for comparing multiple speakers' articulatory on Hungarian and English datasets

DTW using UTI, demonstration samples

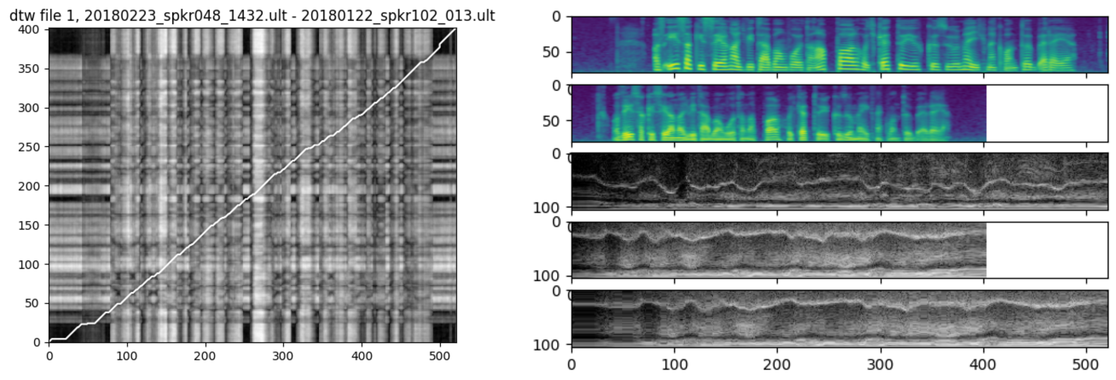

Fig. 5: Left: DTW sample based on the same sentence (*„Az északi szél nagy vitában volt a Nappal, hogy kettőjük közül melyiknek van több ereje."*) by two Hungarian speakers, calculated from speech MFCC. Right: speech spectrogram and temporal change of the midline of the ultrasound tongue images.

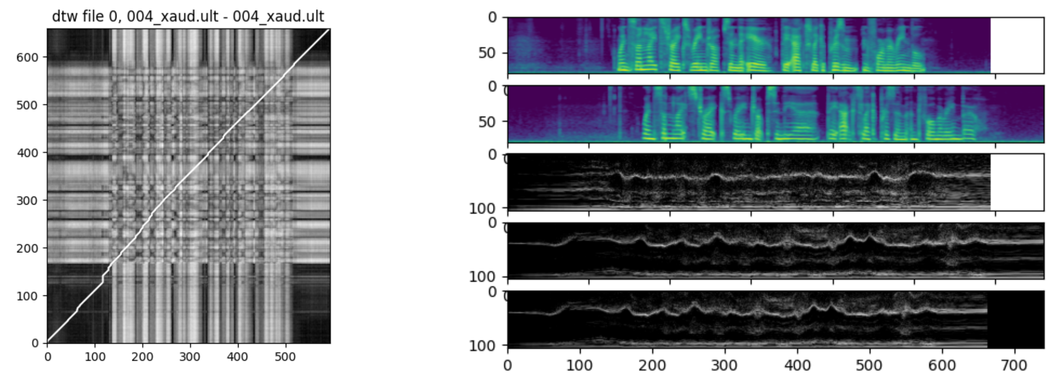

Fig. 6: Left: DTW sample based on the same sentence (*„When sunlight strikes raindrops in the air, they act like a prism and form a rainbow."*) by two English speakers, calculated from speech MFCC. Right: speech spectrogram and temporal change of the midline of the ultrasound tongue images.

4. Discussion and Conclusions

Answer to the title

- **YES** [1], but this was just a feasibility study. *„Although the manuscript contains no experimental results (I mean, not even a single one), this is surprisingly fine..."*

Future work

- objective quantification of location of articulatory inflection points
- align the audio recordings along the resulted DTW path to examine the acoustic difference
- follow-up: Interspeech [2], Special session, *Neural Processing of Speech and Language*
- **happy to discuss to hear your thoughts!**
- planned application for speech-based brain-computer interfaces to supplement the brain signal (measured with EEG, ECoG or sEEG) with ultrasound tongue image based articulatory information [3, 4, 5] (+ Momentum grant, `https://neurart.tmit.bme.hu`)

References

[1] T. G. Csapó, "Is Dynamic Time Warping of speech signals suitable for articulatory signal comparison using ultrasound tongue images?" in *WINS 2023*, 2023.

[2] ——, "Cross-speaker speech articulatory movement comparison with Ultrasound Tongue Image and Dynamic Time Warping," in *submitted to Interspeech*, 2023.

[3] T. G. Csapó, F. V. Arthur, P. Nagy, and Á. Boncz, "A beszéd artikulációs mozgásának predikciója agyi jel alapján - kezdeti eredmények," in *MSZNY 2023*, 2023.

[4] ——, "Towards Ultrasound Tongue Image prediction from EEG during speech production," in *submitted to Interspeech*, 2023.

[5] T. G. Csapó et al, "OTKA FK-22, Analysis of articulation and brain signals for speech-based brain-computer interfaces," 2022. [Online]. Available: http://nyilvanos.otka-palyazat.hu/index.php?menuid=930&lang=EN&num=142163