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

Tamás Gábor Csapó
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Budapest, Hungary
csapot@tmit.bme.hu

***Abstract*—In speech technology, the examination of speaker dependency is vital – that is, whether methods developed for one speaker can be adapted to another speaker or not. In the case of text-to-speech synthesis, well-usable speaker adaptation methods are already available, but they cannot be used directly for articulatory data (movement of the tongue, lips, etc, during speech production). In this research, we investigate the above question and analyze the speaker dependency of the articulatory movement, using audio signal and ultrasound tongue imaging (UTI) recorded in parallel during speech production. For the comparison, we use the well-known Dynamic Time Warping (DTW) procedure of speech technology. DTW of the speech signal has already been successfully applied 1) with UTI, for within-speaker comparisons, 2) with electromagnetic articulography (EMA), for the analysis of inter-speaker differences, 3) with EMA and electrocortocography (ECoG), also for inter-speaker comparisons. However, there has been no previous research yet on the application of DTW on speech signals with ultrasound tongue images for different speakers. In the present research, we examine the applicability of DTW for comparing speakers' speech and articulatory data on a few Hungarian and English examples, and visually analyze them. In the long term, we plan to use the results for speech-based brain-computer interfaces, so that we can supplement the brain signal with ultrasound-based articulation information.**

***Index Terms*—speech technology, dynamic time warping, signal processing**

I. Introduction

Research on human-computer interaction is important in the information society. Speech technology research fits into this process – speech is one of the most complex human biological signals, but we do not yet understand all the aspects of speech production and articulation. Digital applications using speech technology could significantly help the everyday communication of speech impaired people.

A. The relationship between articulatory movement and speech signal

The relationship between articulation (the coordinated movement of the speech production organs) and acoustics (the speech signal itself) has been a topic of interest to speech researchers since the 1700s [1]. In order to study the movement of speech organs (e.g., vocal fold, tongue, lips), special instruments are needed, as most of these organs are not visible continuously during speech. Of the articulatory organs, the tongue is a relatively large and important organ, but measuring and quantifying its function is challenging, in part because it is located within the oral cavity [2].

The relationship between the articulatory movement and the speech signal can be studied in many ways; one example is Articulatory-to-Acoustic Mapping (AAM); also known as Silent Speech Interface (SSI) [3]. SSI systems represent a revolutionary direction in speech technology, where silent articulatory movements are captured by some device and from this speech is automatically generated while the original speaker does not make a sound [3]. In most previous research on SSI, only a few speakers have been studied [3]–[9]. Although the results of these studies are encouraging, further research is needed to develop session- and speaker-independent SSI systems [10].

B. Ultrasound tongue imaging

Ultrasound has been used for speech research and articulation since the early 1980s [11]. Depending on the position (orientation) in which the transducer is placed under the jaw, the tongue can be examined from several different orientations; of which the most common is the mid-sagittal orientation. In mid-sagittal imaging, the transducer is placed under the chin; thus, the greatest change in the ultrasound signal is caused by the upper surface of the tongue muscles, ideally resulting in a clearly visible white line on the ultrasound images. To prevent the ultrasound transducer from moving during speech, some form of probe fixing helmet is typically used.

The advantage of ultrasound over other articulatory recording techniques is that it is easy to use, non-invasive, affordable, and can be used to record at a relatively high resolution (up to 800 x 600 pixels) and high speed (up to 100–150 frames per second) [12]. A good spatial resolution is important to obtain a more accurate picture of the shape of the tongue; while a good temporal resolution is necessary to study rapid changes during the production of speech sounds (e.g. stop bursts; coarticulation). To some extent, the use of ultrasound has the disadvantage that it only provides information from the middle part of the tongue; often the root and/or tip of the tongue is not visible. In addition, if the surface of the tongue

is nearly parallel to the ultrasound beam, information about the middle part may be incomplete.

C. Speaker dependence of the articulatory movement and machine learning models

In speech technology, it is vital to examine speaker dependency – that is, whether or not methods developed for one speaker can be adapted to another speaker. For example in text-to-speech synthesis, there are already good speaker adaptation methods available [13], [14], but they cannot be used directly with articulatory data.

The image quality of the ultrasound tongue images may vary between speakers. The quality of the images is influenced by many factors, such as the anatomy of the speaker or the condition of the tissues of the articulatory organs (e.g., hydration). The variation between speakers may also be due to the fact that the ultrasound transducer is positioned differently (in different orientations) for different head sizes. The recording software usually provides the possibility to adjust the ultrasound hardware parameters (e.g., transducer frequency, field of view, depth, dynamic range, line density, etc.), but this may not be a sufficient solution for all speakers.

Due to differences in the size and shape of the speakers' heads, the ultrasound probe cannot be positioned identically for different speakers, so the possibility of comparing speakers is limited due to potentially different orientations. In addition, recordings of the same speaker taken in different sessions may not be comparable, as it is not possible to adjust the angle of the ultrasound head exactly the same for each occasion. Because of the above speaker and session dependencies, there is no good method yet in either linguistic or technological research for analyzing different UTI recordings together, so typically recordings with different speakers are analyzed separately. Also, when machine learning (e.g., deep neural networks, DNNs) are applied, they are extremely sensitive to the type of input or target data, and therefore, current experiments are typically done in a way that DNNs are trained separately for each individual speaker [15]–[17].

Most previous studies in this area have used point-tracking tools such as electromagnetic articulography (EMA) [18]. For example, for speaker adaptation, the articulation data of different speakers is investigated by combining a 'Procrustes Matching' procedure with voice conversion methods [10]. Since medical imaging methods and point tracking tools produce a very different type of signal, the above speaker adaptation results cannot be used directly for ultrasound tongue imaging.

D. Dynamic time warping estimated on speech, used for articulatory and brain signal analysis

Dynamic Time Warping (DTW) is a long-established method for comparing speech samples of different lengths [19]. However, it has only been used sparsely in the context of articulatory data [20]–[22]. In the work of Yang and colleagues [20], the aim of DTW was to produce a 3D reconstruction of the tongue from 2D ultrasound slices. To do this, the same speakers repeatedly read out given sentences while recording their articulatory movements in different orientations with the 2D ultrasound. Using DTW on the speech signal, they merged recordings from the same speaker and were able to produce 3D visualizations of the positions and movements of the tongue during speech [20]. Jayanthi and his colleagues further developed the classical DTW with a 'divide-and-warp' strategy for speaker-invariant articulatory investigations [21]. Point-tracking EMA is used as articulatory signal, while analyzing data from the MOCHA-TIMIT database [23]. DTW is computed between speech samples and then the alignment of articulatory landmark points (e.g., opening/closing of the lips) is evaluated on the EMA data. The results measured on four speakers showed that the addition of 'divide-and-discard' reduced the unwanted shift of peaks in the articulatory data from 121 ms to 34 ms on average [21]. The aim of Le Godais [22] was to analyze articulatory information in the brain signal (acquired with ECoG) and speech. Since articulatory movements were not recorded in parallel with the brain signal, he estimated indirect articulatory data from other speakers using the BY2014 database, which also used EMA [24]. Since the same sentences but different speakers were used to record the brain signal and the articulatory signals, it is possible to compute EMA-based indirect articulatory information to supplement the ECoG data based on the DTW computed on the speech signal. Thus, ECoG-based speech synthesis has been successfully extended with derived articulatory information; although the synthesized speech samples are not yet intelligible [22].

As mentioned above, DTW of the speech signal has been successfully applied to 1) UTI and intra-speaker comparisons, 2) EMA for analyzing inter-speaker differences, 3) EMA and ECoG, also for inter-speaker comparisons. However, there has been no previous research on the application of DTW for cross-speaker ultrasound tongue image analysis.

E. Goal of the current study

In this research, we will investigate the above questions and analyze the speaker dependency of articulatory movement using ultrasound tongue imaging. For the comparison, we use the dynamic time warping procedure. We investigate the applicability of DTW for comparing multiple speakers' articulatory on Hungarian and English datasets.

II. Methods

Existing databases were used to investigate the ultrasound tongue images; recordings were selected from native speakers of Hungarian and English.

A. Hungarian recordings

The Hungarian recordings were created for the previous research on articulatory-to-acoustic mapping [15]. The mid-sagittal movement of the tongue was recorded using the 'Micro' system (AAA v220.02 software, Articulate Instruments Ltd.) with a 2–4 MHz, 64-element, 20-mm radius convex ultrasound transducer at 81.67 fps, and a probe fixing metal headset

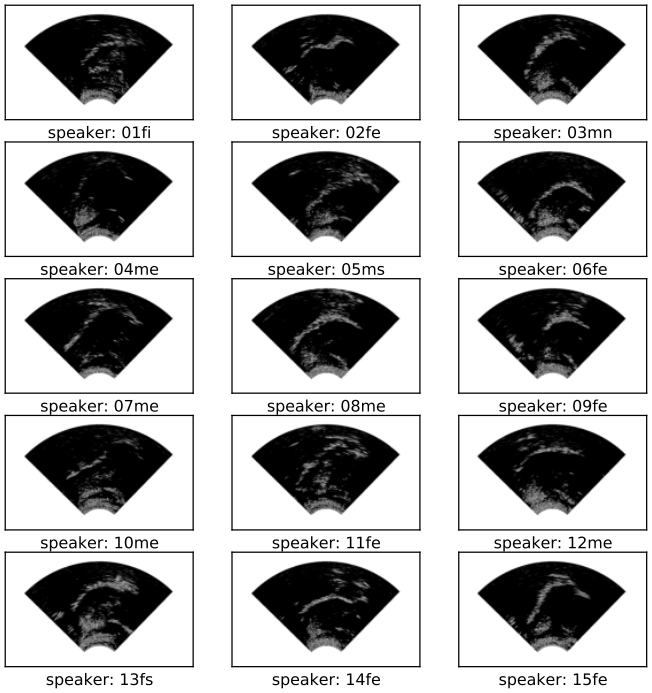

Fig. 1: Examples of the differences in the quality of ultrasound tongue images between speakers from the UltraSuite-Tal80 database.

was also used. Speech was recorded with a Beyerdynamic TG H56c tan omnidirectional condenser microphone clipped to the helmet, 20 cm from the mouth. The sound was digitized at a sampling rate of 22 050 or 44 100 Hz using an M-Audio - MTRACK PLUS sound card. Synchronization of ultrasound data and speech signals was performed using a tool provided by Articulate Instruments Ltd.

B. English recordings

For English data, the UltraSuite-TaL80 database was used [25], downloaded from https://ultrasuite.github.io/data/tal_corpus/. The recordings were made with the same 'Micro' system as for the Hungarian data. Lip video was also recorded in the UltraSuite-TaL80, but this information was not used in the current study.

An example for the cross-speaker differences in the ultrasound tongue image recordings is shown in Fig. 1, presenting ultrasound images of 15 speakers. It can be clearly seen that ultrasound can visualize different sections of the tongue (e.g. '09fe' has a shorter tongue, while '06fe' has a longer tongue), and also different visibility of the tongue contour (e.g. '01fi' has a blurred image, but '02fe' has a clear upper surface of the tongue).

C. Ultrasound tongue image representations

In our experiments we used articulation features calculated from the 'raw' ultrasound data. The 'raw' data means that the intensity information from the ultrasound device was saved directly in binary format (so no data was lost during the image

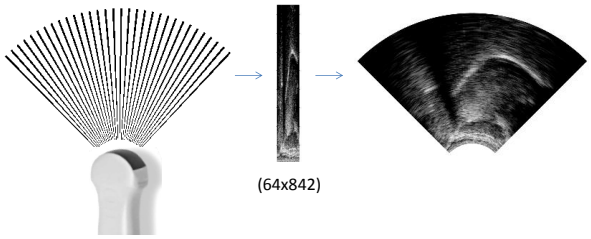

Fig. 2: Ultrasond tongue image representations: raw scanlines during recording (left), array of raw scanline data (middle), and a wedge-formatted image (right).

conversion) and processed as such. Fig. 2 shows how the scan is performed with the "Micro" system: the ultrasound transducer measures intensity (i.e. grayscale) on 64 radial lines, with 842 locations on each line, and stores each intensity value in the raw data in 8 bits. If this is to be converted to the usual ultrasound image, the data can be represented as a grayscale image in a polar coordinate system.

D. Preprocessing the articulation data

The ultrasound tongue images were used as 8-bit grayscale pixels in the raw ultrasound form of the "Micro" system. The images, originally 64x842 pixels, were resized to 64x128 pixels as this does not cause significant loss of information [9]. The ultrasound image is relatively redundant and can therefore be compressed efficiently, which can be an advantage in subsequent processing, as we only need to work with data of smaller dimensions.

III. Experiments and results

A. DTW using UTI, demonstration samples

From two speakers of the Hungarian database ('048' and '102') we selected one sentence for demonstration purposes, which occurred in both speakers' recordings. Based on the speech signal, we computed MFCC (frame shift was chosen to be 12 ms in line with the ultrasound video) with the `librosa` package, and then computed DTW with the `dtw` tool. Fig. 3 shows the DTW path between the two sentences: it can be seen that the speech sample took about 400 frames to be uttered by one speaker and about 520 frames by the other speaker, i.e. their articulatory speeds are different. The spectrograms of the speech samples from the two speakers are shown in Fig. 4 a) and b): here again, the difference in the length of the speech samples is visible. Fig. 4 c) and d) show a 'kymogram' [26, Fig. 8], i.e., a kind of 'articulatory signal over time': the middle slices (midline, c.f. Fig. 2) of the ultrasound tongue images were cut (approximately corresponding to the middle of the tongue) and plotted as a function of time, similarly to a spectrogram. The articulatory landmarks / inflexion points appear at different locations for the two speakers, in subfigures c) and d). Fig. 4. e) shows the result when the speech of the two speakers is DTW-aligned and the articulation data is stretched for the second (lower articulatory speed) speaker. Thus, Figures 4 c) and e) show that the articulatory data of

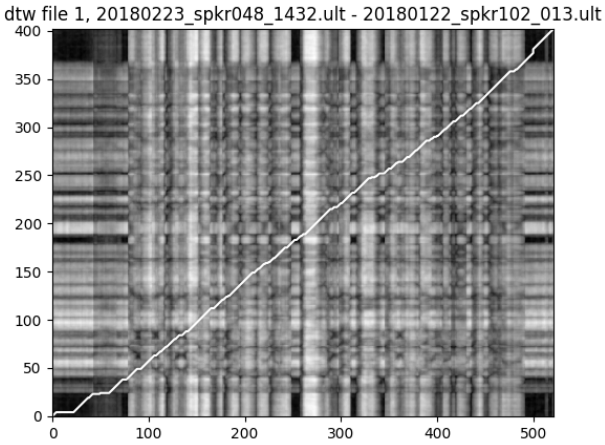

Fig. 3: DTW sample based on the same sentence (*„Az északi szél nagy vitában volt a Nappal, hogy kettőjük közül melyiknek van több ereje."*) by two Hungarian speakers, calculated from speech MFCC.

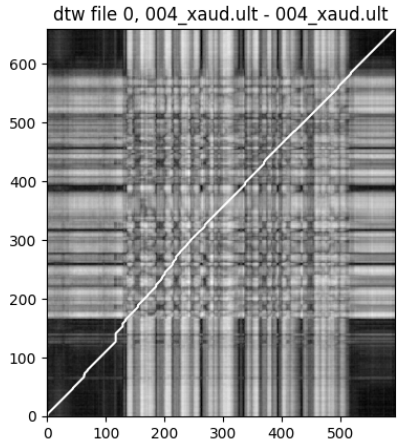

Fig. 5: DTW sample based on the same sentence (*„When sunlight strikes raindrops in the air, they act like a prism and form a rainbow."*) by two English speakers, calculated from speech MFCC.

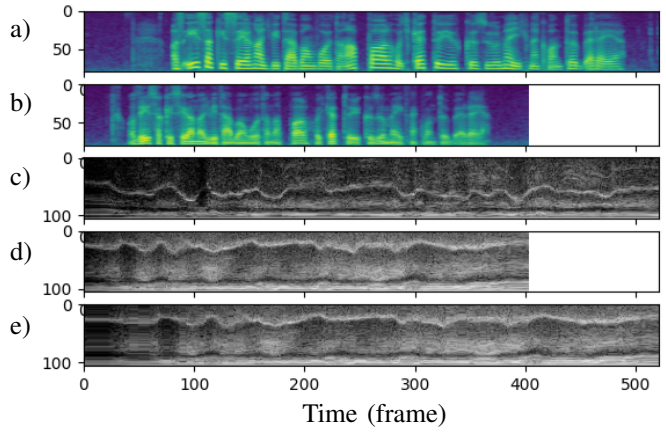

Fig. 4: Results: speech spectrogram and temporal change of the midline of the ultrasound tongue images, based on the sentence of Fig. 3.

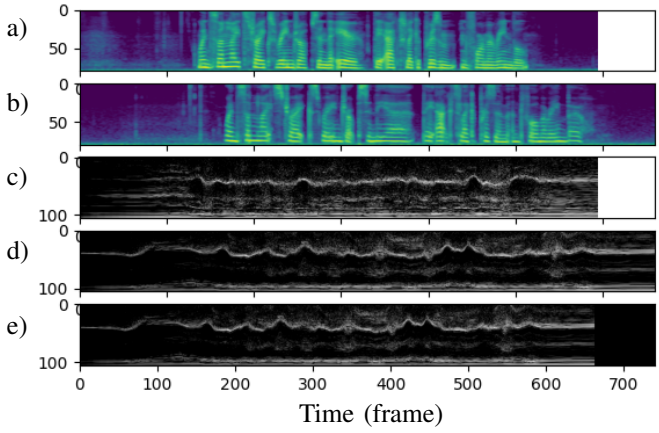

Fig. 6: Results: speech spectrogram and temporal change of the midline of the ultrasound tongue images, based on the sentence of Fig. 5

speaker '048' and speaker '102' were successfully aligned, and the landmarks / inflection points of tongue movement are at similar locations in the DTW-aligned Fig. 4 e) as in Fig. 4 c). For example, the sentence starts with the Hungarian back vowel 'a' and continues with a front vowel 'é', therefore, back-front movement of the tongue is visible on the c) and e) subfigures roughly around 50–100th time frames.

Similarly to the above, we have chosen a sentence from two speakers of the English database ('01fi' and '02fe'). The DTW result, as well as the speech spectrograms and adjusted articulatory data, are shown in Figs. 5 and 6. In this case, the DTW-adjusted articulatory data are less consistent with the reference compared to the example of Hungarian speakers.

B. Objective measures

In order to quantify how good the shifts of the DTW-aligned articulatory data are, it would be useful to first automatically determine the articulatory inflexion points (e.g. front/back position change of the tongue). For EMA data, a method for this has already been developed [21], but is not yet available for ultrasound tongue image signals. The accuracy of DTW-aligned ultrasound data is therefore not quantified in this research; for the time being, we rely on the visual examples above.

In future work, we plan to align the audio recordings along the resulted DTW path to examine the acoustic differences, as an objective quantification of the results.

IV. Discussion and Conclusions

The use of articulatory information in speech technology is less mature than standard speech recognition or speech synthesis; those methods using articulatory signals are currently at the basic research level and do not yet have applications for everyday people. Most related research deals with the way how articulatory information can be used to extend speech technology, for example as input or output of the system, like articulatory-to-acoustic mapping [9], [15], [16] or acoustic-to-articulatory inversion [17], [27].

The research problem is complicated by the fact that the mapping between articulation and acoustics is non-linear and not necessarily unique, i.e. several different articulatory configurations can result in the same speech output [28].

In the present study, we investigated the applicability of DTW for cross-speaker comparison of articulatory data on Hungarian and English recordings. According to the visualized demonstration samples, dynamic time warping seems to be a reasonable choice for such comparisons (and therefore, the most probable answer for the question in the title is YES), but the lack of a more advanced objective quantification of the accuracy of DTW for UTI is a limitation of the current study, which we will be happy to discuss at the WINS workshop with the interested audience of colleagues in speech technology, biomedical engineering, linguistics, and/or computational cognitive neuroscience. For the future, it would be worthwhile to try to align the audio recordings along the resulted DTW path to examine the acoustic difference.

We plan to use the results for speech-based brain-computer interfaces to supplement the brain signal (measured with EEG, ECoG or sEEG) with ultrasound tongue image based articulatory information [29]–[31].

V. Acknowledgements

The research was partially funded by the National Research, Development and Innovation Office of Hungary (FK 142163 grant), by the Bolyai János Research Fellowship of the Hungarian Academy of Sciences and by the ÚNKP-22-5-BME-316 New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund.

References

[1] W. von Kempelen, *Mechanismus der menschlichen Sprache, nebst der Beschreibung seiner sprechenden Maschine*. Vienna, Austria: Degen, 1791.

[2] D. H. Whalen, J. Kang, R. Iwasaki, G. Shejaeya, B. Kim, K. D. Roon, K. Mark, Tiede, J. Preston, E. Phillips, T. McAllister, and S. Boyce, "Accuracy assessments of hand and automatic measurements of ultrasound images of the tongue," in *Proc. ICPhS*, Canberra, Australia, 2019, pp. 542–546.

[3] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[4] T. Hueber, E.-l. Benaroya, B. Denby, and G. Chollet, "Statistical Mapping Between Articulatory and Acoustic Data for an Ultrasound-Based Silent Speech Interface," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 593–596.

[5] T. G. Csapó, T. Grósz, L. Tóth, and A. Markó, "Beszédszintézis ultrahangos artikulációs felvételekből mély neuronhálók segítségével," in *MSZNY 2017*, 2017, pp. 181–192.

[6] T. Grósz, L. Tóth, G. Gosztolya, T. G. Csapó, and A. Markó, "Kísérletek az alapfrekvencia becslésére mély neuronhálós, ultrahang-alapú némabeszéd-interfészekben," in *MSZNY 2018*, Szeged, Hungary, 2018, pp. 196–205.

[7] N. Kimura, M. C. Kono, and J. Rekimoto, "Sottovoce: An ultrasound imaging-based silent speech interaction using deep neural networks," in *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, UK, 2019, pp. 1–11.

[8] L. Tóth, A. H. Shandiz, G. Gosztolya, C. Zainkó, A. Markó, and T. G. Csapó, "3D konvolúciós neuronhálón és neurális vokóderen alapuló némabeszéd-interfész," in *MSZNY 2021*, 2021, pp. 123–137.

[9] T. G. Csapó, G. Gosztolya, L. Tóth, A. H. Shandiz, and A. Markó, "Optimizing the Ultrasound Tongue Image Representation for Residual Network-Based Articulatory-to-Acoustic Mapping," *Sensors*, vol. 22, 2022.

[10] B. Cao, A. Wisler, and J. Wang, "Speaker Adaptation on Articulation and Acoustics for Articulation-to-Speech Synthesis," *Sensors*, vol. 22, no. 16, p. 6056, 2022.

[11] M. Stone, B. Sonies, T. Shawker, G. Weiss, and L. Nadel, "Analysis of real-time ultrasound images of tongue configuration using a grid-digitizing system," *Journal of Phonetics*, vol. 11, pp. 207–218, 1983.

[12] M. Stone, "A guide to analysing tongue motion from ultrasound images," *Clinical Linguistics and Phonetics*, vol. 19, no. 6-7, pp. 455–501, jan 2005.

[13] A. R. Mandeel, M. S. Al-Radhi, and T. G. Csapó, "Investigations on speaker adaptation using a continuous vocoder within recurrent neural network based text-to-speech synthesis," *Multimedia Tools and Applications*, pp. 1–15, oct 2022.

[14] ——, "Speaker adaptation experiments with limited data for end-to-end Text-To-Speech synthesis using Tacotron2," *Infocommunications Journal*, 2022.

[15] T. G. Csapó, M. S. Al-Radhi, G. Németh, G. Gosztolya, T. Grósz, L. Tóth, and A. Markó, "Ultrasound-based Silent Speech Interface Built on a Continuous Vocoder," in *Proc. Interspeech*, Graz, Austria, 2019, pp. 894–898.

[16] T. G. Csapó, "Speaker dependent articulatory-to-acoustic mapping using real-time MRI of the vocal tract," in *Proc. Interspeech*, Shanghai, China, 2020, pp. 2722–2726.

[17] ——, "Speaker dependent acoustic-to-articulatory inversion using real-time MRI of the vocal tract," in *Proc. Interspeech*, Shanghai, China, 2020, pp. 3720–3724.

[18] P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad, "Electromagnetic articulography: use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain and Language*, vol. 31, no. 1, pp. 26–35, may 1987.

[19] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[20] C. Yang and M. Stone, "Dynamic programming method for temporal registration of three-dimensional tongue surface motion from multiple utterances," *Speech Communication*, vol. 38, no. 1-2, pp. 201–209, sep 2002.

[21] S. M. Jayanthi, L. Menard, and C. Laporte, "Divide-and-warp temporal alignment of speech signals between speakers: Validation using articulatory data," in *Proc. ICASSP*, New Orleans, LA, USA, 2017, pp. 5465–5469.

[22] G. Le Godais, "Decoding speech from brain activity using linear methods," Ph.D. dissertation, Université Grenoble Alpes, 2022.

[23] A. A. Wrench, "A Multichannel Articulatory Database and its Application for Automatic Speech Recognition," in *Proc. 5th Seminar on Speech Production: Models and Data*, Kloster Seeon, Bavaria, Germany, 2000, pp. 305–308.

[24] F. Bocquelet, T. Hueber, L. Girin, C. Savariaux, and B. Yvert, "BY2014 articulatory-acoustic dataset," sep 2016.

[25] M. S. Ribeiro, J. Sanger, J.-X. X. Zhang, A. Eshky, A. Wrench, K. Richmond, and S. Renals, "TaL: a synchronised multi-speaker corpus of ultrasound tongue imaging, audio, and lip videos," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China, 2021, pp. 1109–1116.

[26] S. M. Lulich, K. H. Berkson, and K. de Jong, “Acquiring and visualizing 3D/4D ultrasound recordings of tongue motion,” *Journal of Phonetics*, vol. 71, pp. 410–424, 2018.
[27] D. Porras, A. Sepúlveda-Sepúlveda, and T. G. Csapó, “DNN-based Acoustic-to-Articulatory Inversion using Ultrasound Tongue Imaging,” in *International Joint Conference on Neural Networks*, Budapest, Hungary, 2019, pp. N–19 221.
[28] D. Neiberg, G. Ananthakrishnan, and O. Engwall, “The acoustic to articulation mapping: non-linear or non-unique?” in *Proc. Interspeech*, Brisbane, Australia, 2008, pp. 1485–1488.
[29] 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.
[30] ——, “Towards Ultrasound Tongue Image prediction from EEG during speech production,” in *submitted to Interspeech*, 2023.
[31] 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