Interspeech 2022
18-22 September 2022, Incheon, Korea

Relationship between the acoustic time intervals and tongue movements of German diphthongs

Arne-Lukas Fietkau, Simon Stone, Peter Birkholz

Institute of Acoustics and Speech Communication, Technische Universität Dresden, Germany

arne-lukas.fietkau@tu-dresden.de

Abstract

This study investigated the relationship between tongue movements during the production of German diphthongs and their acoustic time intervals. To this end, five subjects produced a set of logatomes that contained German primary, secondary, and peripheral diphthongs in the context of bilabial and labiodental consonants at three different speaking rates. During the utterances, tongue movements were measured by means of optical palatography (OPG), i.e. by optical distance sensing in the oral cavity, along with the acoustic speech signal. The analysis of the movement signals revealed that the diphthongs have s-shaped tongue trajectories that strongly resemble half cosine periods. In addition, acoustic and articulatory diphthong durations have a linear, but not proportional, relationship. Finally, the peak velocity and midpoint between the two targets of a diphthong are reached in the middle of both the acoustic and articulatory diphthong time intervals, regardless of the duration and type of diphthong. These results can help to model realistic tongue movements for diphthongs in articulatory speech synthesis.

Index Terms: diphthong, articulatory synthesis, optical palatography

1. Introduction

Articulatory speech synthesis is a technique to produce artificial speech based on simulations of the human speech production system. A major challenge in articulatory synthesis is the generation of realistic articulatory movements. Currently, most articulatory synthesizers are controlled with low-level input, e.g., with the direct piece-wise specification of the articulatory trajectories [1, 2] or the time functions of muscle commands [3], or with the specification of articulatory targets [4], gestures [5, 6, 7, 8], or acoustic events [9]. These methods of control are very time consuming, as they require the manual specification of a large number of parameters. Thus, for the articulatory synthesis of larger amounts of text, higher-level control is necessary.

In the current version 2.3 of the articulatory synthesizer VocalTractLab (`www.vocaltractlab.de`), we implemented the possibility to synthesize speech from a specification of the intended phoneme sequence and the corresponding phoneme durations, similar to the input commands for the well-known MBROLA synthesizer [10]. The algorithm is based on a set of rules that specify the articulatory gestures underlying each phoneme as a function of phoneme duration and phonetic context. These rules in turn are loosely based on previous studies on articulatory-acoustic relations (e.g. [11, 12, 13]). Synthetic German sentences created in this way are already of high quality, but are not yet up to state-of-the-art commercial synthesizers [14]. One reason for this is that the diphthongs in the synthesized sentences do not sound quite natural in some cases. This is because the diphthongs were mapped to articulatory gestures on a purely heuristic basis, since there are significantly fewer studies on their articulatory-acoustic relations compared to monophthongs. As there are about 20 different diphthongs in German (including secondary diphthongs) with quite a high frequency in the language, more research is needed in this direction.

Therefore, the aim of this study was to characterize the articulation of diphthongs in relation to their acoustic start and end times. Here we focussed on the transitions of the *tongue* from the initial to the final targets of the diphthongs. In particular, this study addressed the following questions:

1. When does the articulatory transition of a diphthong begin and end relative to its acoustic start and end times?
2. When does the movement reach its peak velocity and cross the midway point between the two diphthong targets?
3. How does the peak velocity relate to the distance between the targets and the transition duration?
4. How do 1) and 2) vary with the diphthong duration (which ranges from about 70 ms to 340 ms, according to our data) and the type of diphthong?

2. Method

To address the above questions, five subjects uttered a set of logatomes that contained the diphthongs at three different speaking rates (to cover a wide range of diphthong durations), while the speech audio signal and the tongue movements were recorded. The diphthongs were then segmented at both the acoustic and articulatory levels, the measures of interest were extracted, and the sought relations between the measures were determined.

2.1. Recording setup

The articulatory data were recorded with Electro-Optical Stomatography (EOS) [16, 17] – a method that is developed at the TU Dresden and combines the measurement of tongue-palate contact patterns as in electropalatography (EPG, [18, 19]) with the measurement of tongue-palate distances in the midsagittal plane as in optical palatography (OPG, [20, 21]). The contact sensors and the distance sensors are mounted on the same artificial palate, each of which must be adapted to the shape of the subject's palate. The system used here had 32 contact sensors, 5 optical distance sensors for measuring tongue position, and 2 optical distance sensors for measuring lip position [16]. The complete set of sensor data was captured at a rate of 100 Hz. In this study, only the tongue distance measurements were analyzed, i.e. the system was used as a pure OPG system. The OPG sensors were equally spaced along the midline of the artificial palate as illustrated in Figure 1, and the distances were measured along their optical axes indicated by the red lines. The

10.21437/Interspeech.2022-73

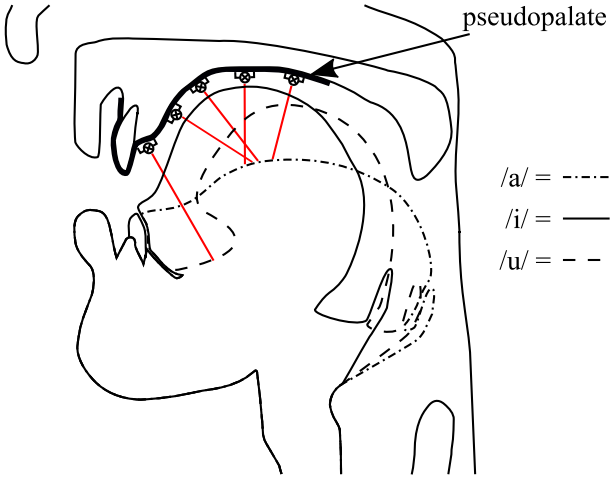

Figure 1: *Midsagittal view of the vocal tract with tongue shapes for the vowels /a/, /i/, and /u/ (based on the MRI data in [15]) showing the optical axes of the five OPG distance sensors.*

distance measurements were individually calibrated for each subject, as described in [22]. Thus, the articulatory measurements provided a set of 5 distance values (in cm) sampled at a rate of 100 Hz.

Audio recordings of the speech signals were made with a studio microphone (M930 by Microtech Gefell) connected to a computer via an external audio interface (MOTU 896 HD). The audio signals were sampled at a rate of 22050 Hz and with 16 bit quantization. The simultaneous recording of both the OPG and audio data was controlled using a custom-made software running on the computer. All recordings were performed in a soundproofed audio studio.

Table 1: *Recorded corpus of logatomes. The logatomes with the diphthongs that were not suitable for analysis are marked with an asterisk.*

1–10	11–20	21–30	31–40
/baɪpə/	/bɔɐpə/*	/faɪfə/	/fɔɐfə/*
/baʊpə/*	/bøːɐpə/	/faʊfə/*	/føːɐfə/
/bɔɪpə/	/bœɐpə/	/fɔɪfə/	/fœɐfə/
/baːɐpə/*	/buːɐpə/*	/faːɐfə/*	/fuːɐfə/*
/beːɐpə/	/bʊɐpə/*	/feːɐfə/	/fʊɐfə/*
/bɛːɐpə/	/byːɐpə/	/fɛːɐfə/	/fyːɐfə/
/bɐɐpə/	/bʏɐpə/	/fɐɐfə/	/fʏɐfə/
/biːɐpə/	/bʊɪpə/	/fiːɐfə/	/fʊɪfə/
/bɪɐpə/	/bɛɪpə/	/fɪɐfə/	/fɛɪfə/
/boːɐpə/*	/bɔʊpə/*	/foːɐfə/*	/fɔʊfə/*

2.2. Subjects, corpus and experimental procedure

Five healthy male native German subjects (27–42 years), who gave informed consent, participated in the experiment. Each subject produced a set of 40 different logatomes at three different speaking rates, while the OPG and audio data were recorded.

The logatomes are shown in Table 1. Each logatome contains one of 20 German primary, secondary, and peripheral diphthongs, once in the context of the bilabial consonants /b, p/ and once in the symmetric context of the labiodental consonant /f/. The context consonants were specifically chosen to minimize their effect on the tongue movement, i.e. the tongue was supposed to make the diphthongal transition with as little consonantal influence as possible. To ensure a uniform prosodic realization, each logatome was embedded in the carrier phrase: “Ich habe ... bestellt” - /ɪç habə ... bəʃtɛlt/ (English: “I have ordered ...”).

Each subject produced the phrases in three runs for the slow, normal and fast speaking rates, respectively. The 40 phrases to utter in each run were successively presented to the subjects on a computer screen, and the recording of each phrase was started with a mouse click. To induce the intended speaking rates, a progress bar was displayed below each phrase that automatically ran from 0% to 100% in 1 s, 2 s, or 3 s for the fast, normal, and slow speaking rates, respectively. It was not necessary that the actual utterance length corresponded exactly to the duration of 1, 2, or 3 seconds. The intention was merely to achieve a wide variation of phrase durations and thus of diphthong durations. To facilitate correct pronunciation of the diphthongs in the logatomes, the two phrases with the same diphthong were always presented consecutively, and a *real* German word with the respective diphthong was presented before them as a practice example (but not used in the later analysis). The order of the pairs of phrases for the individual diphthongs was randomized in each run. In total, 600 logatome instances were recorded (20 diphthongs × 2 consonant contexts × 3 speaking rates × 5 subjects).

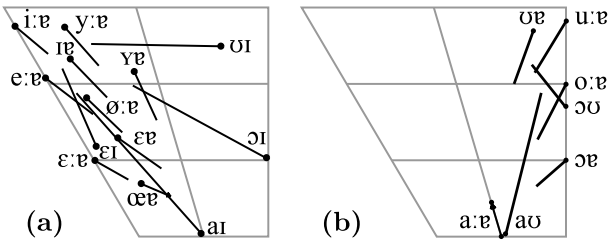

Figure 2: *Vowel charts of the German diphthongs [23, 24]: **(a)** Diphthongs with usable OPG trajectories, and **(b)** omitted diphthongs.*

2.3. Data postprocessing

Although the audio and OPG data recordings were started at the same time by the recording software in each session, the two data streams were not strictly synchronous due to unknown latencies in the operating system and the audio drivers. In order to establish exact synchronicity after the recordings, the OPG and audio data were manually aligned by means of precisely determinable acoustic and articulatory landmarks in the utterance “tatata” that was spoken at the beginning and end of each recording session.

An initial analysis of the OPG data revealed that for some diphthongs, the optical distance sensors did not properly capture the tongue movement. For these diphthongs, there was either no monotonic change of the measured OPG distances from the first to the second target of a diphthong, or the change of the distances was very small compared to the noise. This was the case for the 7 diphthongs shown in Figure 2b, which are characterized by a very posterior or low tongue position. They were excluded from the further analysis. Figure 2a shows the 13 used diphthongs (corresponding to 390 logatomes).

A detailed exploration of the trajectories also showed that the distance measurements of the most anterior and the most posterior OPG sensors did in some cases not reflect the smooth

tongue transition between the two diphthong targets. For example, when the tongue tip moved into or out of the light beam of the most anterior sensor, there was a sudden change of the measured distance (e.g., compare the tongue contours for /i/ and /u/ in Figure 1). Therefore, the most anterior and posterior sensors were excluded from the analysis. The distance curves obtained by the three remaining central sensors were averaged to obtain a single tongue position signal for each recording. These signals were finally smoothed with a (zero-phase) Gaussian low-pass filter with a cutoff frequency of 10 Hz to obtain the signal $d_{\mathrm{art}}(t)$.

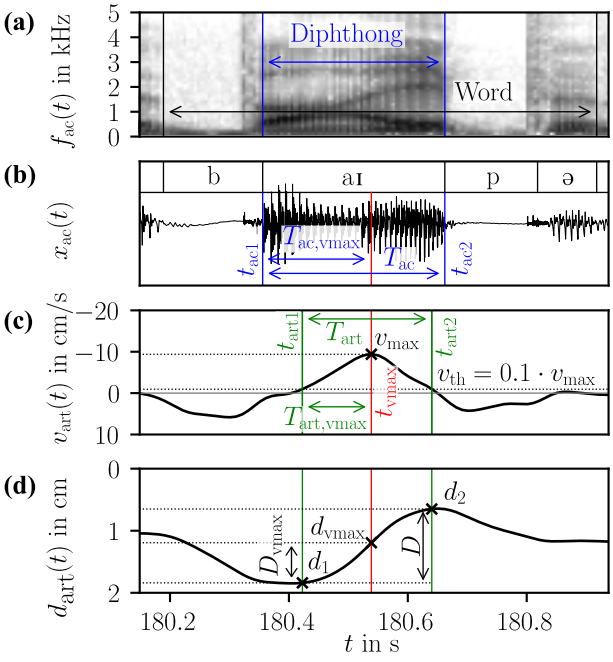

Figure 3: *Illustration of the determined quantities by means of the diphthong /aɪ/ in the logatome /baɪpə/.* ***(a)*** *spectrogram,* ***(b)*** *audio signal* $x_{\mathrm{ac}}(t)$*,* ***(c)*** *tongue velocity* $v_{\mathrm{art}}(t)$ *(first derivative of* $d_{\mathrm{art}}(t)$*),* ***(d)*** *tongue position signal* $d_{\mathrm{art}}(t)$*.*

2.4. Data analysis

Figure 3 illustrates the quantities that were determined for each spoken logatome. From the audio signal, the start time t_{ac1} and the end time t_{ac2} of the acoustic diphthong interval were manually marked. From the tongue velocity signal (first derivative of $d_{\mathrm{art}}(t)$) the time t_{vmax} and the value v_{max} of the (absolute) peak velocity were determined. The start time t_{art1} and the end time t_{art2} of the articulatory transition between the diphthong targets were defined as the times left and right from t_{vmax} where the velocity dropped to 10% of v_{max}. At t_{art1}, t_{art2}, and t_{vmax} the respective tongue positions d_1, d_2, and d_{vmax} were determined. In addition, the following quantities were calculated: the duration of the articulatory transition $T_{\mathrm{art}} = t_{\mathrm{art2}} - t_{\mathrm{art1}}$, the acoustic diphthong duration $T_{\mathrm{ac}} = t_{\mathrm{ac2}} - t_{\mathrm{ac1}}$, the movement range $D = |d_2 - d_1|$, the relative tongue position at the peak velocity $D_{vmax,rel} = |d_{vmax} - d_1|/D$, the time of the peak velocity relative to the acoustic diphthong start time $T_{\mathrm{ac,vmax}} = t_{\mathrm{vmax}} - t_{\mathrm{ac1}}$ and relative to the beginning of the articulatory transition $T_{\mathrm{art,vmax}} = t_{\mathrm{vmax}} - t_{\mathrm{art1}}$.

The above set of quantities could be uniquely determined for 306 out of the 390 diphthong realizations. The remaining 84 items were omitted from the further analysis. The relationships that characterize the tongue movements during the diphthongs relative to their acoustic time intervals were investigated by means of scatterplots and ordinary least squares regression models implemented in the Python module "statsmodels" [25].

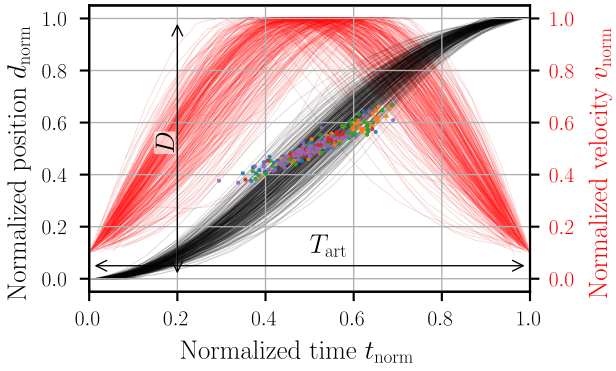

Figure 4: *Time and amplitude normalized diphthong transitions of the 306 evaluated samples (black curves) and their normalized velocities (red curves). The dots in the center show the time and position where the velocity is maximum for each position curve. The different colors indicate the type of diphthong.*

3. Results

An analysis of variance (ANOVA) of the data showed that the factors subject, speaking rate, context consonants, and diphthong type had in most cases no significant effect on the analyzed relations below. Hence, the following results are each based on all 306 samples together. The general characteristics of the tongue movements are presented first.

Figure 4 shows the time and amplitude normalized diphthong transitions for all 306 samples (black curves), together with the normalized velocities (red curves). Note that diphthong curves with an originally decreasing tongue position have been flipped upside down for a uniform analysis. All position curves have a smooth s-shape with a high degree of symmetry. As indicated by the colored dots in Figure 4, the peak velocity is reached around the midpoint of the transition, both temporarily and spatially, at the normalized time $t_{\mathrm{norm}} = 0.517 \pm 0.073$ and at the normalized position $d_{\mathrm{norm}} = 0.511 \pm 0.057$. This is also supported by the high correlation ($R^2 = 0.660$) between the (non-normalized) duration T_{art} of the articulatory transition and the time $T_{\mathrm{art,vmax}}$ of the peak velocity shown in Figure 5a. The regression line had the form

$$T_{\mathrm{art,vmax}} = 0.003\ \mathrm{s} + 0.496 \cdot T_{\mathrm{art}},$$

where only the coefficient for T_{art} had a significant effect ($p < 0.05$). The re-calculation of the regression line without the non-significant intercept ($p = 0.329$) yielded the relation $T_{\mathrm{art,vmax}} = 0.516 \cdot T_{\mathrm{art}}$ with $R_0^2 = 0.977$.

The similarity of the transitions is furthermore shown in Figure 5b, where the peak velocity is plotted as a function of the movement range D divided by the articulatory transition duration T_{art}. The regression line is

$$v_{\mathrm{max}} = -0.208\ \mathrm{mm/s} + 1.61 \cdot D/T_{\mathrm{art}} \approx 1.61 \cdot D/T_{\mathrm{art}}$$

with $R^2 = 0.997$ and a non-significant intercept ($p = 0.506$, $R_0^2 = 0.997$). This relation indicates that the peak velocity is highly proportional to the distance between the two diphthong targets and inversely proportional to the transition duration. The

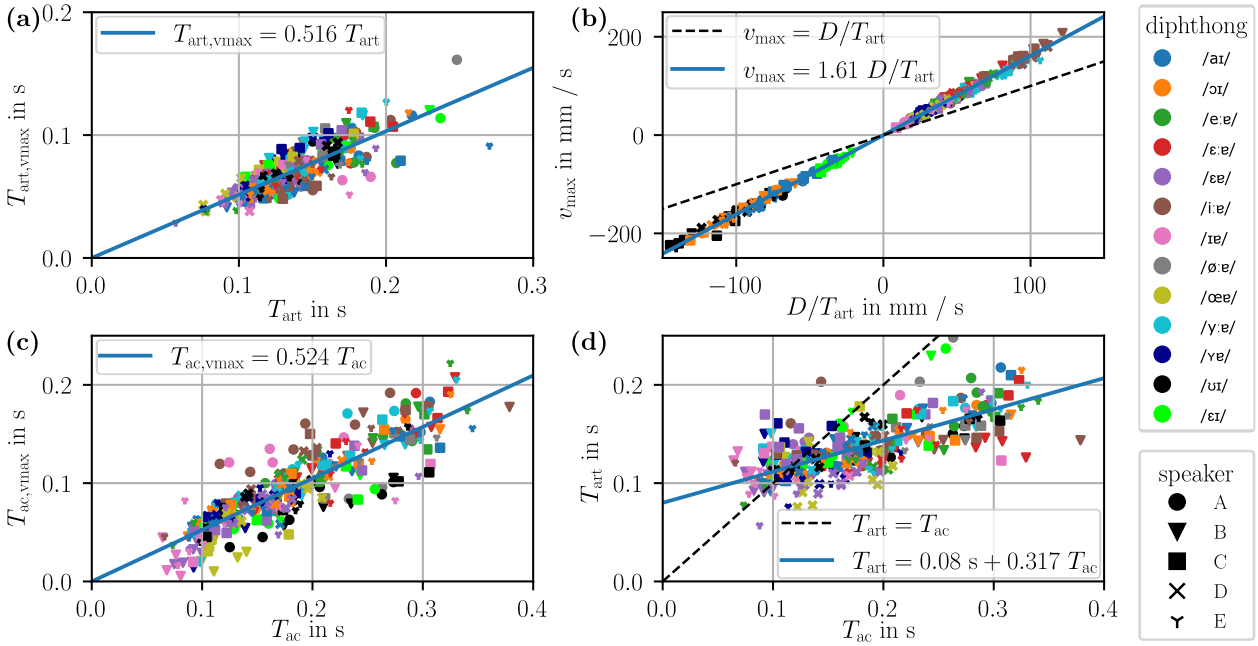

Figure 5: *Relations between the quantities of interest and the corresponding regression lines.* ***(a)*** $T_{\mathrm{art,vmax}}$ *as a function of* T_{art}, ***(b)*** $v_{\max}$ *as a function of* D/T_{art}, ***(c)*** $T_{\mathrm{ac,vmax}}$ *as a function of* T_{ac}, *and* ***(d)*** T_{art} *as a function of* T_{ac}.

proportionality factor of 1.61 is very close to the value of $\pi/2 = 1.57$ that would result if the transitions had the form of a perfect cosine half-wave.

With regard to the temporal alignment between the articulatory transition and the acoustic diphthong interval we found the following two relationships. First, there is a linear dependency between the time $T_{\mathrm{ac,vmax}}$ of the tongue peak velocity (relative to the onset of the acoustic diphthong interval) and the duration T_{ac} of the acoustic diphthong interval, as shown in Figure 5c. The regression line is

$$T_{\mathrm{ac,vmax}} = -0.005\ \mathrm{s} + 0.547 \cdot T_{\mathrm{ac}}$$

with $R^2 = 0.768$. Since there is only a significant effect ($p < 0.05$) of the slope (but not of the intercept with $p = 0.155$), the regression line can be re-calculated without the intercept term as $T_{\mathrm{ac,vmax}} = 0.524 \cdot T_{\mathrm{ac}}$ with $R_0^2 = 0.959$. The slope of 0.524 indicates that the peak velocity occurs very close to the temporal midpoint of the acoustic diphthong interval. Because the peak velocity also occurs at the temporal midpoint of the *articulatory* transition, the temporal midpoints of the acoustic interval and the articulatory transition can be considered to coincide.

To complete the description of the temporal alignment, we analyzed how the articulatory transition duration T_{art} relates to the acoustic diphthong duration T_{ac} (Figure 5d). Again, there is a linear relation between these variables with

$$T_{\mathrm{art}} = 0.080\ \mathrm{s} + 0.317 \cdot T_{\mathrm{ac}}$$

and $R^2 = 0.511$. In this case, both the intercept and the slope significantly affect the result. According to this relation, the articulatory and acoustic durations are equal for $T_{\mathrm{art}} = T_{\mathrm{ac}} = 117$ ms. For acoustic diphthong intervals shorter than 117 ms, the articulatory transition takes longer, and for acoustic diphthong intervals longer than 117 ms, the articulatory transition is shorter. With the derived equations, it is possible to determine the beginning, the end, and the turning point (time of the peak velocity) of the articulatory transition given the acoustic diphthong time interval.

4. Discussion and conclusions

The results of this study suggest that German diphthongs are formed over a wide range of durations essentially with a symmetrical s-shaped tongue movement whose peak velocity is reached in the middle of both the articulatory transition and the acoustic diphthong interval. While the s-shape of vowel-vowel transitions has been shown previously [26, 27], the specific temporal alignment over a wide range of durations has not, to our knowledge, been described before. The peak velocity itself was found to be highly proportional to the articulatory distance between the two diphthong targets, and inversely proportional to the transition duration. This corroborates and extends previous findings of linear relations between peak velocity and displacement for different articulators [28, 29]. A key finding was that the durations of the articulatory transition and the acoustic diphthong interval are linearly related, but not directly proportional. One consequence of this relation is that the overlap between the articulatory gestures for the diphthong and the adjacent phonemes reduces for increasing duration, which leads to increased speech clarity at slower speaking rates [30]. All of the found relations were independent of the type of diphthong and the speaker. The findings of this study will be implemented in the articulatory speech synthesizer VocalTractLab to facilitate the generation of natural-sounding diphthongs [31]. Future work should also extend this study with more subjects, with more languages, and using potentially different measurement techniques like electromagnetic articulography.

5. Acknowledgment

The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany in the programme of “Souverän. Digital. Vernetzt.”. Joint project 6G-life, project identification number: 16KISK001K

6. References

[1] A. J. S. Teixeira, R. Martinez, L. N. Silva, L. M. T. Jesus, J. C. Principe, and F. A. C. Vaz, "Simulation of human speech production applied to the study and synthesis of European Portuguese," *EURASIP Journal on Applied Signal Processing*, vol. 9, pp. 1435–1448, 2005.

[2] H. Nam, C. Mooshammer, K. Iskarous, and D. Whalen, "Hearing tongue loops: Perceptual sensitivity to acoustic signatures of articulatory dynamics," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. 3808–3817, 2013.

[3] J. E. Lloyd, I. Stavness, and S. Fels, "Artisynth: A fast interactive biomechanical modeling toolkit combining multibody and finite element simulation," in *Soft tissue biomechanical modeling for computer assisted surgery*, 2012, pp. 355–394.

[4] S. Prom-on, P. Birkholz, and Y. Xu, "Identifying underlying articulatory targets of Thai vowels from acoustic data based on an analysis-by-synthesis approach," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–11, 2014.

[5] R. Alexander, T. Sorensen, A. Toutios, and S. Narayanan, "A modular architecture for articulatory synthesis from gestural specification," *The Journal of the Acoustical Society of America*, vol. 146, no. 6, pp. 4458–4471, 2019.

[6] P. Birkholz and S. Drechsel, "Effects of the piriform fossae, transvelar acoustic coupling, and laryngeal wall vibration on the naturalness of articulatory speech synthesis," *Speech Communication*, vol. 132, pp. 96–105, 2021.

[7] B. J. Kröger and P. Birkholz, "A gesture-based concept for speech movement control in articulatory speech synthesis," in *Springer Proceedings of the COST 2102 Workshop on Verbal and Nonverbal Communication Behaviours*, Vietri sul Mare, Italy, 2007.

[8] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "TADA: An enhanced, portable Task Dynamics model in MATLAB," *The Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2430–2430, 2004.

[9] B. H. Story and K. Bunton, "A model of speech production based on the acoustic relativity of the vocal tract," *The Journal of the Acoustical Society of America*, vol. 146, no. 4, pp. 2522–2528, 2019.

[10] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. Van der Vrecken, "The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes," in *Proceeding of Fourth International Conference on Spoken Language Processing (ICSLP 1996)*, vol. 3, 1996, pp. 1393–1396.

[11] K. Iskarous, C. H. Shadle, and M. I. Proctor, "Articulatory–acoustic kinematics: The production of American English /s/," *The Journal of the Acoustical Society of America*, vol. 129, no. 2, pp. 944–954, 2011.

[12] A. Löfqvist and V. L. Gracco, "Interarticulator programming in VCV sequences: Lip and tongue movements," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1864–1876, 1999.

[13] K. N. Stevens, *Acoustic Phonetics*. The MIT Press, 1998.

[14] P. K. Krug, S. Stone, and P. Birkholz, "Intelligibility and naturalness of articulatory synthesis with VocalTractLab compared to established speech synthesis technologies," in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 102–107.

[15] P. Birkholz, S. Kürbis, S. Stone, P. Häsner, R. Blandin, and M. Fleischer, "Printable 3D vocal tract shapes from MRI data and their acoustic and aerodynamic properties," *Scientific Data*, vol. 7, no. 1, pp. 1–16, 2020.

[16] S. Stone and P. Birkholz, "Cross-speaker silent-speech command word recognition using electro-optical stomatography," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, Shanghai, China, 2020, pp. 7849–7853.

[17] S. Stone, *A silent-speech interface using electro-optical stomatography*, ser. Studientexte zur Sprachkommunikation. Dresden: TUDpress, 2021, vol. 102.

[18] S. Kelly, A. Main, G. Manley, and C. McLean, "Electropalatography and the Linguagraph system," *Medical Engineering & Physics*, vol. 22, no. 1, pp. 47–58, 2000.

[19] A. A. Wrench, "Advances in EPG palate design," *Advances in Speech-Language Pathology*, vol. 9, no. 1, pp. 3–12, 2007.

[20] C.-K. Chuang and W. S. Wang, "Use of optical distance sensing to track tongue motion," *Journal of Speech and Hearing Research*, vol. 21, pp. 482–496, 1978.

[21] S. G. Fletcher, P. A. Dagenais, and P. Critz-Crosby, "Teaching vowels to profoundly hearing-impaired speakers using glossometry," *Journal of Speech, Language, and Hearing Research*, vol. 34, no. 4, pp. 943–956, 1991.

[22] S. Preuß and P. Birkholz, "Optical sensor calibration for electro-optical stomatography," in *Proc. of the Interspeech 2015*, Dresden, Germany, 2015, pp. 618–622.

[23] K. J. Kohler, "German," *Journal of the International Phonetic Association*, vol. 20, no. 1, pp. 48–50, 1990.

[24] S. Kleiner, Ed., *Duden - Das Aussprachewörterbuch*. Berlin: Dudenverlag, 2015.

[25] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.

[26] R. A. Houde, "A study of tongue body motion during selected speech sounds," Ph.D. dissertation, University of Michigan, 1967.

[27] P. Mermelstein, "Articulatory model for the study of speech production," *The Journal of the Acoustical Society of America*, vol. 53, no. 4, pp. 1070–1082, 1973.

[28] D. P. Kuehn and K. L. Moll, "A cineradiographic study of VC and CV articulatory velocities," *Journal of Phonetics*, vol. 4, no. 4, pp. 303–320, 1976.

[29] J. A. S. Kelso, E. Vatikiotis-Bateson, E. L. Saltzman, and B. Kay, "A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling," *The Journal of the Acoustical Society of America*, vol. 77, no. 1, pp. 266–280, 1985.

[30] S. M. Tasko and K. Greilick, "Acoustic and articulatory features of diphthong production: A speech clarity study," 2010.

[31] S. Stone, Y. Gao, and P. Birkholz, "Articulatory synthesis of vocalized /r/ allophones in German," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.