INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Towards the prediction of the vocal tract shape from the sequence of phonemes to be articulated

Vinicius Ribeiro[1], *Karyna Isaieva*[2], *Justine Leclere*[2,3], *Pierre-André Vuissoz*[2], *Yves Laprie*[1]

[1]Université de Lorraine, CNRS, Inria, LORIA, Nancy, France
[2]Université de Lorraine, INSERM, U1254, IADI, Nancy, France
[3]Service de Médecine Bucco-dentaire, Hôpital Maison Blanche, Reims, France

vinicius.souza-ribeiro@loria.fr, karyna.isaieva@univ-lorraine.fr,
justine.leclere@chu-reims.fr, pa.vuissoz@chru-nancy.fr, yves.laprie@loria.fr

Abstract

In this work, we address the prediction of speech articulators' temporal geometric position from the sequence of phonemes to be articulated. We start from a set of real-time MRI sequences uttered by a female French speaker. The contours of five articulators were tracked automatically in each of the frames in the MRI video. Then, we explore the capacity of a bidirectional GRU to correctly predict each articulator's shape and position given the sequence of phonemes and their duration. We propose a 5-fold cross-validation experiment to evaluate the generalization capacity of the model. In a second experiment, we evaluate our model's data efficiency by reducing training data. We evaluate the point-to-point Euclidean distance and the Pearson's correlations along time between the predicted and the target shapes. We also evaluate produced shapes of the critical articulators of specific phonemes. We show that our model can achieve good results with minimal data, producing very realistic vocal tract shapes.

Index Terms: phoneme-to-articulatory, speech production, neural networks

1. Introduction

Continuous speech is a dynamic and non-stationary process that requires the interaction of several articulators. It is essentially the rapid transitions between vocal tract configurations that allows speech production [1] and the articulation of phonemes is thus very context-dependent [2]. A model of the vocal tract shape during speech is the subject of scientific interest for years. At first, Öhman [3] proposed a numerical model consistent with vocal tract measurements on X-ray images of Sweedish vowel-consonant-vowel utterances. Later, Maeda proposed an articulatory model [4] and then studied compensatory effects [5] showing how different articulatory pairs could produce the same phonetic pattern. Then, [6] incorporated phonetic constraints derived from the common knowledge of inversion to the model.

Nowadays, deep learning dominates the field of speech processing, using extensive amounts of data, new learning schemes and computational power. Even though deep neural networks have an incredible impact in many areas, speech production still presents many challenges due to the diversity of physiological and/or physical phenomena involved and the high intra- and inter-speaker variability. [7] characterized speaker-independent articulatory strategies and inter-speaker variability for 11 French speakers uttering 62 vowels and consonants. The author's model is capable of explaining 66% to 69% of the variance. However, the research suggests that most of the variability is due to anatomical differences instead of speaking strategies.

Current research on the direct problem of phoneme-to-articulatory prediction, or its inverse form – acoustics-to-articulatory inversion – varies not only in the method itself but also in the source of articulatory data. The most common data acquisition modality is electromagnetic articulography (EMA), which has been extensively utilized in research [8, 9, 10, 11].

The disadvantage of EMA is the small number of sensor coils attached to the subject's articulators and the impossibility of measuring articulators located deeper in the vocal tract. Therefore medical imaging techniques are the most promising to account for less accessible articulators. MRI is a non-invasive imaging technique that provides a precise observation of all the vocal tract structures (except bones and teeth) without presenting any hazard for the patient. Current real-time MRI (RT-MRI) enable acquisitions with a high frame rate at a lower resolution [12].

Csapó [13] explored RT-MRI for acoustic-to-articulatory inversion. Midsagittal RT-MRI images of the vocal tract were estimated using MGC-LSP spectral features as input and showed that the LSTMs are the most suitable for the task. Even though it is an excellent concept, the generated images contain several artifacts, and the produced shapes are not sufficiently realistic. We hypothesize that by targeting the complete MRI frame, the model misused the representative capacity to learn features that are not related to speech production.

In this work, we explore an alternative approach. We want to produce a direct mapping between the phonemes to be articulated and four vocal tract articulators – lower and upper lips, tongue, and soft palate – plus the pharyngeal wall which is useful to derive the nasopharyngeal port. Those five mentioned articulators are tracked in training images using a deep convolutional neural network approach (DCNN). The speech sound recorded with an optical microphone, and then denoised, enables ASR "forced" alignment and phonetic segmentation. Each frame is thus labelled with the corresponding articulated phoneme. Figure 1 presents a sample of the delineated contours automatically produced by our method [14]. This papers describes the approach used to predict the position of the articulators from the raw sequence of phonemes to be articulated.

2. Materials

2.1. Corpus

The dataset used corresponds to one adult female French native speaker. The RT-MRI sequences were acquired at the Centre Hospitalier Régional Universitaire de Nancy (CHRU Nancy), France. The recordings have a frame rate of 50 fps and image resolution of 136x136 pixels. The audio recordings are sampled

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-184

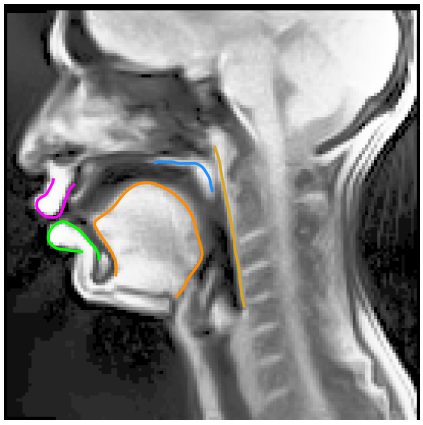

Figure 1: *Sample of MRI image with lower lip (green), upper lip (magenta), tongue (orange), pharynx (dark yellow), and soft palate (blue) and wireframe representation contours. Contours were automatically extracted from MR images.*

at 16 000 Hz. We used Astali [15] to "forced" align the speech with the transcriptions and obtain the phonetic segmentation. An expert manually corrected the phonetic annotations.

The corpus contains 16 acquisitions, with a median acquisition time of 51.5 seconds, a minimum of 43 seconds and a maximum of 79 seconds. The acquisitions have a median of 7 sentences each, a minimum of 4 sentences and a maximum of 9 sentences. Some of the sentences are repeated during the acquisition. The sentences were selected to provide a phonetically balanced coverage of French. The corpus includes 107 spoken sentences, with 36 unique phonemes plus 4 tokens to represent silence, inaudible or non speech noises, hesitations, stop closures and burst releases.

2.2. Automatic tracking of articulators

To track the articulators, we trained a Mask R-CNN model [16] to segment boundaries of the articulators. Then, we post-process the probability map using the graph-based algorithm described by [14]. The algorithm searches for the two most distant points in the probability map and connects them using Dijkstra's shortest path algorithm. For the soft palate, our approach is subtly different. Because it is a thin structure, the algorithm does not perform correctly. We explored other approaches, e.g., active contours models [17], but the results were not satisfactory. Instead, we modified the network's output such that the graph-based approach would generate the center line of the soft palate, which we use as target. In the end, the contours are regularized using b-splines to achieve a smoother curve. For the two cases in which the model could not extract the contour for the lower lip, we did it manually.

3. Generation of the vocal tract shape from the sequence of phonemes

3.1. Architecture

We employed the architecture proposed by [18], with a few modifications. The network depicted in Figure 2 contains two parts. The first includes two layers of bidirectional gated recurrent units (GRU). Then, a linear layer with ReLU activation reshapes the GRU cell's output. The second part is the Articulator Predictor block, which comprises a sequence of layer normalization and linear layers with ReLU activation. The sigmoid activation in the end guarantees the outputed coordinates are between 0 and 1. Our architecture is composed by five Articulator Predictor blocks, one for each articulator.

The input of the network is the phonetic segmentation, i.e. the phonemes to be articulated and their boundaries. In this current work there is no prediction of the sound duration from the sequence of phonemes to be articulated. The network produces the five contours for each 20 ms frame.

3.2. Evaluation

The models were trained by minimizing the Euclidean distance between the predicted spatial coordinates and the target positions. The loss function is given by

$$L(p, \hat{p}) = \frac{1}{N_{art} \times N_{samples}} \sum_{i=1}^{N_{art}} \sum_{j=1}^{N_{samples}} d(p_{ij}, \hat{p_{ij}}) \quad (1)$$

where p and $\hat{p}$ are the ground truth and the predicted points, N_{art} is the number of articulators, $N_{samples}$ is the number of predicted samples, and d is the Euclidean distance. In our case, $N_{art} = 5$ and $N_{samples} = 50$. We also measure Pearson's correlation between the predicted and the target curves' trajectories along time.

It should be stressed the contours used as the ground truth are automatically delineated by the algorithm described in Section 2.2. Despite its very good robustness some contours might be erroneous.

Finally, we make a subjective analysis of the produced shapes. We evaluate specific phonemes that have critical articulators. Critical articulators are those resistant to context and that have a co-articulatory effect on neighboring phones [19]. For example, for the phonemes /p/ and /b/, it is mandatory to have the lips fully closed. The tongue tip must touch the teeth to produce the phonemes /l/ and /t/. For /k/, the tongue dorsum must approach the hard palate, and for nasal vowels, the soft palate must be open. We evaluate our model to check if these physical constraints are satisfied.

3.3. Experimental Setup

We propose two experiments. In the first, we separate three of the 16 acquisitions in our dataset for the test. From the remaining 13 acquisitions, we run a 5-fold cross-validation scheme, keeping 11 for training and two for validation. In the second, we train the model with 14, 11, 8, 6, 4, and 2 acquisitions, keeping validation and test sets fixed with one acquisition each.

We feed each sentence phoneme-by-phoneme to the model. To account for phoneme duration, we repeat the phonemes to match the number of frames of the RT-MRI films. Thus, each token in the input sentence lasts for 20 ms. The model predicts one shape per input token.

We trained our models for 300 epochs, with 20 epochs of patience for early stopping. We used the Adam optimizer [20], with a learning rate of 1e-4, and we reduce the learning rate by a factor of 0.1 after ten epochs without improvements. We implemented the code using Pytorch [21].

4. Results

Table 1 presents the results in the 5-fold cross-validation framework. Figure 3 presents our results in the data efficiency experiment. Figure 5 presents samples of critical articulators. The

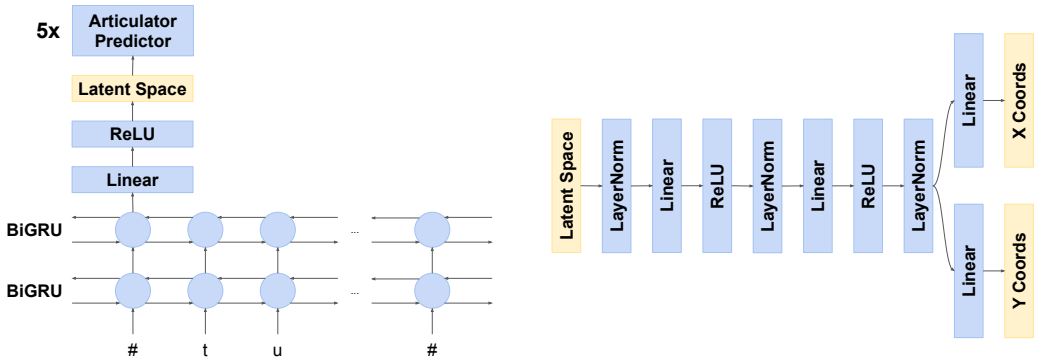

Figure 2: *The proposed network architecture. We present the recurrent encoding path on the left and the articulator predictor decoder on the right. Each RNN cell output is reshaped and fed into the articulator predictor. We draw only the first cell for simplicity.*

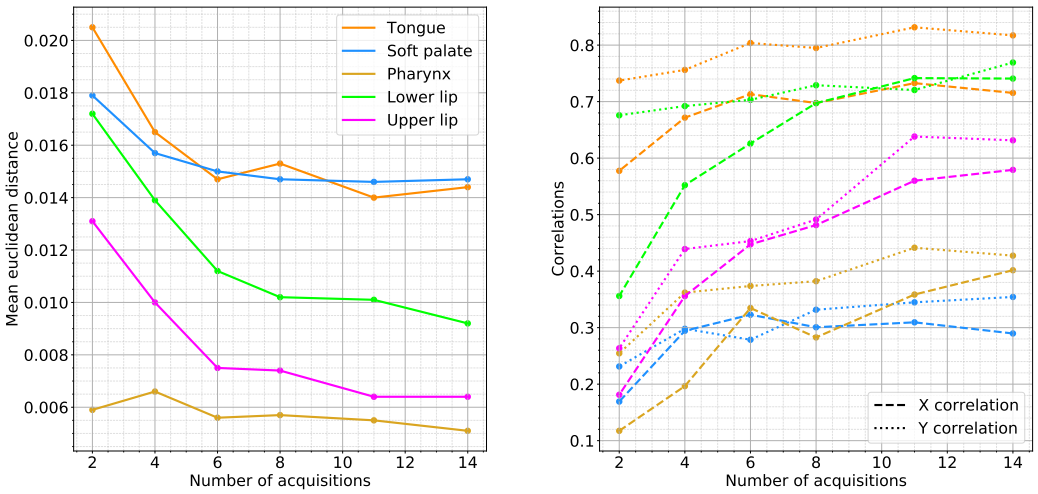

Figure 3: *Results from the data efficiency experiment. On the left, the mean Euclidean distance for each articulator. On the right, the X (dashed line) and Y (dotted line) correlations for each articulator.*

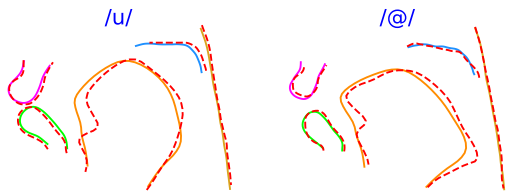

Figure 4: *Samples of the model's output with presence of the sublingual cavity. The solid lines corresponds to the predicted shape and the dashed line corresponds to the ground truth.*

depicted figures present the predicted contours after b-spline regularization, which we apply only for visualization.

In our second experiment, i.e., data-efficiency evaluation, the larger experiment – with more acquisitions during training – fully contains the data from all of the smaller ones. Thus, the number of training sentences increases with the number of acquisitions and phonemes present in the smaller experiments are present in the larger ones as well.

Table 1: *Mean Euclidean distance, X, and Y correlations for each articulator in the 5-fold cross-validation scheme. We omit the mean Euclidean distances' standard deviations to save space, but they are approximately* 0 001

Articulator	MED	ρ_x	ρ_y
Tongue	0.016	0.689 ± 0.065	0.637 ± 0.050
Pharynx	0.007	0.327 ± 0.099	0.396 ± 0.047
Soft palate	0.012	0.407 ± 0.031	0.403 ± 0.010
Lower lip	0.011	0.668 ± 0.050	0.714 ± 0.028
Upper lip	0.008	0.535 ± 0.052	0.512 ± 0.034

5. Discussion

Table 1 and Figure 5 show that the model can produce high-quality shapes, with tiny errors and good correlations between the predicted and target trajectories. Even though the lowest errors occur for the upper lip, the best x- and y-correlations occur for the tongue and the lower lip. The tongue is the largest considered articulator, with the highest number of degrees of freedom. Thus it is expected that it accounts for the most significant errors. The pharynx accounts for the lowest correlations, which

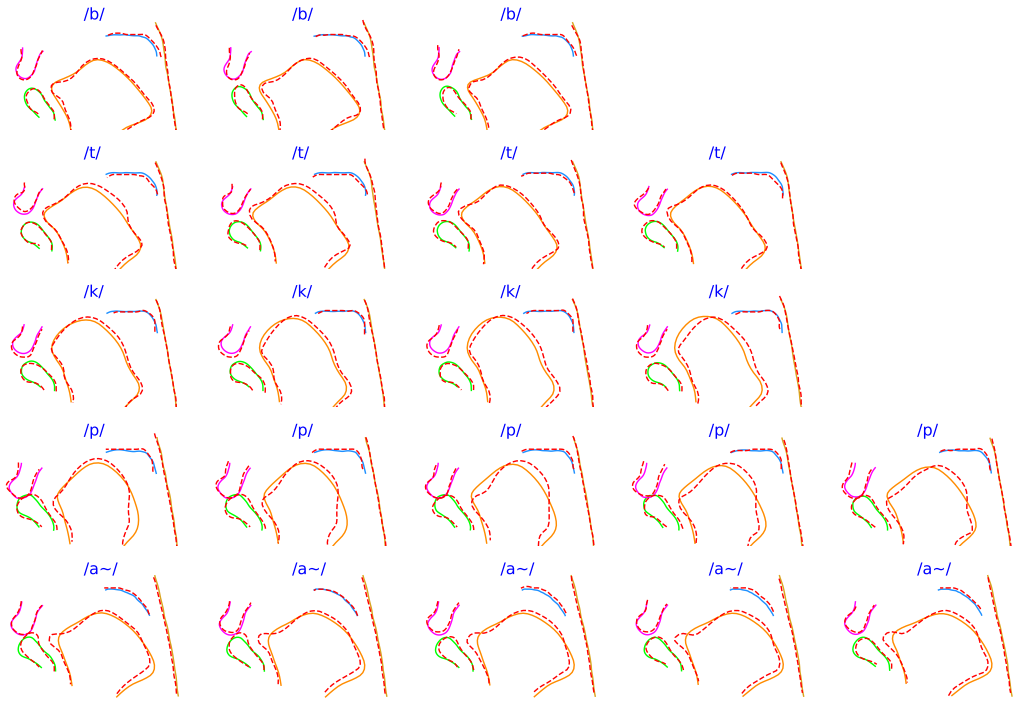

Figure 5: *Samples of our model predictions for the phonemes /b/, /t/, /p/, /k/, and /ã/, in order of appearance from top to bottom. The solid lines represent model predictions. The red dotted line represents the ground truth. Each row represents one phoneme occurrence, and the columns represent the time steps (20 ms).*

is expected since it is a fixed articulator, with almost no degrees of freedom.

Figure 4 shows that the model fails to reconstruct the sublingual cavity when it is present in the ground truth contours. We hypothesize that it is an error propagated from the tracking algorithm which sometimes fails to identify the sublingual cavity. Usually, this is a fuzzy region in the low-resolution image and the probability map is not sufficiently clear.

On the data efficiency experiments, we evaluate the model's capacity of learning from fewer data. Our results show that our model is highly data-efficient. Even though we reach the best results using all the acquisitions during training, the marginal gain with more than six acquisitions is less expressive. This behavior is observable for both Euclidean distance and correlations.

From the critical articulators' perspective, our model can learn some specific physical constraints. As discussed before, lips should be closed for the production of /p/ and /b/. Figure 5 shows that this condition is satisfied for /p/ but not for /b/. Actually there is a closing movement but it is not complete. One reason is probably the fact that the closure for /b/ is shorter than /p/, thus corresponds to very few frames with a complete closure, with the consequence of not training the model with enough relevant data.

The evaluation is trickier for the phonemes /k/, /t/, and nasal vowels because they depend on the interaction with fixed structures. For /t/, we see that the model projects the tongue tip forward, suggesting contact with (virtual) teeth. However, for some cases in the dataset, we expected a more prominent tongue tip. For /k/, a similar effect occurs with the tongue dorsum, which the model projects up, closing the gap with a (virtual) hard palate.

When we compare the last row of Figure 5 with the previous ones, which refers to the /ã/, we observe that the soft palate opens for nasal vowels, and closes for non-nasal phonemes.

When we compare the predicted curves with the ground truths, we see that the critical articulator constraints are mostly satisfied in our model's predictions.

6. Conclusions

In this paper, we present a novel approach for predicting the vocal tract shape from the sequence of phonemes to be articulated. To the best of our knowledge, this work is the first to produce the main articulator contours for continuous speech. Our model can yield promising results for a single speaker with minimal data even if one expects better results with a bigger corpus. In the subjective evaluation, critical articulators' physical constraints are satisfied, which shows that our results are realistic.

However, there are still limitations. First, this research focuses on articulators in the upper part of the vocal tract. An obvious continuation is the addition of other structures, i.e. the epiglottis and larynx, which are essential to get the complete geometry so as to synthesize a speech signal by means of the numerical simulations of the aero-acoustic equations.

Second, we rely on automatic contour tracking which produces reference contours used as ground truth. The presented model has a performance upper bound linked on the capacity of producing an accurate and relevant ground truth.

Last, we work with a single speaker. Our input does not encode any speaker-specific feature, and it only uses the sequence of phonemes to be articulated and their expected duration. Thus, learning from several speakers is a challenge and a sufficient amount of data is a requirement to marginalize anatomical variance.

7. References

[1] P. Saha, P. Srungarapu, and S. Fels, "Towards automatic speech identification from vocal tract shape dynamics in real-time mri," *arXiv preprint arXiv:1807.11089*, 2018.

[2] S. E. Öhman, "Coarticulation in vcv utterances: Spectrographic measurements," *The Journal of the Acoustical Society of America*, vol. 39, no. 1, pp. 151–168, 1966.

[3] ——, "Numerical model of coarticulation," *The Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 131–320, 1967.

[4] S. Maeda, "Un modèle articulatoire de la langue avec des composantes linéaires," in *Actes 10èmes Journées d'Etude sur la Parole*, Grenoble, May 1979, pp. 152–162.

[5] ——, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech production and speech modelling*. Springer, 1990, pp. 131–149.

[6] B. Potard, Y. Laprie, and S. Ouni, "Incorporation of phonetic constraints in acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2310–2323, 2008.

[7] A. Serrurier, P. Badin, L. Lamalle, and C. Neuschaefer-Rube, "Characterization of inter-speaker articulatory variability: a two-level multi-speaker modelling approach based on mri data," *The Journal of the Acoustical Society of America*, vol. 145, no. 4, pp. 2149–2170, 2019.

[8] T. Biasutto-Lervat and S. Ouni, "Phoneme-to-articulatory mapping using bidirectional gated rnn," in *Interspeech 2018-19th Annual Conference of the International Speech Communication Association*, 2018.

[9] A. Illa and P. K. Ghosh, "Low resource acoustic-to-articulatory inversion using bi-directional long short term memory." in *INTERSPEECH*, 2018, pp. 3122–3126.

[10] N. Bozorg and M. T. Johnson, "Acoustic-to-articulatory inversion with deep autoregressive articulatory-wavenet," *Networks (CNNs)*, vol. 22, p. 23, 2020.

[11] A. S. Shahrebabaki, S. M. Siniscalchi, G. Salvi, and T. Svendsen, "Sequence-to-sequence articulatory inversion through time convolution of sub-band frequency signals," *database*, vol. 1, p. 5, 2020.

[12] M. Uecker, S. Zhang, D. Voit, A. Karaus, K.-D. Merboldt, and J. Frahm, "Real-time mri at a resolution of 20 ms," *NMR in Biomedicine*, vol. 23, no. 8, pp. 986–994, 2010.

[13] T. G. Csapó, "Speaker dependent acoustic-to-articulatory inversion using real-time mri of the vocal tract," *arXiv preprint arXiv:2008.02098*, 2020.

[14] K. Isaieva, Y. Laprie, N. Turpault, A. Houssard, J. Felblinger, and P.-A. Vuissoz, "Automatic tongue delineation from mri images with a convolutional neural network approach," *Applied Artificial Intelligence*, pp. 1–9, 2020.

[15] "Astali." [Online]. Available: http://ortolang108.inist.fr/astali/fr/

[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[17] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[18] T. Biasutto, S. Dahmani, S. Ouni *et al.*, "Modeling labial coarticulation with bidirectional gated recurrent networks and transfer learning," in *INTERSPEECH 2019-20th Annual Conference of the International Speech Communication Association*, 2019.

[19] S. Silva and A. J. Teixeira, "Critical articulators identification from rt-mri of the vocal tract." in *INTERSPEECH*, 2017, pp. 626–630.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf