Autoencoder-Based Tongue Shape Estimation During Continuous Speech

Vinicius Ribeiro, Yves Laprie

Autoencoder-Based Tongue Shape Estimation During Continuous Speech

Vinicius Ribeiro[1], *Yves Laprie*[1]

[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France

{vinicius.souza-ribeiro, yves.laprie}@loria.fr

Abstract

Vocal tract shape estimation is a necessary step for articulatory speech synthesis. However, the literature on the topic is scarce, and most current methods lack adequacy to many physical constraints related to speech production. This study proposes an alternative approach to the task to solve specific issues faced in the previous work, especially those related to critical articulators. We present an autoencoder-based method for tongue shape estimation during continuous speech. An autoencoder is trained to learn the data's encoding and serves as an auxiliary network for the principal one, which maps phonemes to the shapes. Instead of predicting the exact points in the target curve, the neural network learns how to predict the curve's main components, i.e., the autoencoder's representation. We show how this approach allows imposing critical articulators' constraints, controlling the tongue shape through the latent space, and generating a smooth output without relying on any postprocessing method.

Index Terms: autoencoder, tongue shape estimation, phoneme-to-articulation

1. Introduction

Correctly predicting the temporal evolution of the vocal tract shape is a requirement of articulatory speech synthesis. Reconstructing the complete shape from the glottis to the lips allows the vocal tract area function to be measured and acoustic simulations to be carried out to synthesize acoustic signals [1, 2, 3]. Even though the quality of generated speech is lower than that of end-to-end synthesis, the advantage of these techniques is the capacity to control a set of vocal tract parameters. In this sense, correctly reproducing the articulator's shape for each phoneme to be articulated and taking coarticulation into account is a substantial step towards understanding and mimicking the speech production dynamics [4, 5] and for motor control research [6, 7].

Previous work [8, 9] proposed an encoder-decoder neural network to predict the vocal tract shape for a sequence of phonemes to be articulated. The former was the first to provide the complete vocal tract shape, including all speech articulators from the glottis to the lips. They predict a fixed number of points for each vocal tract articulator's curve for each phoneme in the target sentence. However, this approach presents some disadvantages. First, estimating all of the points in the curve gives too many degrees of freedom to the models, making them susceptible to overfitting. Second, the predicted samples are highly correlated, making it difficult to control the shape with phonetically relevant parameters and analyze the impact of changes in the synthesis. Third, postprocessing (implemented with regularizing splines) is required to produce a smooth output curve. Last, it is tough to add constraints on critical articulators. Imposing a contact between the tongue and the palate, for instance, leads to moving only one tongue point and consequently to a very artificial and inconsistent tongue form.

This work proposes an approach for 2D-tongue shape prediction to overcome these difficulties. It is divided into two parts. First, an autoencoder is trained to learn the tongue shape encodings. Second, a recurrent network is trained to estimate the autoencoder's representation for each phoneme to be articulated. The pre-trained autoencoder is used as an auxiliary network in the training of the second part. It is important to stress that the autoencoder's weights are frozen during the second part, i.e., the autoencoder is not learning anymore.

Let $X = \{x_1, x_2, ..., x_N\}$ be the sequence of phonemes and $Y = \{y_1, y_2, ..., y_N\}$ be the associated tongue shapes. On the one hand, our previous work was searching for the function $f(\cdot)$ such that

$$\hat{y} = f(x)$$

and $d_1(y, \hat{y})$ is minimal, where d_1 is the point-wise Euclidean distance. On the other hand, this work searches for the functions $g_{\text{enc}}(\cdot)$, $g_{\text{dec}}(\cdot)$ and $h(\cdot)$ such that

$$\hat{y} = g_{\text{dec}}(h(x))$$

and $d_2(g_{\text{enc}}(y), h(x))$ is minimal, where d_2 is the L^2 loss. Here, $g_{\text{enc}}(\cdot)$ corresponds to the autoencoder's encoder network, $g_{\text{dec}}(\cdot)$ is the autoencoder's decoder network, and $h(\cdot)$ is the recurrent encoder-decoder network. Afterward, we can impose constraints in the reconstructed shape ($g_{\text{dec}}(h(x))$) since the encoder-decoder network is limited to exploring the autoencoder's latent space and is not capable of "cheating" by proposing curious tongue shapes.

We focus solely on predicting the tongue shape instead of targeting the complete vocal tract to keep the paper concise while evaluating essential ideas related to the task. The tongue is the largest articulator, with the most degrees of freedom. Also, the tongue is the primary articulator for two out of four tract variables (investigated in [9]), i.e., the Tongue Body Constraint Degree (TBCD) and the Tongue Tip Constraint Degree (TTCD).

2. Methods

2.1. Autoencoder

An autoencoder is a non-linear method for efficiently learning how to encode information. In simpler terms, it is a neural network trained to attempt to copy its input to its output [10]. A general autoencoder architecture contains an encoder network, an information bottleneck (the latent space), and a decoder network. Since the latent space has a much more limited dimensionality than the inputs, the bottleneck would only encode the essential information for the reconstruction performed by the decoder. Therefore, the autoencoder is often seen as a non-linear dimensionality reduction algorithm.

Compared to traditional linear dimensionality reduction approaches, the autoencoder lacks some essential characteristics.

Principal Components Analysis (PCA) is the most popular and well-diffused of such methods [11]. PCA aims to find the set of orthogonal vectors that explains the most variance in the input data. Using orthogonal vectors, PCA guarantees that the encoded features are linearly uncorrelated. A single-layer autoencoder without non-linearities would be comparable to PCA, but these characteristics (orthogonality, zero-covariance, and ranked components) would still be missing.

A PCA-like autoencoder would require changes to the traditional training methodology [12]. It is required to reduce the covariance between the latent features, orthogonalize the weights matrix, and rank the components by explained variance. Our autoencoder training mimics some of these desired ideas while keeping a non-linear structure. Still, we do not cover all of the requirements. Each network (the encoder and the decoder) contains three linear layers. The first two linear layers have `ReLU` activation, while the last encoder layer has `Tanh` activation, and the last decoder layer does not have any activation. This architecture guarantees that the latent space is between -1 and 1, which improves model stability. The autoencoder is trained using weighted L^2 loss plus a covariance minimization on the feature space. We give a weight of 3 to phonemes that have critical articulators attached to the tongue (/l, d, t, n, k, g/), a weight of 0.1 to non-phonetic tokens, such as silence and noise, and other tokens remain with a weight of 1.

In this work, the latent space is denoted by Z, the ith autoencoder component is denoted by z_i, and $\dim Z$ denotes the dimensionality of Z.

2.2. Phoneme to autoencoder's components

The encoder-decoder network that maps phonemes to the autoencoder's latent space is very similar to the one used in [8]. The same GRU-based encoder with a linear reshaping layer is used. The main difference is that the Articulator Predictor head is replaced by a single block composed of layer normalization and linear layers with `ReLU` activation, and a final block of layer normalization and linear layer with `Tanh` activation. The networks' inputs are the sequence of phonemes, with phoneme duration encoded as repetitions. The network's outputs are the components corresponding to each phoneme.

The learning objective is

$$\mathcal{L} = \mathcal{L}_{\text{latent}} + \alpha \cdot \mathcal{L}_{\text{reconstruction}} + \beta \cdot \mathcal{L}_{\text{critical}}$$

where $\mathcal{L}_{\text{latent}}$ is the L^2 loss between the target and the predicted autoencoder's representations, $\mathcal{L}_{\text{reconstruction}}$ is the mean point-wise Euclidean distance between the reconstructed and the target shapes, and $\mathcal{L}_{\text{critical}}$ is the masked critical loss. The critical loss enforces that the critical articulators' targets are achieved. It is calculated as the mean minimal distance between the reconstructed curve and a pre-computed reference, i.e., the critical loss reinforces the place of articulation. The critical loss is masked to consider only phonemes in which the tongue is a critical articulator. In our case, the pre-computed reference is the alveolar region for TTCD and the hard palate for TBCD. We set $\alpha = 1.0$ and $\beta = 0.3$.

2.3. Experimental design

The dataset [13] is the same used in [9]. It comprises real-time MRI (rt-MRI) videos of one male French native speaker. The MRI sequences were recorded at Max Plank Institute, Göttingen, Germany. The recordings have a frame rate of 55 fps, pixel spacing of 1.412 mm, and an image resolution of 136×136 pixels. The audio recordings sampling frequency is 16 000 Hz. It has 38 MRI acquisitions, with several sentences per acquisition. We used 34 acquisitions for training and kept two for validation and two for the test. The tongue's and upper incisor's (alveolar region + hard palate) shapes are described as a 50×2 array, representing the (x, y) coordinates of the 50 samples in the articulator's curve and were extracted using the procedure described in [9]. The phonetic annotations were extracted using forced alignment [14] and then corrected by a specialist in phonetics.

For the autoencoder training, our primary evaluation metric is the reconstruction error, measured in terms of the point-to-closest-point (P2CP) distance [15]. We evaluated the autoencoder's information bottleneck with 8, 10, 12, and 16 components. A secondary metric used is the variance explained by the autoencoder. Finally, we evaluated how changing a single component in the autoencoder's latent space affects the reconstruction while holding all remaining components constant. The former evaluation enables tongue shape control, which is desired.

The main network's results are evaluated in two aspects. The first is the same reconstruction metric as that used for the autoencoder. The second is the measurement of the TBCD and the TTCD. The predicted trajectories are monitored for these two vocal tract variables, and Pearson's correlation with the ground truth trajectory is calculated. It is important to stress that these tract variables are mostly relevant for phonemes in which they are critical, i.e., /l, t, d, n/ for TTCD, and /k, g/ for TBCD.

The dataset for the autoencoder contains 110 231 shapes for training, 7 549 shapes for validation, and 7 509 shapes for testing. The phoneme to autoencoder's components dataset contains 637 utterances for training, 36 utterances for validation, and 34 utterances for testing. Models were trained for 3 000 epochs with 20 epochs of patience for early stopping. The Adam optimizer [16] was used, with a weight decay of 10^{-6} and a learning rate of 10^{-4}, which was reduced by a factor of ten after ten epochs without improvements in the validation loss. We implemented the code with `PyTorch` [17]. We re-ran the experiments from [9] with the same training configuration and train-validation-test splits for a fair comparison.

3. Results

3.1. Autoencoder

Table 1 presents the mean reconstruction error and the variance explained by each setting. We ran a t-test to evaluate the statistical significance between these errors – with a significance level of $p \leq 0.008$. We found that all of the combinations presented statistically significant differences. Figure 1 displays the autoencoder's components covariance matrix and Figure 2 presents a nomogram illustrating how varying a single component at a time affects the reconstructed shape for the eight components configuration.

Table 1: *Mean P2CP distance and explained variance for each of the autoencoder's configurations.*

$\dim Z$	8	10	12	16
Mean P2CP (mm)	1.228	1.410	1.015	1.003
Explained Var.	0.930	0.888	0.955	0.957

Figure 1: *Autoencoder's latent space covariance matrix (*$\dim Z = 8$*).*

3.2. Phoneme to autoencoder's components

Table 2 presents the mean P2CP distance of the tongue reconstructions, and the TBCD and TTCD trajectories correlations. Figure 3 presents the target and predicted TTCD and TBCD trajectories for two sentences in the test dataset.

Table 2: *Mean P2CP distance, x- and y- correlations for the previous and proposed (*$\dim Z = 8$*) works.*

	Previous work [9]	This work
P2CP (mm)	$\mathbf{2.83 \pm 0.86}$	2.95 ± 1.06
ρ_{TBCD}	$\mathbf{0.95 \pm 0.07}$	0.91 ± 0.11
ρ_{TTCD}	$\mathbf{0.95 \pm 0.07}$	0.88 ± 0.14

4. Discussion

4.1. Autoencoder

Table 1 shows that the four settings can provide an adequate tongue shape reconstruction, with high explained variance scores. The small reconstruction error is essential since the autoencoder's performance is a lower bound for the second model. The 12 and 16 components provided similar performances in terms of reconstruction and explained variance. However, the eight components provided a very competitive reconstruction, and we decided to continue with the more compact representation for the follow-up experiments.

The ability to control the vocal tract shape is essential for studying speech production and is valuable for many issues, e.g., the study of compensatory phenomena [18], expressions, and talking heads. From Figure 2, we observe that each component is responsible for meaningfully changing one specific aspect of the tongue. For example, the seventh component moves the tongue upwards, while the fourth component reproduces the tongue tip. These findings agree with the values displayed in Figure 1. First, the main diagonal concentrates 42% of the autoencoder's variance. Second, the covariances between components are tiny – more than $10\times$ lower than the four leading components. Third, we see that the components with the most significant variances are those with the most considerable reconstruction impact.

4.2. Phoneme to autoencoder's components

Our results show that the proposed approach provides an outstanding estimation of the tongue shape, especially compared to the theoretical lower bound imposed by the autoencoder. The tongue movements are correct, as shown by the high correlations between the ground truth and the predicted TBCD and TTCD trajectories. This new approach is very competitive compared to our previous work regarding the reconstruction error.

However, the most significant effect we should observe is the improvement in the critical articulators compared to the previous approach. Adding the tract variables as training objectives could compromise the shape reconstruction and the critical constraints. One exciting effect that we can observe in Figure 3 is that for many phonetic intervals, our model yields a complete closure between the tongue and the reference curve, even when the ground truth does not provide this information. A probable explanation for this effect is that the ground truth is noisy once it is subjected to tracking errors. These tracking errors in the target curve impose a performance upper bound in the previous approach [8]. However, since we enforce phoneme-wise constraints in the reconstruction, the critical loss inputs prior domain knowledge to the model generating a potentially more realistic result than the ground truth, which explains why the ρ_{TBCD} and ρ_{TTCD} are slightly lower than in the previous work. When the ground truth is incorrect, our model deviates from the actual curve, searching for a more physically adequate path.

It is important to note that the Figure 3 indicates all the phonemes for which the tract variable is relevant, even though it might not necessarily be minimal, e.g., TBCD is relevant for /ʃ/ and /ʒ/ but a complete closure between the tongue body and the hard palate is not necessary. Therefore, these phonemes were not included in the minimization procedure.

5. Conclusions

This paper proposed an alternative approach to tongue shape estimation using an autoencoder as an auxiliary network to solve specific issues we found in the past. By building a compact latent space, we limited the number of degrees of freedom, allowing constraints in the reconstructed curve since the model is incapable of "cheating". Additionally, the model output is sufficiently smooth without postprocessing.

Even though we restricted this study to the tongue, nothing prohibits extending to other articulators, data types, and domains. The most significant changes required would be the size of the latent space and the constraints. For example, the associated tract variable for labial phonemes is Lips Aperture (LA). Hence the constraint would be in the distance between the two predicted lips, contrarily to the case presented, in which we require the tongue to approach a pre-computed and fixed curve.

The idea proposed in this work opens up some possibilities for future work. The autoencoder provides a deterministic latent space. One alternative is to explore probabilistic models, e.g., as Variational Autoencoders [19], and Normalizing Flows [20], to learn a distribution over the components. Second, our method allows controlling the set of parameters in the autoencoder latent space, enabling the control of the vocal tract shape, which is highly important for studying speech production.

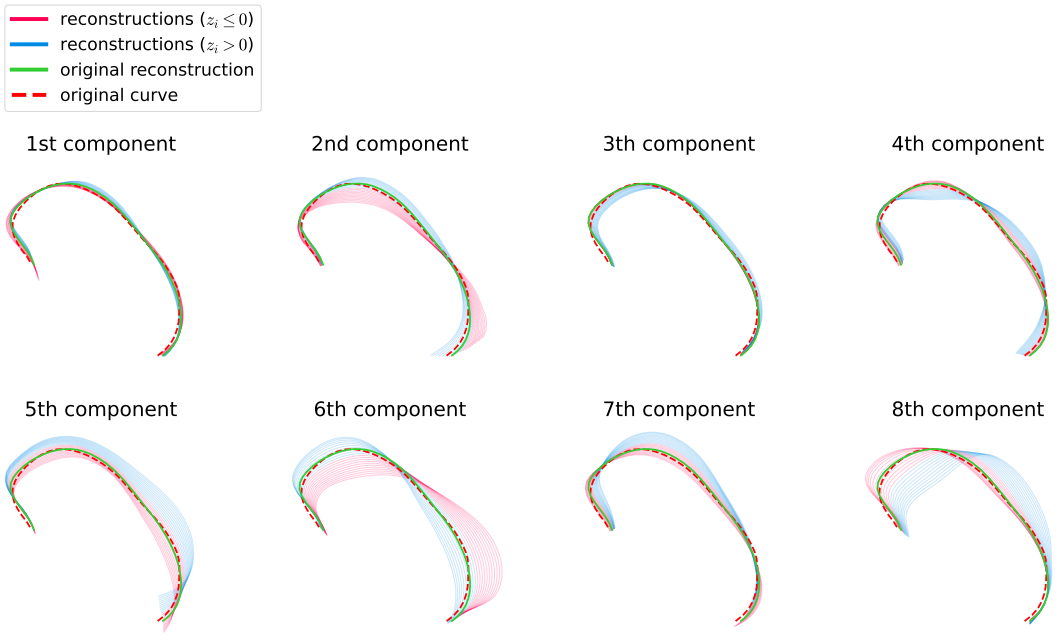

Figure 2: *The nomogram illustrates the reconstruction's effect after varying a single component in the autoencoder's latent space* ($\dim Z = 8$). *The dashed red curve represents the original curve, the solid green curve represents the reconstruction with the original encoding, and the solid pink and blue curves represent the reconstructed curve after varying a single component in the* $[-1, 1]$ *interval.*

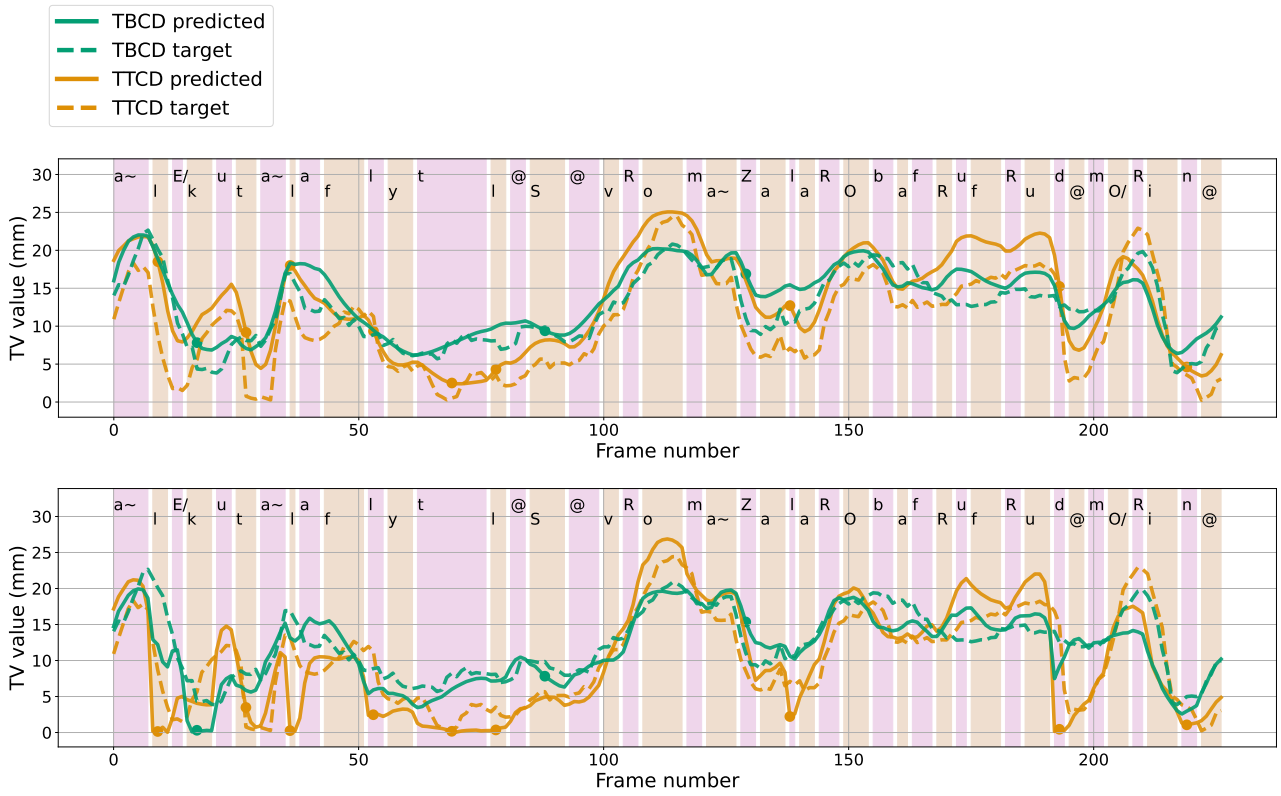

Figure 3: *TBCD and TTCD trajectories for the utterance "En écoutant la flûte, le chevreau mangea la roobe à froufrous de Maurin". The upper image refers to the previous work [9] and the bottom image refers to this work* ($\dim Z = 8$). *The circles indicate when the corresponding tract variable is relevant. The alternating colors are used only for improved visualization of the phonemes' onsets and offsets. Phonetic intervals are annotated with the corresponding phoneme.*

6. References

[1] S. Maeda, "A digital simulation method of the vocal-tract system," *Speech communication*, vol. 1, no. 3-4, pp. 199–229, 1982.

[2] Y. Laprie, M. Loosvelt, S. Maeda, R. Sock, and F. Hirsch, "Articulatory copy synthesis from cine x-ray films," in *InterSpeech-14th Annual Conference of the International Speech Communication Association-2013*, 2013.

[3] A. Toutios, T. Sorensen, K. Somandepalli, R. Alexander, and S. S. Narayanan, "Articulatory synthesis based on real-time magnetic resonance imaging data." in *Interspeech*, 2016, pp. 1492–1496.

[4] Y. Gao, S. Stone, and P. Birkholz, "Articulatory copy synthesis based on a genetic algorithm." in *INTERSPEECH*, 2019, pp. 3770–3774.

[5] Y. Gao, P. Steiner, and P. Birkholz, "Articulatory copy synthesis using long-short term memory networks," *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2020*, pp. 52–59, 2020.

[6] J. A. Tourville and F. H. Guenther, "The diva model: A neural theory of speech acquisition and production," *Language and cognitive processes*, vol. 26, no. 7, pp. 952–981, 2011.

[7] B. Grimme, S. Fuchs, P. Perrier, and G. Schöner, "Limb versus speech motor control: A conceptual review," *Motor control*, vol. 15, no. 1, pp. 5–33, 2011.

[8] V. Ribeiro, K. Isaieva, J. Leclere, P. Vuissoz, and Y. Laprie, "Towards the Prediction of the Vocal Tract Shape from the Sequence of Phonemes to be Articulated," in *Proc. Interspeech 2021*, 2021, pp. 3325–3329.

[9] V. Ribeiro, K. Isaieva, J. Leclere, P.-A. Vuissoz, and Y. Laprie, "Automatic generation of the complete vocal tract shape from the sequence of phonemes to be articulated," *Speech Communication*, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639322000607

[10] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[11] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[12] S. Ladjal, A. Newson, and C.-H. Pham, "A pca-like autoencoder," *arXiv preprint arXiv:1904.01277*, 2019.

[13] I. Douros, J. Felblinger, J. Frahm, K. Isaieva, A. Joseph, Y. Laprie, F. Odille, A. Tsukanova, D. Voit, and P. Vuissoz, "A multimodal real-time mri articulatory corpus of french for speech research," in *INTERSPEECH 2019-20th Annual Conference of the International Speech Communication Association*, 2019.

[14] D. Fohr, O. Mella, and D. Jouvet, "De l'importance de l'homogénéisation des conventions de transcription pour l'alignement automatique de corpus oraux de parole spontanée," 2015.

[15] M. Labrunie, P. Badin, D. Voit, A. A. Joseph, J. Frahm, L. Lamalle, C. Vilain, and L. Boë, "Automatic segmentation of speech articulators from real-time midsagittal mri based on supervised learning," *Speech Communication*, vol. 99, pp. 27–46, 2018.

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

[17] 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

[18] S. Maeda, "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.

[19] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[20] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *International conference on machine learning*. PMLR, 2015, pp. 1530–1538.