PinkVocalTransformer: Neural Acoustic-to-Articulatory Inversion Based on the Pink Trombone

Zhiyuan Xu(✉) and Joshua Reiss

Centre for Digital Music, Queen Mary University of London, London, UK
{zhiyuan.xu,joshua.reiss}@qmul.ac.uk

Abstract. Articulatory synthesis generates speech by modeling vocal tract configurations, but estimating articulatory parameters from audio–the acoustic-to-articulatory inversion (AAI) problem-remains challenging due to data scarcity, ambiguity, and the limitations of optimization-based methods. We propose PinkVocalTransformer, a Transformer framework that reformulates AAI as a sequence-to-sequence classification task over 44-dimensional vocal tract diameter sequences derived from the Pink Trombone physical synthesizer. By modeling complete tract shapes rather than higher-level articulatory trajectories, our approach yields a more interpretable and spatially consistent representation. To enable supervised learning, we generated over four million synthetic audio–parameter pairs under controlled static configurations. HuBERT embeddings improve feature extraction and robustness to real audio inputs. Reformulating regression as classification helps mitigate convergence issues arising from multimodal parameter distributions, leading to more stable predictions. Since ground-truth articulatory data are unavailable for real recordings, we regenerate audio from predicted parameters to indirectly evaluate reconstruction quality. Experiments show PinkVocalTransformer outperforms VAE-based and optimization baselines in vowel reconstruction. Objective ViSQOL metrics and ABX listening tests confirm higher perceptual similarity and listener preference for the regenerated audio compared to baselines. While the model performs strongly on static and simple dynamic segments, future work will focus on extending coverage to more diverse articulatory transitions and adapting the framework to more complex vocal tract models. Overall, this approach provides an efficient, data-driven framework for recovering interpretable articulatory parameters from audio, demonstrating both improved reconstruction quality and perceptual similarity compared to existing baselines.

Keywords: Acoustic-to-articulatory inversion · Transformer · Articulatory synthesis · Pink trombone

1 Introduction

Articulatory synthesis [1] generates speech by simulating vocal tract dynamics. Unlike statistical [2,3] or concatenative methods [4,5], it explicitly models artic-

A. Karpov and G. Gosztolya (Eds.): SPECOM 2025, LNAI 16188, pp. 161–173, 2026.
https://doi.org/10.1007/978-3-032-07959-6_12

ulatory motion, offering better interpretability and control. These advantages benefit linguistic research and have potential to help diagnose speech disorders and vocal tract conditions.

Several foundational models have supported articulatory synthesis, including the Liljencrants-Fant (LF) model [6] and the source-filter theory [7], which offer key insights into speech mechanics. Building on these, physical vocal tract simulators such as Pink Trombone (PT) [8] and VocalTractLab [9] have been developed to study articulatory coordination.

Despite these advances, realistic synthesis remains difficult. Black-box methods rely on optimization but face local minima and high cost. White-box approaches infer parameters analytically but are also costly and inefficient at scale.

To address these challenges, we developed a black-box deep learning method to inversely model the shape of the vocal tract. As shown in Fig. 1, our approach takes acoustic signals as input and outputs the articulatory parameters that best reconstruct the original sound. To establish this mapping, we use PT as the synthesizer and generate a dataset of more than four million static audio parameter pairs. Unlike previous studies [10,11], which often rely on traditional articulatory features, we focus on the diameters of the vocal tract as articulatory parameters, representing them as a 44-dimensional sequence [12]. Consequently, the problem can be framed as mapping acoustic representations onto a spatial sequence of articulatory diameters.

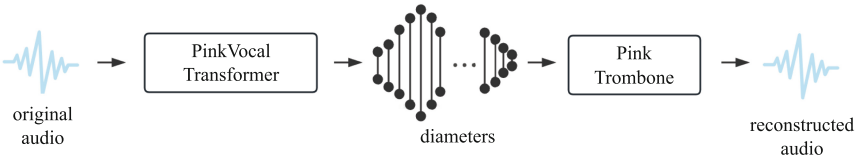

Fig. 1. Work process of the PinkVocalTransformer.

A major challenge in this approach is the instability caused by the multi-peak distribution of articulatory parameters, which complicates training. To address this, we reformulate the regression task as classification to stabilize learning. Since training only on PT data limits generalization to real-world audio, we integrate the pretrained HuBERT model [13] into the embedding layer to enhance feature extraction.

The objectives of this paper are the following.

- Propose a method that formulates acoustic-to-articulatory inversion (AAI) as a sequence-to-sequence problem and
- Evaluate the model's accuracy and robustness through systematic testing on both synthetic and real audio.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces the PT-based dataset and the proposed Transformer-based method. Section 4 presents experimental results, and Sect. 5 discusses and concludes the findings.

2 Related Work

Articulatory synthesis models the vocal tract and controls its motion to simulate articulatory behavior and generate audio. Early efforts include Kempelen's 18th-century mechanical synthesizer [14] and the computational vocal tract model introduced by Kelly and Lochbaum in 1962 [15], which laid the foundation for digital articulatory simulation. With advances in medical imaging technologies, such as magnetic resonance imaging (MRI) and computed tomography (CT), and in modeling methods, researchers have integrated structures such as lips and tongue into simulations, greatly improving precision and expanding applications beyond audio generation.

As modeling techniques [6,7,16] evolved, early work focused on building paired datasets of audio and articulatory parameters, often using codebook-based inversion methods [17,18]. Later, analysis-by-synthesis approaches [1–3] matched model-generated audio to targets using iterative optimization, but these methods were time-consuming and vulnerable to local minima [19], limiting accuracy and scalability.

Deep learning offers a more efficient alternative by directly learning the mapping between acoustic and articulatory features. Prior work typically used vocal tract parameters as targets [10,19,21], with acoustic features such as Mel spectrograms, MFCCs, and related variants. Models including convolutional neural networks (CNNs), long short-term memory networks (LSTMs), and variational autoencoders (VAEs) have all been applied to this task.

Study [11], for example, introduced a two-head VAE architecture where one decoder reconstructs the mel-spectrogram while the other predicts six Pink Trombone control parameters. The authors also explored pretrained encoders such as EnCodec and wav2vec2.0 combined with lightweight projector networks for parameter estimation. While effective, these methods rely on low-dimensional control vectors and predict all target values in a single step, implicitly assuming conditional independence, which may limit the model's ability to capture structured dependencies among articulatory positions.

In contrast, our method uses a full 44-dimensional sequence of vocal tract diameters as articulatory features, providing a more intuitive geometric representation. This formulation supports an autoregressive decoder that captures spatial dependencies across articulatory positions and produces more coherent reconstructions.

3 Datasets and Methods

This section describes our method for static acoustic-to-articulatory inversion (AAI), which maps short audio segments to vocal tract diameter sequences using

a Transformer-based architecture. We first introduce the Pink Trombone (PT) synthesizer and its articulatory parameterization, followed by the dataset construction process and modeling of glottal excitation. Finally, we explain the use of HuBERT-based feature extraction and the reformulation of regression into classification to improve training stability.

3.1 Pink Trombone

PT is a two-dimensional physical model of the human vocal tract that simulates audio production using a compact set of parameters, including constriction location, tongue location, and glottal excitation. While generating audio from these parameters is straightforward, the inverse problem remains computationally challenging. Prior PT-based studies [10] derived parameters from user interactions, with ranges listed in Table 1, but assumed tongue and constriction movements occur simultaneously, whereas they can occur independently. We address this by using 44-dimensional diameter sequences.

Table 1. Ranges of User Interaction Parameters

Parameters	Lower Bound	Upper Bound
pitch (Hz)	75	330
voiceness	0	1
tongue index	14	27
tongue diameter (cm)	1.55	3
lips diameter (cm)	0.6	1.2
constriction index	12	42
constriction diameter (cm)	0.6	1.2
throat diameter (cm)	0.5	1.0

3.2 Data

To construct the experimental dataset, we generated two types of audio based on the user interaction parameters in Table 1: one with constriction interactions and one without. Each sample was paired with its corresponding vocal tract diameter sequence, forming articulatory–acoustic mappings.

To ensure adequate parameter coverage, we used Latin Hypercube Sampling (LHS) [26], which divides each parameter's range into intervals and randomly samples one value per interval. The resulting combinations were input into Pink Trombone to generate the corresponding audio signals.

The resulting dataset contains 4,374,000 pairs, including 4,252,500 with constrictions. To improve robustness under noise, we applied augmentation by adding white noise at signal-to-noise ratios (SNRs) of 10 and 5. This not only

enhances generalization in noisy conditions but also expands the dataset. The final version is denoted as $pt\_data\_exlarge$.

3.3 Glottal Flow Derivative

PT uses the LF model to generate glottal flow derivative (GFD) waveforms, represented by the parameter R_d, which correlates with perceived vocal effort [22] and can be estimated from the GFD spectrum [23]. PT does not directly use R_d as a control parameter, but instead adopts a related parameter, Tenseness (T), defined as $T = 1 - R_d/3$. The prediction of (T) follows the approach in [24], while the fundamental frequency is predicted using the CREPE model [25].

3.4 Pretrained Models

Without pretrained models, deep models trained solely on $pt\_data\_exlarge$ perform well on PT reproduction but generalize poorly to real audio, consistent with prior findings [10,20]. This limitation arises because the training data, which are entirely generated by PT, lack speaker variability. Consequently, the model cannot distinguish speaker-dependent characteristics from the underlying articulatory content when applied to real audio, regardless of whether Mel spectrograms or MFCCs are used.

To address this, we incorporate a pretrained model for feature extraction. Given the static nature and short duration (0.125 s) of the audio, we opted against wav2vec2.0 [30], which relies on contrastive learning over long sequences and is better suited for dynamic audio tasks. In contrast, HuBERT uses unsupervised clustering-based self-supervised learning, making it more effective for capturing phonetic representations in short signals. Its ability to produce contextualized embeddings from limited temporal context enables better feature extraction for static AAI. While HuBERT embeddings improve robustness compared to conventional acoustic features, they still retain some speaker-dependent characteristics, which can introduce variability when applied to recordings with unseen speakers.

3.5 Regression to Classification

We initially implemented a Transformer-based regression model using conditional sequence modeling to predict 44 vocal tract diameters from acoustic features. Despite experimenting with both MSE and Huber loss, the model converged slowly and yielded suboptimal performance. Prior studies [27] have shown that Transformers often struggle with regression tasks due to error accumulation and high data demands. Additionally, standard loss functions that ignore dependencies among the 44 dimensions may further limit model effectiveness.

To address this, we reformulated the task as classification to better leverage Transformer architectures. We analyzed the parameter distributions with the Freedman–Diaconis rule [28], and used the resulting histogram bins as class intervals.

Although discretization introduces quantization error, it significantly improved performance. The multimodal nature of most diameter distributions further supports this approach, as classification can be viewed as a tokenization process that helps Transformers better model multimodal targets [29]. Ultimately, the 44 continuous diameter values were transformed into a classification task with 6,123 discrete categories.

3.6 PinkVocalTransformer

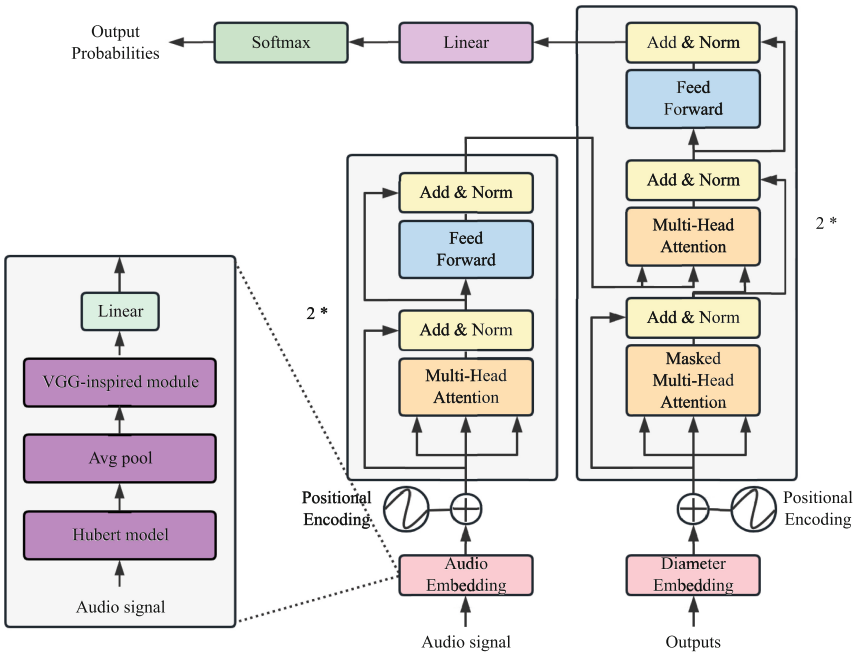

Fig. 2. Structure of PinkVocalTransformer.

Figure 2 illustrates the model architecture. Since our dataset is generated with Pink Trombone, its limited presence in real-world speech may hinder generalization. To address this, we adopt HuBERT as the core for audio feature extraction to enhance robustness.

HuBERT outputs feature vectors of shape $(time\_dimension, 1024)$. Because our dataset mainly consists of static short audio segments, the temporal variation within these sequences is limited. To reformulate the problem as a sequence-to-sequence mapping over articulatory spatial positions, we designed a VGG-inspired module that not only compresses the time dimension to 1 but also extracts higher-level acoustic representations across the HuBERT feature space. The resulting (128, 128) representation summarizes the spectral content of the

input and produces a fixed-length feature sequence along the embedding dimension. This feature sequence aligns with the spatial dependencies of the 44 articulatory diameter values and enables the model to learn their structured relationships effectively. The detailed data flow is shown in Fig. 3.

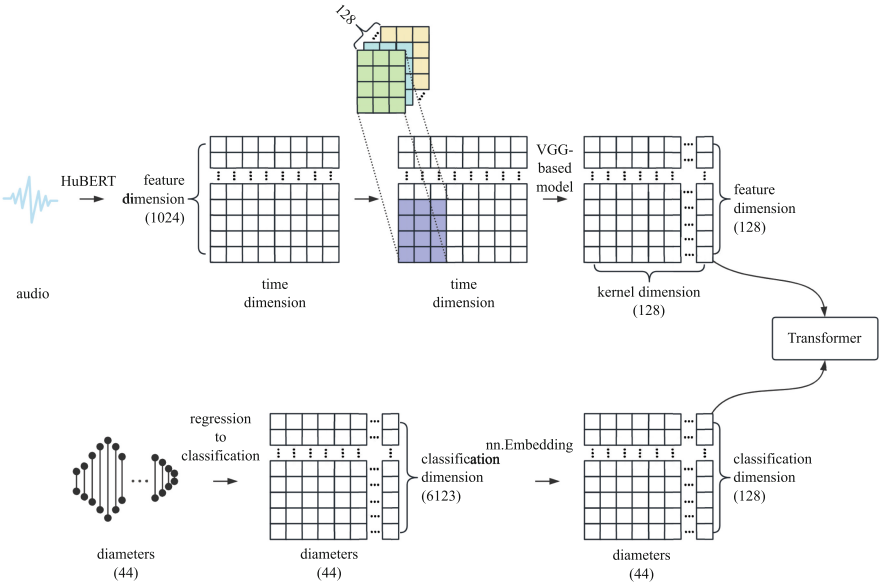

Fig. 3. Detailed data flow of PinkVocalTransformer.

During decoding, the model predicts the articulatory diameter sequence autoregressively. At each prediction step, the decoder is initialized only with a start-of-sequence token and the encoded acoustic features, without access to any ground-truth articulatory values. This prevents any information leakage between the target outputs and ensures that predictions rely solely on the learned dependencies across spatial positions. Causal masking is applied within the decoder so that each predicted position depends only on preceding predictions and not on future targets.

We find that intermediate-layer features outperform the final output for AAI tasks [31–33], offering a better balance between detail and abstraction. These features preserve acoustic and temporal cues essential for modeling articulatory motion.

4 Results

This section presents the experimental results for PinkVocalTransformer on both PT and non-PT tasks. For PT evaluation, we summarize training and validation performance. Because the training data consisted solely of static audio

segments, the evaluation of non-PT audio focuses on the model's ability to reconstruct vowel-to-vowel (VV) transitions. To assess perceptual quality, we employed ViSQOL [34] and conducted a subjective listening test. For these evaluations, each dynamic utterance in the real recordings was segmented into short frames treated as independent static inputs. The predicted articulatory sequences were then concatenated and smoothed to approximate continuous motion without requiring fully dynamic ground-truth labels.

4.1 Model Training Results

The PinkVocalTransformer was trained using the *pt_data_exlarge* dataset, which includes additive noise to improve robustness. The dataset was split into 80% training and 20% validation sets. We used the AdamW optimizer (initial learning rate of 1×10^{-4}, weight decay of 1×10^{-2}) together with a cosine annealing warm restarts scheduler ($T_0 = 10$ epochs, $T_{\text{mult}} = 1$) and early stopping (patience of 10 epochs, monitoring validation loss) to mitigate overfitting. Training was performed with a batch size of 128 and data shuffling at each epoch. The classification task was optimized using cross-entropy loss and evaluated in terms of accuracy, precision, and recall.

To recover regression targets, classification outputs were mapped to continuous diameter values and evaluated using MSE. The loss converged smoothly, indicating high training stability and effective recovery of articulatory parameters. The best model achieved an accuracy of 0.963, precision of 0.947, and recall of 0.944 on the validation set, indicating strong performance in the PT classification task.

4.2 ViSQOL Evaluation

Because our training used only synthetic audio from Pink Trombone, we needed to evaluate the generalization to real recordings. Since no ground- truth articulatory parameters exist for real speech, we adopted an indirect strategy: if the model produces reasonable parameters, the audio regenerated via Pink Trombone should approximate the original content. We therefore used VisQOL in speech mode, which focuses on intelligibility and clarity, to compare our model and baselines under realistic conditions. VisQOL outputs a score between 1 and 5, where higher values indicate greater perceptual similarity to the reference.

We evaluated our model using 24 real human audio samples from a prior AAI study [11], including 11 single-vowel, 7 slow vowel-to-vowel, and 6 complex vowel-dominant samples. The dataset includes 18 male and 6 female samples. All audio was regenerated using multiple baseline methods for comparison. Specifically, we used the outputs of the two-heads decoding VAE proposed in study [11], as well as variants employing Encodec and wav2vec2.0 embeddings as latent representations. Each baseline was tested under *fast* and *slow* configurations reflecting different levels of articulatory dynamics in the training data. We also evaluated optimization-based AAI methods from study [10], but their ViSQOL scores clustered near 2.0, so for clarity we excluded them from the figure.

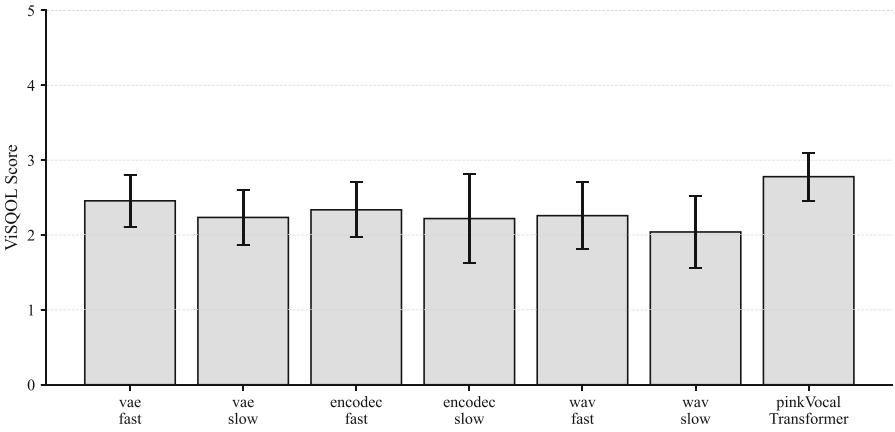

Fig. 4. Mean ViSQOL scores with standard deviations for PinkVocalTransformer and baseline models.

As shown in Fig. 4, PinkVocalTransformer outperformed all baselines in VisQOL scores. Our model achieved the highest average score, with smaller variability across samples compared to most baselines. Although the absolute scores may appear modest, this is expected since our approach targets acoustic-to-articulatory inversion rather than direct waveform synthesis. The regenerated audio is produced solely for evaluation by feeding predicted articulatory parameters into Pink Trombone, which inevitably introduces timbral and speaker-dependent differences relative to the original recordings. These differences can reduce ViSQOL scores even when the articulatory reconstruction is accurate. While other methods showed lower means and larger error bars, our results remained consistently higher and more stable. Selected audio examples are accessible via our GitHub repository [35].

4.3 Listening Test

To further validate the ViSQOL results, we conducted a single ABX discrimination test comparing our model with the strongest baseline vae_fast identified in the objective evaluation. In this test, participants were presented with a reference recording alongside two synthesized candidates and were asked to indicate which synthesized candidate more closely resembled the reference recording in terms of perceived similarity. This procedure captures perceptual similarity rather than overall audio quality. Ten representative samples were selected from the same set of 24 real audio recordings used in the ViSQOL evaluation, ensuring both phonetic diversity and manageable listener effort. In each trial, participants listened to a reference and two synthesized versions, and indicated which one more closely resembled the original. A "neither" option was included to avoid forcing decisions when no clear match was perceived. A total of 21 listeners participated in the test.

Figure 5 presents the results for each listener (L1–L21), showing the number of samples in which they selected the proposed model, the baseline, or neither. Most listeners preferred the audio generated by PinkVocalTransformer, with relatively few neutral responses. For quantitative analysis, we computed ABX accuracy as the proportion of valid trials in which the proposed model was judged closer to the reference. Trials where listeners selected "neither" were excluded from this calculation, as they do not reflect a clear perceptual preference. Based on this criterion, the accuracy reached 81.09%, reinforcing the perceptual advantage of PinkVocalTransformer over the baseline.

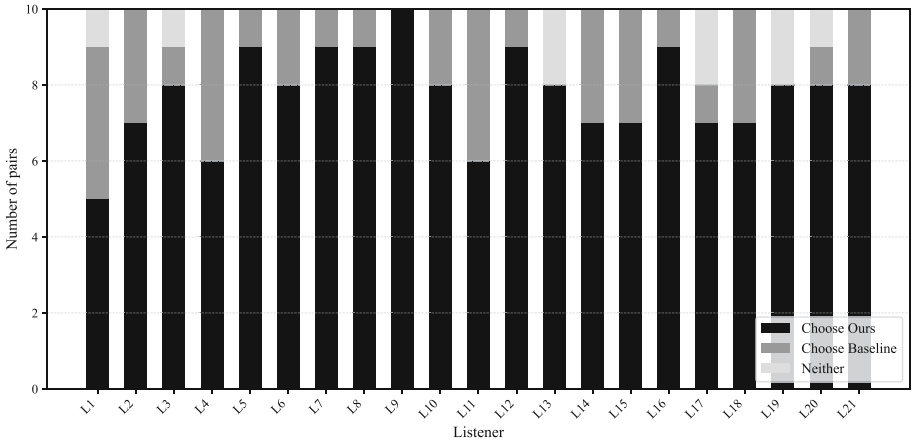

Fig. 5. Listener-level preference distribution in the ABX test.

While the model performed well on most samples, one case showed reduced accuracy. It involved a male-spoken /i/ vowel with dominant low-frequency and stable high-frequency energy. The model emphasized low-frequency cues, causing high-frequency details to be underrepresented and resulting in a dull perceptual quality. In contrast, all other test samples, including those containing /i/ under less extreme conditions, were reconstructed with high perceptual accuracy.

5 Discussion and Conclusion

PinkVocalTransformer shows strong performance in vowel reconstruction and offers a more interpretable formulation of the AAI task. Unlike optimization-based methods that estimate control parameters iteratively, our model is trained once and reused efficiently. In contrast to prior neural approaches that use PT's user-defined interaction parameters, we directly model vocal tract shape using 44 diameters, which we treat as a continuous spatial sequence rather than independent variables. This formulation enables the decoder to autoregressively predict articulatory configurations, providing a physically grounded and spatially explicit articulation model.

However, the model inherits limitations from its training setup. Trained solely on static, short-duration audio, it performs well on vowels and simple consonants but struggles to reconstruct plosives. To address this, incorporating dynamic yet brief audio and modifying the architecture to model temporal variation may help capture articulatory transitions without increasing overall complexity.

Additionally, the reliance on HuBERT embeddings introduces potential variability when applied to real recordings. Although HuBERT improves robustness compared to conventional acoustic features, it does not explicitly disentangle speaker identity from phonetic content. As a result, predictions may be partially influenced by speaker-dependent cues. Exploring more speaker-invariant representations, such as ContentVec, could help mitigate this effect and improve consistency across diverse speakers.

Another constraint arises from the synthesizer itself. Pink Trombone, being a two-dimensional articulatory model, lacks the natural acoustic richness of real vocal tracts. While this framework offers precise articulatory control, it limits the realism of generated audio.

Future work should address these challenges by exploring alternative articulatory parameterizations and more advanced synthesis techniques. Improving generalization across models and signal domains will be key to developing scalable, black-box AAI systems that remain robust and interpretable across diverse audio conditions.

Acknowledgments. The authors thank Prof. Shalom Lappin for his encouragement and valuable insights.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Richmond, K.: Estimating articulatory parameters from the acoustic speech signal. Annexe Thesis Digitisation Project 2017 Block 11 (2002)
2. Tokuda, K., et al.: Speech synthesis based on hidden Markov models. Proc. IEEE **101**(5), 1234–1252 (2013). https://doi.org/10.1109/JPROC.2013.2251852
3. Van den Oord, A., et al.: WaveNet: a generative model for raw audio. In: Proceedings of the 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9), p. 125 (2016)
4. Hunt, A.J., Black, A.W.: Unit selection in a concatenative speech synthesis system using a large speech database. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Atlanta, GA, USA, vol. 1, pp. 373–376 (1996). https://doi.org/10.1109/ICASSP.1996.541110.
5. Dutoit, T., et al.: The MBROLA project: towards a set of high quality speech synthesizers free of use for non-commercial purposes. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP), Philadelphia, PA, USA, vol. 3, pp. 1393–1396 (1996). https://doi.org/10.1109/ICSLP.1996.607874.
6. Fant, G., Liljencrants, J., Lin, Q.: A four-parameter model of glottal flow. STL-QPSR, vol. 4, no. 1985, pp. 1–13 (1985)

7. Fant, G.: Acoustic Theory of Speech Production. The Hague. Mouton, The Netherlands (1960)
8. Thapen, N.: Pink Trombone. https://dood.al/pinktrombone/. Accessed 04 July 2025
9. Birkholz, P.: Modeling consonant-vowel coarticulation for articulatory speech synthesis. PLoS ONE **8**(4), e60603 (2013). https://doi.org/10.1371/journal.pone.0060603
10. Cámara, M., et al.: Optimization techniques for a physical model of human vocalisation. In: 26th International Conference on Digital Audio Effects (DAFx), Copenhagen, Denmark, 4–7 September 2023
11. Cámara, M., et al.: Decoding vocal articulations from acoustic latent representations. In: Proceedings of the AES Europe Convention, Madrid, Spain (2024)
12. Mathur, S., Story, B., Rodriguez, J.: Vocal-tract modeling: Fractional elongation of segment lengths in a waveguide model with half-sample delays. IEEE Trans. Audio Speech Lang. Process. **14**, 1754–1762 (2006)
13. Hsu, W.-N., et al.: HuBERT: self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 3451–3460 (2021). https://doi.org/10.1109/TASLP.2021.3122291
14. Dudley, H., Tarnoczy, T.H.: The speaking machine of Wolfgang von Kempelen. J. Acoust. Soc. Am. **22**(2), 151–166 (1950). https://doi.org/10.1121/1.1906583
15. Kelly, K.L., Lochbaum, C.C.: Speech synthesis. In: Proceedings of the Fourth ICA (1962)
16. Story, B.H.: A parametric model of the vocal tract area function for vowel and consonant simulation. J. Acoust. Soc. Am. **117**(5), 3231–3254 (2005)
17. Chenoukh, S., et al.: Voice mimic system using an articulatory codebook for estimation of vocal tract shape. In: Proceedings of the EuroSpeech-97, Rhodes, pp. 429–432 (1997)
18. Ouni, S., Laprie, Y.: Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. J. Acoust. Soc. Am. **118**(1), 444–460 (2005). https://doi.org/10.1121/1.1921448
19. Sorokin, V.N., Leonov, A.S., Trushkin, A.V.: Estimation of stability and accuracy of inverse problem solution for the vocal tract. Speech Commun. **30**(1), 55–74 (2000)
20. Saha, P., et al.: Learning joint articulatory-acoustic representations with normalizing flows. In: Proceedings of the Interspeech (2020)
21. Pasad, A., Shi, B., Livescu, K.: Comparative layer-wise analysis of self-supervised speech models. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2023, Rhodes Island, Greece, pp. 1–5 (2023). https://doi.org/10.1109/ICASSP49357.2023.10096149
22. Lu, H.-L., Smith, J.O.: Glottal source modeling for singing voice synthesis. In: Proceedings of the ICMC (2000)
23. Fant, G.: The LF-model revisited: transformations and frequency domain analysis. STL-QPSR, vol. 2, no. 3 (1995)
24. Südholt, D., et al.: Vocal tract estimation by gradient descent. In: 26th International Conference on Digital Audio Effects (DAFx), Copenhagen, Denmark, 4–7 September 2023
25. Kim, J.W., et al.: Crepe: a convolutional representation for pitch estimation. In: Proceedings of the ICASSP, pp. 161–165 (2018). https://doi.org/10.1109/ICASSP.2018.8461329
26. Iman, R.L., Davenport, J.M., Zeigler, D.K.: Latin hypercube sampling (program user's guide) (1980)

27. Nath, S., Khadilkar, H., Bhattacharyya, P.: Transformers are expressive, but are they expressive enough for regression? arXiv preprint arXiv:2402.15478 (2024)
28. Freedman, D., Diaconis, P.: On the histogram as a density estimator: L2 theory. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **57**(4), 453–476 (1981). https://doi.org/10.1007/BF01025868
29. Xu, P., Zhu, X., Clifton, D.A.: Multimodal learning with transformers: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **45**(10), 12113–12132 (2023). https://doi.org/10.1109/TPAMI.2023.3275158
30. Baevski, A., et al.: Wav2vec 2.0: a framework for self-supervised learning of speech representations. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS), Red Hook, NY, USA, Article no. 1044, pp. 12449–12460. Curran Associates Inc. (2020)
31. Chang, H.-J., Yang, S., Lee, H.-Y.: DistilHuBERT: speech representation learning by layer-wise distillation of hidden-unit BERT. In: Proceedings of the ICASSP, pp. 7087–7091 (2022). https://doi.org/10.1109/ICASSP43922.2022.9747490
32. Kumar, P., Sukhadia, V.N., Umesh, S.: Investigation of robustness of HuBERT features from different layers to domain, accent and language variations. In: Proceedings of the ICASSP, pp. 6887–6891 (2022). https://doi.org/10.1109/ICASSP43922.2022.9746250
33. Yoon, J.W., Woo, B.J., Kim, N.S.: HuBERT-EE: early exiting HuBERT for efficient speech recognition. In: Proceedings of the Interspeech, pp. 2400–2404 (2024). https://doi.org/10.21437/Interspeech.2024-80
34. Chinen, M., et al.: ViSQOL v3: an open source production ready objective speech and audio metric. In: 2020 QoMEX, pp. 1–6. IEEE (2020)
35. Xu, Z.: PinkVocalTransformer Project Page. https://zhiyuanxu27.github.io/pinkVocalTransformer/. Accessed 04 July 2025