Deep Speech Synthesis from Articulatory Features

Anonymous ACL submission

Abstract

In the articulatory synthesis task, speech is synthesized from input features containing information about the physical behavior of the human vocal tract. This task provides a promising direction for speech synthesis research, as the articulatory space is compact, smooth, and interpretable. Current works have highlighted the potential for deep learning models to perform articulatory synthesis. However, it remains unclear whether these models can achieve the efficiency and fidelity of the human speech production system. To help bridge this gap, we propose a time-domain articulatory synthesis methodology and demonstrate its efficacy with both electromagnetic articulography (EMA) and synthetic articulatory feature inputs. Our model is both computationally efficient and highly intelligible, achieving a transcription word error rate (WER) of 7.14% for the EMA-to-speech task. Through interpolation experiments, we also highlight the generalizability and interpretability of our approach.

1 Introduction

Speech synthesis has seen rapid development in recent years with deep learning based techniques. These models have shown success in tasks like text-to-speech (TTS) (Wang et al., 2017; Hayashi et al., 2021; Prenger et al., 2019), speech-to-speech translation (S2ST) (Tjandra et al., 2019; Jia et al., 2019; Inaguma et al., 2020), voice conversion (VC) (Polyak et al., 2021; Wu et al., 2021a; Sisman et al., 2020), and more (Anumanchipalli et al., 2019; Yu et al., 2019; Gaddy and Klein, 2021). Moreover, this technology has yielded impactful technologies like speech synthesis aids for people with blindness or paralysis (Karmel et al., 2019; Angrick et al., 2019; Anumanchipalli et al., 2019). While speech synthesizers have already shown promising results for assistive tasks in healthcare and other challenging domains, technologies like brain-to-speech devices are still nascent and require new algorithms in order to be deployed as high-fidelity, open-vocabulary synthesizers. To this end, our work focuses on devising a deep speech synthesis methodology that is computationally efficient, real-time, and high-fidelity. We propose a time-domain articulatory synthesis approach that is suitable for attaining these three properties and empirically validate our method on two distinct articulatory modalities, EMA and a synthetic articulatory space. Our deep learning models also exhibit valuable interpretability properties, which we demonstrate through interpolation experiments.

We proceed by discussing speech synthesis in the context of deep learning and articulatory synthesis in Section 2. In Section 3, we describe our deep articulatory models and time-domain methodology. Then, we discuss the two articulatory datasets chosen for our empirical studies and their respective modalities in Section 4. With these datasets, we conduct computational efficiency, interpolation, and synthesis quality studies, discussed in Sections 5, 6, and 7, respectively. We then provide further analyses with respect to phoneme confusability in Section 8. Finally, we summarize our results and propose future directions in Section 9. Audio samples and additional related information are all available at https://articulatorysynthesis.github.io.

2 Speech Synthesis

2.1 Deep Speech Synthesis

Currently, state-of-the-art speech synthesis algorithms use deep learning (Hayashi et al., 2021; Anumanchipalli et al., 2019; Jia et al., 2021; Polyak et al., 2021; Gaddy and Klein, 2021). While existing methods can generate high-fidelity speech, they tend to be computationally expensive and difficult to interpret and generalize (Nekvinda and Dušek, 2020; Zhang et al., 2019). We attribute underspecification to the primary cause of these issues, as speech data is very high dimensional and current algorithms lack sufficient inductive biases. To help bridge this gap, we devise deep articulatory synthesis techniques that exhibit suitable computational efficiency, generalizability, and interpretability properties by behaving more similarly to the human speech production process than existing methods.

2.2 Articulatory Synthesis

Articulatory synthesis generally refers to the task of synthesizing speech from articulatory features, i.e., features containing information about the physical behavior of the human vocal tract (Fant, 1991; Rubin et al., 1981; Scully, 1990). We identify two primary research directions in articulatory synthesis: 1. modelling the human vocal tract (Fant, 1995; Iskarous et al., 2003; Birkholz, 2013a), and 2. learning the mapping from articulatory features to speech through a statistical means (Aryal and Gutierrez-Osuna, 2016; Bocquelet et al., 2014; Chen et al., 2021). The former direction, due to its focus on computational modelling, has yielded articulatory synthesizers that are interpretable and relatively space-efficient but computationally slow. On the other hand, the latter direction has yielded methods that are much faster but have worse interpretability and memory efficiency. Ideally, speech synthesizers should have low space and time complexities, which would enable many impactful real-time applications. For example, such systems could allow patients with paralysis or aphasia to communicate naturally at any moment in time. Thus, we focus on making methods in the second research direction more memory-efficient in this work. Additionally, we highlight how statistical articulatory synthesis methods could also be highly interpretable, thus containing all of the benefits of articulatory synthesizers built using physical modelling.

We also focus on the statistical research direction in this work because of the transferability of our methodology to all forms of speech synthesis. Current state-of-the-art speech synthesis systems rely on an intermediate speech representation, typically a spectrum or a learned representation (Kong et al., 2020; Morrison et al., 2022; Badlani et al., 2021; Kim et al., 2021; Elias et al., 2021). Inductive biases offer one potential way of making these models efficient, generalizable, and interpretable as mentioned in Section 2.1. Constraining these intermediate representations to an articulatory feature space is one way to impose such an inductive bias, especially since there is a limited set of articulator configurations that can completely specify all possible human speech. The resulting model would then need to perform an articulatory-to-speech mapping, of which the behavior is relatively unknown to our knowledge. This work aims to bridge this gap by studying the efficiency, generalizability, interpretability, and fidelity of such a mapping using two distinct articulatory modalities, EMA and a synthetic one generated using a vocal tract model, detailed in Section 4.

While deep EMA-to-speech models have been previously studied, as far as we are aware (Taguchi and Kaburagi, 2018; Stone et al., 2020; Liu et al., 2018), current models are not highly intelligible, achieving a transcription WER of around 30% on open-vocabulary tasks (Taguchi and Kaburagi, 2018). In this work, we build an EMA-to-speech model that achieves a transcription WER of 7.14% and perform detailed error analyses on the synthesized utterances. We also extend this approach to building a speech synthesizer using a synthetic articulatory modality. This model is efficient, high-fidelity, and interpretable, which has previously been unattained to our knowledge. We detail these models and our proposed time-domain articulatory synthesis methodology in Section 3 below.

3 Deep Articulatory Models

3.1 Frequency- and Time-Domain Modeling

Similarly to the state-of-the-art speech synthesis works discussed in Section 2, current deep articulatory synthesis works rely on synthesizing an intermediate spectrum representation, from which waveforms are generated (Csap’o et al., 2020; Georges et al., 2020). Since this behavior is not present in the human speech production process, we propose a model that directly maps articulatory features to waveforms in this work. Since this model does not explicitly rely on a frequency-based intermediate, we refer to this approach as a time-domain one. This modification noticeably improves model efficiency while achieving comparable intelligibility on our two datasets, as discussed in Sections 5 and 7. We proceed to discuss our spectrum-intermediate baseline in Section 3.2 and our two time-domain methods in Sections 3.3 and 3.4.

3.2 Spectrum-Intermediate Baseline

For our baseline deep learning model, we build on the state-of-the-art articulatory synthesis architecture proposed by Gaddy and Klein (Gaddy and Klein, 2021). Namely, we map articulatory features to spectrums using a six-layer Transformer (Vaswani et al., 2017) prepended with three residual convolution blocks. To map spectrums to waveforms, we use HiFi-GAN (Kong et al., 2020), which havs been shown to perform better than the WaveNet vocoder used by Gaddy and Klein (Gaddy and Klein, 2021). For our spectrum representation, we use Mel spectrograms instead of MFCCs, as done in the HiFi-GAN paper and most deep speech synthesis works (Kong et al., 2020; Wang et al., 2017; Hayashi et al., 2021).

We also modify the loss function used by Gaddy and Klein (Gaddy and Klein, 2021). To avoid requiring phoneme annotations to train the model, we omit the phonemic loss. We instead improve model performance by adding the adversarial loss used by HiFi-GAN (Kong et al., 2020). Since our data in this work has sequences of articulatory features that are pre-aligned with waveforms, we also do not need the dynamic time warping loss. We refer to this resulting baseline as the spectrum-intermediate (Spec.-Int.) model below.

In all of our experiments, we train the Transformer model using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of $3.0 * 10^{-5}$ for both the generator and the discriminators, a batch size of 32, and loss balancing coefficients matching those used with the original HiFi-GAN model (Kong et al., 2020). Our discriminator architectures and HiFi-GAN spectrum-to-speech vocoder parameters also match those of Kong et al. (Kong et al., 2020), and our Transformer has a hidden dimension of 1024 and a dropout rate of 0.2.

3.3 Time-Domain HiFi-GAN

For our first time-domain model, we feed our articulatory input features directly into HiFi-GAN (Kong et al., 2020), keeping the architecture and loss functions the same while changing the input modality. To our knowledge, directly feeding articulatory inputs into a deep vocoder architecture has not yielded any successful results previously. However, we observe that this model is comparable to our baseline, as discussed in Section 7. Moreover, removing the need for an articulatory-to-spectrum architecture noticeably improves computational efficiency, as discussed in Section 5. For all of our experiments, we optimize this model using the same hyperparameters as the HiFi-GAN spectrum-to-speech vocoder used in the Section 3.2 baseline above.

3.4 NSF-CAR Model

For our second time-domain model, we build on the neural source-filter (NSF) architecture (Wang et al., 2019). Since articulatory features can be divided into source- and filter-related attributes (Birkholz, 2013a), we experiment with this architecture in order to study whether explicitly modelling this separation could improve articulatory synthesis performance.

Similarly to our baseline, we use the loss function from HiFi-GAN to improve synthesis fidelity. We also leverage autoregression to improve the pitch and periodicity of model outputs and make our model a streaming-based one. Namely, we incorporate the autoregressive encoder from CARGAN (Morrison et al., 2022) into our model, concatenating its output with each vector in the condition module input sequence. We replace the convolutions in the NSF condition module with GBlock layers (Morrison et al., 2022), which we found to further improve model performance. Figure 7 in the Appendix depicts the architecture of our generator.

To our knowledge, neural source filter models are currently only used for building vocoders that map spectrums to speech (Wang et al., 2019; Georges et al., 2020). In this work, we leverage source-filter modelling to perform articulatory synthesis without relying on an intermediate spectrum representation.

3.5 WSOLA

As observed by Morrison et al. (Morrison et al., 2022), simply concatenating the output chunks generated through an autoregressive process yields artifacts at the concatenation points. Thus, during evaluation, we join outputs using an approach based on WSOLA. Namely, we overlap-and-add adjacent output chunks at intersections with maximum cross-correlation, sliding the chunks up to a distance of one pitch period. We calculate a pitch period by multiplying the sampling rate with the reciprocal of the last F0 value in the first chunk input. Figure 1 depicts one such WSOLA operation.

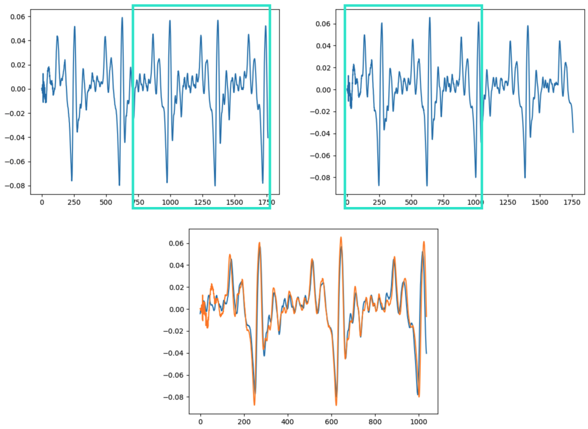

Figure 1: WSOLA-based method for concatenating waveforms.

4 Datasets

4.1 Electromagnetic Articulography (EMA)

For our first task, we perform EMA-to-speech using the MNGU0 dataset (Richmond et al., 2011), which contains 67 minutes of single-speaker speech recorded at 16 kHz annotated with 12-dimensional EMA features recorded at 200 Hz. We use the train-test split provided in the original work, which has 1,129 utterances for training and 60 for testing. Among the 1,129 training utterances, we set off a random size-60 subset for validation. Since EMA on its own does not contain voicing information, we concatenate estimated F0 sequences extracted using CREPE (Kim et al., 2018; Morrison et al., 2022) to the EMA features, forming a 13-dimensional input feature.

4.2 Synthetic Articulatory Features

Since EMA data does not contain enough manner information to perfectly reconstruct the original speech, we also experiment with synthetic articulatory data that does. Namely, we use the vocal tract model from Birkholz et al. (Birkholz, 2013a) to create a single-speaker corpus of pseudo-words, each composed of two to three vowel and consonant sounds. Our training set has 10,000 such utterances, and our validation set has 250, totaling a few hours of speech. For our evaluation set, we use the Birkholz vocal tract model outputs corresponding to the first 99 phoneme sequences in the CMU US KAL Diphone database (Lenzo and Black, 2000). All waveforms have a sampling rate of 44100 Hz and articulatory features are recorded every 110 samples. We refer to this dataset as the Birkholz-Pseudoword (Birk.-Pseudo.) dataset below. In this dataset, our articulatory features are 30-dimensional.

5 Computational Efficiency

Computational efficiency during training is essential for low-resource speech synthesis tasks like brain-to-speech and other articulatory synthesis tasks where data collection is expensive. During inference, computational efficiency is essential for building real-time speech synthesizers, e.g., for brain-to-speech. We observe that our time-domain articulatory synthesis model has some suitable computational efficiency properties compared to the frequency-domain baseline. As shown in Table 1, our model is able to train twice as fast as the baseline on a single RTX 2080 Ti GPU for the task with synthetic articulatory data. While our model synthesizes utterances slower than the baseline due to the nature of autoregression (Morrison et al., 2022), we observe that generation on a CPU is still faster than real-time.

Compared to the baseline, our time-domain models are much more memory efficient, as detailed in Table 2. Our models are able to use over 8 to 20 times less number of parameters than the baseline due to their ability to directly map articulatory features to speech. Namely, while current articulatory synthesis models like our baseline rely on two components, one to output spectrums and another to convert spectrums to waveforms, our time-domain models only contain one. We note that the real-time and memory efficient properties of our time-domain models make them a viable choice for streaming, on-device tasks.

Data	**Birk.-Pseudo.**	**EMA-MGNU0**
NSF-CAR	34	81
HiFi-GAN	8	9
Spec.-Int.	68	80

Table 1: Total training time for each model in hours.

Model	**Birk.-Pseudo.**	**EMA-MGNU0**
NSF-CAR	$4.4 * 10^6$	$4.2 * 10^6$
HiFi-GAN	$14.2 * 10^6$	$12.6 * 10^6$
Spec.-Int.	$98.7 * 10^6$	$94.0 * 10^6$

Table 2: Number of parameters of each model.

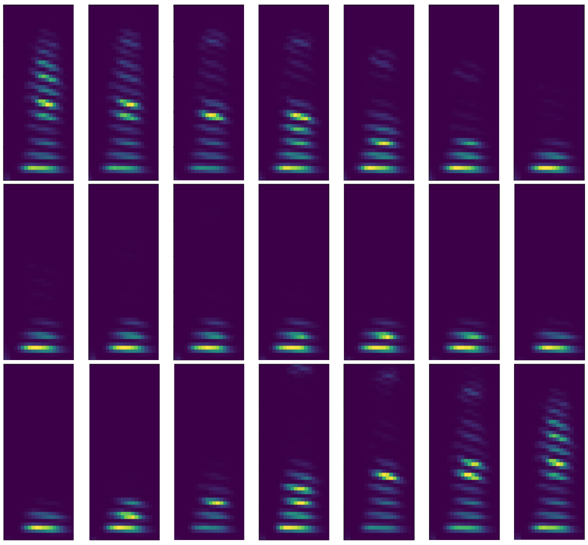

Figure 2: Vowel interpolation. The top row contains the synthesized samples between the "ta" and "tu" sounds, the middle row "tu" and "ti", and the bottom row "ti" and "ta".

6 Interpolation

6.1 Vowel Interpolation

To study the generalizability of our time-domain model, we perform interpolation experiments. First, to analyze how well our model generalizes across vowel sounds, we perform vowel interpolation. Namely, we interpolate between the "ta" and "tu" sounds, "tu" and "ti", and "ti" and "ta" using the synthetic articulatory data. We generate the articulatory features for "ta", "tu", and "ti" using the code provided by Birkholz et al., similarly to our approach for creating the synthetic articulatory dataset described above. For each of the three pairs of sounds, we perform a linear interpolation between the two articulatory features, generating seven evenly spaced weighted combinations. The figures below are generated using outputs from our NSF-CAR model, and we observe similar trends with our time-domain HiFi-GAN as well, which we include in the supplementary website linked in Section 1.

Figure 2 contains the mel-spectrograms of the generated speech from our model for each of these combined articulatory features. Our model is able to generalize to the unseen articulatory features between the three sounds. Moreover, the transitions between spectrum values in each interpolation are smooth, suggesting that our network is able to model the continuity of articulator movements, at least with respect to vowels.

6.2 Consonant Interpolation

We also study the generalizability of our model with respect to consonants. To study how well our model generalizes across types of consonant sounds, we fix the place of articulation and interpolate between consonant types. Namely, we interpolate between the alveolar consonants "ra", "na", and "la", using the same methodology as our vowel interpolation experiment in Section 6.1.

Figure 3 depicts the mel-spectrograms of synthesized interpolation samples from our time-domain articulatory synthesis model. Similarly to our vowel interpolation results, we observe that our model generalizes to the unseen samples between the three consonants and exhibits smooth generation. Specifically, these results indicate that our model can smoothly transition between nasal, approximant, and lateral approximant consonants, similarly to the human speech production process.

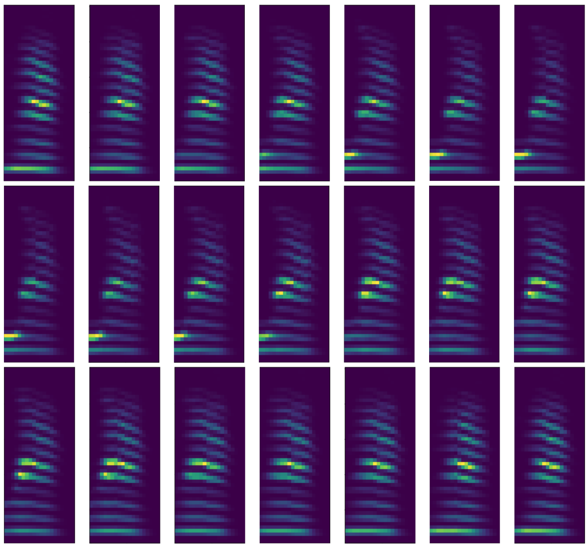

Figure 3: Alveolar consonant interpolation. The top row contains the synthesized samples between the "ra" and "na" sounds, the middle row "na" and "la", and the bottom row "la" and "ra".

To study how well our model generalizes across place of articulation, we fix the consonant type and interpolate between two places. Namely, we interpolate between the approximant consonants "ra" and "ja", using the same aforementioned methodology. Figure 4 depicts these results. As with our alveolar consonant interpolation results, we observe that our model generalizes to unseen samples and produces smooth transitions between synthesized interpolation samples here.

To quantify how the synthesized utterances change across the interpolation, we create two plots studying changes in the magnitudes of different bands of the mel-spectrogram. Namely, our first graph plots the magnitude of each mel-spectrogram frequency vector across the seven utterances, going left to right in the interpolation. Our second plot does the same with time vectors, i.e., columns in the mel-spectrograms. We compute the magnitude of each vector using the L1 norm, which is just a sum here since mel-spectrogram values are non-negative. To improve readability in both plots, we omit vectors that on average change less than 0.3 in magnitude between adjacent interpolation samples.

As shown in the bottom row of Figure 4, the vector magnitude lines are generally monotonic and almost linear in many cases when going left to right in the interpolation. This supports our hypothesis that our model has learnt to transition smoothly between consonants when synthesizing articulatory features.

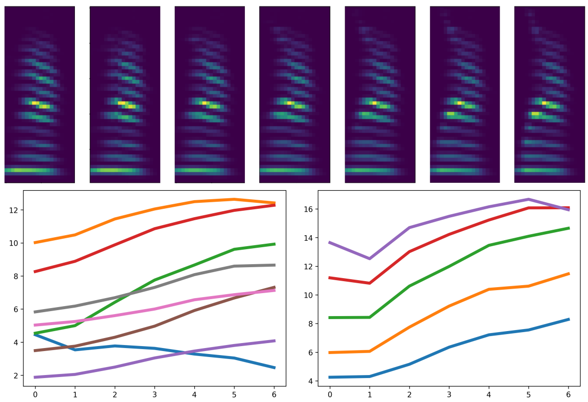

Figure 4: Approximate consonant interpolation. *Top row:* synthesized samples between the "ra" and "ja" sounds. *Bottom row left:* frequency vector magnitudes for each spectrum. *Bottom row right:* time vector magnitudes for each spectrum.

6.3 Interpretability

We note that these interpolation results also highlight the interpretability of articulatory features. Namely, we are able to simply take an element-wise weighted sum of two same-length sequences of articulatory features in order to create the utterance corresponding to articulator movements in between the two gestures. For example, to create the "tɛ" sound, we would just need to synthesize the average of the articulatory feature sequences for "ti" and "ta". To our knowledge, this degree of interpretability is not supported by other speech representations like spectrums or deep-learning-based ones.

7 Synthesis Quality

7.1 Fidelity

Since MCD serves as an objective measure of synthesis quality (Black, 2019), we first measure synthesis fidelity using this metric. As detailed in Table 3, we observe that our time-domain articulatory synthesis approach achieves performance comparable to the frequency-domain baseline. Namely, our approach performs noticeably better than the baseline on the synthetic articulatory dataset and slightly worse on the EMA-to-speech task. Given these results, we attribute the performance drop of our model on the EMA task to information loss within in the input data. Namely, the model appears to confuse phonemes due to the lack of manner information in the EMA inputs, which can be heard in the accompanying samples. We discuss this phoneme confusion in more detail below.

Model	**MCD**	
	Birk.-Pseudo	**EMA-MGNU0**
NSF-CAR	3.36 ± 0.28	5.44 ± 0.67
HiFi-GAN	$\mathbf{2.90 \pm 0.22}$	4.81 ± 0.76
Spec.-Int.	5.15 ± 0.48	$\mathbf{4.75 \pm 0.81}$

Table 3: MCD for each model on Birkholz and EMA data.

7.2 Automatic Speech Recognition

To evaluate the intelligibility of our synthesis approach, we conduct open-vocabulary transcription experiments for the EMA-to-speech task with our time-domain HiFi-GAN model described in Section 3.3. First, we perform an objective evaluation using deep automatic speech recognition (ASR) models. Specifically, we use DeepSpeech[1] (Hannun et al., 2014) as done by Gaddy and Klein (Gaddy and Klein, 2021) as well as the ESPnet Conformer ASR model trained on LibriSpeech[2] (Guo et al., 2021; Panayotov et al., 2015). We use these models to transcribe the synthesis outputs of our model on the entire MNGU0 evaluation set described in Section 4.1 and calculate the average word error rates (WERs) and character error rates (CERs). Since some utterances in the evaluation set contain proper nouns, we also compute ASR

[1] https://github.com/mozilla/DeepSpeech
[2] https://zenodo.org/record/4604066#.YeNA0i2z2CM

metrics on all of the evaluation set utterances composed entirely of common nouns, which form a 32-utterance subset.

Table 4 summarizes our ASR results. On the common-noun subset, our model achieves a character error rate of 10.7% with the ESPnet ASR model, indicating that our model is able to synthesize intelligible speech. The consistent differences between the WER and CER values as well as the entire set and common-noun subset performances suggests that these ASR metrics may be underestimating intelligibility, as also observed by Gaddy and Klein (Gaddy and Klein, 2021). Thus, we also evaluate the intelligibility of our model though human evaluations, as discussed in Section 7.3 below.

ASR Model	WER		CER	
	All	Com.	All	Com.
ESPnet	32.9	19.2	17.9	10.7
DeepSpeech	41.3	32.9	20.2	15.5

Table 4: ASR. entire evaluation set (All) and common noun subset (Com.).

7.3 Human Evaluation

To further understand the intelligibility of our time-domain articulatory synthesis approach, we also perform open-vocabulary transcription tests with human listeners, evaluating our same time-domain HiFi-GAN model (Section 3.3) used in our Section 7.2 ASR experiments above. Namely, we randomly select ten utterances from our EMA corpus evaluation set, choosing among the 32 sentences without proper nouns. Based on the transcriptions from six English-speaking listeners, our model achieves an average WER of 7.14%, indicating that our model is able to produce intelligible speech. To our knowledge, this value is noticeably lower than prior results, which are around 30.1% (Taguchi and Kaburagi, 2018). This suggests that our time-domain articulatory synthesis methodology is a suitable approach for efficiently performing speech synthesis while achieving high intelligibility.

8 Phoneme Confusion

To further study the phonological errors made by our model, we analyze the phonemes that our EMA-to-speech model confused during synthesis. Namely, we study phoneme confusability for our time-domain HiFi-GAN model (Section 3.3) through the transcriptions, both from the ASR ones described in Section 7.2 and the human ones described in Section 7.3. For each transcribed utterance, we convert the graphemes to a phoneme sequence using Phonemizer[3] (Bernard and Titeux, 2021) and their eSpeak NG backend,[4] and repeat this grapheme-to-phoneme conversion with the ground truth texts. We identify the phoneme confusion pairs using sclite,[5] which aligns each predicted sequence with the respective ground truth and then records the substitution errors.

For our human evaluation analysis, we use all of the transcripts from the six listeners, i.e., 60 utterances. Figure 5 depicts the resulting phoneme confusion pairs. We plot these confusion pairs on an International Phonetic Alphabet (IPA) chart that extends the one from Gaddy and Klein to more phonemes (Gaddy and Klein, 2021), indicating pairs with a higher frequency of substitution errors using darker lines. We also populate this IPA chart with our confusion pairs from the ASR transcriptions in Figure 6, for which we use the texts transcribed by the ESPnet model for the entire MNGU0 evaluation set, as discussed in Section 7.2. We omit the phoneme pairs that are only confused once in Figure 6 in order to improve readability.

From these two IPA charts, we observe that the most of the word substitution errors are due to plosive or vowel confusions. Since the primary vowel confusions in Figure 5 differ from those in Figure 6, we hypothesize that vowel confusability for human evaluators mainly resulted from the substitution of vowels to form logical, grammatically correct words and phrases. The automatic transcribers may not have as much of such bias and we observe that the primary confused vowel pairs are relatively close to each other with our ASR-based results, reinforcing this hypothesis. One potential reason for the plosive substitutions is that plosives generally have a shorter duration than other consonant types like fricatives (Alwan et al., 2011) and thus may be more readily confusable. Among the plosives, "p", "b", "t", and "d" may have been easier to confuse than "k" and "g" for the human evaluators because the latter two plosives have longer voice onset times, a pattern also observed by Birkholz (Birkholz, 2013b). From Figure 6, we also observe that multiple voiced-unvoiced pairs are confused. We hypothesize that this is because the only voic-

[3] https://github.com/bootphon/phonemizer
[4] https://github.com/espeak-ng/espeak-ng
[5] https://github.com/usnistgov/SCTK

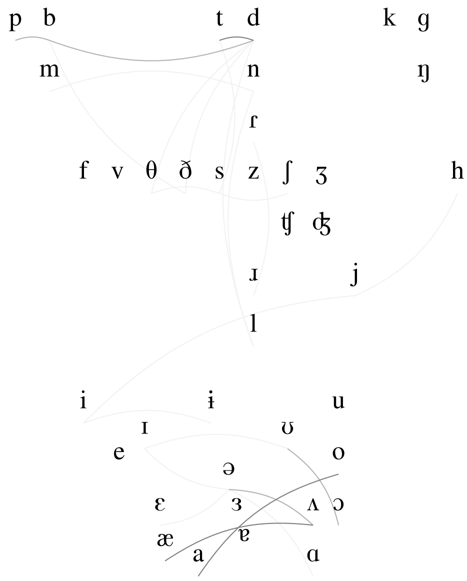

Figure 5: Phoneme confusability based on human transcriptions. Phoneme pairs that are confused more frequently have darker lines.

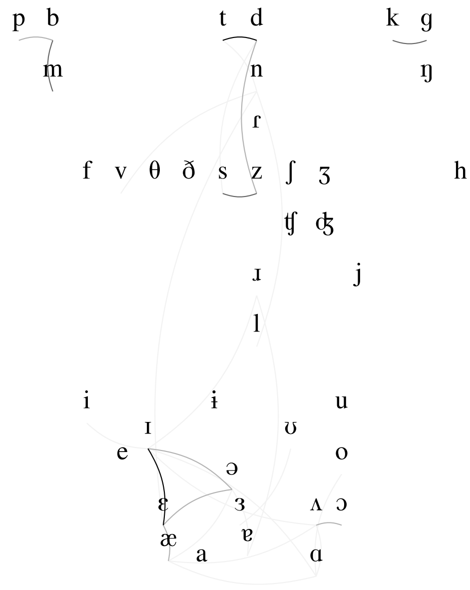

Figure 6: Phoneme confusability based on ASR transcriptions. Phoneme pairs that are confused more frequently have darker lines.

ing information that our EMA-to-speech model receives as input is the estimated F0 sequence, as described in Section 4.1.

9 Conclusion and Future Directions

In this work, we study ways to build deep articulatory synthesizers that are efficient and high-fidelity. Based on computational efficiency evaluations, we observe that our proposed time-domain methodology is suitable for achieving time and space complexities that are noticeably lower than the baseline spectrum-intermediate approach. Our interpolation study also highlights the generalizability and interpretability of our approach. Through MCD, ASR, and human transcription experiments, we demonstrate that our model is also highly intelligible, achieving a transcription word error rate (WER) of 7.14% for the EMA-to-speech task. Moving forward, we plan to test our methodology on other modalities like electromyography (EMG) (Gaddy and Klein, 2021) and real-time magnetic resonance imaging (RT-MRI) (Lim et al., 2021). We also plan to extend our approach to multi-speaker and multilingual settings (Richmond et al., 2011; Lim et al., 2021; Wu et al., 2021b).

References

Abeer Alwan, Jintao Jiang, and Willa Chen. 2011. Perception of place of articulation for plosives and fricatives in noise. *Speech communication*, 53(2):195–209.

Miguel Angrick, Christian Herff, Emily Mugler, Matthew C Tate, Marc W Slutzky, Dean J Krusienski, and Tanja Schultz. 2019. Speech synthesis from ecog using densely connected 3d convolutional neural networks. *Journal of neural engineering*, 16(3):036019.

Gopala K Anumanchipalli, Josh Chartier, and Edward F Chang. 2019. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498.

Sandesh Aryal and Ricardo Gutierrez-Osuna. 2016. Data driven articulatory synthesis with deep neural networks. *Computer Speech & Language*, 36:260–273.

Rohan Badlani, Adrian Łancucki, Kevin J Shih, Rafael Valle, Wei Ping, and Bryan Catanzaro. 2021. One tts alignment to rule them all. *arXiv preprint arXiv:2108.10447*.

Mathieu Bernard and Hadrien Titeux. 2021. Phonemizer: Text to phones transcription for multiple languages in python. *Journal of Open Source Software*, 6(68):3958.

Peter Birkholz. 2013a. Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PloS one*, 8(4):e60603.

Peter Birkholz. 2013b. Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PloS one*, 8:e60603.

Alan W Black. 2019. CMU wilderness multilingual speech dataset. In *ICASSP*, pages 5971–5975. IEEE.

Florent Bocquelet, Thomas Hueber, Laurent Girin, Pierre Badin, and Blaise Yvert. 2014. Robust articulatory speech synthesis using deep neural networks for bci applications. In *15th Annual Conference of the International Speech Communication Association (Interspeech 2014)*.

Yu-Wen Chen, Kuo-Hsuan Hung, Shang-Yi Chuang, Jonathan Sherman, Wen-Chin Huang, Xugang Lu, and Yu Tsao. 2021. Ema2s: An end-to-end multimodal articulatory-to-speech system. In *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE.

Tam'as G'abor Csap'o, Csaba Zaink'o, L. Viktor T'oth, Gábor Gosztolya, and Alexandra Mark'o. 2020. Ultrasound-based articulatory-to-acoustic mapping with waveglow speech synthesis. In *Interspeech*.

Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Jia Ye, R. J. Skerry-Ryan, and Yonghui Wu. 2021. Parallel tacotron 2: A non-autoregressive neural tts model with differentiable duration modeling. *ArXiv*, abs/2103.14574.

Gunnar Fant. 1991. What can basic research contribute to speech synthesis? *Journal of Phonetics*, 19(1):75–90.

Gunnar Fant. 1995. The lf-model revisited. transformations and frequency domain analysis. *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm*, 2(3):40.

David Gaddy and Dan Klein. 2021. An improved model for voicing silent speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 175–181, Online. Association for Computational Linguistics.

Marc-Antoine Georges, Pierre Badin, Julien Diard, Laurent Girin, Jean-Luc Schwartz, and Thomas Hueber. 2020. Towards an articulatory-driven neural vocoder for speech synthesis. In *International Seminar on Speech Production*.

Pengcheng Guo, Florian Boyer, Xuankai Chang, Tomoki Hayashi, Yosuke Higuchi, Hirofumi Inaguma, Naoyuki Kamo, Chenda Li, Daniel Garcia-Romero, Jiatong Shi, et al. 2021. Recent developments on espnet toolkit boosted by conformer. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5874–5878. IEEE.

Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Gregory Frederick Diamos, Erich Elsen, Ryan J. Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and A. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition. *ArXiv*, abs/1412.5567.

Tomoki Hayashi, Ryuichi Yamamoto, Takenori Yoshimura, Peter Wu, Jiatong Shi, Takaaki Saeki, Yooncheol Ju, Yusuke Yasuda, Shinnosuke Takamichi, and Shinji Watanabe. 2021. Espnet2-tts: Extending the edge of tts research. *arXiv preprint arXiv:2110.07840*.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. 2020. ESPnet-ST: All-in-one speech translation toolkit. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 302–311, Online. Association for Computational Linguistics.

Khalil Iskarous, Louis Goldstein, Douglas H Whalen, Mark Tiede, and Philip Rubin. 2003. Casy: The haskins configurable articulatory synthesizer. In *International Congress of Phonetic Sciences, Barcelona, Spain*, pages 185–188.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. 2021. Translatotron 2: Robust direct speech-to-speech translation. *arXiv preprint arXiv:2107.08661*.

Ye Jia, Ron Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. In *Interspeech*, pages 1123–1127.

A Karmel, Anushka Sharma, Muktak pandya, and Diksha Garg. 2019. Iot based assistive device for deaf, dumb and blind people. *Procedia Computer Science*, 165:259–269. 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION , 2019 November 11-12, 2019.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5530–5540. PMLR.

Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. 2018. Crepe: A convolutional representation for pitch estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 161–165. IEEE.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033. Curran Associates, Inc.

Kevin Lenzo and Alan Black. 2000. Diphone collection and synthesis. *ICSLP*.

Yongwan Lim, Asterios Toutios, Yannick Bliesener, Ye Tian, Sajan Lingala, Colin Vaz, Tanner Sorensen, Miran Oh, Sarah Harper, Weiyi Chen, Yoonjeong Lee, Johannes Töger, Mairym Llorens Monteserin, Caitlin Smith, Bianca Godinez, Louis Goldstein, Dani Byrd, Krishna Nayak, and Shrikanth Narayanan. 2021. A multispeaker dataset of raw and reconstructed speech production real-time mri video and 3d volumetric images. *Scientific Data*, 8.

Zheng-Chen Liu, Zhen-Hua Ling, and Li-Rong Dai. 2018. Articulatory-to-acoustic conversion using blstm-rnns with augmented input representation. *Speech Communication*, 99:161–172.

Max Morrison, Rithesh Kumar, Kundan Kumar, Prem Seetharaman, Aaron Courville, and Yoshua Bengio. 2022. Chunked autoregressive gan for conditional waveform synthesis. In *Submitted to ICLR 2022*.

Tomáš Nekvinda and Ondřej Dušek. 2020. One Model, Many Languages: Meta-Learning for Multilingual Text-to-Speech. In *Interspeech*, pages 2972–2976.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In *Interspeech*.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 2019. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621.

Korin Richmond, Phil Hoole, and Simon King. 2011. Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. In *Interspeech*, pages 1505–1508.

Philip Rubin, Thomas Baer, and Paul Mermelstein. 1981. An articulatory synthesizer for perceptual research. *The Journal of the Acoustical Society of America*, 70(2):321–328.

Celia Scully. 1990. Articulatory synthesis. In *Speech production and speech modelling*, pages 151–186. Springer.

Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. 2020. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Simon Stone, Philipp Schmidt, and Peter Birkholz. 2020. Prediction of voicing and the f0 contour from electromagnetic articulography data for articulation-to-speech synthesis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7329–7333. IEEE.

Fumiaki Taguchi and Tokihiko Kaburagi. 2018. Articulatory-to-speech conversion using bi-directional long short-term memory. In *Interspeech*, pages 2499–2503.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2019. Speech-to-speech translation between untranscribed unknown languages. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 593–600.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xin Wang, Shinji Takaki, and Junichi Yamagishi. 2019. Neural source-filter-based waveform model for statistical parametric speech synthesis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5916–5920.

Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Z. Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Robert A. J. Clark, and Rif A. Saurous. 2017. Tacotron: Towards end-to-end speech synthesis. In *Interspeech*.

Peter Wu, Paul Pu Liang, Jiatong Shi, Ruslan Salakhutdinov, Shinji Watanabe, and Louis-Philippe Morency. 2021a. Understanding the tradeoffs in client-side privacy for downstream speech tasks. In *APSIPA ASC*.

Peter Wu, Jiatong Shi, Yifan Zhong, Shinji Watanabe, and Alan W Black. 2021b. Cross-lingual transfer for speech processing using acoustic language similarity. In *ASRU*.

Chengzhu Yu, Heng Lu, Na Hu, Meng Yu, Chao Weng, Kun Xu, Peng Liu, Deyi Tuo, Shiyin Kang, Guangzhi Lei, et al. 2019. Durian: Duration informed attention network for multimodal synthesis. *arXiv preprint arXiv:1909.01700*.

Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, RJ Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. 2019. Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. *arXiv preprint arXiv:1907.04448*.

A Appendix

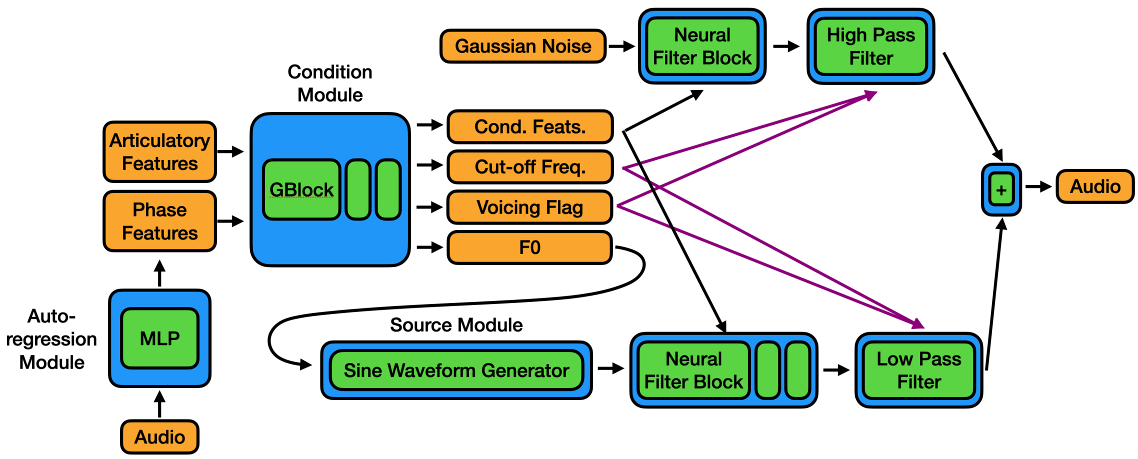

Figure 7: Model architecture of our NSF-CAR generator.